COMPASS: A Formal Framework and Aggregate Dataset for Generalized Surgical Procedure Modeling

Kay Hutchinson[1*], Ian Reyes[2,3], Zongyu Li[1] and Homa Alemzadeh[1,2]

[1*]Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, 22903, VA, USA.
[2]Department of Computer Science, University of Virginia, Charlottesville, 22903, VA, USA.
[3]IBM, RTP, 27709, NC, USA.

*Corresponding author(s). E-mail(s): kch4fk@virginia.edu;
Contributing authors: ir6mp@virginia.edu; zl7qw@virginia.edu;
ha4d@virginia.edu;

Abstract

Purpose: We propose a formal framework for the modeling and segmentation of minimally-invasive surgical tasks using a unified set of motion primitives (MPs) to enable more objective labeling and the aggregation of different datasets. **Methods:** We model dry-lab surgical tasks as finite state machines, representing how the execution of MPs as the basic surgical actions results in the change of surgical context, which characterizes the physical interactions among tools and objects in the surgical environment. We develop methods for labeling surgical context based on video data and for automatic translation of context to MP labels. We then use our framework to create the COntext and Motion Primitive Aggregate Surgical Set (COMPASS), including six dry-lab surgical tasks from three publicly-available datasets (JIGSAWS, DESK, and ROSMA), with kinematic and video data and context and MP labels. **Results:** Our context labeling method achieves near-perfect agreement between consensus labels from crowd-sourcing and expert surgeons. Segmentation of tasks to MPs results in the creation of the COMPASS dataset that nearly triples the

amount of data for modeling and analysis and enables the generation of separate transcripts for the left and right tools. **Conclusion:** The proposed framework results in high quality labeling of surgical data based on context and fine-grained MPs. Modeling surgical tasks with MPs enables the aggregation of different datasets and the separate analysis of left and right hands for bimanual coordination assessment. Our formal framework and aggregate dataset can support the development of explainable and multi-granularity models for improved surgical process analysis, skill assessment, error detection, and autonomy.

Keywords: minimally invasive surgery, robotic surgery, surgical context, surgical gesture recognition, surgical process modeling

1 Introduction

Surgical process models (SPMs) [17, 25] decompose surgical operations into smaller units of activity such as steps, tasks, and gestures as shown in Figure 1. While modeling procedures with phases and steps enables standardization and supports teaching [24], finer-grained modeling is needed for automated assistance, skill assessment, and autonomy in robot-assisted surgery. Gestures, defined as intentional activities with meaningful outcomes [6], are an important analytical unit for skill evaluation [31, 37] and error detection [13, 14, 19]. Finer-grained activities called actions or motions [11, 26, 30] have also been proposed to improve the understanding of tool-tissue interactions and their relationship to different granularity levels. Recognition of motions can increase the explainability of gesture-level analyses, improve error detection by identifying the exact erroneous parts of gestures [13], and enable autonomy or error recovery through the execution of motion primitives [5, 8].

The decomposition of tasks into gestures has been done with models such as graphs [1, 34], statecharts [5], hybrid automata [4], and behavior trees [10] for cooperative, autonomous, and supervisory systems in robotic surgery [22]. However, prior work has not explicitly modeled or formalized surgical context, which is characterized by the status and interactions among surgical tools and tissues/objects, and their relationship to motions, gestures, and surgical workflow.

Additionally, while research has focused on standardized surgical ontologies [7], labeling methods [21], and action triplets [26], there is still a need for a common surgical language and larger multimodal datasets to support comparative analysis of activity recognition and error detection models [31]. The **Online Resource** provides a detailed summary of related work on gesture and action definitions and datasets. As shown in Tables 1 and 2 in the Online Resource, the definitions, numbers, and types of activity labels vary in the existing datasets. The most commonly used dataset is JIGSAWS [6], which contains kinematic data, videos, gesture labels, and surgical skill scores for three dry-lab surgical tasks. However, only two of its tasks are labeled with

similar sets of gestures. Other recently developed datasets such as DESK [20] and V-RASTED [23] have defined their own sets of gestures while ROSMA [29] is not labeled. Datasets such as these with both kinematic and video data from a surgical robot/simulator are small and contain only a handful of trials of a few simulated or dry-lab training tasks performed by a limited number of subjects. This scarcity of data hinders training and generalization of machine learning models. Also, most datasets on finer-grained actions have focused on only video data from real surgery. While video data is required for labeling, inclusion of kinematic data is valuable for safety analysis [13, 19], improved recognition accuracy through multimodal analysis [28, 33], or when video data is not available or is noisy due to smoke or occlusions [36].

Furthermore, annotating surgical workflow is costly and requires guidance from expert surgeons [15], and the resulting labels may contain errors and inconsistencies such as those identified in JIGSAWS [13, 32]. Label quality and inter-rater agreement has not been examined when creating the existing datasets. Also, the labels in these datasets do not differentiate between activities performed by the left and right hands, which is important for detailed skill assessment and analysis of bimanual coordination [2].

We address these challenges by making the following contributions:

- Proposing a novel formal framework for modeling surgical dry-lab tasks with finite state machines using a standardized set of motion primitives whose execution leads to changes in important state variables that make up the surgical context. Context characterizes the physical interactions among surgical tools and objects, and motion primitives represent basic surgical actions across different surgical tasks and procedures.
- Developing a method for labeling surgical context based on video data of dry-lab tasks that achieves near-perfect agreement between crowd-sourced labels and expert surgeon labels, higher agreement among annotators than using existing gesture definitions, and such that the context labels can be automatically translated into motion primitive labels.
- Applying our framework and labeling method to create an aggregate dataset, called COMPASS (COntext and Motion Primitive Aggregate Surgical Set), which includes kinematic and video data as well as context and motion primitive labels for a total of six dry-lab tasks from the JIGSAWS [6], DESK [20], and ROSMA [29] datasets.

The tools for labeling surgical context based on video data, automated translation of context to motion primitive labels, and the aggregated dataset with context and motion primitive labels are publicly available at https://github.com/UVA-DSA/COMPASS.

2 Methods

Our framework models surgical procedures as a language with a grammar dictating how motion primitives (MPs) are combined to perform gestures and
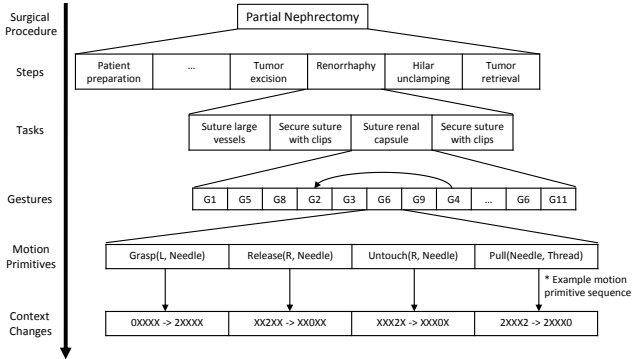
Fig. 1 Surgical Hierarchy. Adapted from [13]

tasks, thus bridging the gap between semantic-less motions [25], and intent-based gestures [6]. Our framework formally defines MPs, their relation to surgical context and task progress, and their combination to perform dry-lab tasks. We develop methods to objectively label surgical context and translate context labels to MPs, and apply them to three publicly available datasets to create the aggregated COMPASS dataset. We consider Suturing (S), Needle Passing (NP), and Knot Tying (KT) from JIGSAWS [6]; Peg Transfer (PT) on the da Vinci surgical robot from DESK [20]; and Pea on a Peg (PoaP) and Post and Sleeve (PaS) from ROSMA [29] (see Figure 5).

2.1 Modeling Framework

2.1.1 Surgical Hierarchy

Surgical procedures follow the hierarchy of levels defined in [25] which provides context for actions during the procedure, 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) as shown in grammar graphs [1]. These **gestures** are made of basic **motion primitives** such as moving an instrument or closing the graspers, which effect changes in important states that comprise the overall surgical **context**.

2.1.2 Surgical Context

Surgical environments (in dry-lab or real surgical procedures) can be modeled with state variables that characterize the status and interactions among surgical instruments (e.g., graspers, scissors, electro-cautery) and objects (e.g., needles, threads, blocks) or anatomical structures (e.g., organs, tissues) over time. Changes in surgical context happen as the result of performing a set of basic MPs with the robot (either controlled by the surgeon or autonomously). We focus on dry-lab training tasks and manually construct
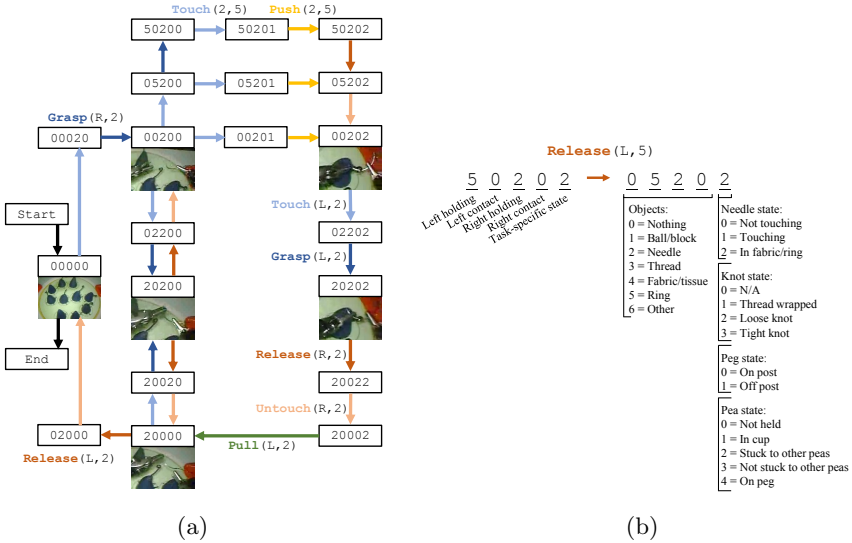
Fig. 2 a Finite state machine model representing the ideal performance of a Needle Passing trial with context and MPs. In this task, the surgeon threads the needle through four of the rings. b An example of the "Release" MP in Needle Passing resulting in the change of state variables

finite state machines (FSMs) to model each task after reviewing the videos to understand the general activities in the task. In these models, states represent context and transitions represent MPs. Figure 2a shows an example of the FSM for NP. This representation of surgical tasks incorporates surgical context into procedure modeling which is missing from previously proposed models such as grammar graphs and Hidden Markov models [1] where hidden states obscure lower-level actions.

