2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

2021 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9077-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRA48506.2021.9561770

Aggregating Long-Term Context for Learning Laparoscopic and Robot-Assisted Surgical Workflows

Yutong Ban[1,2], Guy Rosman[1,3], Thomas Ward[2], Daniel Hashimoto[2]
Taisei Kondo[4], Hidekazu Iwaki[4], Ozanan Meireles[2], Daniela Rus[1]

***Abstract*— Analyzing surgical workflow is crucial for surgical assistance robots to understand surgeries. With the understanding of the complete surgical workflow, the robots are able to assist the surgeons in intra-operative events, such as by giving a warning when the surgeon is entering specific keys or high-risk phases. Deep learning techniques have recently been widely applied to recognizing surgical workflows. Many of the existing temporal neural network models are limited in their capability to handle long-term dependencies in the data, instead, relying upon the strong performance of the underlying per-frame visual models. We propose a new temporal network structure that leverages task-specific network representation to collect long-term sufficient statistics that are propagated by a sufficient statistics model (SSM). We implement our approach within an LSTM backbone for the task of surgical phase recognition and explore several choices for propagated statistics. We demonstrate superior results over existing and novel state-of-the-art segmentation techniques on two laparoscopic cholecystectomy datasets: the publicly available Cholec80 dataset and MGH100, a novel dataset with more challenging and clinically meaningful segment labels.**

Keywords: laparoscopic surgery, robot-assisted surgery, work flow recognition, temporal context aggregation.

I. Introduction

The future of robot-assisted laparoscopic surgery relies upon a strong automated understanding of surgical workflow from laparoscopic video. In order for robot systems to assist surgeons during surgery, they need a fundamental understanding of the surgical process. Surgical video represents an invaluable source of information as it is sufficient for surgical situational awareness and plentiful in the modern medical environment. While significant works has been performed for natural video analysis [1], [2], [3], [4], [5], [6], works also try to improve the understanding of video [7], [8], [9], [10], [11] and producing better annotation and supervision cues [7], [12], [13], [14] for both laparoscopic and robot-assisted surgeries [15], existing models still fall short of a complete and automatic interpretation of surgery. A key cause of this performance gap is the manner in which surgery is interpreted from videos — surgery is an inherently temporal process with a partially-observable state, and long-term temporal patterns. This is in contrast to other fields which have seen improved performance such as interpreting images from Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) [16], where the patient state is more completely observed, or surgical technique analysis, where short-term interactions are sufficient.

Current computer vision efforts in surgical workflow analysis have addressed operative phase recognition, as well as related tasks such as tool usage detection, tool segmentation and prediction of remaining surgery duration. In recent years, with the rise of deep learning methods, Convolutional Neural Networks (CNN) with SVM and Hierarchical HMM being introduced in [7] for surgical phase recognition. The majority of approaches now use a CNN with a Long-Short Term Memory [17] (LSTM) backbone, including [18], [19], [20]. However, understanding surgical workflows requires reasoning about events across highly varied temporal scales, from a few seconds to hours, exceeding the capabilities of existing models. For example, in an endoscopic setting such as colonoscopy, identification of a polyp early in the procedure during insertion of the scope can influence the decision to perform polypectomy later during withdrawal. In laparoscopic cholecystectomy, "Dissection of Calot's triangle" involves removing the lower portion of the gallbladder from the liver bed (i.e. clearing the cystic plate). This phase can be visually indistinct from "Removal of the Gallbladder from the Liver Bed" later in the case and requires knowledge that key phases (occurring minutes later) have not yet occurred to accurately infer the current surgical phase. In such cases, information extracted by LSTM remains local compared to the total duration of the surgery and fails to improve classification performance of these cases.

