Analysis of Executional and Procedural Errors in Dry-lab Robotic Surgery Experiments

**Kay Hutchinson*[1] · Zongyu Li*[1] · Leigh A. Cantrell[2] ·
Noah S. Schenkman[3] · Homa Alemzadeh[1]**

Abstract

Background We aim to develop a method for automated detection of potentially erroneous motions that lead to sub-optimal surgeon performance and safety-critical events in robot-assisted surgery.
Methods We develop a rubric for identifying task and gesture-specific Executional and Procedural errors and evaluate dry-lab demonstrations of Suturing and Needle Passing tasks from the JIGSAWS dataset. We characterize erroneous parts of demonstrations by labeling video data, and use distribution similarity analysis and trajectory averaging on kinematic data to identify parameters that distinguish erroneous gestures.
Results Executional error frequency varies by task and gesture and correlates with skill level. Some predominant error modes in each gesture are distinguishable by analyzing error-specific kinematic parameters. Procedural errors could lead to lower performance scores and increased demonstration times but also depend on surgical style.
Conclusions This study provides preliminary evidence that automated error detection can provide context-dependent and quantitative feedback to surgical trainees for performance improvement.

1 Introduction

With advances in sensing and computing technology, artificial intelligence, and data science, the next generation of Robot-Assisted Surgery (RAS) systems is envisioned to benefit from new capabilities for context-specific monitoring [31] and virtual coaching during simulation training as well as decision support and cognitive assistance during actual surgery to improve safety, efficiency, and quality of care [28]. State-of-the-art RAS systems and simulators are designed with data logging mechanisms to collect system logs, kinematics, and video data from surgical procedures. The recorded data has been mostly used for offline surgical skill evaluation [7, 20, 21], with the aim of improving surgeons' performance and making evaluations objective and scalable.

Current methods for objective assessment of robotic technical skills can be classified into two general categories: manual assessment and automated assessment. Manual skill evaluation is usually performed globally, assessing performance over an entire demonstration using frameworks such as OSATS (Objective Structured Assessment of Technical Skills) [15], GOALS (Global Operative Assessment of Laparoscopic Skills) [30], GEARS [22], and R-OSATS [27]. However, manual assessment methods are subjective, cognitively demanding, and prone to errors [7]. In response, automated assessment methods utilizing kinematic, video, and system event data [19] are being developed to provide objective and quantitative metrics [7] and [5], and explainable feedback [8]. Automated methods also allow the subdivision of demonstrations into subtasks or gestures, and to base performance assessment and technical skill evaluation on the quality and/or sequence of these components

[*] Co-first-authors: Contributed equally to the paper.
[1] Department of Electrical and Computer Engineering,
[2] Department of Obstetrics and Gynecology,
[3] Department of Urology,
University of Virginia, Charlottesville, VA 22903
E-mail: [1]{kch4fk, zl7qw, ha4d}@virginia.edu, [2,3]{lac6vz, nss2f}@hscmail.mcc.virginia.edu

as proposed in [1], [26], and [21]. Further, some gestures are more indicative of skill level than others [29].

The metrics used for surgical skill assessment can be classified into three broad categories of: i) efficiency (e.g., path length, completion time), ii) safety (e.g., instrument collisions [18], instruments out of view, excessive force, needle drops, tissue damage [6]), and iii) task/procedure specific metrics (e.g., task outcome metrics, camera movement, energy activation [11]).

While most previous works focused on skill evaluation for distinguishing between expertise levels, less attention has been paid to identifying specific erroneous surgical motions that contribute to sub-optimal performance and potential safety-critical events. The closest related work is [16] which proposed an objective gesture-based checklist for laparoscopic suturing and validated it with measures of time, path length, needle positioning, and knot quality. Others have proposed general and custom rubrics for evaluation of human errors [12] and technical errors [3] in laparoscopic surgery. Related works on errors in RAS mainly focused on analyzing adverse events and system malfunctions as reported by the surgical teams and institutions [2].

Our goal is to augment RAS systems and simulators with mechanisms to monitor the progress of surgical tasks, and provide early and context-specific feedback to surgeons on potentially sub-optimal or unsafe motions that might lead to low performance scores in training or safety-critical events during surgery [31]. In this study, we take a step towards this goal by analyzing recorded dry-lab demonstrations of two common tasks (Suturing and Needle Passing) performed on the da Vinci Research Kit (dVRK) [14]. We focus on identifying which parts of a trajectory (spanning one or more gestures) are potentially erroneous (sub-optimal) versus error-free (optimal). We then characterize the erroneous trajectories by identifying the most common types of errors for each task and gesture, and the kinematic parameters and surgeon-specific signatures that distinguish between optimal and sub-optimal performance. The results from this study can aid in designing more efficient training modules, curricula, and simulation tools that reinforce optimal performance by providing more detailed and quantitative feedback to surgeons. The key contribution of this paper is proposing a novel technique for objective evaluation of RAS procedures with a focus on:

- A task and gesture-specific rubric for errors in RAS for manual or automated annotation of errors using video data collected from real or simulated demonstrations.
- Errors specific to gestures, as gestures are the building blocks of surgical tasks and can characterize the context of the procedure and unique surgeon's signatures
- Deviations from standard acceptable surgical trajectories that are potentially safety-critical and, if not detected and corrected, can lead to adverse events.
- Quantitative analysis methods for automated detection of gesture-specific Executional and Procedural errors using kinematic data collected from task demonstrations.

2 Methods

Sources of errors in RAS are diverse and domain-specific, including faults in the robotic system software and hardware, or human errors [2]. In this study, we focus on errors in the execution of procedures that can be observed in video recordings and detected in kinematic data. Surgical procedures follow the hierarchy of levels defined in [17] which provides context [31] for actions during the operation, as shown in Figure 1. A surgical **operation** can involve multiple **procedures** which are divided into **steps**. Each **step** is subdivided into **tasks** comprised of **gestures** (also called sub-tasks or surgemes) which are made of **motions** such as moving an instrument or closing the graspers. Errors can occur at any level of this hierarchy and can propagate and cause errors at other

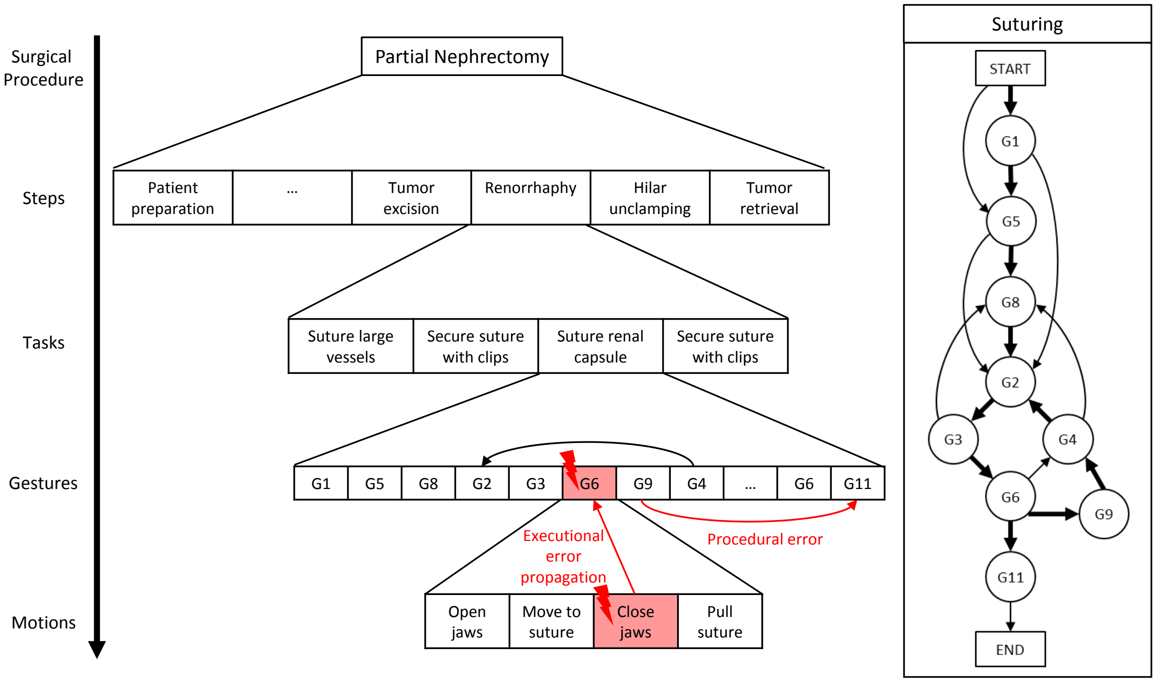


Fig. 1: Surgical hierarchy (adopted from [17]) for an example urological procedure of partial nephrectomy (based on [25] and [13]) along with example gesture-specific Executional and Procedural Errors

levels. We specifically focus on studying the quality of the task demonstrations at the *gesture* level to answer the following research questions:

RQ1: Which tasks and gestures are most prone to errors?