Surgical context is defined using two sets of variables that can be observed or measured using kinematic and/or video data from a surgical scene: (i) general state variables relating to the contact and hold interactions between the tools and objects in the environment, and (ii) task-specific state variables describing the states of objects critical to the current task. We also define independent state variables for the left and right tools to enable the generation of separate label sets for each side. There are four general state variables and one task-specific state variable describing progress in the task as shown in Figure 2b. S involves throwing four sutures and NP involves passing a needle through four rings, so the needle, if held, can be "not touching", "touching", or "in" the fabric or ring. KT involves tying two knots, so the thread can be "wrapped" around the opposite grasper, in a "loose" knot, or a "tight" knot. PT and PaS involve picking up a block and placing it on another post, so the block can be "on" or "off" the peg. PoaP involves picking up a pea and placing it on a post, so the pea, if held, can be "in the cup", "stuck to other peas",

Table 1 General motion primitives for changes in context: 'L' and 'R' represent the left and right graspers as tools, 'a' is a generic object as listed in Fig 2b, and 'X' can be any value.

Motion Primitive	Context Change
Touch(L, a)	X0XX → XaXX
Touch(R, a)	XXX0 → XXXa
Grasp(L, a)	0aXX → aXXX
Grasp(R, a)	XX0a → XXaX
Release(L, a)	aXXX → 0aXX
Release(R, a)	XXaX → XX0a
Untouch(L, a)	XaXX → X0XX
Untouch(R, a)	XXXa → XXX0

"not stuck to other peas", or "on the peg". In Figure 2b, the state 50202 indicates that the left grasper is holding a ring, the right grasper is holding the needle, and the needle is in the ring.

2.1.3 Motion Primitives

We define a unified set of six modular and programmable surgical MPs to model basic surgical actions that lead to changes in the physical context. As shown in Equation 1, each MP is characterized by its type (e.g., Grasp), the specific tool which is used (e.g., left grasper), the object with which the tool interacts (e.g., block), and a set of constraints that define the functional (e.g., differential equations characterizing typical trajectory [8]) and safety requirements (e.g., virtual fixtures and no-go zones [3, 35]) for the execution of the MPs:

$$MP(tool, object, constraints) \tag{1}$$

Tools and objects are considered classes as in object-oriented programming with attributes such as the specific type of tool and current position. MPs can be further decomposed into the fundamental transformations of move/translate, rotate, and open/close graspers which characterize low level kinematic commands. These can be used for programming and execution of motions on a robot for semi-autonomous surgery [5], which is the subject of future work.

Segmenting tasks into MPs allows the separation of actions performed by the left and right hands and the generation of separate sets of labels. This can support more detailed skill assessment, analysis of bimanual coordination, and surgical automation [31].

Table 1 shows the set of universal MPs and corresponding changes to surgical context applicable to all tasks which enables the generalizability of this framework and allows activity recognition models to leverage these similarities across tasks. Table 2 shows the sets of MPs and corresponding changes to surgical context applied to specific dry-lab tasks. We focus on dry-lab tasks where the tools are graspers, but do not model or analyze the MP-specific functional and safety constraints here.

Table 2 Task-specific motion primitives for changes in context: 'L' and 'R' represent the left and right graspers as tools, objects are encoded as in Fig 2b, 'b' is a value greater than 0, and 'X' can be any value.

Motion Primitive	Context Change
Suturing/Needle Passing	
Touch(2, 4/5)	2XXX0 → 2XXX1
Touch(2, 4/5)	XX2X0 → XX2X1
Push(2, 4/5)	2XXX1 → 2XXX2
Push(2, 4/5)	XX2X1 → XX2X2
Pull(2, 3)	2XXX2 → 2XXX0
Pull(2, 3)	XX2X2 → XX2X0
Knot Tying	
Pull(L, 3)	3XXX0 ↔ 3XXX1
Pull(R, 3)	XX3X0 ↔ XX3X1
Pull(L, 3) Pull(R, 3)	3X3X1 → 3X3X2
Pull(L, 3) Pull(R, 3)	3X3X2 → 3X3X3
Peg Transfer and Post and Sleeve	
Untouch(1, Post)	XXXX0 → XXXX1
Touch(1, Post)	XXXX1 → XXXX0
Pea on a Peg	
Grasp(L, 1)	0XXX0 → 1XXX1
Grasp(R, 1)	XX0X0 → XX1X1
Pull(L, 1)	1XXX1 → 1XXX2
Pull(R, 1)	XX1X1 → XX1X2
Pull(L, 1)	1XXX1 → 1XXX3
Pull(R, 1)	XX1X1 → XX1X3
Touch(1, 1)	XXXX3 → XXXX2
Untouch(1, 1)	XXXX2 → XXXX3
Touch(1, Peg)	XXXX3 → XXXX4
Untouch(1, Peg)	XXXX4 → XXXX3
Release(L, 1)	1XXXb → 0XXX0
Release(R, 1)	XX1Xb → XX0X0
Push(L, 1)	1XXX2 → 1XXX1
Push(R, 1)	XX1X2 → XX1X1

The definition of MPs based on the changes in the surgical context could enable the translation of context and MPs to existing gesture labels and facilitate aggregation of different datasets labeled with different gesture definitions. However, translation from context and MP labels to existing gesture definitions is complicated. This is because executional and procedural errors in gestures as defined by [13] can effect the MP sequences for each gesture. Additional modeling is needed to develop and evaluate the MP to gesture translation which is beyond the scope of this paper.
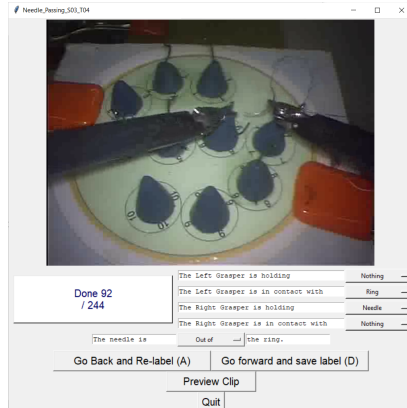
Fig. 3 App for Context Labeling based on Video Data

2.2 Labeling of Context and Motion Primitives

2.2.1 Context Labeling

Gesture recognition models using supervised learning require a large number of annotated video sequences [27]. However, manual labeling of gestures is subjective and can lead to labeling errors [32]. Thus, we developed a tool for manually annotating surgical context (states of objects and instruments) based on video data.

Labeling video data for surgical context provides a more objective way of recognizing gestures and can lead to higher agreement among annotators. As noted in [15], labels for surgical workflow require guidance from surgeons while annotations for surgical instruments do not. Since context labels document objects held by or in contact with the left and right tools, they rely less on surgical knowledge than gestures which require anticipating the next activities in a task to mark when a gesture has ended. Figure 3 shows a snapshot of the tool for manually labeling context based on video data. Annotators indicate the value of different state variables for frames in the video data and may copy over values until a change in context is observed. This differs from other labeling methods where annotators mark the start and end of each segment and assign it a label.

2.2.2 Context to Motion Primitive Translation

Context to MP translation allows us to leverage high quality context labels for creating surgical workflow annotations and aggregating different surgical datasets. Context labels are translated automatically into MPs using the FSMs for each task.

For each change of context in a sequence of context labels, the sequence of MP labels are generated by identifying specific state variable changes that correspond to the transitions in Tables 1 and 2 and are visualized in the FSMs. If multiple states change between labeled frames, then Grasp and Release MPs

Frame	Context		Start	Stop	Motion Primitive
⋮			⋮		
1380	05202		1380	1399	Untouch(L, Ring)
1400	00202	➜	1400	1499	Grasp(L, Needle)
1500	20202		1500	1509	Release(R, Needle)
1510	20002		1510	1559	Pull(L, Needle)
1560	20000		1560	1769	Grasp(R, Needle)
1770	20200		⋮		
⋮					

Fig. 4 Example sequence of context translated into motion primitives

have a higher priority than Touch and Untouch MPs (if they are performed on the same object by the same tool). Otherwise, all MPs are listed in the MP transcript so that separate MP transcripts for the left and right sides could be generated. Figure 4 shows an example of a sequence of context translated into MPs. This rule-based translation method assumes that changes in context can be completely described by the definitions in Tables 1 and 2. Alternatively, data-driven and learning from demonstration approaches can be used for more realistic and personalized modeling of the tasks and label translations.

2.3 COMPASS Dataset

We create the COMPASS dataset by aggregating data from 39 trials of S, 28 trials of NP, and 36 trials of KT performed by eight subjects from the JIGSAWS dataset; 47 trials of PT performed by eight subjects from the DESK dataset; and 65 trials of PaS, and 71 trials of PoaP performed by 12 subjects from the ROSMA dataset (see Figure 5). Thus, COMPASS contains a total of 286 trials by 28 different subjects which is about three times the number of trials and subjects in JIGSAWS.

Tables 3, 4, and 5 list the numbers of MPs and gestures of each type in each task and dataset. By using standardized definitions, COMPASS has fewer classes and more examples than other datasets.

Videos are at 30 fps for the stereoscopic JIGSAWS and DESK tasks and 15 fps for the single camera ROSMA tasks. The kinematic data have been down-sampled to 30 Hz and contain position, velocity, orientation (in quaternions), and gripper angle variables. Since linear velocity data was not available for all tasks, it was derived from the position data using a rolling average over five samples. ROSMA did not contain gripper angle, so a separate round of manually labeling video data was performed to approximate the gripper angle as open or closed.

