International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-023-02922-1

ORIGINAL ARTICLE

# COMPASS: a formal framework and aggregate dataset for generalized surgical procedure modeling

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

Received: 17 December 2022 / Accepted: 14 April 2023
© CARS 2023

**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

## Introduction

Surgical process models (SPMs) [17, 25] decompose surgical operations into smaller units of activity such as steps, tasks, and gestures as shown in Fig. 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

✉ Kay Hutchinson
kch4fk@virginia.edu

Ian Reyes
ir6mp@virginia.edu

Zongyu Li
zl7qw@virginia.edu

Homa Alemzadeh
ha4d@virginia.edu

[1] Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903, USA

[2] Department of Computer Science, University of Virginia, Charlottesville, VA 22903, USA

[3] IBM, RTP, Durham, NC 27709, USA

Published online: 05 May 2023

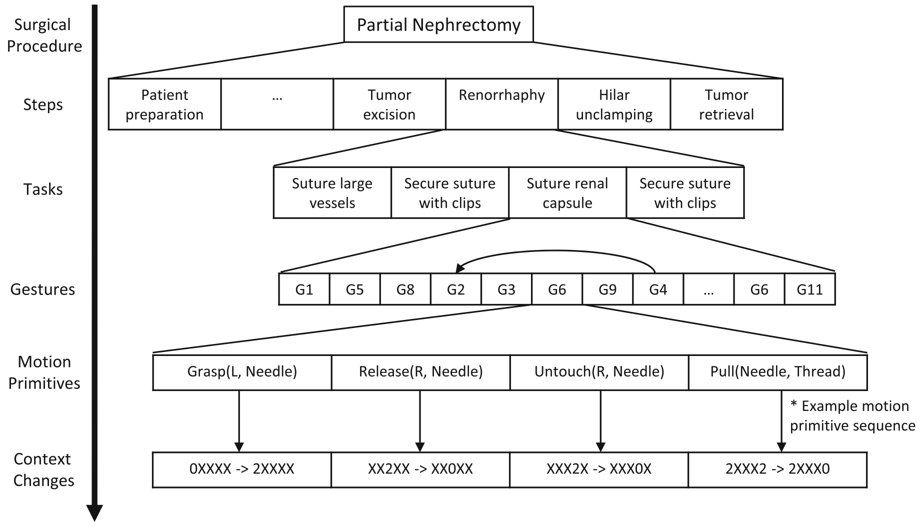


**Fig. 1** Surgical hierarchy. Adapted from [13]

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 are 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 are not available or are 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 have 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.

## Methods

Our framework models surgical procedures as a language with a grammar dictating how motion primitives (MPs) are combined to perform gestures and 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 Fig. 5).

### Modeling framework

#### Surgical hierarchy

Surgical procedures follow the hierarchy of levels defined in [25] which provides context for actions during the procedure, as shown in Fig. 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.

#### 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 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 Fig. 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,” “not stuck to other peas,” or “on the peg.” In Fig. 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.