Two main approaches have been recently proposed to ameliorate LSTM's shortcomings. Temporal and Dilated convolutional approaches take a multiscale approach and handle low-frequency processes well [21]. Attention based models aimed at matching specific points in the distant past have been transformative for natural language processing and have been recently applied to surgical videos as well [22], [23]. Yet, neither of these approach capture more subtle phenomena such as aggregation of partial evidence from a segment in the distant past or order constraints between specific phases in the surgery that require measurements of interval lengths. While specific approaches have been attempted to cover specific priors on phase lengths and order [24], [15], a general framework that successfully handles long-term reasoning under uncertainty is still needed for

[1] Computer Science and Artificial Intelligence Laboratory, 32 Vassar St, Cambridge, MA 02139, US. {yban, rosman, rus}@csail.mit.edu
[2] SAIIL, Department of Surgery, Massachusetts General Hospital. 55 Fruit Street, Boston, MA 02114, US. {tmward, dahashimoto, ozmeireles}@mgh.harvard.edu
[3] Toyota Research Institute, Cambridge, MA 02139, US.
[4] Olympus Corporation, Shinjuku Monolith, Shinjuku-ku, Tokyo, Japan. {hidekazu.iwaki, taisei.kondo2}@olympus.com

surgical video understanding and remains an open problem.

In this paper, we address this problem by aggregating the long-term temporal context through sufficient statistic models (SSM) and combining them with visual cues to fed into an LSTM. The contribution of the paper is summarized as follows:

(i) A novel SSM-LSTM framework that aggregates the long-term temporal context of different time scales to augment LSTM inference.
(ii) An exploration of the various SSM feature choices and evaluation of their contributions.
(iii) Validation of the proposed model on two large laparoscopic video datasets – the publicly available Cholec80, and MGH100, a novel large-scale laparoscopic cholecystectomy video dataset. The proposed model obtains superior performance compared to state of the art methods on both datasets, with an advantage in clinically meaningful phenomena.

II. Related works

Surgical video analysis addresses several applications, such as phase recognition of laparoscopic surgery [18], [19], [20] and estimation of remaining surgery duration [20]. An additional line of work explores videos from robotic-assisted surgery [25] and integrates kinematics and robotic system events [13], [26]. Significant efforts have been made on surgical instrument analysis [27], [7], [28]. [29] integrate a priori information derived from motion flow into a temporal attention pyramid network for automatic instrument segmentation. [30] extended this model by using a dual motion based semi-supervised framework, which leverages the self-supervised sequential cues in surgical videos. [31] uses a Bayesian network for anticipating the use of surgical instruments for context-aware assistance. Progress in the field is limited by data availability, which requires experts in surgery to label the data. [32] tries to overcome the problem by using transfer learning, which prevents the need to re-learn how to segment similar sub-tasks.

CNN-LSTM has become a standard architecture choice for processing video sequences with [18], [19], [20] focusing on laparoscopic surgeries for phase recognition, and [20] estimating a surgery's remaining time by adding a regression head. Moreover, [33] used CNN-LSTM jointly to predict the future state and recognizing different surgery types. Meanwhile, [34] applied the architecture on cataract surgeries, and [35] used it for Per-Oral Endoscopic Myotomy (POEM) surgeries. Some variants of CNN-LSTM with modifications are proposed such as Prior Knowledge Inference (PKI) (hand-crafted knowledge of operative phase workflow to limit incorrect inferences of already finished phases) [24] or ”multitask” architectures that incorporate multiple streams of inference in addition to operative phase labels to improve identification (e.g. tool prediction [12]).

In many of the methods above, the underlying temporal model is limited in its ability to analyze temporal information across time scales. HMMs are Markovian, while LSTMs are limited by their ability to propagate gradients in time [36], [37], [22]. Models such as dilated temporal convolutions and hidden semi-Markov models (HSMMs) are limited in their ability to efficiently train information flow across long periods of time, with the latter limited by its inference computational efficiency. [21] used a temporal convolution network (TCN) for action segmentation. [38] applied TCN in surgery and combined it with reinforcement learning (RL) for surgical gesture recognition. [39] improved upon this using an uncertainty-aware tree search.