To ensure reliable and high quality annotations, three full sets of context labels were obtained using our context labeling tool for all the trials. Two of the authors, with extensive experience with the datasets and the dry-lab robotic surgery tasks, each produced a full set of labels for all trials. The third set of labels was crowd-sourced to 22 engineering students, who had no previous experience but were given a training module on the definitions of context, MPs, and their relationship, and how to use the labeling tool. The "Consensus" labels were then created using majority voting for each state variable.

COMPASS includes these consensus context labels at 3 Hz and the automatically generated MP labels interpolated to 30 Hz for both arms of the robot so that every kinematic sample has an MP label. The original gesture label files from JIGSAWS and DESK are renamed and included to promote comparisons between data and label sets. The dataset is organized into different tasks with directories for the kinematic and video data, and each type of label. The subject and trial numbers from the original datasets are retained so that the LOSO and LOUO setups from [1] can be extended to COMPASS.

(a) Suturing (S)

(d) Pea on a Peg (PoaP)

(b) Needle Passing (NP)

(e) Post and Sleeve (PaS)
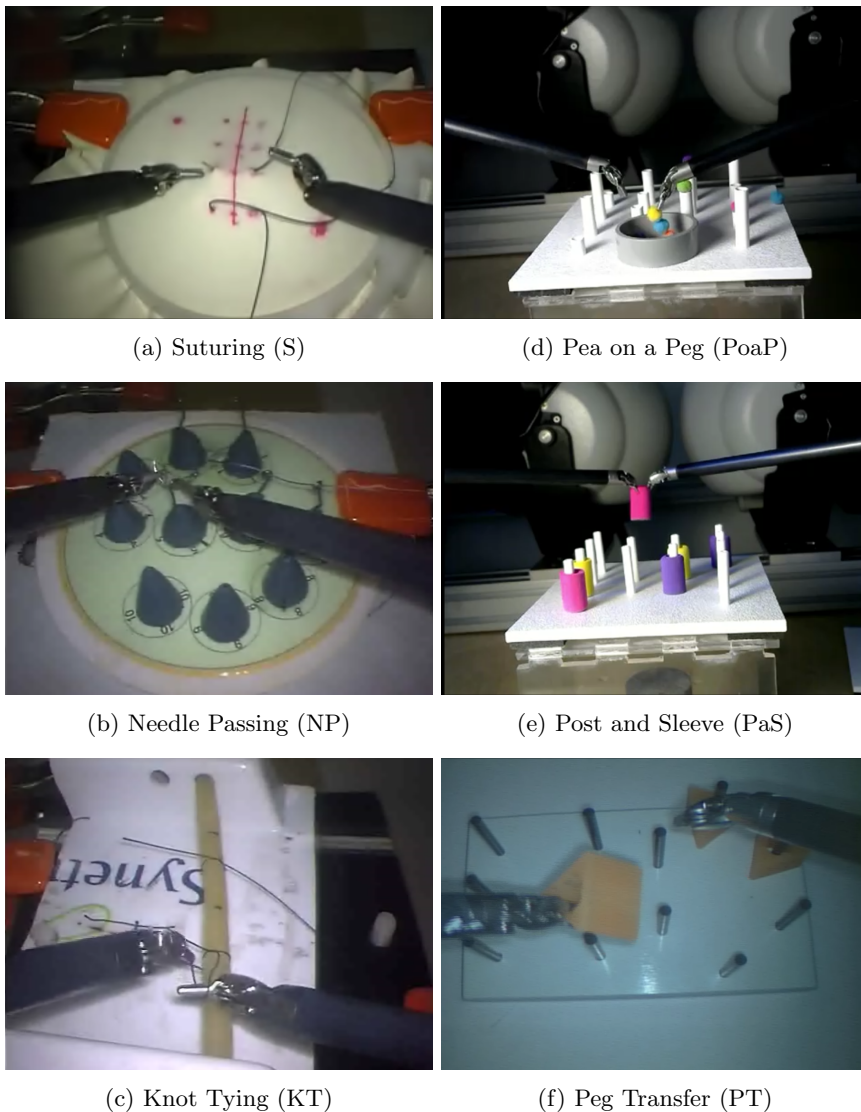

(c) Knot Tying (KT)

(f) Peg Transfer (PT)

Fig. 5 Tasks included in the COMPASS dataset: S, NP, and KT from JIGSAWS [6]; PoaP and PaS from ROSMA [29]; and PT from DESK [20]

Table 3 Number of Motion Primitives (MPs) in each task and the COMPASS dataset: Suturing (S), Needle Passing (NP), and Knot Tying (KT) from JIG-SAWS [6]; Peg Transfer (PT) from DESK [20]; and Pea on a Peg (PoaP) and Post and Sleeve (PaS) from ROSMA [29].

	JIGSAWS			DESK	ROSMA		
MP	S	NP	KT	PT	PoaP	PaS	COMPASS
Grasp	471	373	283	323	577	824	2851
Release	441	365	247	313	556	776	2698
Touch	518	330	135	539	1782	1598	4902
Untouch	314	206	111	364	1261	1131	3387
Pull	194	114	235	0	525	0	1068
Push	179	119	0	0	2	0	300

Table 4 Number of gestures in each JIGSAWS task and dataset.

Gesture	Suturing	Needle Passing	Knot Tying	JIGSAWS
G1	29	30	19	78
G2	166	117	0	283
G3	164	111	0	275
G4	119	83	0	202
G5	37	31	0	68
G6	163	112	0	275
G8	48	28	0	76
G9	24	1	0	25
G10	4	1	0	5
G11	39	25	36	100
G12	0	0	70	70
G13	0	0	75	75
G14	0	0	98	98
G15	0	0	73	73

3 Evaluation

To evaluate our framework, we obtain MP graphs (similar to grammar graphs but with nodes representing MPs) from two expert robotic surgeons describing the execution of S, NP, and KT and compare our proposed models to expert knowledge. To evaluate our context labeling and context to MP translation methods, we obtain two more sets of labels, in addition to the "Consensus" set. Two trios of independent annotators labeled subsets of trials in the JIGSAWS and DESK tasks, respectively, for context, MPs, and gestures to assess and compare the different labeling methods and the context to MP translation. We refer to these labels as the "Multi-level" set and their gesture labels could be compared to the gesture labels from JIGSAWS and DESK. The surgeons also labeled a set of six trials (one from each task to capture task-diversity, and overlapping the trials in the "Multi-level" set) for context, referred to as the "Surgeon" set, against which we evaluate label quality.

Table 5 Number of gestures in DESK for Peg Transfer performed on the da Vinci surgical robot.

Gesture	Peg Transfer
S1	146
S2	147
S3	137
S4	146
S5	146
S6	135
S7	135

3.1 Task Modeling

We evaluate our framework by comparing the MP graphs for S, NP, and KT to MP graphs defined by expert surgeons using graph edit score. Surgeons may not be available to verify future models, so it is important to check that those for surgically-relevant tasks represent expert knowledge. In order to compare the FSMs with the MP graphs defined by expert surgeons, the sequence of MP transitions from the FSMs were converted to MP graphs. Touch and Untouch MPs that immediately preceded or followed Grasp and Release MPs, respectively, were removed since the surgeons assumed that combination when creating their MP graphs. Figure 2a corresponds to Figure 6b with additional tasks in the Online Resource.

Graph Edit Score (GES) is the normalized graph edit distance (GED) calculated using Equation 2 by dividing the minimum cost of transforming A to B by the maximum GED (cost of deleting all nodes and edges in A and inserting all nodes and edges in B where C represents an empty graph). GED is implemented using networkx [9] with the Start nodes as the root and a timeout of 18 hours.

$$\text{GES} = (1 - \frac{GED(A,B)}{(GED(A,C) + GED(B,C))}) \times 100 \qquad (2)$$

GES was lowest for S in Table 6 because it was the most complex task and the surgeons had additional MPs to represent passing the needle from left to right while the proposed model represented that as the inverse of passing the needle from right to left. Although physically possible, several transitions in our proposed graph for S were not in the surgeons' graph since they may not represent an efficient execution of the task. Comparatively, KT was a simpler task, and overall the proposed models are good representations of expert knowledge.

3.2 Context Labeling

We assess context label quality by measuring the agreement among annotators in the "Consensus" set and the agreement between the "Surgeon" and "Consensus" sets of context labels using Krippendorff's Alpha [27]. Then, we
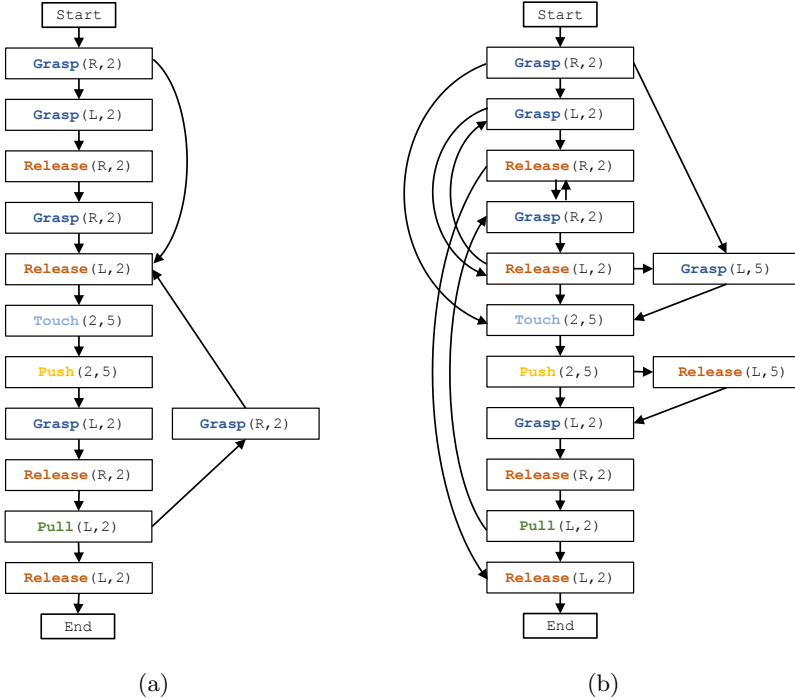
(a)   (b)

Fig. 6 Surgeon-defined (**a**) and proposed (**b**) MP graph models for the Needle Passing task

Table 6 Graph edit scores between the proposed and surgeon-defined MP graphs.