#### 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 Eq. (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

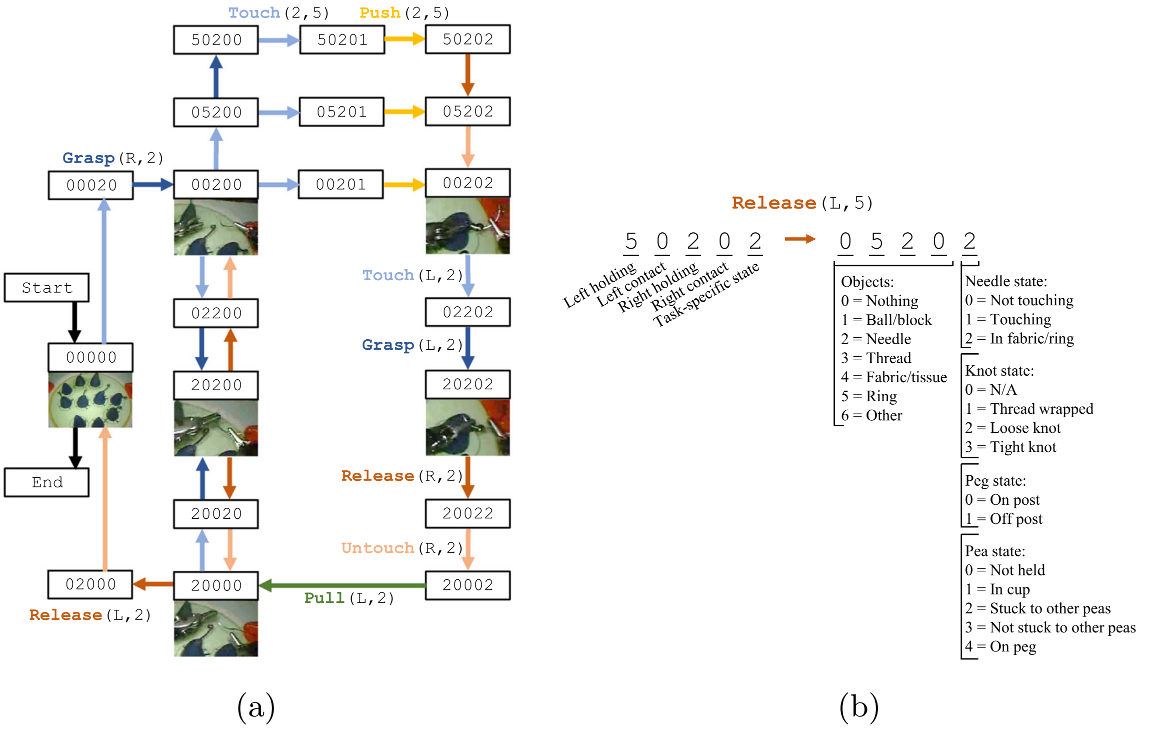


**Fig. 2** **a** Finite-state machine model representing the ideal performance of a Needle Passing trial with context and MPs. In this task, the surgeon threads the needle through four of the rings. **b** An example of the “Release” MP in Needle Passing resulting in the change of state variables

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.

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.

**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 |

## Labeling of context and motion primitives

### Context labeling

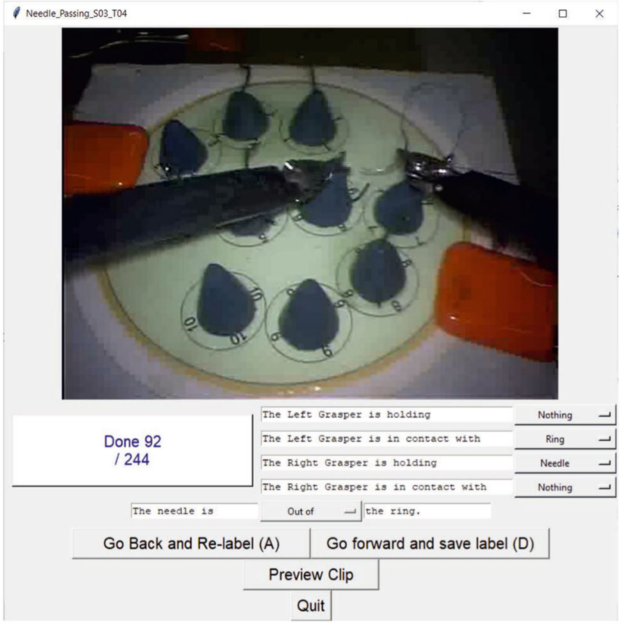
Gesture recognition models using supervised learning require a large number of annotated video sequences [27]. However, manual labeling of gestures is subjective and can lead to labeling errors [32]. Thus, we developed a tool for manually annotating surgical context (states of objects and instruments)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

Labeling video data for surgical context provides a more objective way of recognizing gestures and can lead to higher

**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 |

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.



Done 92 / 244

The Left Grasper is holding: Nothing
The Left Grasper is in contact with: Ring
The Right Grasper is holding: Needle
The Right Grasper is in contact with: Nothing
The needle is Out of the ring.

