SAGES consensus recommendations on an annotation framework for surgical video

As Published: https://doi.org/10.1007/s00464-021-08578-9

Publisher: Springer US

Persistent URL: https://hdl.handle.net/1721.1/136860

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

SAGES consensus recommendations on an annotation framework for surgical video

Cite this article as: Ozanan R. Meireles, Guy Rosman, Maria S. Altieri, Lawrence Carin, Gregory Hager, Amin Madani, Nicolas Padoy, Carla M. Pugh, Patricia Sylla, Thomas M. Ward, Daniel A. Hashimoto<InstitutionalAuthor><InstitutionalAuthorName>the SAGES Video Annotation for AI Working Groups</InstitutionalAuthorName></InstitutionalAuthor>, SAGES consensus recommendations on an annotation framework for surgical video, Surgical Endoscopy https://doi.org/10.1007/s00464-021-08578-9

Title: SAGES Consensus Recommendations on an Annotation Framework for Surgical Video

Short Title: SAGES Video Annotation Consensus

Authors: Ozanan R. Meireles MD[1], Guy Rosman PhD[1,2], Maria S. Altieri MD[3], Lawrence Carin PhD[4], Gregory Hager PhD[5], Amin Madani. MD PhD[6], Nicolas Padoy PhD[7,8], Carla M. Pugh MD PhD[9], Patricia Sylla MD[10], Thomas M. Ward MD[1], Daniel A. Hashimoto MD[1] and the SAGES Video Annotation for AI Working Groups*

[1]Department of Surgery, Massachusetts General Hospital
[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
[3]Department of Surgery, East Carolina University
[4]Department of Electrical and Computer Engineering, Duke University
[5]Department of Electrical and Computer Engineering, Johns Hopkins University
[6]Department of Surgery, University Health Network
[7]ICube, University of Strasbourg
[8]IHU Strasbourg
[9]Department of Surgery, Stanford University
[10]Department of Surgery, Mount Sinai Medical Center

*Includes all members from Table 1

Corresponding Authors:
Ozanan R. Meireles, MD FACS
15 Parkman Street, WAC460
Boston, MA 02114
617-724-5530
ozmeireles@mgh.harvard.edu

Daniel A. Hashimoto, MD MS
15 Parkman Street, WAC460
Boston, MA 02114
617-724-5530
dahashimoto@mgh.harvard.edu

Abstract

Background. The growing interest in analysis of surgical video through machine learning has led to increased research efforts; however, common methods of annotating video data are lacking. There is a need to establish recommendations on the annotation of surgical video data to enable assessment of algorithms and multi-institutional collaboration.

Methods. Four working groups were formed from a pool of participants that included clinicians, engineers, and data scientists. The working groups were focused on four themes: 1) temporal models, 2) actions and tasks, 3) tissue characteristics and general anatomy, and 4) software and data structure. A modified Delphi process was utilized to create a consensus survey based on suggested recommendations from each of the working groups.

Results. After three Delphi rounds, consensus was reached on recommendations for annotation within each of these domains. A hierarchy for annotation of temporal events in surgery was established.

Conclusions. While additional work remains to achieve accepted standards for video annotation in surgery, the consensus recommendations on a general framework for annotation presented here lay the foundation for standardization. This type of framework is critical to enabling diverse datasets, performance benchmarks, and collaboration.

Background

From laparoscopy to flexible endoscopy and robotics, video-based surgery has evolved substantially. As video capture and processing technology has improved and the complexity of video-based operations increased, the practice of recording and sharing surgical videos for educational and quality improvement purposes has become a growing part of minimally invasive surgeons' practices, with applications in surgical training [1], continuing medical education [2], and clinical dissemination of knowledge [3]. More recently, there has been a growing interest in utilizing video data to train machine learning algorithms for novel applications of artificial intelligence in medicine [4].

Computer vision refers to machine understanding of images and videos, and the majority of its applications to surgery have been through supervised learning, where annotations generated by humans are used to teach machines to recognize surgical phenomena. Videos are composed of still images (i.e. frames) played over time and provide both spatial and temporal information, including the nature of interaction between subjects and objects within a frame. In a surgical video, the surgeon -- through their instruments -- acts as a subject motivated by goals to perform actions to alter the surgical environment [5]. While such a concept seems hard to dispute, the details of how such actions are performed on what objects can be difficult to define in a precise manner. A recent review of automated phase recognition highlights this difficulty as each paper reviewed had different definitions and structures for the phases identified, even if the operation was the same [6]. Inconsistency in definitions of the objects of annotations - whether spatial (e.g. anatomy, tools) or temporal (e.g. operative phases, steps, actions) - can limit the opportunities to combine datasets and compare results across studies.

Previous research in developing a common ontology for surgical workflows has suggested that such work can improve the translation of results and facilitate multi-institutional research efforts [5]. The challenges of video annotation have previously been described [7]. At the moment, however, there is no universally agreed-upon method of annotation for surgical video, making it very difficult to compare results between research groups, combine heterogeneous datasets, and rapidly scale to multi-institutional data and harness global knowledge. By creating and adopting consensus recommendations for surgical video annotation, the individual elements of a video (like building blocks) could be annotated independently in a more uniform fashion to allow for the combination or concatenation of heterogeneous datasets from different sources. The importance of consolidating datasets for AI-driven research has been highlighted in the recent years in computer vision[8], natural language processing [9–11], robotic perception[12], and other fields.