Task	Graph Edit Score
Suturing	76.4
Needle Passing	83.6
Knot Tying	93.3

compare context, MP, and gesture level labeling methods using labels in the "Multi-level" set.

Krippendorff's Alpha (α) is a commonly used statistical measure of inter-rater reliability. It indicates how much the data from two or more methods can be trusted to represent the real phenomenon [16].

Krippendorff's Alpha is calculated using Equation 3 by considering the probability D_e that two labelers produced the same annotation due to chance rather than agreement on the data to label, and the observed disagreement D_o between each labeler's annotations:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{3}$$

α takes a value between -1 and 1, with $\alpha > 0.8$ indicating near-perfect agreement, a value between 0.6 and 0.8 indicating substantial agreement, and smaller values indicating less agreement (Table 7). $\alpha = 0$ indicates no agreement other than by chance and negative values reflect more pronounced disagreement.

Table 7 Interpretation of Krippendorff's Alpha (α) from [12].

Range	Interpretation
$\alpha > 0.8$	Near-perfect
$0.6 < \alpha \leq 0.8$	Substantial
$0.4 < \alpha \leq 0.6$	Moderate
$0.2 < \alpha \leq 0.4$	Fair
$\alpha \leq 0.2$	Slight

Each of the labelers annotated a sequence of states encoded as numbers representing categorical data, so the nominal distance or difference function is best suited to quantify agreement between labelers. The nominal distance or difference function in Equation 4 is used to calculate D_e and D_o [12] in Equation 5, where n_l is the number of labelers and n_u is the total number of frames which two or more labelers annotated.

$$d_{nominal}(\text{label}_1 \ \text{label}_2) = \begin{cases} 0 & \text{if label}_1 = \text{label}_2 \\ 1 & \text{if label}_1 \neq \text{label}_2 \end{cases} \tag{4}$$

$$D_o = \frac{1}{2n_u n_l(n_l - 1)} \sum_{l_1,l_2 \in \text{all labels}} d_{nom}(l_1, l_2)$$

$$D_e = \frac{1}{2n_u n_l(n_u n_l - 1)} \sum_{l_1,l_2 \in \text{all labels}} d_{nom}(l_1, l_2) \tag{5}$$

3.2.1 Consensus Context Labels

Table 8 shows near-perfect agreement among annotators in four tasks and substantial agreement in labeling two tasks. The average for all tasks was 0.84, weighted for the number of frames for each task, indicating near-perfect agreement in context labeling overall. Long segments of near-perfect agreement are punctuated by disagreements at the transitions between context. However, disagreement is limited to a few context states instead of the gesture label for a specific frame which results in much greater agreement between annotators when labeling for context than gestures. This shows that our method for labeling context results in a high quality set of fine-grained labels.

Between the Consensus and Surgeon context labels, all tasks had an α of at least 0.8. The average α for all tasks (weighted for the number of frames for each task) was 0.92, indicating near-perfect agreement between crowd-sourced context labels and those given by surgeons.

Table 8 Krippendorff's Alpha among annotators and between Consensus and Surgeon context labels.

Task	Among annotators	Between Consensus and Surgeon
S	0.69	0.86
NP	0.85	0.90
KT	0.79	0.94
PT	0.90	0.94
PoaP	0.83	0.93
PaS	0.89	0.97

3.2.2 Multi-level Labels

Table 9 shows the least agreement among annotators using descriptive gesture definitions. Directly labeling MPs is also difficult, likely due to their short durations. But context annotations have the greatest agreement because labels are based on well-defined interactions among surgical tools and objects observed in video data.

There is also much higher agreement when labeling for context than for gestures and the existing JIGSAWS labels are difficult to reproduce (smallest α in last column). This might be because JIGSAWS labels were generated more subjectively by only one annotator by watching the videos and consulting with a surgeon [6]. We again see that crowd-sourcing context labels results in high quality annotations comparable to those from surgeons and are thus representative of expert knowledge.

3.3 Context to Motion Primitive Translation

To assess the performance of the context to MP translation, we translate the context labels in the "Multi-level" annotations set and compare the resulting translated MP transcripts to the ground truth MP labels for each annotator using accuracy and edit score.

Accuracy: Given the sequence of predicted labels, P, and the sequence of ground truth labels, G, the accuracy is the ratio of correctly classified samples divided by the total number of samples in a trial.

Edit Score: We report the edit score as the normalized Levenshtein edit distance, $edit(G, P)$, by calculating the number of insertions, deletions, and

Table 9 Krippendorff's Alpha among annotators for Multi-level labels.

Task	Multi-Level			Multi-Level vs. Surgeon Context	Multi-Level vs. JIGSAWS/DESK Gestures
	Context	MPs	Gestures		
S	0.72	0.33	0.24	0.86	0.34
NP	0.91	0.41	0.08	0.90	0.04
KT	0.89	0.26	0.20	0.89	0.06
PT	0.89	0.72	0.72	0.89	0.62

replacements needed to transform P to match G [18]. The edit score is normalized by the maximum length of P and G, as shown in Equation 6, where 100 is the best and 0 is the worst.

$$\text{Edit Score} = (1 - \frac{edit(G, P)}{max(len(G), len(P))}) \times 100 \tag{6}$$

3.3.1 Quality of Multi-level Labels

The context labels and the ground truth MP labels show variability with task and annotator skill, both of which can affect the resulting translated MP labels and their evaluation. We first assessed the quality of each annotator based on their agreement with the "Surgeon" context labels and JIGSAWS/DESK gesture labels with respect to the label for each frame and the overall label sequence. Table 10 shows α, and Table 11 shows accuracies and edit scores for each annotator when labeling for context (compared to Surgeons) and gestures (compared to JIGSAWS/DESK). For JIGSAWS, annotator 3 was the most reliable annotator overall with annotator 2 almost as reliable for context labels. Less variation was seen among the annotators for DESK.

Table 10 Krippendorff's Alpha of Multi-level labels compared to Surgeon context and JIGSAWS/DESK gesture labels.

Task	Context			Gestures		
	Annotator 1	Annotator 2	Annotator 3	Annotator 1	Annotator 2	Annotator 3
S	0.87	0.89	0.91	0.06	0.05	0.76
NP	0.88	0.88	0.85	0.09	0.07	0.47
KT	0.72	0.85	0.87	0.08	0.39	0.48
PT	0.93	0.89	0.93	0.40	0.41	0.45

Table 11 Accuracy and edit score of Multi-level labels compared to Surgeon context and JIGSAWS/DESK gesture labels.

Task	Context						Gestures					
	Annotator 1		Annotator 2		Annotator 3		Annotator 1		Annotator 2		Annotator 3	
	Acc	Edit	Acc	Edit	Acc	Edit	Acc	Edit	Acc	Edit	Acc	Edit
S	67.6	69.2	72.3	73.0	77.3	78.0	25.5	49.1	26.1	42.0	85.6	88.7
NP	74.1	74.9	76.0	76.6	74.7	74.9	34.0	56.8	27.7	49.5	47.3	54.5
KT	32.5	32.5	60.7	60.7	62.2	62.2	24.0	51.5	52.3	61.0	62.5	64.1
PT	83.3	84.1	83.7	87.2	83.7	84.0	49.8	52.7	50.6	52.3	54.1	55.8

3.3.2 Translation Accuracy

Table 12 shows the context to MP translation accuracy was higher for annotators 1 and 3 for JIGSAWS, and annotators 2 and 3 for DESK. There is

inter-rater variability across tasks, where KT and PT generally had higher metrics while S had lower metrics likely due to task complexity. However, the ground truth MP labels used in this evaluation had very low agreement among annotators compared to context labels (Table 9) and assessing their reliability is beyond the scope of this paper. Future work will use multi-level labeling methods to better evaluate and improve this translation.

Table 12 Accuracy and edit score between ground truth and translated MPs for Multi-level and Surgeon labels.

	Annotator 1		Annotator 2		Annotator 3		Surgeon	
Task	Acc	Edit	Acc	Edit	Acc	Edit	Acc	Edit
S	27.4	27.9	19.0	26.9	23.7	30.1	28.6	30.8
NP	64.9	67.4	27.2	45.1	45.2	46.4	21.5	25.4
KT	50.5	53.8	31.9	56.1	53.6	56.8	48.6	50.3
PT	29.2	30.7	49.5	50.7	50.6	56.8		

4 Discussion and Conclusion

We present a framework for modeling dry-lab surgical tasks as finite state machines where MPs cause changes in surgical context, characterized by tool and object/tissue interactions. We apply our framework to three publicly available datasets to create an aggregate dataset of kinematic and video data along with context and MP labels. Our method for labeling context achieves substantial to near-perfect agreement among annotators and expert surgeons. Using MPs, we aggregate data from different datasets, tasks, and subjects and nearly triple the amount of data with consistent label definitions.

Future work includes extending the MP framework to tasks from real surgical procedures by defining task-specific state variables to augment the context labels and their associated MPs (e.g., "Cut" for scissors like in [26]). Our framework enables the generalized modeling and comparison of surgical activities between datasets and tasks. This supports the development of datasets and models for surgical automation [8] by providing examples of fine-grained motions, and multi-granularity models for improved fine-grained activity recognition [11].

This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at:

Declarations

Competing Interests

The authors declare that they have no conflict of interest.

Ethics Approval and Informed Consent

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

References

[1] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

[2] Jacob R Boehm, Nicholas P Fey, and Ann Majewicz Fey. Online recognition of bimanual coordination provides important context for movement data in bimanual teleoperated robots. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6248–6255. IEEE, 2021.

[3] Stuart A Bowyer, Brian L Davies, and Ferdinando Rodriguez y Baena. Active constraints/virtual fixtures: A survey. *IEEE Transactions on Robotics*, 30(1):138–157, 2013.