**Fig. 3** App for context labeling based on video data

## 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 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.

## COMPASS dataset

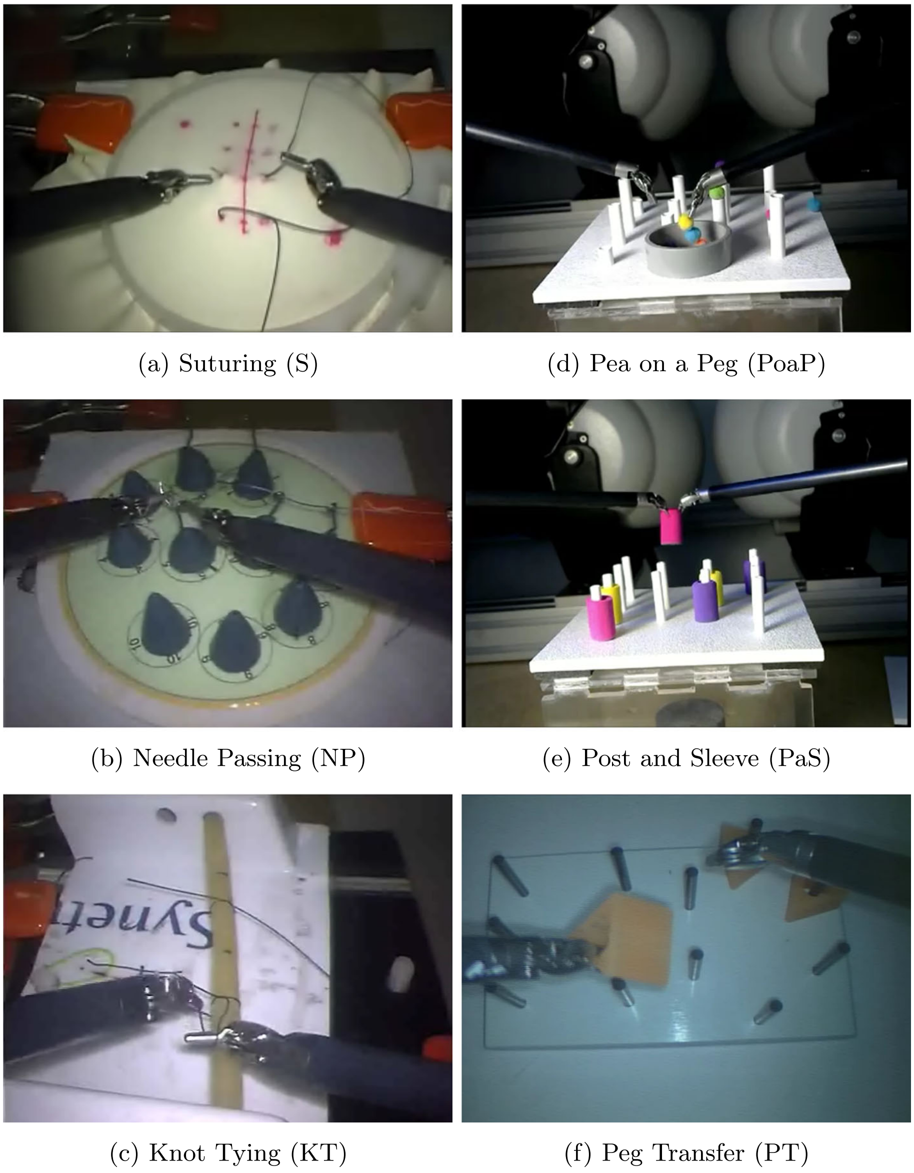
We create the COMPASS dataset by aggregating data from 39 trials of S, 28 trials of NP, and 36 trials of KT performed by eight subjects from the JIGSAWS dataset; 47 trials of PT performed by eight subjects from the DESK dataset; and 65 trials

**Fig. 4** Example sequence of context translated into motion primitives

| Frame | Context |
|---|---|
| ⋮ | |
| 1380 | 05202 |
| 1400 | 00202 |
| 1500 | 20202 |
| 1510 | 20002 |
| 1560 | 20000 |
| 1770 | 20200 |
| ⋮ | |

| Start | Stop | Motion Primitive |
|---|---|---|
| ⋮ | | |
| 1380 | 1399 | Untouch(L, Ring) |
| 1400 | 1499 | Grasp(L, Needle) |
| 1500 | 1509 | Release(R, Needle) |
| 1510 | 1559 | Pull(L, Needle) |
| 1560 | 1769 | Grasp(R, Needle) |
| ⋮ | | |

**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]



(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)

of PaS, and 71 trials of PoaP performed by 12 subjects from the ROSMA dataset (see Fig. 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 downsampled to 30 Hz and contain position, velocity, orientation (in quaternions), and gripper angle variables. Since linear velocity data were not

**Table 3** Number of motion primitives (MPs) in each task and the COMPASS dataset: suturing (S), needle passing (NP), and knot tying (KT) from JIGSAWS [6]; peg transfer (PT) from DESK [20]; and Pea on a Peg (PoaP) and Post and Sleeve (PaS) from ROSMA [29]

| MP | JIGSAWS | | | DESK | ROSMA | | COMPASS |
|---|---|---|---|---|---|---|---|
| | S | NP | KT | PT | PoaP | PaS | |
| 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 |

**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 |

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.