RQ2: Are there common errors modes or patterns across gestures and tasks?

RQ3: Are erroneous gestures distinguishable from normal gestures?

RQ4: What kinematic parameters can be used to distinguish between normal and erroneous gestures?

RQ5: Do errors impact the duration of the trajectory?

RQ6: Are there any correlations between errors and surgical skill levels?

2.1 Rubric for Objective Assessment of Errors in Robotic Surgery

Our goal is to define an error rubric that can be used for automatic detection of errors using quantitative measures, such as instrument position, amount of force, traveling distance, and system events. We adopt a previous categorization of human errors in laparoscopic surgery from [12] and define two types of errors in our rubric: **Procedural errors** and **Executional errors**. Procedural errors involve “the omission or re-arrangement of correctly undertaken steps within the procedure,” while Executional errors are “the failure of a specific motor task within the procedure.” **Technical errors** are the “failure of a planned action to achieve a goal”, including inadequate (too much/too little) use of force or distance, inadequate visualization and wrong orientation of instruments or dissection plane [4], and are considered a subtype of Executional errors that can be quantified with thresholds.

In order to generalize these definitions to different procedures and tasks, we define Executional and Procedural errors at the gesture level. More specifically, we define a set of Executional error modes for each gesture as listed in the rubric in Table 1. Some errors are gesture-specific such as “Needle orientation” which is only defined for G4 and G8 as those gestures specifically manipulate

Gesture Description		Error Mode	Suturing		Needle Passing	
			Total No. Errors	Erroneous Gestures (%)	Total No. Errors	Erroneous Gestures (%)
G1	Reaching for needle with right hand	Multiple attempts	7	8/29 (28%)	N/A	11/30 (37%)
		Needle drop	0		2	
		Out of view	1		10	
G2	Positioning needle	Multiple attempts	21	22/166 (13%)	51	55/117 (47%)
		Needle drop	0		0	
		Out of view	1		6	
G3	Pushing needle through tissue	Not moving along the curve/ Multiple attempts	80	82/164 (51%)	17	17/111 (15%)
		Needle drop	0		0	
		Out of view	2		0	
G4	Transferring needle from left to right	Multiple attempts	19	71/119 (60%)	15	23/83 (28%)
		Needle orientation	53		9	
		Needle drop	0		0	
		Out of view	14		3	
G5	Moving to center with needle in grip	Needle drop	1	2/37 (5%)	0	3/31 (10%)
		Out of view	1		3	
G6	Pulling suture with left hand	Multiple attempts	8	121/163 (74%)	14	46/112 (41%)
		Needle drop	2		0	
		Out of view	120		37	
G8	Orienting needle	Multiple attempts	18	28/48 (58%)	1	3/28 (11%)
		Needle orientation	22		1	
		Needle drop	0		0	
		Out of view	4		2	
G9	Using right hand to help tighten suture	Multiple attempts	3	11/24 (46%)	1	1/1 (100%)
		Needle drop	0		1	
		Out of view	11		0	
All gestures	Total number of errors across all gestures	Multiple attempts	156	345/750 (46%)	99	159/513 (31%)
		Needle drop	3		3	
		Needle orientation	75		10	
		Out of view	154		61	

Table 1: Gesture-specific Executional errors for Suturing and Needle Passing in the JIGSAWS dataset. Example videos for each error mode can be found in the supplementary video files.

the needle in preparation for positioning the needle (G2) and throwing the next suture (G3), as shown in the grammar graph of Figure 1 (adopted from [1]). The standard acceptable practice for those gestures is to hold the needle in the grasper 1/2 to 2/3 of the way from the tip of the needle and with the needle perpendicular to the jaws of the grasper [16]. Other gestures that do not purposely alter the orientation of the needle in the grasper cannot have this error mode. For G3, the definition of a “Multiple attempts” error also includes “Not moving along the curve” of the needle (from [16]) since these two errors are very difficult to distinguish and often happen simultaneously. Other error modes, including “Multiple attempts”, “Needle drop”, and “Out of view”, could occur at any time during a task and are considered for every gesture.

We define Procedural errors as any deviation in the sequence of gestures performed in a demonstration from the standard accepted gesture sequences defined for that task and shown in the grammar graphs in Figures 1 and 2. [12] defined several sub-categories for Procedural errors, including adding an unexpected step, skipping a step, out of order transition, and repetition of steps. These subcategories are included in our analysis of Procedural errors as discussed in Section 2.4.

2.2 JIGSAWS Dataset

The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [9] is a publicly available dataset, collected using the da Vinci Research Kit [14] from eight surgeons of varying skill levels performing three dry-lab surgical tasks: Suturing, Knot Tying, and Needle Passing. These tasks are among the standard modules in most surgical skills training curricula.

The JIGSAWS dataset includes kinematic and video data from up to 39 demonstrations (or trials) per task along with manually annotated gesture **transcripts** (indicating the sequence of