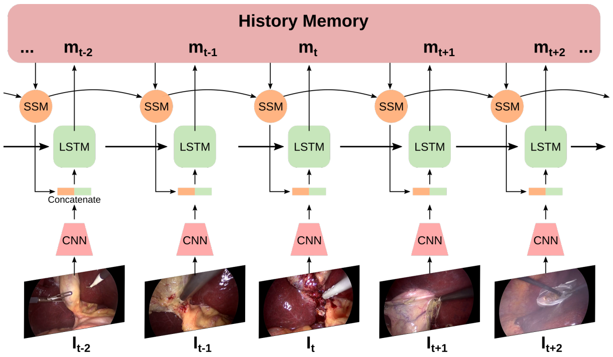


Fig. 1. SSM network architecture. Information from the network phase estimation head is processed as a multi-channel temporal signal. The resulting statistics are concatenated with the visual embedding and passed to the LSTM.

Recently, several approaches tried to incorporate long-term temporal information by expanding the receptive field of traditional architectures. [23] used a non-local operation block on top of a CNN-LSTM framework to capture the long-range temporal dependencies. It acts similarly to a temporal attention network, to compute the similarity between the features of past and current frames. The past features are then linked to the current frame by using skip connections. [40] proposed a Multi-Stage Temporal Convolutional Network (MS-TCN) that performs hierarchical prediction refinement for surgical phase recognition. The dilated convolutions used in the model allow for a larger receptive field. Instead of increasing the receptive field, our approach summarizes the global information from the surgery, as we aggregate main statistical features from the beginning of the surgery, as a side channel.

III. Methods

We introduce an architecture to better leverage long-term temporal information in surgical phase recognition via approximate sufficient statistic features. We then proceed to detail a set of approximate sufficient statistic features included within the proposed architecture.

A. Model Architecture

Surgical phase recognition attempts to classify the correct surgical phase label given video frames $\mathbf{I}_t, t = 0 \dots T$. We denote the ground truth label for frame $\mathbf{I}_t$ by $y_t \in 1 \dots N$,

Algorithm 1: Forward estimate of surgery phase y_{t+1}.

```
input : Past LSTM hidden states $L_0 \cdot L_t$, past
        sufficient statistic $s_0, \cdots, s_t$, next frame $I_{t+1}$
output: next frame hidden state $L_{t+1}$, next frame
        sufficient statistic $s_{t+1}$, next frame phase
        estimate.
Compute visual model output $f_V(I_{t+1})$;
Compute LSTM updated state $L_{t+1}$ given
    $f_V(I_{t+1}), s_t, L_t$;
Compute current phase estimate $y_{t+1}$ from $L_{t+1}$;
Compute frame statistics $s_F(L_{t+1})$;
Compute new sufficient statistics $s_{t+1}$ given
    $s_F(L_{t+1})$ and $s_t$.
```

where N is the number of different surgical phases. We process individual frames via a CNN visual model (based on a ResNet module[41]), encoding the visual content as a single vector $\boldsymbol{v}_t$, which is then fed to the LSTM, forming a standard CNN-LSTM structure.

In analysis of temporal processes, recurrent neural networks such as LSTMs perform well when inference relies mainly on recent information. However, performance suffers when long-term temporal information is required for inference. To address the lack of long-term information, dilated convolutions [42] have been suggested, but they fail to leverage several phenomena involved in the interpretation of surgeries:

1) Correct classification of phase transitions depends on propagation of low-dimensional information that coincides with the actual phases being detected.
2) Short events that are from the distant past can significantly affect interpretation of the current observations. (e.g. clearing the cystic plate is a visually identical task necessary in both "Dissection of Calot's Triangle" and "Removal of Gallbladder from Liver Bed" phases. The correct phase is identified from prior knowledge that the cystic structures have already been clipped and divided).
3) Some of the temporal evidence collection occurs over a long period of time (consider priors about the length of each phase).

While extracting a perfect sufficient statistic of the past is hard due to estimator dimensionality and the need to capture uncertainty about phase and perception limitations, the above phenomena make it possible to define a family of approximate sufficient statistics that can be computed from the data based on temporal aggregation of some transformation of the LSTM hidden state L_t. This makes it easier for the network to do both short-range temporal reasoning (such as change detection and visual processing), as well as medium- and long-range reasoning (such as counting past frames of each phase). The overall approach is presented in Algorithm 1 and illustrated in Figure 1.