## 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 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.

### 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

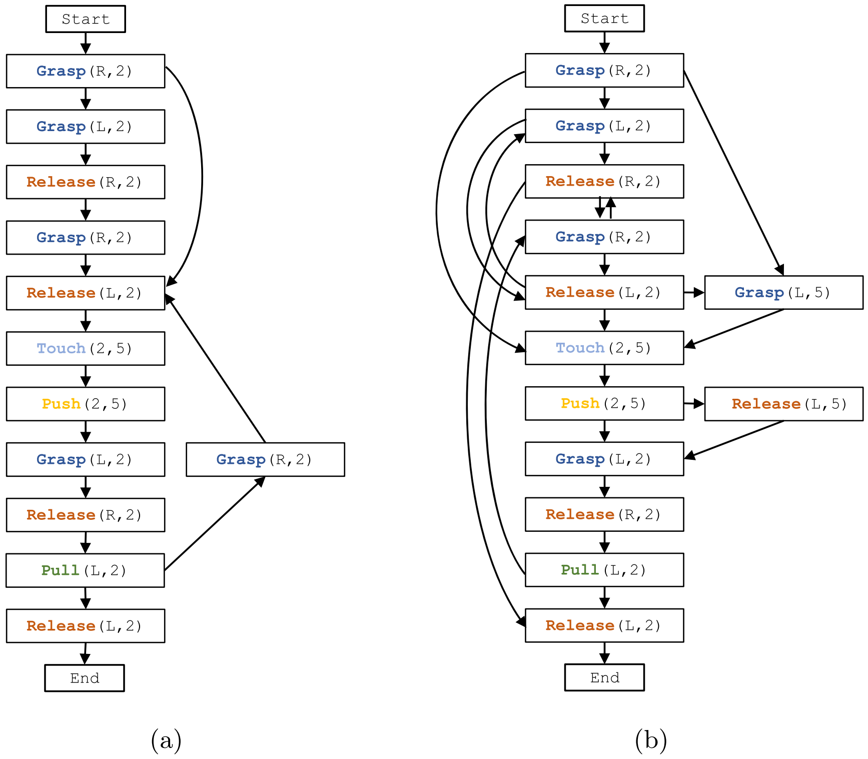


**Fig. 6** Surgeon-defined (**a**) and proposed (**b**) MP graph models for the Needle Passing task

**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 |

that combination when creating their MP graphs. Figure 2a corresponds to Fig. 6b with additional tasks in the Online Resource.

Graph Edit Score (GES) is the normalized graph edit distance (GED) calculated using Eq. (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 h.

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

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.

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

*Krippendorff's Alpha* ($\alpha$) 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 Eq. (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:

$$\alpha = 1 - \frac{D_o}{D_e} \quad (3)$$

**Table 7** Interpretation of Krippendorff's Alpha ($\alpha$) 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 |

**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 |

$\alpha$ 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.

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 Eq. (4) is used to calculate $D_e$ and $D_o$ [12] in Eq. (5), where $n_l$ is the number of labelers and $n_u$ is the total number of frames which two or more labelers annotated.

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

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

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

### 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 $\alpha$ of at least 0.8. The average $\alpha$ 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.

### 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 $\alpha$ 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 result in high-quality annotations comparable to those from surgeons and are thus representative of expert knowledge.

## 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 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 Eq. (6), where 100 is the best and 0 is the worst.

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

**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 |

**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 |

### 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 $\alpha$, 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.

### 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

| Task | Annotator 1 | | Annotator 2 | | Annotator 3 | | Surgeon | |
|---|---|---|---|---|---|---|---|---|
| | 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 | | |

## 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].

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11548-023-02922-1.