In this context, the Society of Gastrointestinal and Endoscopic Surgeons (SAGES) Artificial Intelligence Task Force, as part of its long-term plan for scalability and sustainability of artificial intelligence in surgery, developed the Video Annotation Consensus project to address the lack of uniform methods to perform surgical video annotation for machine learning.

Purpose and Scope

The aims of the consensus project were to identify current practices in surgical video annotation and to propose recommendations for the basic components of the annotation process with the goal of facilitating a more uniform method of annotating video to improve cross-institutional research efforts. These recommendations were intended to assist physicians and engineers engaged in research and clinical implementation of surgical video-based artificial intelligence by facilitating annotation that allows for the comparison of results between research groups, the combination of heterogeneous datasets, and the cross-validation of algorithmic results.

The scope of this consensus was limited to surgical video for minimally invasive surgery, including laparoscopic, thoracoscopic, endoscopic, and robotic-assisted procedures of the chest, abdomen, and pelvis. The goal was to generate a general framework for video annotation to inform future, more specific, methods of annotating surgical video for training and testing algorithms.

Methods

A steering group of 11 experts in the domains of surgical artificial intelligence and data science was assembled (Table 1), including both clinicians and engineers. The subject of video annotation for surgical AI was divided into the following four domains: 1) temporal models, 2) actions and tasks, 3) tissue characteristics and general anatomy, and 4) software and data structure. Four working groups covering each of these domains were formed.

Participant selection criteria and eligibility

Drawing from SAGES membership and from authors who had published in one or more of the four domains listed above, additional experts from clinical practice and academic engineering were invited to participate in the working groups.

As the steering group recognized that industry may be active in the research and development of applications of surgical AI, approval was obtained from the SAGES Executive Committee of the Board of Governors to include members of industry in the consensus. Industry participation was based on sponsorship of the in-person Video Annotation Consensus Conference. Each company was allowed to appoint up to two individuals who met participation eligibility as outlined below. Appointment of industry participants was reviewed by the chairs of the steering group.

Participation eligibility

Clinician participants needed to be practicing general surgeons, gastrointestinal surgeons, thoracic surgeons, or general surgery residents (Post-graduate Year 3 and above). A minimum qualification of enrollment in or completion of a surgical residency with board eligibility was required. Additional experience in research related to machine learning, computer vision, surgical decision making, minimally invasive surgery (laparoscopic, endoscopic, robotic), or surgical education was required.

Engineers and data scientists needed to be actively involved in research on artificial intelligence, computer vision, machine learning, surgical video annotation, or

related software platforms Completion of an undergraduate degree was necessary as well as was active technical research in machine learning, computer vision, surgical decision-making, or surgical education.

Industry participants were required to have a primary role in research and development related to machine learning and/or computer vision and have no primary role in the marketing, sales, or public relations of their company. Non-researchers from industry, including executives and individuals from marketing and sales without necessary clinical or technical qualifications as noted above, were not eligible to participate in this project.

All participants were required to disclose industry ties and potential conflicts of interest. The steering group reserved the right to exclude individuals who reported or exhibited behaviors suggestive of substantial commercial bias and relevant, significant conflicts of interest.

Working Group Meetings

From December 2019 to January 2020, working groups met weekly online to conduct brainstorming, research, and discussion on the assigned domains. The working groups were tasked to consider drafting a series of recommendations the goal of the providing a basic framework that can be used to annotate any aspect of any operation, regardless of geographic location, surgeon-specific differences, institutional-differences, and/or cultural differences. In late January 2020, each working group presented via a web conference a summary of their findings and suggested recommendations to all members of the consensus working groups. The presentations were recorded to allow members to access the videos for future reference as needed.

Modified Delphi Survey

A modified Delphi process was utilized to create a consensus survey based on suggested recommendations from each of the working groups. This was performed in three rounds. Participation in the modified Delphi process was contingent on participant agreement to have viewed the final working group presentations on suggested recommendations that was held in January 2020. Round 1 was conducted online using REDCap (Research Electronic Data Capture) electronic data capture tools hosted at Massachusetts General Hospital (Boston, MA). REDCap is a secure, web-based software platform designed to support data capture for research studies. Supplementary File 1 demonstrates the questions asked in Round 1 and their results.

The *a priori* criteria were set for each round of the modified Delphi survey. Round 1 was conducted online with statements for consideration based on initial key points that were raised by each of the working groups and was meant to inform in-person discussion held in February 2020. The following criteria were set regarding adoption of each statement: 1) ≥90% agreement would result in statement adoption with no further revision needed; 2) 80-89% agreement would result in statement adoption but with the option for discussion and revision among in-person attendees; 3) <80% agreement

would require group discussion, revision of the statement, and a revote regarding inclusion of the revised statement into Round 3 of the modified Delphi process.

A combined in-person and online discussion was held at the SAGES Video Annotation Consensus Meeting in Houston, TX, USA, on February 22, 2020. Attendees could participate in-person or via a web conferencing solution. Participants were shown the results from Round 1 and, when applicable, discussion, revision, and revoting held. Voting was performed anonymously using Poll Everywhere (San Francisco, CA). Supplementary File 2 contains the voting results, where applicable, from Round 2.

Round 3 was conducted online using REDCap with the following *a priori* criteria: 1) only participants who participated in Round 1 were able to participate in Round 3; 2) statements with ≥80% agreement were adopted; 3) statements with <80% agreement required additional discussion. Supplementary File 3 contains items discussed and revised during Round 3 as well as their voting results.

The final adopted statements and voting percentages were shared with all participants.