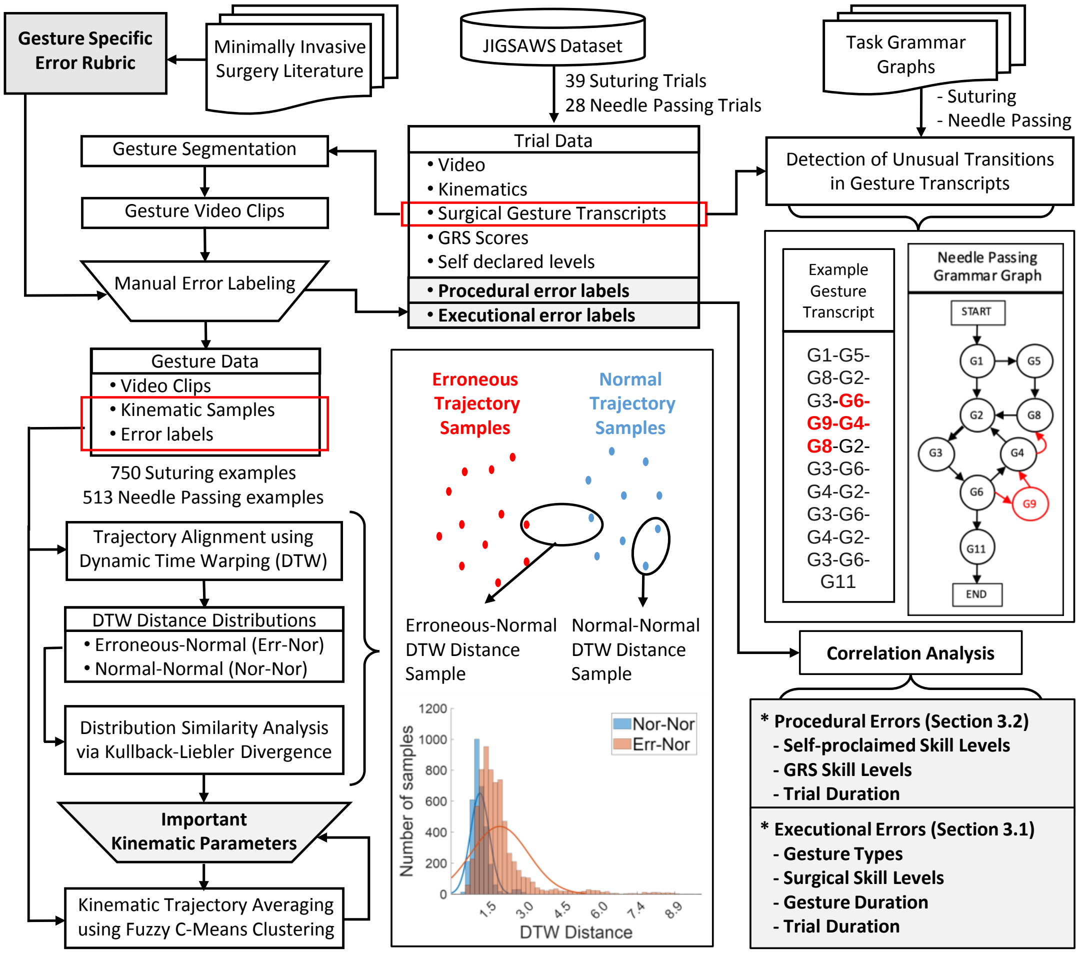

Gesture Specific Error Rubric
Minimally Invasive Surgery Literature
JIGSAWS Dataset
39 Suturing Trials
28 Needle Passing Trials
Task Grammar Graphs
- Suturing
- Needle Passing
Gesture Segmentation
Trial Data
• Video
• Kinematics
• Surgical Gesture Transcripts
• GRS Scores
• Self declared levels
• Procedural error labels
• Executional error labels
Detection of Unusual Transitions in Gesture Transcripts
Gesture Video Clips
Manual Error Labeling
Example Gesture Transcript
G1-G5-G8-G2-G3-G6-G9-G4-G8-G2-G3-G6-G4-G2-G3-G6-G4-G2-G3-G6-G11
Needle Passing Grammar Graph
START
G1
G5
G2
G8
G3
G4
G6
G9
G11
END
Gesture Data
• Video Clips
• Kinematic Samples
• Error labels
750 Suturing examples
513 Needle Passing examples
Erroneous Trajectory Samples
Normal Trajectory Samples
Trajectory Alignment using Dynamic Time Warping (DTW)
DTW Distance Distributions
• Erroneous-Normal (Err-Nor)
• Normal-Normal (Nor-Nor)
Erroneous-Normal DTW Distance Sample
Normal-Normal DTW Distance Sample
Correlation Analysis
Distribution Similarity Analysis via Kullback-Liebler Divergence
Number of samples
1200
1000
800
600
400
200
0
Nor-Nor
Err-Nor
1.5
3.0
4.5
6.0
7.4
8.9
DTW Distance
* Procedural Errors (Section 3.2)
- Self-proclaimed Skill Levels
- GRS Skill Levels
- Trial Duration
* Executional Errors (Section 3.1)
- Gesture Types
- Surgical Skill Levels
- Gesture Duration
- Trial Duration
Important Kinematic Parameters
Kinematic Trajectory Averaging using Fuzzy C-Means Clustering

Fig. 2: Overall Methodology for Analysis of Executional and Procedural Errors

gestures, with the beginning and end of each gesture and its type) and surgical skill levels for each demonstration. The vocabulary of surgical gestures used for labeling is shown in Table 1. Surgical skills were characterized using both self-proclaimed expertise levels and Global Rating Scale (GRS) score for each demonstration. Self-proclaimed (SP) expertise levels were based on the number of hours of robotic surgical experience, divided into: **SP-Expert** (>100 hrs), **SP-Intermediate** (10-100 hrs), and **SP-Novice** (<10 hrs). GRS scores were given using a modified Objective Structured Assessments of Technical Skills (OSATS) approach based on 6 elements (on a rating-scale of 1-5 per element): Respect for tissue, Suture/needle handling, Time and motion, Flow of operation, Overall performance and Quality of final product [9]. We also classified the demonstrations into three groups based on the GRS scores: **GRS-Novice** ($0 \leq$GRS ≤ 9), **GRS-Intermediate** ($10 \leq$GRS ≤ 19), and **GRS-Expert** ($20 \leq$GRS ≤ 30).

Figure 2 shows our overall pipeline for the analysis of Executional and Procedural errors in the JIGSAWS dataset. Due to the limited number of demonstrations for the Knot Tying task in the dataset, our analysis only focused on Suturing and Needle Passing.

2.3 Executional Error Analysis

Kinematic and video data for each trial were first segmented into gestures based on the gesture transcript annotations. The video clip for each gesture was then reviewed and labeled by two to

Index	Description of variables	Parameter name
39-41	Right PSM1 tool tip position (xyz)	R Pos
42-50	Right PSM1 tool tip rotation matrix (R)	R Rot Mat
51-53	Right PSM1 tool tip linear velocity (x' y' z')	R Lin Vel
54-56	Right PSM1 tool tip rotational velocity (α' β' γ')	R Rot Vel
57	Right PSM1 gripper angle (Θ)	R Grip Ang
58-60	Left PSM2 tool tip position (xyz)	L Pos
61-69	Left PSM2 tool tip rotation matrix (R)	L Rot Mat
70-72	Left PSM2 tool tip linear velocity (x' y' z')	L Lin Vel
73-75	Left PSM2 tool tip rotational velocity (α' β' γ')	L Rot Vel
76	Left PSM2 gripper angle (Θ)	L Grip Ang

Table 2: Kinematic variables in the JIGSAWS dataset (adopted from [9])

three independent annotators as normal or erroneous for each error mode. Final labels for each error mode were obtained by taking the consensus among annotators. A gesture example that exhibited one or more errors was marked as erroneous, otherwise, it was labeled as normal. We then proceeded with the analysis of the patient-side manipulator (PSM) kinematic data corresponding to each gesture for all the normal and erroneous demonstrations of each task.

2.3.1 Dynamic Time Warping

We used Dynamic Time Warping (DTW) to measure the similarity between normal and erroneous trajectories for each gesture. DTW is an effective method for aligning two temporal sequences, independent of the non-linear variations in time, by minimizing the Euclidean distance between the two signals. In our analysis, we performed independent DTW on each variable before summing the returned distances for each parameter listed in Table 2. We found no significant difference between this method and dependent DTW where all variables in each parameter group were warped together yielding a single distance instead of a sum of distances (similar observations were made in [24]). DTW was performed on every combination of two example trajectories for each gesture. From this, we obtained comparisons of normal examples to other normal examples ("Nor-Nor") and comparisons of erroneous examples to normal examples ("Err-Nor"). The DTW distance samples represented a distribution of distances for the "Nor-Nor" and "Err-Nor" subsets as shown in the histogram of Figure 2. This resulted in two sets of distance samples for each parameter, each representing a DTW distribution for a comparison subset.

2.3.2 Kullback-Liebler Divergence

Kullback-Liebler (KL) Divergence, also called relative entropy, is a non-symmetric measure of the difference between two probability distributions. The KL Divergence between two identical distributions is zero. As shown in Equation 1, KL Divergence was used to compare the "Err-Nor" and "Nor-Nor" DTW distance distributions for each gesture to determine which parameters had a significant difference between the two distributions.

$$D_{KL}(DTW_{Err-Nor}||DTW_{Nor-Nor}) = -\Sigma DTW_{Err-Nor} log(\frac{DTW_{Nor-Nor}}{DTW_{Err-Nor}}) \quad (1)$$

2.3.3 Trajectory Averaging

We examined the kinematic data for important parameters to verify differences between normal and erroneous gestures using a method based on [10]. Each signal was time-normalized by downsampling the signal by 3 (keeping only every third sample) and then linearly interpolated to stretch it to the average duration of the normal or erroneous gesture examples of that task (supported by our

Algorithm 1: Procedural Error Detection Algorithm

```
Input:
  - A grammar graph $G(V, E)$ for a surgical task which is a digraph with each vertex in $V$ representing the
  entry point START or one of the common gesture types $G_i$ in the task, and each $edge(G_i, G_j) \in E$
  representing a common transition between gestures $G_i$ and $G_j$.
  - A set of $m$ task transcripts $T = \{T_1, T_2, ..., T_m\}$, where $T_k \in T$ is an ordered sequence of $n$ gestures
  $T_k = [G_1, G_2, ..., G_n]$
  Output:
  - A list of erroneous gesture transitions $error\_seq$ for each transcript
  for $T_k \in T$ do
    $error\_seq = \emptyset$
    $val \leftarrow G.successors(START)$
    for $G_i \in T_k$ do
      if $G_i \in V$ then
        if $G_i \notin val$ then
          $error\_seq$.append($[G_{i-1}, G_i]$)
        $val \leftarrow G.successors(G_i)$
      else $error\_seq$.append($[G_i]$)
        $val \leftarrow [G_{i+1}]$
    end
end
```

analysis of gesture durations in Section 3.1.4). Then, fuzzy c-means clustering was performed on each variable and its normalized time index to obtain the average normal and erroneous trajectories (represented by 15 cluster centers), shown with blue (normal) and red (erroneous) dots in Figure 4b.

2.4 Procedural Error Analysis

Previous works proposed modeling the standard acceptable gesture sequences for a task using a grammar graph that shows the relationship, order, and flow of gestures [29], [1]. [26]. The grammar graph of a task is a digraph with each vertex representing the set of gestures for the task and each edge representing a common transition between two gestures. We adopted the grammar graphs for Suturing and Needle Passing from [1] and included an additional directed link from G1 to G2 in Suturing (see Figures 1 and 2).