**Acknowledgements** This work was supported in part by the National Science Foundation grants DGE-1842490, DGE-1829004, and CNS-2146295 and by the Engineering-in-Medicine center at the University of Virginia. We thank the volunteer labelers, Keshara Weerasinghe, Hamid Roodabeh, and Dr. Schenkman, Dr. Cantrell, and Dr. Chen.

## 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. Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, Hager GD (2017) A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans Biomed Eng 64(9):2025–2041
2. Boehm JR, Fey NP, Fey AM (2021) 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), pp 6248–6255. IEEE
3. Bowyer SA, Davies BL, Baena FRY (2013) Active constraints/virtual fixtures: a survey. IEEE Trans Rob 30(1):138–157
4. De Rossi G, Minelli M, Roin S, Falezza F, Sozzi A, Ferraguti F, Setti F, Bonfè M, Secchi C, Muradore R (2021) A first evaluation of a multi-modal learning system to control surgical assistant robots via action segmentation. IEEE Trans Med Robot Bionics
5. Falezza F, Piccinelli N, De Rossi G, Roberti A, Kronreif G, Setti F, Fiorini P, Muradore R (2021) Modeling of surgical procedures using statecharts for semi-autonomous robotic surgery. IEEE Trans Med Robot Bionics 3(4):888–899
6. Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, Tao L, Zappella L, Béjar B, Yuh DD, Chen CCG, Vidal R, Khudanpur S, Hager GD (2014) 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
7. Gibaud B, Forestier G, Feldmann C, Ferrigno G, Gonçalves P, Haidegger T, Julliard C, Katić D, Kenngott H, Maier-Hein L, März K, de Momi E, Nagy DÁ, Nakawala H, Neumann J, Neumuth T, Balderrama JR, Speidel S, Wagner M, Jannin P (2018) Toward a standard ontology of surgical process models. Int J Comput Assist Radiol Surg 13(9):1397–1408
8. Ginesi M, Meli D, Roberti A, Sansonetto N, Fiorini P (2020) Autonomous task planning and situation awareness in robotic surgery. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp. 3144–3150. IEEE
9. Hagberg A, Swart P, Chult DS (2008) Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab. (LANL), Los Alamos, NM (USA)
10. Hu D, Gong Y, Hannaford B, Seibel EJ (2015) Semi-autonomous simulated brain tumor ablation with ravenii surgical robot using behavior tree. In: 2015 IEEE international conference on robotics and automation (ICRA). IEEE, pp 3868–3875
11. Huaulmé A, Sarikaya D, Le Mut K, Despinoy F, Long Y, Dou Q, Chng C-B, Lin W, Kondo S, Bravo-Sánchez L, Arbeláez P, Reiter W, Mitsuishi M, Harada K, Jannin P (2021) Micro-surgical anastomose workflow recognition challenge report. Comput Methods Programs Biomed 212:106452
12. Hughes J (2021) krippendorffsalpha: an R package for measuring agreement using Krippendorff's alpha coefficient. R Journal 13(1):413–425
13. Hutchinson K, Li Z, Cantrell LA, Schenkman NS, Alemzadeh H (2022) Analysis of executional and procedural errors in dry-lab robotic surgery experiments. Int J Med Robot Comput Assist Surg 18(3):e2375
14. Inouye DA, Ma R, Nguyen JH, Laca J, Kocielnik R, Anandkumar A, Hung AJ (2022) Assessing the efficacy of dissection gestures in robotic surgery. J Robot Surg, pp 1–7
15. Kitaguchi D, Takeshita N, Hasegawa H, Ito M (2021) Artificial intelligence-based computer vision in surgery: recent advances and future perspectives. Ann Gastroenterol Surg 6:10
16. Krippendorff K (2011) Computing Krippendorff's alpha-reliability
17. Lalys F, Jannin P (2014) Surgical process modelling: a review. Int J Comput Assist Radiol Surg 9(3):495–511
18. Lea C, Vidal R, Reiter A, Hager GD (2016) Temporal convolutional networks: a unified approach to action segmentation. In: European conference on computer vision, pp 47–54. Springer
19. Li Z., Hutchinson K., Alemzadeh H (2022) Runtime detection of executional errors in robot-assisted surgery. In: 2022 International conference on robotics and automation (ICRA), pp 3850–3856. IEEE Press
20. Madapana N, Rahman MM, Sanchez-Tamayo N, Balakuntala MV, Gonzalez G, Bindu JP, Vishnunandan Venkatesh LV, Zhang X, Noguera JB, Low T, et al (2019) 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), pp 6928–6934. IEEE
21. Meireles OR, Rosman G, Altieri MS, Carin L, Hager G, Madani A, Padoy N, Pugh CM, Sylla P, Ward TM et al (2021) Sages consensus recommendations on an annotation framework for surgical video. Surg Endosc 35(9):4918–4929
22. Meli D, Fiorini P (2021) Unsupervised identification of surgical robotic actions from small non-homogeneous datasets. IEEE Robot Autom Lett 6(4):8205–8212
23. Menegozzo G, Dall'Alba D, Zandonà C, Fiorini P (2019) Surgical gesture recognition with time delay neural network based on kine-