[4] Giacomo De Rossi, Marco Minelli, Serena Roin, Fabio Falezza, Alessio Sozzi, Federica Ferraguti, Francesco Setti, Marcello Bonfè, Cristian Secchi, and Riccardo Muradore. A first evaluation of a multi-modal learning system to control surgical assistant robots via action segmentation. *IEEE Transactions on Medical Robotics and Bionics*, 2021.

[5] Fabio Falezza, Nicola Piccinelli, Giacomo De Rossi, Andrea Roberti, Gernot Kronreif, Francesco Setti, Paolo Fiorini, and Riccardo Muradore. Modeling of surgical procedures using statecharts for semi-autonomous robotic surgery. *IEEE Transactions on Medical Robotics and Bionics*, 3(4):888–899, 2021.

[6] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın

Béjar, David D Yuh, Chi Chiung Grace Chen, René Vidal, Sanjeev Khudanpur, and Gregory D. Hager. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI*, volume 3, page 3, 2014.

[7] Bernard Gibaud, Germain Forestier, Carolin Feldmann, Giancarlo Ferrigno, Paulo Gonçalves, Tamás Haidegger, Chantal Julliard, Darko Katić, Hannes Kenngott, Lena Maier-Hein, Keno März, Elena de Momi, Dénes Ákos Nagy, Hirenkumar Nakawala, Juliane Neumann, Thomas Neumuth, Javier Rojas Balderrama, Stefanie Speidel, Martin Wagner, and Pierre Jannin. Toward a standard ontology of surgical process models. *International journal of computer assisted radiology and surgery*, 13(9):1397–1408, 2018.

[8] Michele Ginesi, Daniele Meli, Andrea Roberti, Nicola Sansonetto, and Paolo Fiorini. Autonomous task planning and situation awareness in robotic surgery. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3144–3150. IEEE, 2020.

[9] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[10] Danying Hu, Yuanzheng Gong, Blake Hannaford, and Eric J Seibel. Semi-autonomous simulated brain tumor ablation with ravenii surgical robot using behavior tree. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3868–3875. IEEE, 2015.

[11] Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, Pablo Arbeláez, Wolfgang Reiter, Manoru Mitsuishi, Kanako Harada, and Pierre Jannin. Micro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*, 212:106452, 2021.

[12] John Hughes. krippendorffsalpha: An R Package for Measuring Agreement Using Krippendorff's Alpha Coefficient. *The R Journal*, 13(1):413–425, 2021.

[13] Kay Hutchinson, Zongyu Li, Leigh A Cantrell, Noah S Schenkman, and Homa Alemzadeh. Analysis of executional and procedural errors in dry-lab robotic surgery experiments. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 18(3):e2375, 2022.

[14] Daniel A Inouye, Runzhuo Ma, Jessica H Nguyen, Jasper Laca, Rafal Kocielnik, Anima Anandkumar, and Andrew J Hung. Assessing the efficacy of dissection gestures in robotic surgery. *Journal of Robotic Surgery*,

pages 1–7, 2022.

[15] Daichi Kitaguchi, Nobuyoshi Takeshita, Hiro Hasegawa, and Masaaki Ito. Artificial intelligence-based computer vision in surgery: Recent advances and future perspectives. *Annals of Gastroenterological Surgery*, 6, 10 2021.

[16] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[17] Florent Lalys and Pierre Jannin. Surgical process modelling: a review. *International journal of computer assisted radiology and surgery*, 9(3):495–511, 2014.

[18] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pages 47–54. Springer, 2016.

[19] Zongyu Li, Kay Hutchinson, and Homa Alemzadeh. Runtime detection of executional errors in robot-assisted surgery. In *2022 International Conference on Robotics and Automation (ICRA)*, page 3850–3856. IEEE Press, 2022.

[20] Naveen Madapana, Md Masudur Rahman, Natalia Sanchez-Tamayo, Mythra V Balakuntala, Glebys Gonzalez, Jyothsna Padmakumar Bindu, LN Vishnunandan Venkatesh, Xingguang Zhang, Juan Barragan Noguera, Thomas Low, et al. Desk: A robotic activity dataset for dexterous surgical skills transfer to medical robots. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6928–6934. IEEE, 2019.

[21] Ozanan R Meireles, Guy Rosman, Maria S Altieri, Lawrence Carin, Gregory Hager, Amin Madani, Nicolas Padoy, Carla M Pugh, Patricia Sylla, Thomas M Ward, et al. Sages consensus recommendations on an annotation framework for surgical video. *Surgical endoscopy*, 35(9):4918–4929, 2021.

[22] Daniele Meli and Paolo Fiorini. Unsupervised identification of surgical robotic actions from small non-homogeneous datasets. *IEEE Robotics and Automation Letters*, 6(4):8205–8212, 2021.

[23] Giovanni Menegozzo, Diego Dall'Alba, Chiara Zandonà, and Paolo Fiorini. Surgical gesture recognition with time delay neural network based on kinematic data. In *2019 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2019.

[24] T Nazari, EJ Vlieger, MEW Dankbaar, JJG van Merriënboer, JF Lange, and T Wiggers. Creation of a universal language for surgical procedures using the step-by-step framework. *BJS open*, 2(3):151–157, 2018.

[25] Dayana Neumuth, Frank Loebe, Heinrich Herre, and Thomas Neumuth. Modeling surgical processes: A four-level translational approach. *Artificial intelligence in medicine*, 51(3):147–161, 2011.

[26] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[27] Sunghyun Park, Gelareh Mohammadi, Ron Artstein, and Louis-Philippe Morency. Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 29–34, 2012.

[28] Yidan Qin, Seyedshams Feyzabadi, Max Allan, Joel W Burdick, and Mahdi Azizian. davincinet: Joint prediction of motion and surgical state in robot-assisted surgery. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2921–2928. IEEE, 2020.

[29] Irene Rivas-Blanco, Carlos J Pérez-del Pulgar, Andrea Mariani, Claudio Quaglia, Giuseppe Tortora, Arianna Menciassi, and Víctor F Muñoz. A surgical dataset from the da vinci research kit for task automation and recognition. *arXiv preprint arXiv:2102.03643*, 2021.

[30] Natalia Valderrama, Paola Ruiz Puentes, Isabela Hernández, Nicolás Ayobi, Mathilde Verlyck, Jessica Santander, Juan Caicedo, Nicolás Fernández, and Pablo Arbeláez. Towards holistic surgical scene understanding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 442–452. Springer, 2022.

[31] Beatrice van Amsterdam, Matthew Clarkson, and Danail Stoyanov. Gesture recognition in robotic surgery: a review. *IEEE Transactions on Biomedical Engineering*, 2021.

[32] Beatrice van Amsterdam, Matthew J Clarkson, and Danail Stoyanov. Multi-task recurrent neural network for surgical gesture recognition and progress prediction. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1380–1386. IEEE, 2020.

[33] Beatrice Van Amsterdam, Isabel Funke, Eddie Edwards, Stefanie Speidel, Justin Collins, Ashwin Sridhar, John Kelly, Matthew J Clarkson, and Danail Stoyanov. Gesture recognition in robotic surgery with multimodal attention. *IEEE Transactions on Medical Imaging*, 2022.

[34] S Swaroop Vedula, Anand O Malpani, Lingling Tao, George Chen, Yixin Gao, Piyush Poddar, Narges Ahmidi, Christopher Paxton, Rene Vidal, Sanjeev Khudanpur, Gregory D. Hager, and Chi Chiung Grace Chen.

Analysis of the structure of surgical activity for a suturing and knot-tying task. *PloS one*, 11(3):e0149174, 2016.

[35] Mohammad Samin Yasar, David Evans, and Homa Alemzadeh. Context-aware monitoring in robotic surgery. In *2019 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2019.

[36] Nathanael Yong, Philippe Grange, and David Eldred-Evans. Impact of laparoscopic lens contamination in operating theaters: a study on the frequency and duration of lens contamination and commonly utilized techniques to maintain clear vision. *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, 26(4):286–289, 2016.

[37] Dandan Zhang, Zicong Wu, Junhong Chen, Anzhu Gao, Xu Chen, Peichao Li, Zhaoyang Wang, Guitao Yang, Benny Ping Lai Lo, and Guang-Zhong Yang. Automatic microsurgical skill assessment based on cross-domain transfer learning. *IEEE Robotics and Automation Letters*, 2020.

COMPASS: A Formal Framework and Aggregate Dataset for Generalized Surgical Procedure Modeling

Supplementary Information

Kay Hutchinson[1*], Ian Reyes[2,3], Zongyu Li[1] and Homa Alemzadeh[1,2]

[1*]Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, 22903, VA, USA.
[2]Department of Computer Science, University of Virginia, Charlottesville, 22903, VA, USA.
[3]IBM, RTP, 27709, NC, USA.

*Corresponding author(s). E-mail(s): kch4fk@virginia.edu;
Contributing authors: ir6mp@virginia.edu; zl7qw@virginia.edu; ha4d@virginia.edu;

Related Work

The "Language of Surgery" project [13] models surgical procedures as a language and uses grammar to dictate how gestures are combined to perform tasks. A hierarchical framework has been proposed to model surgical procedures [17]. Available datasets primarily focus on the task, gesture, and action levels as summarized in Table 1 and Table 2. Within this hierarchical framework, tasks consist of a sequence of gestures, and actions are defined below the gesture level.