We acquired the gesture sequences performed for Suturing and Needle Passing from the JIGSAWS transcripts. Then we developed a method for checking if each gesture sequence follows the standard acceptable sequence of gestures in the grammar graph. As shown in Algorithm 1, for each gesture we check if it is in the grammar graph for the task and if it is a valid successor of the previous gesture, otherwise it is marked as a procedural error. Each transcript can have multiple, possibly sequential, procedural errors. This algorithm, combined with a gesture segmentation algorithm, can be used for automated detection of procedural errors in real-time.

Deviations from the grammar graph might also happen because of variations in surgical style and expertise, as discussed in Section 3.2.

3 Results

This section presents a summary of results and observations from our analysis of the JIGSAWS dataset.

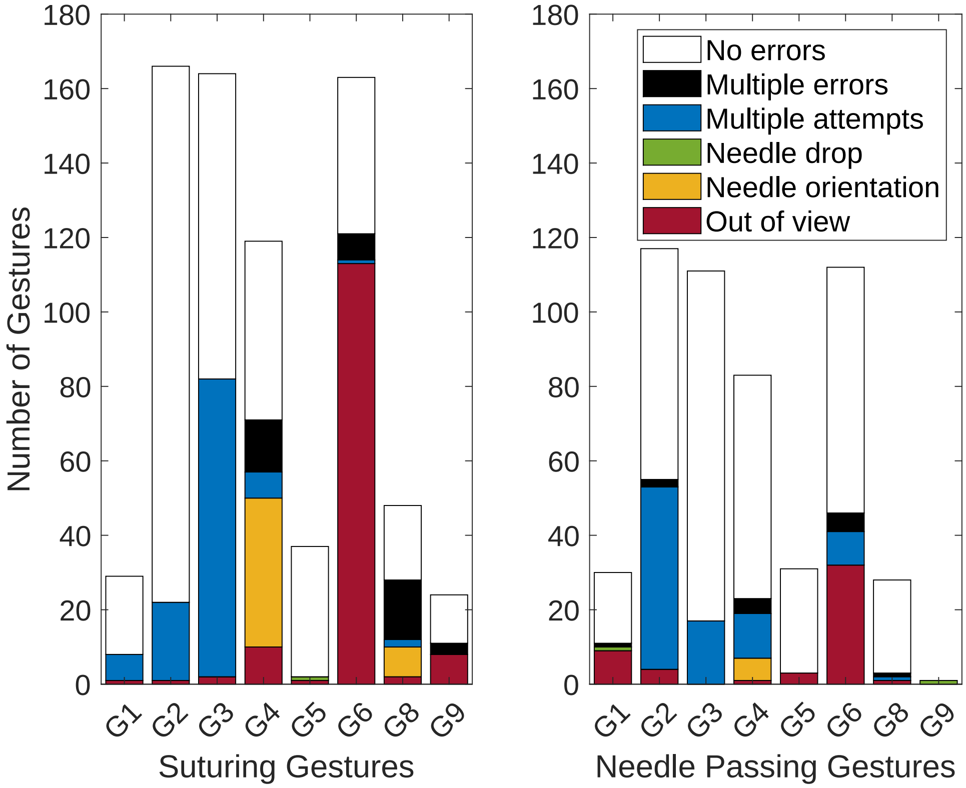


Fig. 3: Distribution of Errors for Each Gesture

3.1 Executional Errors

Table 1 lists the number of examples of each error mode as well as the total number of erroneous examples for each gesture. Note that a gesture example could exhibit multiple error modes, so the sum of the total number of errors does not necessarily equal the number of erroneous gestures.

3.1.1 Distribution of Executional Errors Among Gestures

Figure 3 shows the distribution of errors of each type for each gesture from Table 1. If a gesture example had more than one error label, it was counted under the "Multiple errors" category. We made the following observations:

- G5 for both tasks and G1, G8, and G9 for Needle Passing did not have enough examples of executional errors, so further analysis was not performed on these gestures. G8 from Needle Passing and G5 from both tasks had the lowest percentage of errors because they may be less challenging than other gestures.
- G2 and G3 have the most "Multiple attempts" errors in both tasks because they require a high level of accuracy in positioning and driving the needle though the tissue, respectively. G2 has more errors in Needle Passing because the eye of the ring is a smaller target than the dot on the fabric. G3 has more errors in Suturing because surgeons often tried multiple times to align the tip of the needle with the exit point while the needle was not visible beneath the fabric. Comparatively in Needle Passing, the needle only had to pass through one point and was always visible.
- G4 and G6 from both tasks, and G8 from Suturing have the most gestures with multiple errors. G4 and G8 both involve manipulating the needle between the graspers and the predominant error modes were "Needle orientation" and "Multiple attempts" likely due to issues with hand

Task	Gesture	Parameters
Suturing	G1	Right Gripper Angle Right Linear Velocity Right Position
	G3	Right Linear Velocity Right Rotational Velocity Right Gripper Angle
	G6	Left Position
	G8	Right Position Left Gripper Angle Left Linear Velocity Right Gripper Angle
	G9	Left Gripper Angle
Needle Passing	G2	Left Rotational Velocity Left Linear Velocity
	G3	Left Rotational Velocity Right Rotation Matrix Right Gripper Angle

Table 3: Kinematic Parameters with the Greatest KL Divergence Distinguishing Errors in Different Gestures

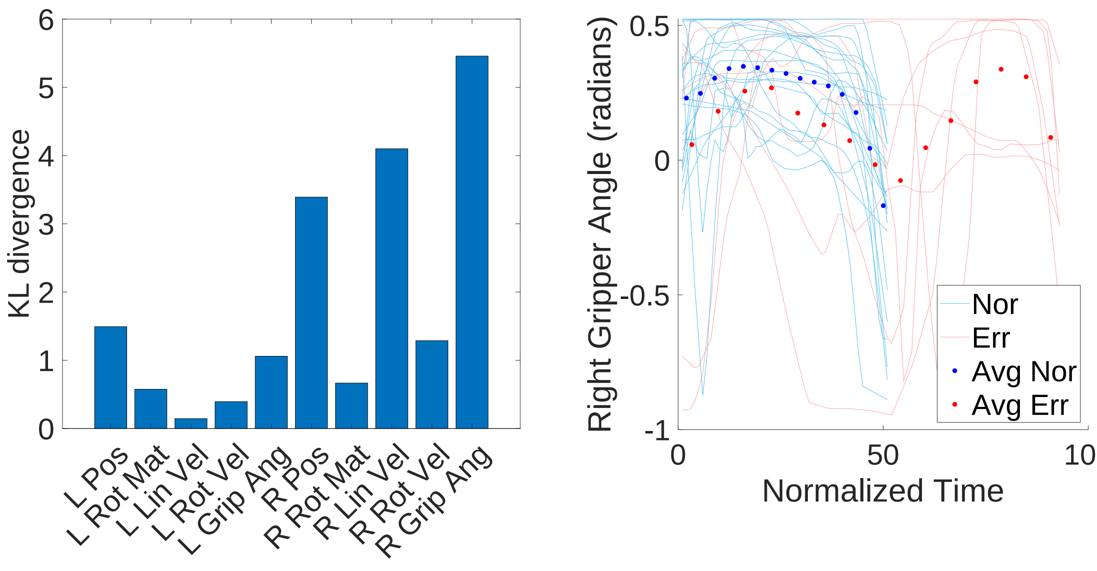
coordination. For G6, the main error modes were "Out of view" and "Multiple attempts" due to multiple attempts at grasping the needle and pulling it through the ring or tissue and then moving off-camera to pull the suture through.

- G6 has a large number of "Out of view" errors especially in Suturing possibly because surgeons could not move the camera for the trials in the JIGSAWS data set. However, a different technique to pull the suture could have been used such as hand-over-hand or the pulley method that would have kept the tools within view.