Our architecture takes the past hidden LSTM layer, and passes it through a transform (the phase recognition module), to get a vector m_t, conceptualized as a temporal vector signal $\mathbf{M}_t = \{\boldsymbol{m}_1 \ldots \boldsymbol{m}_t\}$. It then computes aggregate statistics of the transformed signal, resulting in a sufficient statistics feature stream $\mathbf{S} = \{\boldsymbol{s}_1 \ldots \boldsymbol{s}_t\}$. By concatenating $\boldsymbol{v}_t$, it then feeds them to the current time phase LSTM inference as an augmented feature $\boldsymbol{c}_t$. After concatenation, an LSTM is applied taking $\boldsymbol{c}_t$ as input to output the likelihood for each phase. Note that for both training and testing, the history memory $\mathbf{M}_t$ is initialized with zeros.

We note that several existing models fall within the family of functions described by this model, including temporal causal convolution (TCN) networks, PKI[24], and LSTMs. Furthermore, several novel sufficient statistic features are detailed in Section III-B. The LSTM output space captures an approximate sufficient statistic on the past. The SSM module extracts from past information a reduced set of approximate sufficient statistics to make inference in the current time-point more efficient. Tailoring the choice of sufficient statistics can make it much easier for the network to learn specific dependencies and cues. While we describe this family regardless of resource constraints, in practice, many of the SSM features can be calculated in either $O(1)$ for computing-only or $O(1)$ for both computing and memory, as we will demonstrate.

B. Sufficient Statistic Features

Different choices of summarization S can make it easy for the network to learn long-term interactions. A few of the approaches explored in our experiments are described below.

1) Hidden Markov Model (HMM): HMM is a well-known statistical method for temporal filtering of discrete states. It is thus intuitive to use HMM as a feature to encode temporal information. There are two main advantages of using HMM: 1) it provides smoother inference results, which provides an additional timescale of reasoning; 2) it can filter out impossible phase transitions using a state transition matrix, discouraging illogical phase transition inferences.

2) Cumulative Sum Likelihood (CSL): Temporal information can also be propagated by accumulating the LSTM inference likelihood in time:

$$\boldsymbol{f}_t = \log(\sum_{t'=1}^{t}(\mathcal{I}_l(\boldsymbol{m}_{t'})) + 1), \tag{1}$$

where $\boldsymbol{f}_t$ represents the CSL at time step t, and $\mathcal{I}_l$ represents thresholding of the elements of $\boldsymbol{m}$ with a set of threshold levels l, with respect to the maximum probability phase at time t. The CSL feature enhances the network's understanding of some global contexts and allows the network to capture both maximum-probability and probable interpretation of the phase at time t. It should have the capability to answer the question "where we are" in a surgery, including if certain phases have or have not already occurred. e.g. CSL can

	MGH100		**Cholec80**
Index	Phase Name	Description	Phase Name
0	Port placement	Placement of ports for tool access to abdominal cavity	Preparation
1	Fundus retraction	Retraction of the GB fundus in preparation for dissection	Calot Triangle Dissection
2	Release GB peritoneum	Dissection of peritoneal lining of GB infundibulum	Clipping and Cutting
3	Dissection of Calot's triangle	Dissection of the hepatocystic triangle to expose the cystic duct & artery	Gallbladder Dissection
4	Checkpoint 1	Inactive period prior to the first clipping of a structure	Gallbladder Packaging
5	Clip Cystic Artery	Application of surgical clips to the cystic artery	Cleaning and Coagulation
6	Divide Cystic Artery	Division of the cystic artery	Gallbladder Retraction
7	Clip Cystic Duct	Application of surgical clips to the cystic duct	-
8	Divide Cystic Duct	Division of the cystic duct	-
9	Checkpoint 2	Inactive period between the end of the division & the start of GB removal	-
10	Remove GB from liver bed	Dissection of the GB from the liver	-
11	Bagging	Placing the GB in a laparoscopic retrieval bag	-
12	Other step	Any other undefined step (free text annotation)	-