Surgical gestures: Gestures are defined as "the smallest surgical motion gesture that encapsulates a specific intent, (e.g., insert needle through tissue)" with semantic meaning. The JIGSAWS dataset [6], provides gesture level labels

1

Table 1 Datasets and definitions for gesture recognition

Paper	Dataset	Tasks	Gestures	
Gao 2014 [6] [1]	JIGSAWS • Video • Kinematics	Suturing Needle Passing Knot Tying	G1 - Reaching for needle with right hand G2 - Positioning needle G3 - Pushing needle through tissue G4 - Transferring needle from left to right G5 - Moving to center with needle in grip G6 - Pulling suture with left hand G7 - Pulling suture with right hand G8 - Orienting needle G9 - Using right hand to help tighten suture G10 - Loosening more suture G11 - Dropping suture at end and moving to end points G12 - Reaching for needle with left hand G13 - Making C loop around right hand G14 - Reaching for suture with right hand G15 - Pulling suture with both hands	
DiPietro 2019 [4]	MISTIC-SL • Video • Kinematics	Suturing Needle Passing Knot Tying	G1-G12 & G14 G13 - Grab suture using 2nd needle driver G15 - Rotate suture twice using 1st needle driver around 2nd needle driver G16 - Grab suture tail using 2nd needle driver in knot tying G17 - Pull suture tail using 2nd needle driver through knot G18 - Pull ends of suture taut G19 - Rotate suture once using 2nd needle driver around 1st needle driver G20 - Grab suture tail using 1st needle driver in knot tying G21 - Pull suture tail using 1st needle driver through knot G22 - Grab suture using 1st needle driver	
Gonzalez 2020 [8]	DESK • Video • Kinematics	Peg Transfer	S1 - Approach peg S2 - Align & grasp S3 - Lift peg S4 - Transfer peg - Get together	S5 - Transfer peg - Exchange S6 - Approach pole S7 - Align & place
Menegozzo 2019 [16]	V-RASTED • Video • Kinematics	Pick and Place	1 – Collecting ring 2 – Passing ring R to L 3 – Posing ring on pole	4 – Failing 1 5 – Failing 2 6 – Failing 3
Goldbraikh 2022 [7]	own (open) • Video • Kinematics	Suturing (not robotic)	No gesture Needle passing Pull the suture	Instrument tie Lay the knot Cut the suture
Qin 2020 [22]	RIOUS • Kinematics • Video • Events	Ultrasonic prob-ing	S1 Probe released, out of view S2 Probe released, in view S3 Reaching for probe S4 Grasping probe	S5 Lifting probe up S6 Carrying probe to tissue surface S7 Sweeping S8 Releasing probe

for the Suturing, Needle Passing, and Knot Tying tasks in a dry lab experiment setting. Recent works have introduced new datasets as shown in Table 1, but differing gesture definitions limit comparisons between them as well as their generalizability to other tasks. Specifically, [4], [16], and [7] all performed gesture recognition based on kinematic data, but used different datasets and gesture definitions making comparisons difficult. [21] fused kinematic, video, and event data to recognize and predict gestures. But previous works have not combined data from multiple sets since the gesture labels were incompatible.

Action triplets: Action triplets, <surgical tool /instrument, action verb, target anatomy>, are used to describe tool-tissue interactions (TTI) in surgical process modeling [19]. One of the early works formalizes Laparoscopic Adreanectomies, Cholecystectomies and Pancreatic Resections [10] with surgical activities in the form of action triplets for surgical phase inference. [28] annotates two robotic surgery datasets of MICCAI robotic scene segmentation and Transoral Robotic Surgery (TORS) in the form of action triplets to generate surgical reports. In the SARAS Endoscopic Surgeon Action Detection (ESAD) challenge [23], instead of action triplets, actions are described by both the verb and the anatomy. In the CholecTriplet2021 benchmark challenge

Table 2 Datasets and models for surgical actions

Paper	Dataset	Tasks	Actions		
Nwoye 2022 [20]	CholecT50 • Video	Laparoscopic cholecystectomy	Aspirate Clip Coagulate Cut	Dissect Grasp Irrigate Null	Pack Retract
Li 2022 [12]	EndoVis2018 • Video	Nephrectomy	Cauterization Suction Staple Ultrasound sensing	Looping Idle Tool manipulation Suturing	Clipping Retraction
Meli 2021 [15]	own (dVRK) • Video • Kinematics	Ring Transfer	Move Grasp Release Extract		
Forestier 2012 [5]	own (open) • Video	Lumbar disk herniation	Right: Sew Install Hold Remove Coagulate Swab Irrigate	Left: Hold Install Remove	
Wagner 2021 [26]	EndoVis 2019 • Video	Laparoscopic cholecystectomy	Grasp Hold Cut Clip		
De Rossi 2021 [3]	own (dV and SARAS) • Video	Pick and place (semi-autonomous and cooperative)	A01 – MS moves to ring A02 – MS picks ring A03 – MS moves ring to exchange area A04 – AS moves to ring A05 – AS grasps ring and MS leaves ring A06 – AS moves ring to delivery area A07 – AS drops ring on target A08 – AS moves to starting position		
Valderrama 2022 [25]	PSI-AVA • Video	Radical prostate-ctomy	Cauterize Close Close Something Cut Grasp Hold	Open Open Something Pull Push Release Staple	Still Suction Travel Wash
Ma 2021 [14]	own (dV) • Video	Renal hilum dissection (Partial nephrec-tomy)	Single blunt dissection: Spread Peel/push Hook	Single sharp dissection: Cold cut Hot cut Burn dissect	Combination: Pedicalize 2-hand spread Coagulate then cut
Huaulmè 2021 [9]	MISAW • Kinematic • Video	Suturing Knot Tying	Catch Give slack Hold Insert	Loosen completely Loosen partially Make a loop	Pass through Position Pull

for surgical action triplet recognition [18], the challenge dataset, CholecT50, consists of 50 video recordings of laparoscopic cholecystectomy labeled for 100 action triplet classes composed from 6 instruments, 10 verbs, and 15 targets. Despite being more descriptive of the surgical scene, the number of action triplets in the form of verbs, instruments, and targets can grow exponentially compared to a more limited number of gestures.

Surgical Actions: Surgical actions are generally referred to as the level of the surgical hierarchy below gestures. Motions, motion primitives, and the action verb in action triplets are surgical actions. In surgical process modeling, [17] and [11] define motions as an activity performed by only one hand and without semantic meaning. Many other works define actions as atomic units as listed in Table 2.

The gesture and action label datasets are mostly proposed for surgical workflow segmentation [2, 24–27], and gesture [4, 7, 16, 21], action, or action triplet recognition [12, 20, 26]. However, different datasets have varying definitions of gestures and actions. This makes combining data from multiple

sets challenging. Besides varying definitions, prior datasets on surgical actions mainly contain video data, and none of the prior datasets look into the process of labeling and labeling agreement. Prior datasets also do not differentiate actions performed by the left and right hands, which are important for detailed skill assessment and analysis of bimanual coordination. In our framework, we formally define motion primitives for the left and right hands. Our motion primitives are sets of action definitions that are generalizable across different datasets. We also look into how motion primitives relate to surgical context and task progress. Our proposed dataset contains both kinematic and video data along with context and motion primitives labels for a total of six dry-lab tasks. Our dataset will facilitate the development of recognition, skill assessment, and error detection models using both vision and kinematic data.

Context and Motion Primitive Models of Tasks

The ideal performance of each task is modeled using context and MPs as a finite state machine (FSM) where the states are specific contexts and the transitions between states are MPs. The FSM models for the tasks are shown in Figures 1, 2, 3, 4, and 5.

Motion Primitive Graphs of Tasks

The Suturing, Needle Passing, and Knot Tying tasks were also modeled as MP graphs and compared to surgeon-defined MP graphs as shown in Figures 6, 7, and 8.

References

[1] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

[2] Yutong Ban, Guy Rosman, Thomas Ward, Daniel Hashimoto, Taisei Kondo, Hidekazu Iwaki, Ozanan Meireles, and Daniela Rus. Surgical prediction gan for events anticipation. *arXiv preprint arXiv:2105.04642*, 2021.

[3] Giacomo De Rossi, Marco Minelli, Serena Roin, Fabio Falezza, Alessio Sozzi, Federica Ferraguti, Francesco Setti, Marcello Bonfè, Cristian Secchi, and Riccardo Muradore. A first evaluation of a multi-modal learning system to control surgical assistant robots via action segmentation. *IEEE Transactions on Medical Robotics and Bionics*, 2021.
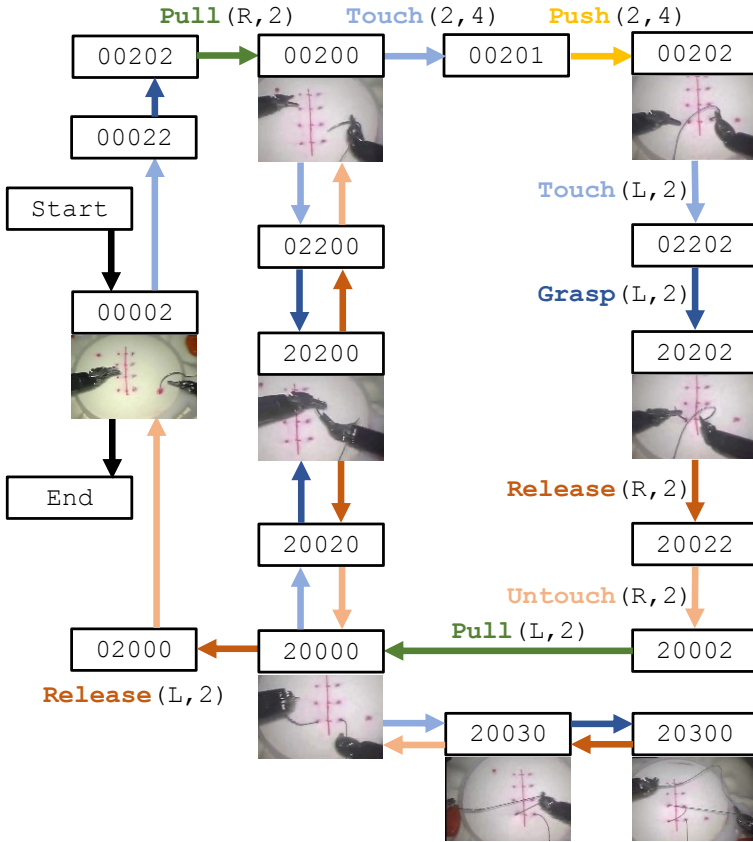
Fig. 1 Finite state machine model representing the ideal performance of a Suturing trial with context and MPs

[4] Robert DiPietro, Narges Ahmidi, Anand Malpani, Madeleine Waldram, Gyusung I Lee, Mija R Lee, S Swaroop Vedula, and Gregory D Hager. Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. *International journal of computer assisted radiology and surgery*, 14(11):2005–2020, 2019.