3.1.2 Kinematic Parameters for Distinguishing Errors in Each Gesture

We performed a comparative analysis of KL Divergence values for parameters in each gesture and identified the kinematic parameters that are associated with error occurrence as listed in Table 3. The following are key observations from this analysis:

- For G1 in Suturing, the predominant error mode was "Multiple attempts" at picking up the needle. Figures 4b and 5 show that erroneous gestures exhibited a second opening and closing of the grasper and a large difference in Y Position trajectories. This explains the large KL Divergences for those right hand parameters in Figure 4a.
- For G2 in Needle Passing, Figure 6 shows a large difference in KL Divergence for Left Rotational and Linear Velocities which may be due to the active role the left hand plays in stabilizing the ring unlike in Suturing. This is an important contextual difference between tasks.
- The main error mode for G3 was "Not moving along the curve/ Multiple attempts". Erroneous gestures in Suturing were caused by lateral, instead of characteristically rotational, movements of the needle while in the fabric. In surgery, lateral movements may tear tissue and contribute to a safety-critical event. This explains the high KL Divergences for the parameters listed in Table 3 and shown in Figure 7a and is consistent with [23] who found that the rate of orientation change during needle insertion (i.e. Rotational Velocity during G3) was higher for experienced surgeons. However, Needle Passing shows nearly the opposite result in Figure 7b. Upon reviewing the gesture clips for both tasks, we noticed that clips for Suturing showed the right grasper driving the needle through the fabric and the left grasper pulling it through, but clips for Needle Passing began with the needle halfway through the ring and only showed the left grasper pulling the needle through. Due to the large difference in KL Divergences between the two tasks, we see that the part of G3 that involves driving the needle with the right grasper is important to this gesture's correct execution.



(a) (b)

Fig. 4: G1 in Suturing: (a) KL Divergence of Kinematic Parameters, (b) Right Gripper Angle Trajectories for Normal and Erroneous Gestures

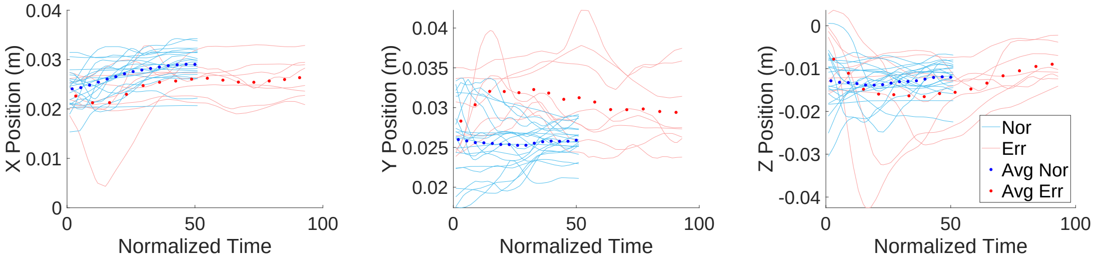
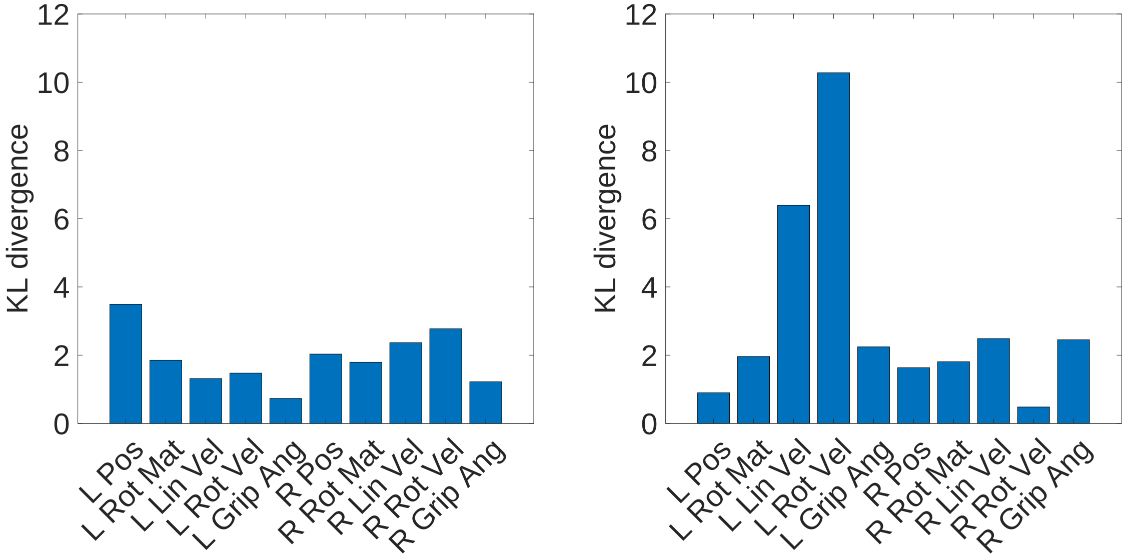


Fig. 5: Right Tooltip XYZ Position for Normal and Erroneous G1 in Suturing

- In both tasks, G4 had KL Divergences below 0.6 for all parameters meaning normal and erroneous examples have very similar kinematics.
- G6 in Suturing had the most errors with primarily "Out of view" errors. Figure 9 shows that final Y and Z positions for the left grasper were much larger for erroneous gestures as the left grasper exceeded the threshold for visibility while pulling the suture. This explains the large KL Divergence for the Left Position parameter in Figure 8.
- There were two main error modes for G8 in Suturing: "Multiple attempts" and "Needle orientation". Figure 10 shows a comparison of DTW and KL Divergence analysis for G8 from Suturing for all errors, for "Multiple attempts" versus all other examples, and for "Needle orientation" versus all other examples. The "Needle orientation" error alone had the greatest KL Divergence and contributed the most to the results for all errors. For the "Multiple attempts" error, both the Left and Right Position parameters had the highest KL Divergence which suggests that hand coordination is important in this gesture. Since this gesture includes the right gripper moving to grasp the needle, we see that Right Position is an important parameter in the "Multiple attempts" error both in G1 and G8.

3.1.3 Executional Errors and Skill Levels

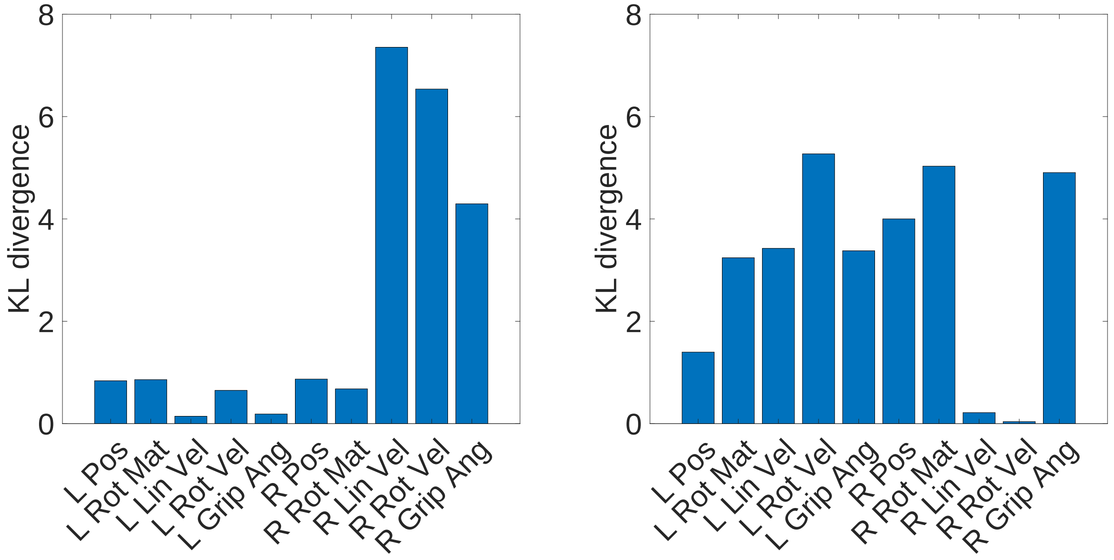
We analyzed the relationship between executional errors and surgical skill levels. Based on self-proclaimed expertise levels, Figure 11a shows a clear difference in the number of errors across



(a) Suturing

(b) Needle Passing

Fig. 6: KL Divergence of Kinematic Parameters for G2



(a) Suturing

(b) Needle Passing

Fig. 7: KL Divergence of Kinematic Parameters for G3

different self-proclaimed expertise groups for Suturing. However, no similar pattern was seen in Needle Passing. This might be because Suturing is a more difficult task so the number of executional errors is more reflective of self-proclaimed skill levels in Suturing. For GRS-defined skill levels, the total number of executional errors per trial was larger for GRS-Novices than for GRS-Experts in Needle Passing (Figure 11b), which is consistent with our expectation that experts with high GRS scores make fewer executional errors than novices. However, since there was only one GRS-Novice trial for Suturing, we did not observe clear differences.