TABLE I

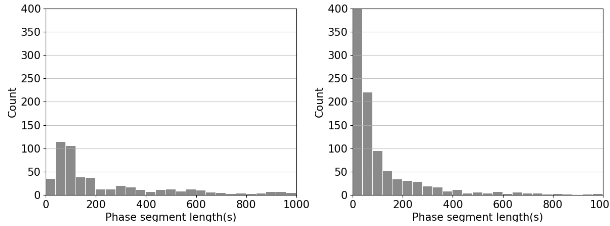
Left: Phases in MGH100 and corresponding phase descriptions Right: Phases in Cholec80

indicate that the phase "Divide Cystic Duct" has already been achieved; since it is a non-repeated event, we know that future frames cannot be classified as such.

3) Wavelets Transform: To capture temporal events at various time scales, we used a wavelet transform to summarize temporal information. We chose Gabor filter as a standard wavelet decomposition [43] which is defined as the modulation product of a Gaussian envelope and a complex sinusoidal wave, where the 1-dimensional Gabor filter can be formally written as: We collect a filter bank with Gabor filters of different Gaussian envelope sizes to directly apply to the likelihood space along the time axis. The filtered results are then concatenated to gather the temporal information of different time scales.

4) Causal vs. acausal features: As the information can be propagated through time both in forward and backward directions, each of the features above can be built either in a causal manner or acausal manner, leveraging information from the future signal. We show the acausal SSM features as a proof of concept for offline analysis purposes [7]. For certain phases (e.g. checkpoint 1 in MGH100 dataset), the inference of such phases may benefit from the future information.

IV. Experiments



(a) Cholec80 (b) MGH100

Fig. 2. Phase segment length distribution in the different datasets used in our experiments (x axis in seconds). Our dataset with phases as shown in Table 1 has a much larger distribution of small segments due to fragmentation of the phases of the surgery.

In this section, we introduce the details of the experimental setups, in addition to the results and discussions.

1) Datasets: We evaluated our method on two datasets:

Cholec80 dataset[7] contains 80 cholecystectomy videos each with 25 frames per second (fps). It provides annotations of tool presence and surgical phases (Table I). The dataset is divided into a 40-40 split for training and testing.

MGH100 dataset contains 100 cholecystectomy videos each with sampling rate of 30 fps. The phases in MGH100 (see Table I) are more granular and clinically meaningful compared to Cholec80 (e.g., separating the broad category of "clipping and cutting" into separate tasks of clipping/cutting specific structures). We also added Checkpoints 1 and 2 to capture the decision points prior to clipping and removal of gallbladder, respectively. 80 videos are used for training while 20 are used for testing.

Prior work in surgical phase recognition has largely utilized public datasets such as Cholec80, which is annotated into long, visually distinct phases with linear progression. Clinically-meaningful phases, however, are often visually indistinct, of variable length, non-linearly progressing, repeating, and may be influenced by prior phases over the short and long-term – such characteristics are reflected in the annotation structure of phases in MGH100 (Table I). These phases can be considered clinically actionable (i.e. phases at which influencing a surgeon's actions could modify risk of complications) and align more closely with surgical decision-making [8].

2) Evaluation Metrics: We use standard metrics to evaluate our algorithms' overall performance. The phased-averaged recall and precision, and F1-score across phases, in addition total video per-frame accuracy (frames correctly inferred/total frames) [44] are used.

3) Model Parameters: We down-sampled all the videos to 1 fps. We kept parameters identical for the SSM-LSTM and LSTM for a fair comparison. The LSTM hidden state was 64-dimensions. During training, batch size was set to 32. An Adam [45] optimizer was applied with a learning rate of 0.0025. The number of training epochs was 20. We first trained the CNN model for 20 epochs then fine-tuned the CNN model during temporal model training. For Gabor feature calculation, 10 different scales, σ, ranging from 10 to 30 frames, were applied. After that, features of different sizes were concatenated together.