Results

Participant Demographics

Fifty-two individuals participated in the overall consensus process. Of these, 37 agreed to participate in the online voting rounds of the modified Delphi process, and 35 (94.6%) completed both online Rounds 1 and 3 of the survey with no dropouts between rounds. Fifty-six individuals attended the February 2020 session where the Round 2 discussion was held live. Forty-eight percent (48%) of participants were engineers or data scientists; 46% were surgeons or surgical trainees. Six percent were non-engineer or non-clinical researchers serving in academia or industry. Seventy-one (71%) were from academia or clinical sites (i.e. hospital-based) with the remaining 29% from industry.

Sixty percent (60%) of the participating surgeons reported having some experience in annotation of surgical video for machine learning purposes. Figure 1 demonstrates the types of annotations they had performed. Table 2 reports the annotation software used by the participants. Figure 2 demonstrates reporting of use of surgical video amongst the surgeons who participated.

Delphi Process

During working group meetings prior to the first Delphi round, the majority of participants agreed to a conceptual framework based on the concept of the temporal and the spatial features of a surgical video having hierarchical structure. This decision was based on prior work that has been performed on the use of “surgemes” (atomic surgical gestures) and low-level activities to help identify or define high-level surgical phases [13–15]. Since new supervision targets arise often in rapidly developing fields, not everything fits within a single hierarchy; however, participants felt structure should be used in models, data and tools, when possible. Thus, we will use “hierarchical” in the remainder of the text to describe the preference for hierarchical relationships.

In Delphi Round 1, 27 candidate statements were included for consideration and received votes (Supplementary File 1). In Round 2, discussion around candidate statements 10-13 regarding the annotation of temporal models led to their exclusion from

the consensus recommendations in favor of combining those recommendations into one recommendation regarding the hierarchical architecture for the annotation of temporal models. Consensus agreement meeting *a priori* criteria was met after Round 3 on 24 consensus statements (Supplementary File 3).

Consensus statements fell into the categories of 1) Need for video annotation consensus recommendations, 2) General considerations for video annotation, 3) Annotator considerations, 4) Lexicon and vocabulary, 5) Temporal annotations, 6) Spatial annotations, and 7) Annotation software.

The recommendations and the discussion surrounding each category of recommendations is presented below.

Consensus Statements, Recommendations, and Discussion

Need for video annotation consensus recommendations

Statement. As of January 2020, it is difficult to perform multi-institutional studies involving surgical video due to the lack of well-defined annotation standards.

- 91% Strongly Agree or Agree.

Recommendation. To promote ongoing research and development for analysis of surgical video, a well described set of guidelines on annotation of surgical video is necessary.

- 97.1% Strongly Agree or Agree.

Discussion. While a formal needs assessment was not undertaken, the steering committee wanted to determine the extent to which working group members felt a consensus around annotations was needed. As of January 2020, no published, well-defined standards or consensus recommendations for the annotation of surgical video had been identified though prior work had established the importance of having a common ontology for the annotation of surgical video [5]. The group reviewed literature that suggested that many research groups collect their own datasets and label them as needed for their research questions. Systematic reviews examining applications of deep learning in surgery have subsequently confirmed significant heterogeneity in definitions of annotations across research studies, even when the same phenomenon (e.g. workflow) within the same operation is under investigation [6, 16]. Thus, the majority of participants agreed that a well-described set of guidelines on the annotation of surgical video was necessary.

General considerations for video annotation

Recommendation. A video annotation framework should be universal, i.e. scalable to all general surgery, including minimally invasive (laparoscopic, robotic, endoscopic) procedures.

- 91.4% Strongly Agree or Agree

Recommendation. A video annotation framework should be both machine readable and clinically applicable.

- 97.1% Strongly Agree or Agree

Recommendation. A video annotation framework should be flexible, i.e. applicable to variations in operations, including rare events and multiple use cases.

- 100% Strongly Agree or Agree

Discussion. Given a goal of setting recommendations for annotation was facilitate to comparison of results between research groups, the combination of heterogeneous datasets, and the cross-validation of algorithmic results, a portion of discussion surrounded general considerations regarding video annotation that could enable such collaborative or cross-institutional work. The consensus was that it should be universal and flexible to be applicable to variations in operations, including rare events. This is to account for differences in technique, style, or preferences that may occur. Furthermore, annotations should be applicable (i.e. interpretable) to clinical uses without the need to apply them to machine learning while maintaining usability by machines. This was felt to be important so that clinicians could evaluate the clinical significance of annotations (e.g. is bleeding clinically significant due to physiologic impact on the patient or due to obscuring visualization of the operative field?) and so that annotations could potentially be used for alternative purposes that may be identified in the future.

There are potential limitations of universal frameworks. For example, defining universal annotations for any context could be overly broad and lead to limited granularity. Conversely, if it is too detailed, it could become impractical. As such, general annotation frameworks would likely be better suited for high level annotation whereas more specific frameworks (e.g. specific definitions for varying grades of inflamed gallbladder or density of adhesions) would likely be needed for specific use cases, highlighting the importance of a flexible hierarchical scalability within the framework (i.e. allow the ability to move up or down a hierarchy of specificity in the description of annotations).

Annotators' experience

Statement. A video annotation framework will yield variable inter-annotator variability, depending on the annotation task. Concrete concepts such as tool annotations should yield low variability, whereas concepts and phenomena that require interpretation and inference prior to annotation may have higher inter-annotator variability.