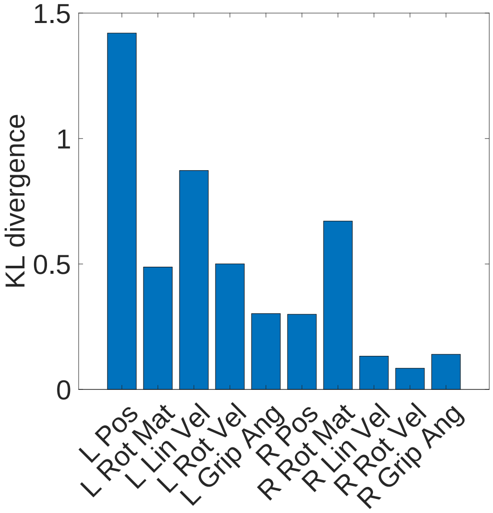


Fig. 8: KL Divergence of Kinematic Parameters for G6 in Suturing

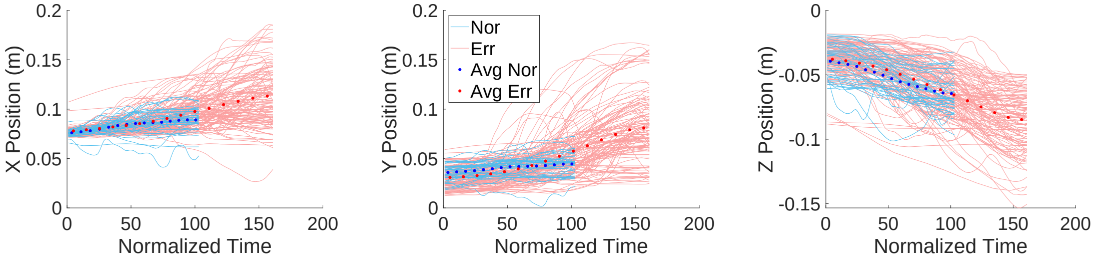
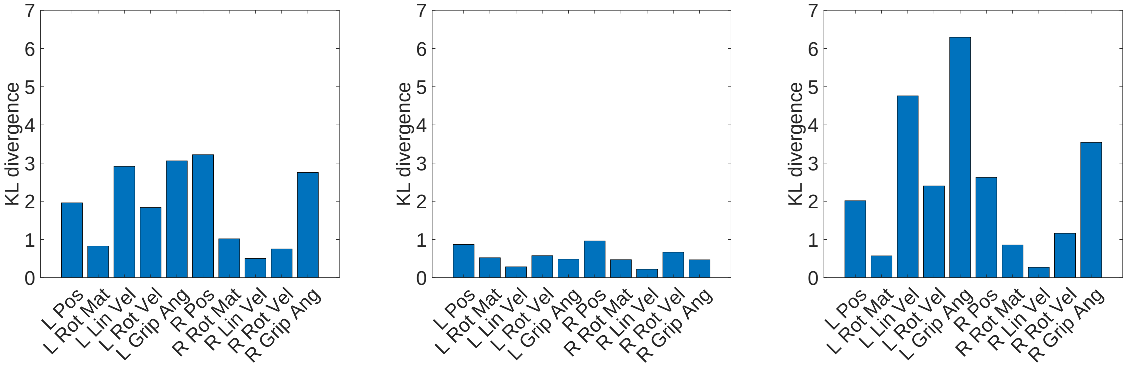


Fig. 9: Left Tooltip XYZ Position for Normal and Erroneous G6 in Suturing



(a) All Errors

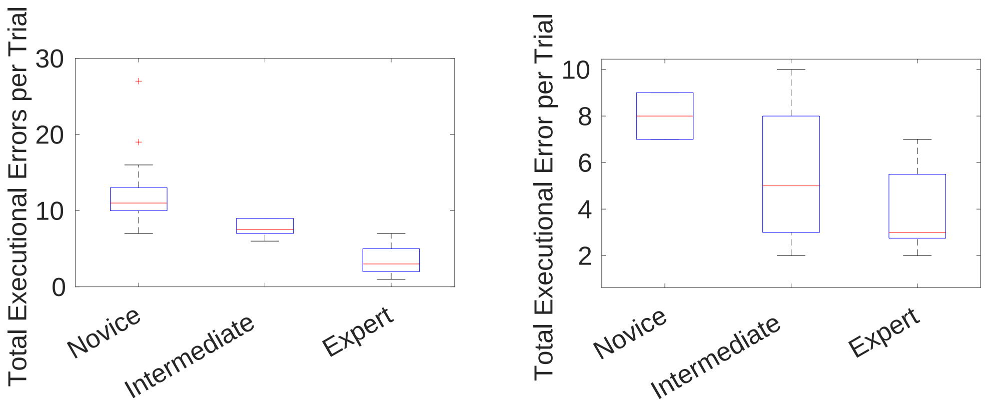
(b) Multiple Attempts Error

(c) Needle Orientation Error

Fig. 10: KL Divergence of Kinematic Parameters for G8 in Suturing

3.1.4 Executional Errors and Gesture Duration

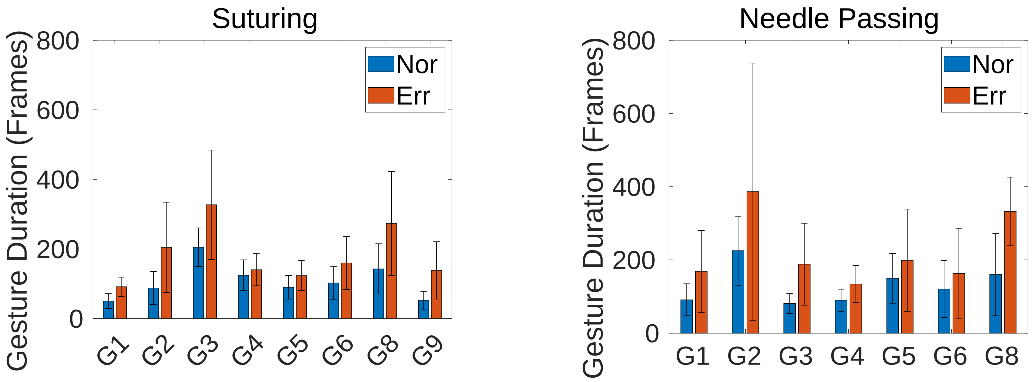
We compared erroneous and normal gesture durations using a one-tailed t-test. The null hypothesis is that normal and erroneous gestures have similar durations. The alternative hypothesis is that erroneous gestures are longer than normal gestures. Figures 12 and 13 respectively show average durations and several examples of differences in durations (along with the p-values from the hypothesis test) for normal and erroneous gestures in both tasks.



(a) Self-Proclaimed Skill Levels

(b) GRS Skill Levels

Fig. 11: Total Number of Executional Errors across Surgical Skill Levels: (a) Self-proclaimed Skill Levels in Suturing, (b) GRS Skill Levels in Needle Passing



(a) Suturing

(b) Needle Passing

Fig. 12: Average Normal and Erroneous Gesture Durations

We observed that some error types increase the gesture duration, e.g., “Multiple attempts” for G1, G2, G3, and G8 in Suturing, and G2 and G3 in Needle Passing; and “Out of view” for G6 and G9 for Suturing, and G4 in Needle Passing. Erroneous gestures with “Out of view” errors are longer because the distance traveled by the tool is larger, while the speed is similar. We rejected the null hypothesis and found that erroneous gestures are longer than normal gestures for all gestures of both tasks. There is a relatively large p-value (p=0.308) for G4 compared to other p-values. This could be because “Needle orientation” is the primary error mode in G4 and an erroneous needle orientation takes comparable time as a normal needle orientation.

3.1.5 Executional Errors and Trial Duration

For each trial, we summed the executional errors of all gestures in the trial. Then we analyzed the correlation between the total number of executional errors per trial and the duration of the trial (in number of frames). Figure 14 shows that there is a significant positive correlation for Suturing (r=0.837, p=6.18e-12), but no significant correlation for Needle Passing. This is likely due to the limited number of trials and fewer errors for Needle Passing in the JIGSAWS dataset (see Table 1).