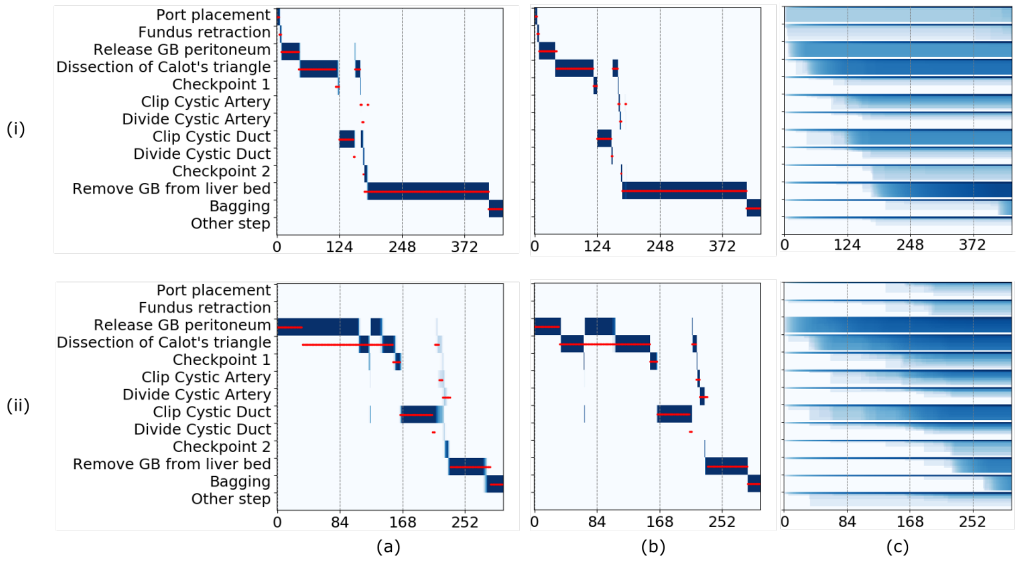


Fig. 3. Example result in MGH100 dataset (i): video 6, (ii): video 17. The x-axis is the time in seconds (s); y-axis rows represent a phase. The red line indicates the Ground truth trajectories. (a) LSTM (b) SSM-LSTM. The blue bar indicates the estimation of the algorithm, with saturation proportional to certainty. (c) Cumulative Sum Likelihood (CSL) SSM features. The saturation of the blue bar indicates the accumulated values. In video 6, thanks to the SSM features, SSM-LSTM is able to accurately detect 'Clip/Divide Cystic Artery/Duct'. At the moment of 'Clip Cystic Artery', from (c) CSL feature the, 'Clip Cystic Duct' has already accomplished'. The current phase is therefore high likely to be 'Clip Cystic Artery'. Similar phenomenons in (ii).

4) Training Strategy: To facilitate memory efficiency during training, we detached the computation of the sufficient statistics from its effect on recent past analysis via the LSTM, which was trained with temporal sequences of length 8. We numerically stabilized convergence via a lagged-step iteration, encouraging outputs to close to the previous epoch [46].

5) Results: The experiment results are shown both quantitatively and qualitatively. The results for Cholec80 dataset are in Table II, benchmarking the SSM-LSTM model against several state-of-the-art models. Despite only leveraging phase labels without incorporation of other features (tool, kinematics) like *EndoNet* [7] or *MTRCNet-CL* [12], our proposed SSM-LSTM model has the best accuracy of 90.0% among the different models, with similar performance across precision, recall and F1 score. The result of our proposed method is then followed by an HMM for further smoothing [7]. The results of combining causal and acausal features are shown as a demonstration of offline applications, to gauge the effect of acausal information.

The results on the MGH100 dataset are in Table III. LSTM is a baseline of CNN-LSTM structure similar to [24]. An ablation study of the proposed model using individual SSM features (e.g. Gabor, CSL) is also shown. The model's accuracy and F1 score benefited from multiple SSM features. With a combination of different SSM features, denoted as "SSM", the proposed model achieves the best performance in all the four metrics, significantly improving on LSTM.

We also evaluated the performance of the model with phases of different lengths (Table IV). We noticed the LSTM has a substantial performance drop in short phases ($<30s$). However our SSM features helped ameliorate in such cases (cf. Accuracy). Our approach was able to improve significantly performance on short phases in this more challenging dataset, since the SSM provides additional information on the workflow structure. We also show the curve of the accuracy of the phases of different lengths in Fig 4, where we can see the SSM-LSTM (Purple) is significantly better than (LSTM) in short phases.