- 85.7% Strongly Agree or Agree

Recommendation. A video annotation framework should balance annotation effort with granularity to account for varying project needs.

- 88.6% Strongly Agree or Agree

Discussion. A video annotation framework should yield variable inter-annotator variability, depending on the annotation context. For example, labeling surgical tools should yield little variability, given the concrete nature of tools. However, when the annotation process involves a higher level of inference or interpretation prior to labeling a complex phenomenon (e.g. a surgical task, tissue characteristic, or abnormal anatomy), the framework should allow a higher degree of inter-annotator variability. Some degree of inter-annotator variability for such phenomena was considered potentially important as inter-annotator variability may provide a clue into areas where no consensus is possible

or where current clinical practice may not be well established. That is, variability can provide a clue that these phenomena are clinically not well-defined. These cases are important to study and understand because further human input may be necessary or machine learning could be used to cluster elements of phenomena to lead to better definitions.

A universal framework should balance annotation effort with granularity by providing the structure to account for varying project needs so that an annotator can choose the necessary amount of effort to invest. At early stages, too much granularity may not be needed and may be too time-intensive, negatively impacting the efficiency of the annotation process. However, in the future, granularity may be important to ensure collaboration among institutions with diverse datasets.

Lexicon and Vocabulary

Recommendation. A video annotation framework should be nonredundant, i.e. vocabulary used to describe phenomena should minimize multiple categorizations/terms to describe the same things.

- 94.3% Strongly Agree or Agree

Recommendation. A video annotation framework should allow hierarchical categorization with common, fixed vocabulary and the ability to add free text to expand the vocabulary where needed.

- 94.3% Strongly Agree or Agree

Discussion. The use of specific vocabulary for the annotation of phenomena relies on a shared conceptual model of what comprises a clinical phenomenon. Other ontologies to describe clinical phenomena include SNOMED CT (for clinical documentation and billing) [17] and OntoSPM (surgical processes) [5]. Lessons learned from these ontologies influenced recommendations made by the working groups. A hierarchical categorization allows for different levels of granularity or specificity with which to describe a phenomenon. Eliminating or minimizing redundancy in vocabulary prevents multiple terms or phrases that could describe the same phenomenon; thus, improving data capture by preventing the “splitting” of data that would otherwise be considered under one category across multiple categories. Furthermore, a common, fixed vocabulary could ease the process through which researchers identify the appropriate vocabulary to annotate their phenomena of interest. However, we recognized that a purely fixed vocabulary could limit the work of researchers to generate new ideas or capture a more diverse set of phenomena, especially as technology improves and additional phenomena are captured. Thus, we also recommend the ability to add free text to expand or further specify phenomena within the fixed vocabulary when needed. Ultimately, the vocabulary should be both specific enough yet generalizable enough to encompass broad categorizations of clinically meaningful surgical elements while also providing the ability to capture clinically unique aspects of individual operations. It should be easy to use and search with minimal free text required (but available when desired).

Temporal Annotations

Recommendation. A temporal annotation framework should provide the means to annotate events as either time-points or segments of time. A time-point suggests a discrete single moment in time versus a segment of time that lasts seconds to minutes.

- 97.2% Strongly Agree or Agree

Recommendation. A hierarchy of surgical phases, steps, tasks, and actions should be used to annotate temporal events in surgical video as further defined in Table 3.

Recommendation. A temporal annotation framework should provide the means to define relationships between various parts of a particular video (e.g. step C was caused by events that occurred in step A)

- 91.4% Strongly Agree or Agree

Recommendation. The temporal annotation framework should provide the means to annotate times when no visible surgical activity occurs.

- 88.5% Strongly Agree or Agree

Discussion: For the purposes of this consensus, the temporal component of the video referred to the time that elapses from the beginning of the operation until the end of the operation. Extensive discussion was held across the working groups regarding the organization of temporal events in operative videos in a manner that allows for a breakdown of an operation into its component parts. First, the decision was made to recommend that events be annotated either as single time-points (i.e. point estimate events) or as segments of time. While we noted that all events occur over some segment of time (even milliseconds), clinical perception of an event may best be represented by a single time point (e.g. perforating the gallbladder during cholecystectomy). Furthermore, in some projects, it may be preferable to mark a point in time after which the context of an operation changes (e.g. tumor or stool spilled at time t_x and the case is considered contaminated thereafter). Having both options offers researchers greater flexibility.

Table 3 details the four types of events (Phase, Step, Task, Action), their hierarchical relationship, and their definitions. Phases are the first level temporal component of an operation. They must occur sequentially in the following order: Access, Execution of Surgical Objectives, Closure. Access refers to the act of gaining access to the surgical field. Execution of Surgical Objectives refers to the phase that includes steps that must be achieved during the course of an operation to yield the outcome of choice (e.g. what steps are necessary -- independent of access and closure -- to perform a sleeve gastrectomy). Closure refers to the phase of the operation in which the surgical field is exited.

Within and across the three phases, there are surgical steps, the second level temporal component. Steps are procedure-specific (e.g. steps of a laparoscopic gastric bypass) and represent specific segments of an operation to accomplish a clinically meaningful goal -- without which the procedure cannot be completed (e.g. mobilization of the right colon). Steps do not need to be performed in a specific order, they can be interrupted, and they do not have to be unique to that operation alone (e.g. mobilization of the stomach

can be a step in a sleeve gastrectomy for weight loss or in a wedge resection for a gastric gastrointestinal stromal tumor).