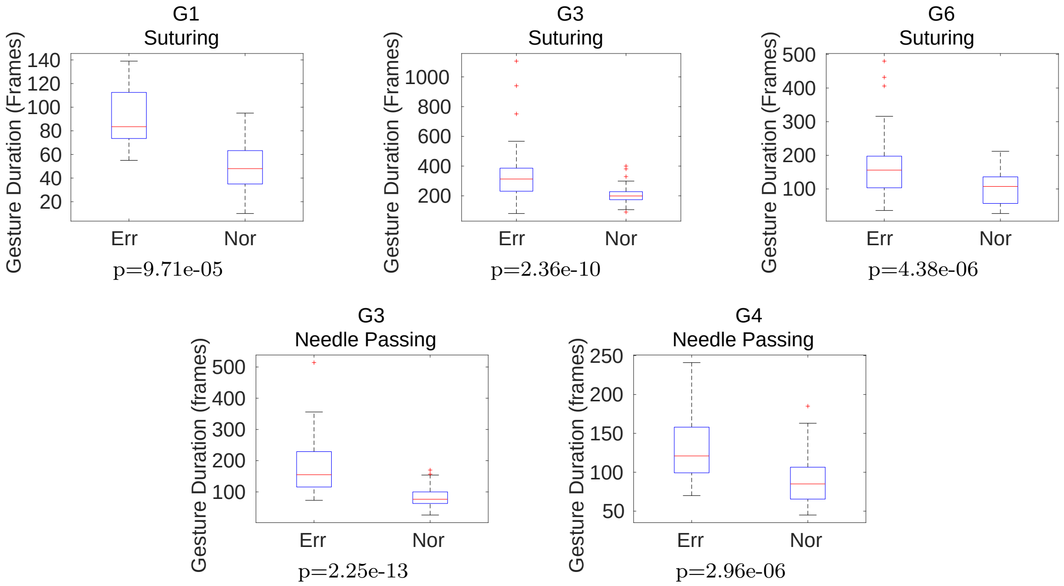


Fig. 13: Erroneous vs. Normal Gesture Durations for Suturing and Needle Passing

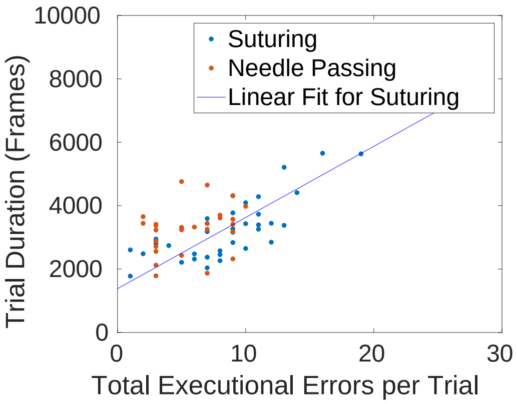


Fig. 14: Correlation between Executional Errors and Durations of Trials

3.2 Procedural Errors

We analyzed the numbers and patterns of procedural errors by task, skill level, and subject. We hypothesize that the number of procedural errors is inversely proportional to surgical experience and negatively correlated with the demonstration duration.

3.2.1 Procedural Errors and Self-Proclaimed Skill Levels

We compared the percentage of erroneous trials for SP-Novice, SP-Intermediate and SP-Expert groups. As shown in Table 4, we observe that for both tasks, SP-Expert surgeons on average had more procedural errors compared to SP-Intermediate surgeons. For Needle Passing, SP-Intermediate surgeons made more errors than SP-Novice surgeons. This could be due to variations in surgical style especially in more experienced surgeon groups. For example, our analysis of error patterns by subject showed that one of the SP-Expert subjects consistently made G9-G11 transitions in different trials of Suturing (see Figure 1). This is a unique non-safety-critical pattern that was not observed in the trials by other subjects. However, procedural errors by SP-Novice subjects were more random and did not follow specific patterns.

Task - Skill Level	Total Number of Procedural Errors	Percentage of Erroneous Trials	Longest Erroneous Gesture Sequences
Suturing - SP-Expert	11	6/10	G9-G6-G2
Suturing - SP-Intermediate	2	2/10	G3-G11
Suturing - SP-Novice	23	10/19	G4-G5-G6-G2
Needle Passing - SP-Expert	11	6/9	G2-G6-G10
Needle Passing - SP-Intermediate	9	5/8	G6-G8-G6
Needle Passing - SP-Novice	7	4/11	G6-G5-G6
Total	63	33/67	-

Table 4: Procedural Errors and GRS Skill Levels

	Suturing		Needle Passing	
GRS Sub-score	Correlation Coefficient	p-value	Correlation Coefficient	p-value
Respect for Tissue	-0.41	0.009	-0.12	0.528
Suture & Needle Handling	-0.50	0.001	-0.26	0.184
Time & Motion	-0.55	<0.001	-0.11	0.594
Flow of Operation	-0.43	0.006	-0.22	0.268
Overall Performance	-0.62	<0.001	-0.16	0.412
Quality of Final Product	-0.26	0.115	-0.02	0.920
GRS Score	-0.51	<0.001	-0.15	0.434

Table 5: Correlation between Number of Procedural Errors and GRS sub-scores for Suturing and Needle Passing

Task	r	p-value
Suturing	0.71	<0.001
Needle Passing	0.17	0.399

Table 6: Correlation between Procedural Errors and Trial Durations

Of the two tasks, the longest erroneous gesture sequence is G4-G5-G6-G2 in Suturing performed by an SP-Novice surgeon. Upon review of the video, G5 may be a typo in the transcript.

3.2.2 Procedural Errors and GRS Skill Levels

We analyzed the correlation between the number of procedural errors and GRS score (Table 5). The strongest negative correlation between the number of procedural errors, GRS score, and GRS sub-scores is in Suturing. Among the sub-scores of Suturing, Overall Performance has the strongest negative correlation with procedural errors. This could happen because an inefficient procedure has the greatest impact on Overall Performance in Suturing. Needle Passing has a weaker negative correlation between procedural errors and GRS score. The Needle Handling sub-score has the highest negative correlation with the number of procedural errors. This is expected as Needle Handling is the main component of the Needle Passing task and poor performance due to procedural errors will lead to a lower score.

3.2.3 Procedural Errors and Trial Duration

In Suturing, there is a significant positive correlation between procedural errors and the duration of the trials, so more procedural errors lead to longer trials. However, there is no significant correlation in Needle Passing possibly because Needle Passing is an easier task (Table 6).

4 Discussion

We used our insights from the analysis of executional and procedural errors in the JIGSAWS dataset to answer the research questions posed in Section 2:

RQ1: Which tasks and gestures are most prone to errors? More challenging gestures in each task that require a high level of accuracy and hand coordination were more prone to executional errors. As shown in Table 1, Suturing is more difficult than Needle Passing and had a greater number of executional errors. G6, G3, and G4 had the greatest number of executional errors in Suturing while G2 and G6 had the greatest number of executional errors in Needle Passing. However, procedural errors were almost equally likely in both tasks. 18/39 Suturing trials and 15/28 Needle Passing trials contained procedural errors (Table 4).

RQ2: Are there common error modes or patterns across gestures and tasks? Within each task, each gesture had a different predominant error mode that correlated with the challenging aspects of performing the gesture. For both tasks, G2 and G3 had a large number of "Multiple attempts" errors, G5 had the fewest errors, G6 had the largest number of "Out of view" errors, and G4 and G6 had the greatest number of gestures with Multiple Errors. Thus, executional errors are context-specific and their type and frequency depend on both task and gesture.

RQ3: Are erroneous gestures distinguishable from normal gestures? KL Divergence magnitude provides insight into which gestures have the greatest difference between normal and erroneous examples. We found that G9 from Suturing, G2 from Needle Passing, and G3 from Suturing had the three greatest KL Divergences for any parameter. However, upon examination of kinematic data for the Left Gripper Angle of G9 from Suturing, the large KL Divergence for this gesture could be due to the effect of three outlying gestures on an already relatively small sample of only 24 examples.

RQ4: What kinematic parameters can be used to distinguish between normal and erroneous gestures? Table 3 lists the parameters with the greatest KL Divergence for each gesture and task which can be used for automated error detection. Our KL divergence analysis approximated the DTW distance distributions as Gaussian, which might not be always accurate. Future work will focus on further refining our analysis method to address this limitation.

RQ5: Do errors impact the duration of the trajectory? Executional and procedural errors often lead to lengthier trials, especially during more complicated tasks such as Suturing. Timely detection and correction during training or surgery will enable more efficient and safer patient care, and aid in reducing learning curves and time to certification.

RQ6: Are there any correlations between errors and surgical skill levels? The total number of executional errors made per trial could help differentiate skill levels. We found this to be true for self-proclaimed skill levels in Suturing and GRS skill levels in Needle Passing.

There was a significant negative correlation between overall GRS scores and sub-scores and the total number of procedural errors made per trial in Suturing meaning a greater number of procedural errors contributes to a lower GRS score. After inspecting the procedural error patterns for different self-proclaimed skill levels, we noticed that procedural error analysis using a grammar graph may be most effective for novice surgeons as they tend to closely follow the graph, but experts have unique signatures that deviate from the graph. Further verification of the correlation between errors and skill levels requires access to larger datasets representing more tasks and surgeons. Additionally, the grammar graphs cannot completely capture all possible valid gesture sequences and surgeon-specific signatures. Manual labeling may introduce errors in the gesture transcripts (e.g., incorrectly adding or missing some gestures) that might lead to incorrect detection of errors. This motivates developing automated gesture identification and surgeon signature modeling methods using computer vision and kinematic analysis techniques.