[5] Germain Forestier, Florent Lalys, Laurent Riffaud, Brivael Trelhu, and Pierre Jannin. Classification of surgical processes using dynamic time warping. *Journal of biomedical informatics*, 45(2):255–264, 2012.

[6] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, Chi Chiung Grace Chen, René Vidal, Sanjeev
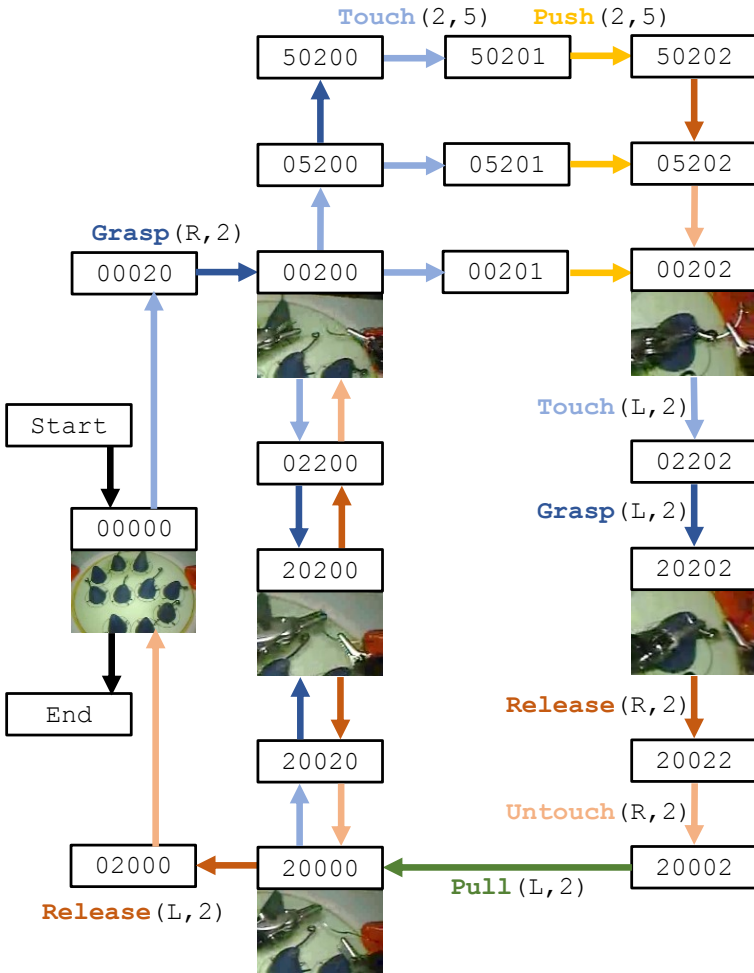
Fig. 2 Finite state machine model representing the ideal performance of a Needle Passing trial with context and MPs

Khudanpur, and Gregory D. Hager. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI*, volume 3, page 3, 2014.

[7] Adam Goldbraikh, Tomer Volk, Carla M Pugh, and Shlomi Laufer. Using open surgery simulation kinematic data for tool and gesture recognition. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–15, 2022.
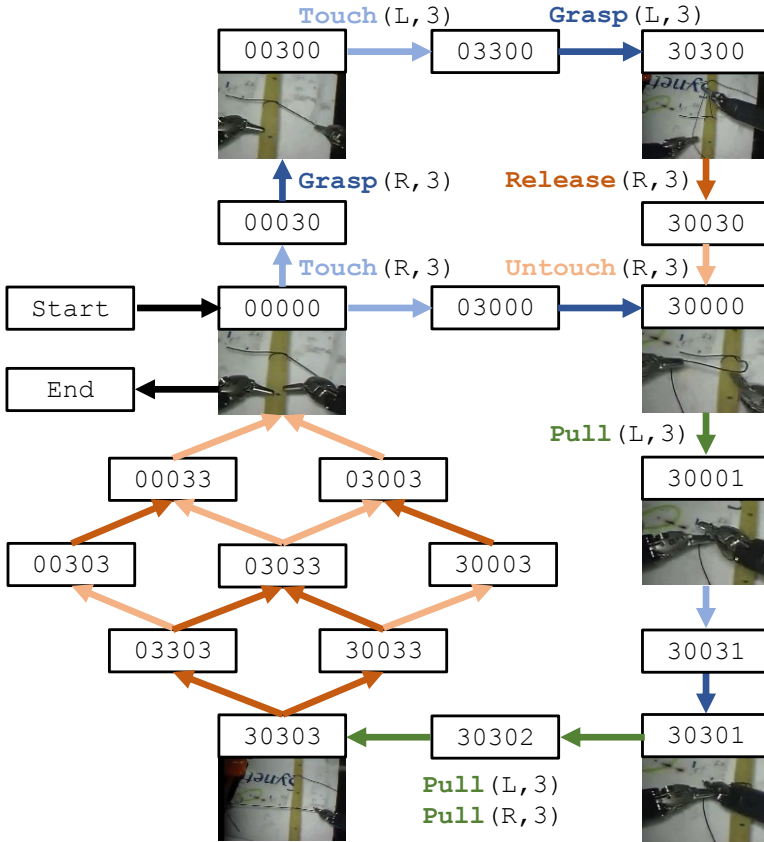
Fig. 3 Finite state machine model representing the ideal performance of a Knot Tying trial with context and MPs
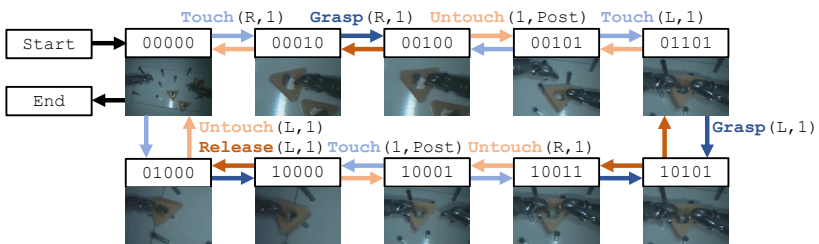
Fig. 4 Finite state machine model representing the ideal performance of a Peg Transfer (or Post and Sleeve) trial with context and MPs
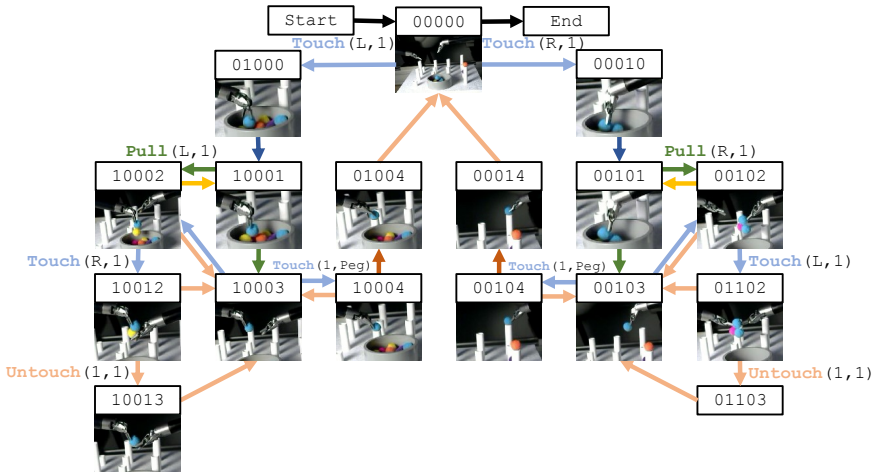
Fig. 5 Finite state machine model representing the ideal performance of a Pea on a Peg trial with context and MPs
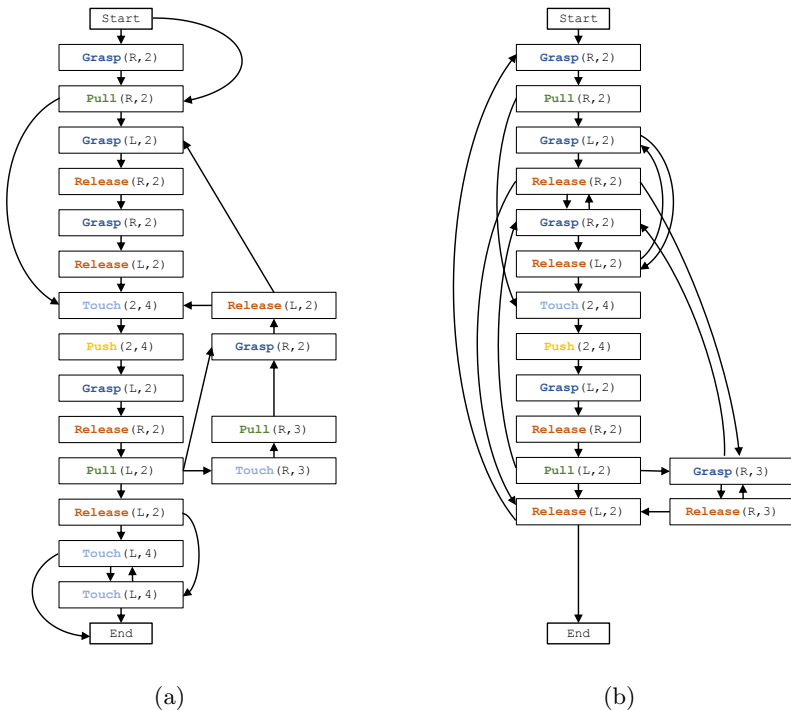

(a)   (b)