matic data. In: 2019 International symposium on medical robotics (ISMR), pp 1–7. IEEE

24. Nazari T, Vlieger EJ, Dankbaar MEW, van Merriënboer JJG, Lange JF, Wiggers T (2018) Creation of a universal language for surgical procedures using the step-by-step framework. BJS Open 2(3):151–157
25. Neumuth D, Loebe F, Herre H, Neumuth T (2011) Modeling surgical processes: a four-level translational approach. Artif Intell Med 51(3):147–161
26. Nwoye CI, Yu T, Gonzalez C, Seeliger B, Mascagni P, Mutter D, Marescaux J, Padoy N (2022) Rendezvous: attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med Image Anal 78:102433
27. Park S, Mohammadi G, Artstein R, Morency L-P (2012) Crowdsourcing micro-level multimedia annotations: the challenges of evaluation and interface. In: Proceedings of the ACM multimedia 2012 workshop on crowdsourcing for multimedia, pp 29–34
28. Qin Y, Feyzabadi S, Allan M, Burdick JW, Azizian M (2020) davincinet: Joint prediction of motion and surgical state in robot-assisted surgery. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 2921–2928. IEEE
29. Rivas-Blanco I, Pérez-del Pulgar CJ, Mariani A, Quaglia C, Tortora G, Menciassi A, Muñoz VF (2021) A surgical dataset from the da vinci research kit for task automation and recognition. arXiv preprint arXiv:2102.03643
30. Valderrama N, Puentes PR, Hernández I, Ayobi N, Verlyck M, Santander J, Caicedo J, Fernández N, Arbeláez P (2022) Towards holistic surgical scene understanding. In: International conference on medical image computing and computer-assisted intervention, pp 442–452. Springer
31. van Amsterdam B, Clarkson M, Stoyanov D (2021) Gesture recognition in robotic surgery: a review. IEEE Trans Biomed Eng
32. van Amsterdam B, Clarkson MJ, Stoyanov D (2020) Multi-task recurrent neural network for surgical gesture recognition and progress prediction. In: 2020 IEEE international conference on robotics and automation (ICRA), pp 1380–1386. IEEE
33. Van Amsterdam B, Funke I, Edwards E, Speidel S, Collins J, Sridhar A, Kelly J, Clarkson MJ, Stoyanov D (2022) Gesture recognition in robotic surgery with multimodal attention. IEEE Trans Med Imaging
34. Vedular SS, Malpani AO, Tao L, Chen G, Gao Y, Poddar P, Ahmidi N, Paxton C, Vidal R, Khudanpur S, Hager GD, Chen CCG (2016) Analysis of the structure of surgical activity for a suturing and knot-tying task. PLoS ONE 11(3):e0149174
35. Yasar MS, Evans D, Alemzadeh H (2019) Context-aware monitoring in robotic surgery. In: 2019 International symposium on medical robotics (ISMR), pp 1–7. IEEE
36. Yong N, Grange P, Eldred-Evans D (2016) 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. Surg Laparosc Endosc Percutaneous Tech 26(4):286–289
37. Zhang D, Wu Z, Chen J, Gao A, Chen X, Li P, Wang Z, Yang G, Lo BPL, Yang G-Z (2020) Automatic microsurgical skill assessment based on cross-domain transfer learning. IEEE Robot Autom Lett 5(3):4148–4155

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.