5 Conclusion

We presented a new rubric and method for objective evaluation of RAS procedures with a focus on gesture and task-specific executional and procedural errors. We used the proposed rubric to evaluate dry-lab demonstrations of Suturing and Needle Passing tasks. Our analysis identified the most common error modes and their correlations with skill levels and demonstration times as well as important error-specific kinematic parameters that distinguish erroneous gestures. This study is a step towards developing methods for automated error detection during procedures and providing real-time context-dependent feedback for performance improvement.

Acknowledgements This work was partially supported by an award from the Engineering-in-Medicine center at the University of Virginia and the National Science Foundation NRT program (Grant No. 1829004). We thank Samin Yasar for his contributions to the initial drafts of the article, and Asha Maran, Devin Gardner, Rohan Chandra, and Ian Reyes for their help with labeling the data.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies involving human participants performed by any of the authors.

References

1. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D.: A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Transactions on Biomedical Engineering **64**(9), 2025–2041 (2017)
2. Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z., Iyer, R.K.: Adverse events in robotic surgery: a retrospective study of 14 years of fda data. PloS one **11**(4), e0151470 (2016)
3. Bonrath, E.M., Dedy, N.J., Zevin, B., Grantcharov, T.P.: Defining technical errors in laparoscopic surgery: a systematic review. Surgical endoscopy **27**(8), 2678–2691 (2013)
4. Bonrath, E.M., Zevin, B., Dedy, N.J., Grantcharov, T.P.: Error rating tool to identify and analyse technical errors and events in laparoscopic surgery. The British Journal of Surgery **100**(8), 1080–1088 (2013). DOI 10.1002/bjs.9168
5. Chen, J., Cheng, N., Cacciamani, G., Oh, P.J., Lin-Brande, M., Remulla, D., Gill, I., Hung, A.: Objective assessment of robotic surgical technical skill: A systematic review. The Journal of Urology **201**, 461–469 (2019)
6. Chowriappa, A.J., Shi, Y., Raza, S.J., Ahmed, K., Stegemann, A., Wilding, G., Kaouk, J., Peabody, J.O., Menon, M., Hassett, J.M., et al.: Development and validation of a composite scoring system for robot-assisted surgical training—the robotic skills assessment score. journal of surgical research **185**(2), 561–569 (2013)
7. Fard, M.J., Ameri, S., Darin Ellis, R., Chinnam, R.B., Pandya, A.K., Klein, M.D.: Automated robot-assisted surgical skill evaluation: Predictive analytics approach. The International Journal of Medical Robotics and Computer Assisted Surgery **14**(1), e1850 (2018)
8. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. International journal of computer assisted radiology and surgery **14**(9), 1611–1617 (2019)
9. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2CAI, vol. 3, p. 3 (2014)
10. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE transactions on pattern analysis and machine intelligence **28**(9), 1450–1464 (2006)
11. Hung, A.J., Chen, J., Jarc, A., Hatcher, D., Djaladat, H., Gill, I.S.: Development and validation of objective performance metrics for robot-assisted radical prostatectomy: a pilot study. The Journal of urology **199**(1), 296–304 (2018)
12. Joice, P., Hanna, G., Cuschieri, A.: Errors enacted during endoscopic surgery - a human reliability analysis. Applied ergonomics **29**(6), 409–414 (1998). DOI https://doi.org/10.1016/S0003-6870(98)00016-7. URL http://www.sciencedirect.com/science/article/pii/S0003687098000167

13. Kaouk, J.H., Khalifeh, A., Hillyer, S., Haber, G.P., Stein, R.J., Autorino, R.: Robot-assisted laparoscopic partial nephrectomy: step-by-step contemporary technique and surgical outcomes at a single high-volume institution. European urology **62**(3), 553–561 (2012)
14. Kazanzides, P., Chen, Z., Deguet, A., Fischer, G.S., Taylor, R.H., DiMaio, S.P.: An open-source research kit for the da vinci® surgical system. In: 2014 IEEE international conference on robotics and automation (ICRA), pp. 6434–6439. IEEE (2014)
15. Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. British journal of surgery **84**(2), 273–278 (1997)
16. Moorthy, K., Munz, Y., Dosis, A., Bello, F., Chang, A., Darzi, A.: Bimodal assessment of laparoscopic suturing skills. Surgical Endoscopy And Other Interventional Techniques **18**(11), 1608–1612 (2004)
17. Neumuth, D., Loebe, F., Herre, H., Neumuth, T.: Modeling surgical processes: A four-level translational approach. Artificial intelligence in medicine **51**(3), 147–161 (2011)
18. Poursartip, B., LeBel, M.E., Patel, R.V., Naish, M.D., Trejos, A.L.: Analysis of energy-based metrics for laparoscopic skills assessment. IEEE Transactions on Biomedical Engineering **65**(7), 1532–1542 (2018)
19. Qin, Y., Pedram, S.A., Feyzabadi, S., Allan, M., McLeod, A.J., Burdick, J.W., Azizian, M.: Temporal Segmentation of Surgical Sub-tasks through Deep Learning with Multiple Data Sources, p. 371–377. IEEE (2020). DOI 10.1109/ICRA40945.2020.9196560
20. Reiley, C.E., Hager, G.D.: Decomposition of robotic surgical tasks: an analysis of subtasks and their correlation to skill. In: M2CAI workshop. MICCAI, London (2009)
21. Rosen, J., Brown, J.D., Chang, L., Sinanan, M.N., Hannaford, B.: Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model. IEEE Transactions on Biomedical engineering **53**(3), 399–413 (2006)
22. Sánchez, R., Rodríguez, O., Rosciano, J., Vegas, L., Bond, V., Rojas, A., Sanchez-Ismayel, A.: Robotic surgery training: construct validity of global evaluative assessment of robotic skills (gears). Journal of robotic surgery **10**(3), 227–231 (2016)
23. Sharon, Y., Jarc, A.M., Lendvay, T.S., Nisky, I.: Rate of orientation change as a new metric for robot-assisted and open surgical skill evaluation. IEEE Transactions on Medical Robotics and Bionics (2021)
24. Shokoohi-Yekta, M., Wang, J., Keogh, E.: On the non-trivial generalization of dynamic time warping to the multi-dimensional case. In: Proceedings of the 2015 SIAM international conference on data mining, pp. 289–297. SIAM (2015)
25. Sukumar, S., Rogers, C.G.: Robotic partial nephrectomy: surgical technique. BJU international **108**(6b), 942–947 (2011)
26. Tao, L., Elhamifar, E., Khudanpur, S., Hager, G.D., Vidal, R.: Sparse hidden markov models for surgical gesture classification and skill evaluation, *Lecture notes in computer science*, vol. 7330, p. 167–177. Springer Berlin Heidelberg (2012). DOI 10.1007/978-3-642-30618-1_17
27. Tarr, M.E., Rivard, C., Petzel, A.E., Summers, S., Mueller, E.R., Rickey, L.M., Denman, M.A., Harders, R., Durazo-Arvizu, R., Kenton, K.: Robotic objective structured assessment of technical skills: a randomized multicenter dry laboratory training pilot study. Female pelvic medicine & reconstructive surgery **20**(4), 228–236 (2014). DOI 10.1097/SPV.0000000000000067
28. Taylor, R.H., Menciassi, A., Fichtinger, G., Fiorini, P., Dario, P.: Medical robotics and computer-integrated surgery. In: Springer handbook of robotics, pp. 1657–1684. Springer (2016)
29. Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.: Data-derived models for segmentation with application to surgical assessment and training. Medical Image Computing and Computer-Assisted Intervention **12**(Pt 1), 426–434 (2009). DOI 10.1007/978-3-642-04268-3_53
30. Vassiliou, M.C., Feldman, L.S., Andrew, C.G., Bergman, S., Leffondré, K., Stanbridge, D., Fried, G.M.: A global assessment tool for evaluation of intraoperative laparoscopic skills. The American journal of surgery **190**(1), 107–113 (2005)
31. Yasar, M., Alemzadeh, H.: Real-time context-aware detection of unsafe events in robot-assisted surgery. In: 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), pp. 385–397 (2020). DOI 10.1109/DSN48063.2020.00054