Tasks are the third level temporal component and are considered generic (i.e. not procedure-specific). Tasks are a series of actions to accomplish a goal. More than one task must be completed to carry out a step. For example, dissection of the hepatocystic triangle may require the actions of exposing and skeletonizing the cystic duct and artery.

Actions are the fourth level temporal component. They are primitive components of a task and most often represented by a verb. For example, actions include irrigating, suctioning, suturing.

To provide an example of how such a hierarchy would be utilized, consider laparoscopic cholecystectomy. Access involves the steps of port placement and visualization of the gallbladder. Execution of Surgical Objectives includes steps such as release of the gallbladder peritoneum and dissection of the hepatocystic triangle. Tasks may include clearing fat and fibrous tissue to achieve the goal of visualizing the critical view of safety. Actions are those such as coagulating fatty tissue or stripping peritoneum.

Defining relationships between various parts of an operation is an important element of temporal annotation as it can help to annotate temporal events with a causal relationship and should be incorporated into an annotation framework. For example, during the task of dissection of the hepatocystic triangle, inadvertent application of electrical current could result in a thermal injury to the colon. This might then require a new step -- repair of thermal injury -- composed of tasks such as placement of Lembert sutures to repair the injury.

While surgemes (atomic surgical gestures that have previously been defined in kinematic data) were recognized by the working groups as an important element of understanding actions [13], the decision was made to stop with actions as the lowest level of segmentation within this annotation framework as gestures have not been extensively investigated in surgical video. However, this was not to the exclusion of surgemes; rather, it was in consideration of the annotation effort required to annotate surgemes and recognition of the paucity of previously published work describing the annotation and use of surgemes in surgical video [13, 18, 19].

The temporal annotation framework should encompass documentation of idle time. Idle time being defined as one when there is no action visualized within the view from the camera.
For example, when activities are occurring outside the body cavity, such as preparing a mash, troubleshooting instruments, surgeon and assistant switching roles. The rationale behind this recommendation was to exclude those frames in the particular step and task of the operation to prevent inaccurate machine learning training, when necessary.

Spatial Annotations

Recommendation. An annotation framework for tissues and anatomy should focus on a general categorization that can be further appended and refined into specific subsets.

- 94.3% Strongly Agree or Agree

Statement. When annotating anatomy, there are different levels of granularity that can be labeled: 1) Anatomic region (e.g. upper or lower abdomen, pelvis, retroperitoneum), 2) General anatomy (e.g. veins, arteries, muscle, etc.), 3) Specific anatomy (e.g. liver, gallbladder, stomach, cystic artery, common bile duct, etc.)

- 97.1% Strongly Agree or Agree

Statement. Annotation of normal versus altered (e.g. inflamed, infiltrated, distended) anatomy is important to define for an annotation framework.

- 94.2% Strongly Agree or Agree

Recommendation. Annotation of tools should include, at a minimum, general instrument types (e.g. grasper, vessel sealer, monopolar, shears) and allow for further specification of instrument type and functions (e.g. hook vs spatula electrode, Maryland dissector, etc.)

- 97.2% Strongly Agree or Agree

Discussion: The hierarchical characterization of surgical tools, general anatomy, and tissue types -- along with patient factors and intraoperative events that influence visual perception through alteration of normal tissue characteristics -- was considered for annotation.

A basic hierarchical organization of anatomic spatial features (including tissue) was reported and agreed upon as anatomic region (e.g. left or right chest, upper or lower abdomen, retroperitoneum), general anatomic structures (e.g. veins, arteries, muscle), and specific anatomic organs and structures (e.g. liver, stomach, common bile duct). The broad categories are in keeping with the additional recommendation that general categorization of spatial annotations could then allow for more specific categories to be appended to the hierarchy -- either within the existing hierarchy (e.g. right upper quadrant as an anatomic region) or as additional level below specific anatomy. We further recommend the definition and annotation of normal and abnormal tissue characteristics such as edema, inflammation, infection, calcification, neoplastic changes, etc. The addition of “expected” and “unexpected” spatial findings were discussed, but no consensus recommendation was agreed upon. Additional research and discussion was felt to be necessary for such an addition, but the hierarchy would allow for inclusion of such annotations at the level of specific anatomy or tissue characteristics. Ultimately a hierarchy spatial structures relating to anatomy would be structured as follows:

1. Anatomic region (e.g. upper or lower abdomen, pelvis, retroperitoneum, mediastinum, pleural cavity, etc).
2. General anatomy (e.g. veins, arteries, muscle.)
3. Specific anatomy (e.g. liver, gallbladder, stomach, cystic artery, common bile duct, etc.)

4. Tissue characteristics -- both normal and abnormal (eg., edema, tumors, inflammation, infection, metal deposits, etc).

For anatomic spatial features, annotations would most likely benefit from anchoring on specific anatomy with appended labels for region, general anatomy, and tissue characteristics that could provide greater context or detail without unnecessarily expanding the number of possible classes in a given frame. Such an organization would be in keeping with the prior recommendation to minimize redundancy in annotation.

Separate from anatomic features, surgical instruments also require spatial annotation. Instruments differ based on their function, manufacturer, intended and possible uses, and sometimes the manner in which they interact with the surrounding environment. We recommend hierarchical annotation with general instrument type (e.g. scissors, dissector, grasper) and specific instrument (e.g. Maryland grasper, hook vs. spatula monopolar electrode). The focus on specific instruments could help future research in identifying device-related complications, whether instrument choice affects outcome, and how instruments are used in specific scenarios. Comments from participants also noted that a given instrument could be used for different purposes (e.g. a hook electrode used for blunt dissection or for application of electrosurgical energy); thus, researchers could consider appending use labels to their instrument annotations.