Fig. 6 Surgeon-defined (**a**) and proposed (**b**) MP graph models for the Suturing task
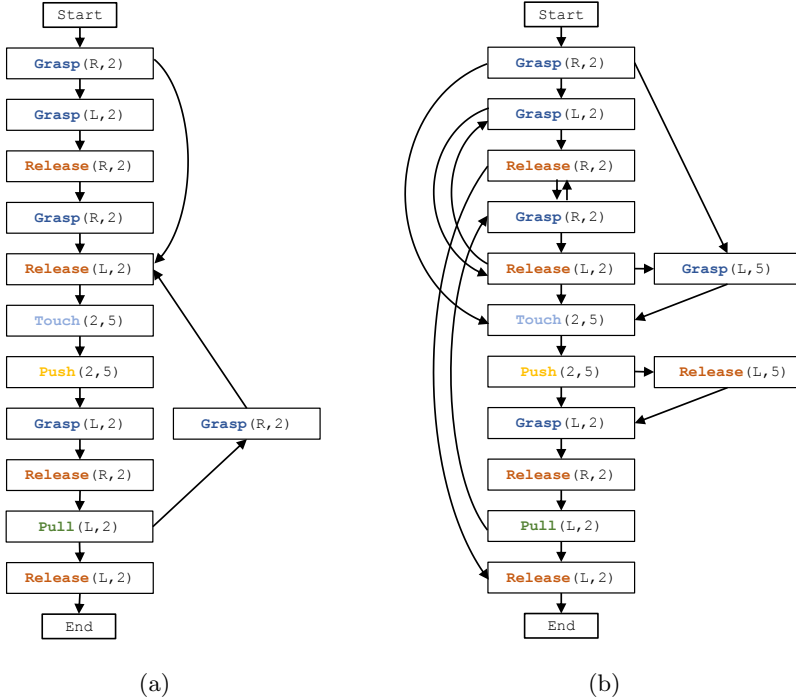
(a)   (b)

Fig. 7 Surgeon-defined (**a**) and proposed (**b**) MP graph models for the Needle Passing task

[8] Glebys T Gonzalez, Upinder Kaur, Masudur Rahma, Vishnunandan Venkatesh, Natalia Sanchez, Gregory Hager, Yexiang Xue, Richard Voyles, and Juan Wachs. From the desk (dexterous surgical skill) to the battlefield–a robotics exploratory study. *arXiv preprint arXiv:2011.15100*, 2020.

[9] Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, Pablo Arbeláez, Wolfgang Reiter, Manoru Mitsuishi, Kanako Harada, and Pierre Jannin. Micro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*, 212:106452, 2021.

[10] Darko Katić, Anna-Laura Wekerle, Fabian Gärtner, Hannes Kenngott, Beat Peter Müller-Stich, Rüdiger Dillmann, and Stefanie Speidel. Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance. In *Information Processing in Computer-Assisted Interventions: 5th International Conference, IPCAI 2014, Fukuoka, Japan, June 28, 2014. Proceedings 5*, pages 158–167. Springer, 2014.
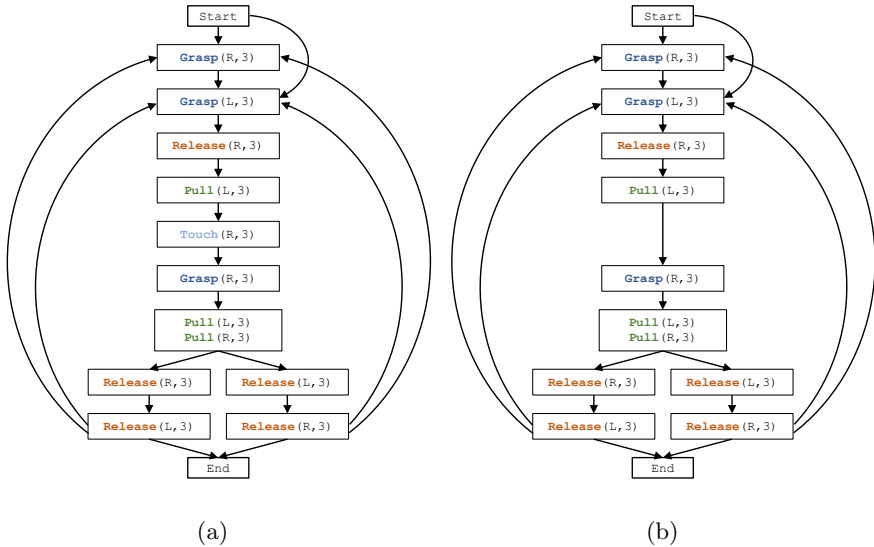
(a)   (b)

Fig. 8 Surgeon-defined (**a**) and proposed (**b**) MP graph models for the Knot Tying task. The "Touch(R, 3)" in **a** refers to a different thread than that of the following "Grasp(R, 3)"

[11] Florent Lalys and Pierre Jannin.   Surgical process modelling: a review. *International journal of computer assisted radiology and surgery*, 9(3):495–511, 2014.

[12] Ling Li, Xiaojian Li, Shuai Ding, Zhao Fang, Mengya Xu, Hongliang Ren, and Shanlin Yang. Sirnet: Fine-grained surgical interaction recognition. *IEEE Robotics and Automation Letters*, 2022.

[13] Henry C Lin. *Structure in surgical motion*. The Johns Hopkins University, 2010.

[14] Runzhuo Ma, Erik B Vanstrum, Jessica H Nguyen, Andrew Chen, Jian Chen, and Andrew J Hung. A novel dissection gesture classification to characterize robotic dissection technique for renal hilar dissection. *The Journal of Urology*, 205(1):271–275, 2021.

[15] Daniele Meli and Paolo Fiorini. Unsupervised identification of surgical robotic actions from small non-homogeneous datasets. *IEEE Robotics and Automation Letters*, 6(4):8205–8212, 2021.

[16] Giovanni Menegozzo, Diego Dall'Alba, Chiara Zandonà, and Paolo Fiorini. Surgical gesture recognition with time delay neural network based on kinematic data. In *2019 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2019.

[17] Dayana Neumuth, Frank Loebe, Heinrich Herre, and Thomas Neumuth. Modeling surgical processes: A four-level translational approach. *Artificial intelligence in medicine*, 51(3):147–161, 2011.

[18] Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, Armine Vardazaryan, Fangfang Xia, Zixuan Zhao, Tong Xia, Fucang Jia, Yuxuan Yang, Hao Wang, Derong Yu, Guoyan Zheng, Xiaotian Duan, Neil Getty, Ricardo Sanchez-Matilla, Maria Robu, Li Zhang, Huabin Chen, Jiacheng Wang, Liansheng Wang, Bokai Zhang, Beerend Gerats, Sista Raviteja, Rachana Sathish, Rong Tao, Satoshi Kondo, Winnie Pang, Hongliang Ren, Julian Ronald Abbing, Mohammad Hasan Sarhan, Sebastian Bodenstedt, Nithya Bhasker, Bruno Oliveira, Helena R. Torres, Li Ling, Finn Gaida, Tobias Czempiel, João L. Vilaça, Pedro Morais, Jaime Fonseca, Ruby Mae Egging, Inge Nicole Wijma, Chen Qian, Guibin Bian, Zhen Li, Velmurugan Balasubramanian, Debdoot Sheet, Imanol Luengo, Yuanbo Zhu, Shuai Ding, Jakob-Anton Aschenbrenner, Nicolas Elini van der Kar, Mengya Xu, Mobarakol Islam, Lalithkumar Seenivasan, Alexander Jenke, Danail Stoyanov, Didier Mutter, Pietro Mascagni, Barbara Seeliger, Cristians Gonzalez, and Nicolas Padoy. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. *arXiv preprint arXiv:2204.04746*, 2022.

[19] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 364–374, Cham, 2020. Springer International Publishing.

[20] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

[21] Yidan Qin, Seyedshams Feyzabadi, Max Allan, Joel W Burdick, and Mahdi Azizian. davincinet: Joint prediction of motion and surgical state in robot-assisted surgery. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2921–2928. IEEE, 2020.

[22] Yidan Qin, Sahba Aghajani Pedram, Seyedshams Feyzabadi, Max Allan, A Jonathan McLeod, Joel W Burdick, and Mahdi Azizian. Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 371–377. IEEE, 2020.

[23] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Inna Skarga-Bandurova, Elettra Oleari, Alice Leporini, Carmela Landolfo, Pengfei Zhao, Xi Xiang, Gongning Luo, Kuanquan Wang, Liangzhi Li, Bowen Wang, Shang Zhao, Li Li, Armando Stabile, Francesco Setti, Riccardo Muradore, and Fabio Cuzzolin. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. *arXiv e-prints*, pages arXiv–2104, 2021.

[24] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[25] Natalia Valderrama, Paola Ruiz Puentes, Isabela Hernández, Nicolás Ayobi, Mathilde Verlyck, Jessica Santander, Juan Caicedo, Nicolás Fernández, and Pablo Arbeláez. Towards holistic surgical scene understanding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 442–452. Springer, 2022.

[26] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Annika Reinke, Tong Yu, Armine Vardazaryan, Chinedu Innocent Nwoye, Nicolas Padoy, Xinyang Liu, Eung-Joo Lee, Constantin Disch, Hans Meine, Tong Xia, Fucang Jia, Satoshi Kondo, Wolfgang Reiter, Yueming Jin, Yonghao Long, Meirui Jiang, Qi Dou, Pheng Ann Heng, Isabell Twick, Kadir Kirtac, Enes Hosgor, Jon Lindström Bolmgren, Michael Stenzel, Björn von Siemens, Hannes G. Kenngott, Felix Nickel, Moritz von Frankenberg, Lena Mathis-Ullrich, Franziska ad Maier-Hein, Stefanie Speidel, and Sebastian Bodenstedt. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *arXiv preprint arXiv:2109.14956*, 2021.

[27] Thomas M Ward, Daniel Hashimoto, Yutong Ban, Elan R Witkowski, Keith D Lillemoe, Guy Rosman, and Ozanan R Meireles. Training with pooled annotations from multiple surgeons has no effect on a deep learning artificial intelligence model's performance. *J Am Coll Surg*, 231(4):e203, 2020.

[28] Mengya Xu, Mobarakol Islam, Chwee Ming Lim, and Hongliang Ren. Learning domain adaptation with model calibration for surgical report generation in robotic surgery. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12350–12356, 2021.