Annotation Software

Statement. Potential users of a video annotation tool include: Surgeons (i.e., clinical expert), Clinical non-experts, Clinical researchers, Non-clinical researchers, Data engineers/scientists.

- 94.3% Strongly Agree or Agree

Recommendation. A video annotation tool should be able to account for both spatial and temporal annotations rather than needing two separate tools to perform each type of annotation.

- 85.7% Strongly Agree or Agree

Recommendation. Recommended functions of an annotation software platform are reported in Table 4 along with percent agreement for each function.

Discussion: The execution of annotations ultimately occurs on software platforms, and we felt it was important to discuss basic needs that researchers may have when dealing with annotation of surgical video. While software for ML production systems has been thoroughly investigated both in unpublished and published literature [20], the specific application to surgical video has not been as thoroughly investigated and does not currently have standards or best practices to follow. Thus, in our discussions, we focused on the software aspects that affect the surgical research community working with surgical video and machine learning.

Several of the end users of annotation software [21] and their potential use cases were discussed (Table 5). Our main focus was on surgeons, clinical/technical researchers, and the tools for collecting annotation data for diverse ML tasks.

There were many considerations for software to enable annotation, including:

- Support of different downstream uses, such as model training, exploration of the data for new phenomena, pruning and quality control of the data and human training tasks
- Support of collaboration, experimentation with data and different ML training tasks, and matching to other data sources (labeled, raw, or otherwise processed)
- Platform independence and an open architecture including operating systems support, input/output file formats, ability to work with different web and cloud platforms, and support for downstream software interfaces to maximize use and collaboration
- Configurability and ease of use

A comprehensive survey of annotation tools was conducted, including the tools presented in Table 2. The main characteristics of annotation software platforms that are either in use today or could be used in the near future were discussed (Table 6). One limitation that stood out was the lack of a common format for annotation outputs that can be leveraged by different ML training tasks downstream.

Beyond surveying the existing tools, we defined a loose interoperable proto buffer file format for annotations that can accommodate the annotation types discussed in the workshop. We have created a software developers kit (SDK) to use it, demonstrating temporal segmentation neural network training, available at [22]. The data in this case was converted from the Cholec80 dataset [23], but we have also tested conversion from the MGH Surgical AI & Innovation Laboratory's annotations of sleeve gastrectomy [24], peroral endoscopic myotomy [25], and laparoscopic cholecystectomy to demonstrate the framework's diversity. We intend to explore the use of this file format to support additional tasks as a basis for other people to use and to bridge different tools in surgical data science. Given the more technical aspects of the work surrounding software and file formats, reporting of these recommendations will be done separately in the future.

Next Steps

The working groups were able to generate a general framework of recommendations for annotation of surgical video to enable research to meet minimum considerations that move toward standardization of annotation processes. While this enables work to begin on creating datasets that are comparable, we recognize that more specific guidance will be necessary moving forward to truly enable collaborative research as well as comparisons and benchmarking of performance. Additional steps will require more specific definitions and guidance to fill out the ontology of temporal and spatial annotation. This work is underway by the working groups and has the benefit of drawing from existing work in medical and surgical ontologies [5, 17]. In this process, additional granularity can be tested through pilot cases across multiple institutions to determine inter-institutional and inter-annotator variability when using this framework. Importantly,

validation studies will need to be performed to assess the reproducibility of annotations under these recommendations.

We anticipate that these consensus recommendations on the annotation of surgical video will enable multiple additional lines of research that would otherwise not have been possible. As the annotation framework is put into place, research can begin to evaluate the expertise necessary to annotate certain aspects (i.e. should lay/crowd workers annotate elements such as tools while experienced surgeons annotate more nuanced concepts such as inflammation). Understanding the appropriateness of annotators could lower the cost of annotation as expertise can be utilized selectively. Furthermore, comparisons of model performance or combination of datasets -- such as through federated learning -- could be performed.

Conclusion

The standardization of surgical video annotation is necessary. While much additional work remains to achieve accepted standards for video annotation in surgery, the consensus recommendations on a general framework for annotation made by an interdisciplinary group of experienced surgeons, engineers, and data scientists from academia and industry lay the foundation for standardization. This type of framework is critical to enabling diverse datasets, performance benchmarks, and collaboration. Subsequent work among the working groups is currently in progress and is expected to provide more granular guidelines for spatial and temporal annotations, and software requirements. Validation studies will be necessary to help structure and substantiate these recommendations.

Acknowledgements: The authors thank SAGES staff Sallie Matthews, Jillian Kelly, Jason Levine, and Shelley Ginsberg for their administrative support in this work. We also thank Dr. Aurora Pryor for her support as SAGES leadership.

Disclosures: This work was supported by the SAGES Foundation, Digital Surgery, Imagestream, Intuitive Surgical, Johnson & Johnson CSATS, Karl Storz, Medtronic, Olympus, Stryker, Theator, and Verb Surgical. Ozanan Meireles is a consultant for Olympus and Medtronic and has received research support from Olympus. Guy Rosman is an employee of Toyota Research Institute (TRI); the views expressed in this paper do not reflect those of TRI or any other Toyota entity. He has received research support from Olympus. Amin Madani is a consultant for Activ Surgical. Gregory Hager is a consultant for theator.io and has an equity interest in the company. Nicolas Padoy is a consultant for Caresyntax and has received research support from Intuitive Surgical. Thomas Ward has received research support from Olympus. Daniel Hashimoto is a consultant for Johnson & Johnson and Verily Life Sciences. He has received research support from Olympus and the Intuitive Foundation.

Table and Figure Legends

Table 1. Steering group (Leads in each working group) and working group members. * denotes individuals who were representing industry at the time of the consensus meeting.

Table 2. Annotation software platforms that have been used by the participants in the modified Delphi process.

Table 3. Hierarchy of events for segmentation of events in surgical video

Table 4. The percentage of participants who strongly agree or agree that these functions should be present in an annotation software platform.

Table 5. Anticipated users of annotation software for surgical video and examples of their potential use cases

Table 6. Software characteristics for annotation tools

Figure 1. Experience of surgeons with different types of annotations

Figure 2. Experience of surgeons in using surgical video

References

1. McKinley SK, Hashimoto DA, Mansur A, Cassidy D, Petrusa E, Mullen JT, Phitayakorn R, Gee DW (2019) Feasibility and Perceived Usefulness of Using Head-Mounted Cameras for Resident Video Portfolios. J Surg Res 239:233–241

2. Greenberg CC, Byrnes ME, Engler TA, Quamme SP, Thumma JR, Dimick JB (2021) Association of a Statewide Surgical Coaching Program with Clinical Outcomes and Surgeon Perceptions. Ann Surg. https://doi.org/10.1097/SLA.0000000000004800

3. Manabe T, Takasaki M, Ide T, Kitahara K, Sato S, Yunotani S, Hirohashi Y, Iyama A, Taniguchi M, Ogata T, Shimizu S, Noshiro H (2020) Regional education on endoscopic surgery using a teleconference system with high-quality video via the internet: Saga surgical videoconferences. BMC Med Educ 20:329

4. Hashimoto DA, Rosman G, Rus D, Meireles OR (2018) Artificial Intelligence in Surgery: Promises and Perils. Ann Surg 268:70–76

5. 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, Rojas Balderrama J, Speidel S, Wagner M, Jannin P (2018) Toward a standard ontology of surgical process models. Int J Comput Assist Radiol Surg 13:1397–1408

6. Garrow CR, Kowalewski K-F, Li L, Wagner M, Schmidt MW, Engelhardt S, Hashimoto DA, Kenngott HG, Bodenstedt S, Speidel S, Müller-Stich BP, Nickel F (2020) Machine Learning for Surgical Phase Recognition: A Systematic Review. Ann Surg. https://doi.org/10.1097/SLA.0000000000004425

7. Thomas M. Ward Danyal M. Fer Yutong Ban Guy Rosman Ozanan R. Meireles Daniel A. Hashimoto (2021) Challenges in Surgical Video Annotation. Computer Assisted Surgery Accepted:

8. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp 248–255

9. Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. arXiv [cs.CL]

10. Gokaslan A, Cohen V (2019) Openwebtext corpus. urlhttp://Skylion007 github io/OpenWebTextCorpus

11. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision. pp 19–27

12. Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: The KITTI dataset. Int J Rob Res 32:1231–1237

13. Varadarajan B, Reiley C, Lin H, Khudanpur S, Hager G (2009) Data-derived models for segmentation with application to surgical assessment and training. Med Image Comput

Comput Assist Interv 12:426–434

14. Katić D, Wekerle A-L, Gärtner F, Kenngott H, Müller-Stich BP, Dillmann R, Speidel S (2014) Knowledge-Driven Formalization of Laparoscopic Surgeries for Rule-Based Intraoperative Context-Aware Assistance. In: Information Processing in Computer-Assisted Interventions. Springer International Publishing, pp 158–167

15. Ahmadi S-A, Sielhorst T, Stauder R, Horn M, Feussner H, Navab N (2006) Recovery of surgical workflow without explicit models. Med Image Comput Comput Assist Interv 9:420–428

16. Anteby R, Horesh N, Soffer S, Zager Y, Barash Y, Amiel I, Rosin D, Gutman M, Klang E (2021) Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surg Endosc. https://doi.org/10.1007/s00464-020-08168-1

17. Bhattacharyya SB (2015) Introduction to SNOMED CT. Springer

18. van Amsterdam B, Clarkson M, Stoyanov D (2021) Gesture Recognition in Robotic Surgery: a Review. IEEE Trans Biomed Eng PP: . https://doi.org/10.1109/TBME.2021.3054828

19. Reiley CE, Hager GD (2009) Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. Med Image Comput Comput Assist Interv 12:435–442

20. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, Crespo J-F, Dennison D (2015) Hidden technical debt in machine learning systems. Adv Neural Inf Process Syst 28:2503–2511

21. Cockburn A (2001) Writing effective use cases. Pearson Education India

22. SAIIL_public. Github

23. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE Trans Med Imaging 36:86–97

24. Hashimoto DA, Rosman G, Witkowski ER, Stafford C, Navarette-Welton AJ, Rattner DW, Lillemoe KD, Rus DL, Meireles OR (2019) Computer Vision Analysis of Intraoperative Video: Automated Recognition of Operative Steps in Laparoscopic Sleeve Gastrectomy. Ann Surg 270:414–421

25. Ward TM, Hashimoto DA, Ban Y, Rattner DW, Inoue H, Lillemoe KD, Rus DL, Rosman G, Meireles OR (2020) Automated operative phase identification in peroral endoscopic myotomy. Surg Endosc. https://doi.org/10.1007/s00464-020-07833-9

Chair: Ozanan R. Meireles, MD **Co-Chairs**: Daniel A. Hashimoto, MD MS and Guy Rosman, PhD	
Domain	**Working Group Members**
Temporal Models	**Leads**: Amin Madani, MD PhD; Nicolas Padoy, PhD **Members**: Yutong Ban, Fillipo Filicori, Pietro Mascagni, John Mellinger, Christopher Schlacta, Stefanie Speidel, Thorsten Juergens*, Pablo Garcia-Kilroy*, Dotan Asselman*
Actions and Tasks	**Leads**: Carla Pugh, MD PhD; Gregory Hager, PhD **Members**: Jordan Bohnen, Rachel Ballantyne Draelos, Hans Fuchs, Ricardo Henao, Duygu Sarikaya, Christopher Boyle*, Danyal Fer*, Zhen Li*, Arvind Ramadorai*, Danail Stoyanov*, Andrew Yoo*
Tissue Characteristics and General Anatomy	**Leads**: Patricia Sylla, MD; Lawrence Carin, PhD **Members:** Cristians Gonzalez, Dmitry Oleynikov, Janey Pratt, Danny Scott, Swaroop Vedula, Elan Witkowski, Takayuki Shimizu*, Mark Tousignant*
Software and Data Structure	**Leads**: Maria Altieri, MD; Guy Rosman, PhD; Thomas Ward, MD **Members**: Dan Azagury, Flavien Bridault, Brian Dunkin, Teodor Grantcharov, Pierre Jannin, Anand Malpani, Silvana Perretta, Steven Schwaitzberg, Anthony Jarc*, Kurt Landfors*, Amit Mahadik*, Holly Nguyen*

Table 1. Steering group (Leads in each working group) and working group members. * denotes individuals who were representing industry at the time of the consensus meeting.

ANVIL b<>com Via Think Like a Surgeon Theator Indexity Surgical Safety Technologies	Digital Surgery CVAT Supervisely Figure 8 Verb LabelMe Other custom software

Table 2. Annotation software platforms that have been used by the participants in the modified Delphi process.

Event	**Definition**
Phase (generic)	Highest level temporal component of an operation for segmentation purposes; phases are divided into Access, Execution of Surgical Objectives, Closure
Step (procedure-specific)	Procedure-specific segment to accomplish a clinically meaningful goal, without which the procedure cannot be completed. Steps need not be performed in a specific order. Steps can be interrupted. Steps do not have to be unique to that operation alone (i.e. a step can be present across similar procedures).
Task (generic)	Sub-component of a step. Composed of a series of actions to accomplish a goal. More than one task must be completed to carry out a step.
Action (generic)	A primitive component of a task. A series of actions are required to complete a task. Most often represented by a verb.

Table 3. Hierarchy of events for segmentation of events in surgical video

Function	Strongly Agree or Agree
Ability to revise prior annotations	100%
Support interaction via mouse and keyboard	100%
Allow for new types of data in the future	97.1%
Define new annotation schema and edit prior schemas	94.3%
Merge annotations on one video from multiple	94.3%

annotators	
Support cross-platform compatibility	94.3%
Export annotations in a variety of supported formats	91.4%
Import/export format and capabilities	91.4%
Variable playback speed	91.4%
Open-source extensible	85.7%

Table 4. The percentage of participants who strongly agree or agree that these functions should be present in an annotation software platform.

Surgeon
Annotate a temporal event (i.e., a duration in time) and apply a textual label. Annotate a time-point event (i.e., an instance in time) and apply a textual label. Annotate the relationships between phases, steps, tasks, and actions. Create video clips of each phase, step, or task that was annotated. Annotate which operator is performing which tasks and examine/compare statistics.
Clinical researcher
Evaluate the accuracy of a trained model (temporal or spatial) by displaying the video and model-derived results using the video annotation tool software.
Clinical non-expert
Inspect a video that was previously annotated by an expert to test how accurately they can identify the temporal or spatial annotations.
Data scientist / non-clinical researcher
Provide annotated training data, specify a particular model to train on the supplied data, and receive performance statistics and a trained model as output. Do analysis based on the annotations, construct new models of the data.

Table 5. Anticipated users of annotation software for surgical video and examples of their potential use cases

Characteristic	**Additional details**

Supported input file formats	Video/image types for input, annotation project save format
Supported output file formats	Temporal annotation file formats, spatial annotation file formats
Platform support	OS support, Database backend, browser support, connection to hospital systems, hardware requirements
User interface	Support temporal and spatial annotations as well as other labels. Support additional language. Support annotation with minimal effort. Include support for active labeling. Support hierarchy and temporal patterns to be annotated where available. For spatial annotation, support multiple label types such as boxes, ovals, scribbles, polygons.
Security	Make sure data and communications are secure.
Privacy	Cater to privacy by design so as to not save personally identifiable information in an unsafe manner.
Annotation actions support	Annotation from picklist/freeform, update labels, create tracks for labels, add temporal labels and modify their lengths/types. Add relationships between labels. Add spatial annotations/update the region or label.
Video control	Change speed, skip time, skip to annotated points, skip according to active labeling approaches.
Smart features support	Provide suggestions on where/what to annotate, predictive tracking on spatial annotations of next frame, allow models to suggest labels for spatial/temporal annotations, support tools for consistency checking and error checking, support external modules for active learning.
General software considerations	Limited memory / computation requirements, open/extendible software, support for easy remote annotation of data.

Table 6. Software characteristics for annotation tools