We qualitatively show how applying SSM features assists the model to understand the temporal structure of the surgery in Fig 3. For both examples, the SSM-LSTM model can accurately detect the short phases "Clip Cystic Artery" and "Clip Cystic Duct." which are more clinically meaningful phase labels, the ground truth labels included alternating patterns, which differ from traditional linear workflows [7], [15], [9], [18], [10]. They are two phases which are hard to distinguish based only on the visual model, since the same surgical tool (clip applier) is used. Moreover, artery and duct are spatially close to each other, making the problem even harder. We can see from Fig 3 (i-b) that SSM-LSTM has a better performance. That is due to the fact when the first 'Clip Cystic Duct' has accomplished, the network gets the 'Duct has been clipped' information from the SSM feature. Therefore, when the model sees the clip applier appear the second time in the video, it is high likely the surgeon is going to "Clip Cystic Artery". Contrast this with the LSTM, which, lacking long-term context, was not able to identify the two phases correctly. Similar phenomena are observed in video 17 in Fig 3 (ii).

We also analyzed SSM-LSTM performance on the individual phases, shown in Fig. 5. On the MGH100 dataset, the algorithm had good performance on long phases such as "Release GB Peritoneum" and "Dissection of Calot's Triangle", with accuracy over 90%. Short phase performance was worse, as some of the short phases are likely harder to

	Model	Tool detection	Accuracy	Precision	Recall	F1 Score
Online	Binary Tool[12]	●	47.5	54.4	60.2	57.2
	EndoNet[7]	●	81.7	73.7	79.6	76.5
	SV-RCNet[24]	-	85.3	80.7	83.5	82.1
	MTRCNet-CL[12]	●	89.2	86.9	88.0	87.4
	SSM-LSTM (Proposed)	-	90.0	87.0	83.0	84.9
Offline	Causal + Acausal SSM-LSTM (Proposed)	-	90.8	85.3	82.7	84.0

TABLE II

MODEL PERFORMANCE ON CHOLEC80 DATASET. ● DENOTES JOINT TRAINING OF TOOLS AND PHASES. OUR METHOD OUTPERFORMS PHASE-ONLY APPROACHES (-), AND REACHES A SLIGHTLY BETTER ACCURACY THAN TOOI

	Models	Accuracy	Precision	Recall	F1 Score
Online	LSTM	83.3	50.2	51.6	50.9
	Gabor	84.1	51.2	57.4	54.1
	CSL	85.4	58.8	60.0	59.4
	SSM	85.6	59.4	61.5	60.4
Offline	Acausal SSM	86.8	61.9	65.4	63.6

TABLE III

MGH100 DATASET RESULTS. (SSM: THE COMBINATION OF ALL CAUSAL SSM FEATURES). USING SSM FEATURES, WE IMPROVE BOTH ACCURACY AND F1 PERFORMANCE.

	Models	1-3s	4-10s	11 -30s	31 - 60s	>60s
Online	LSTM	12.5	40.6	47.0	63.1	91.7
	Gabor	15.6	44.1	53.4	66.4	90.2
	CSL	20.3	45.6	54.7	63.5	93.0
	SSM	31.2	51.6	58.2	64.4	93.0
Offline	Acausal SSM	35.9	52.4	64.3	64.7	93.0

TABLE IV

ACCURACY FOR DIFFERENT PHASE DURATIONS. OUR ALGORITHM SIGNIFICANTLY IMPROVES SHORT AND CHALLENGING PHASE SEGMENTS (DURATION $<$ 30S).

infer due to the lack of data variability (e.g. 37% of accuracy for phase checkpoint 2). However, with SSM, short phase performance exceeded that of LSTM, shown in Table IV.

In addition, we evaluated the accuracy of phase transitions and phase midpoints as per-segment statistics. For transition accuracy, a transition estimated within 10s of ground truth was considered correct. The SSM-LSTM achieved transition accuracy of 48.1% vs. LSTM at 39.0%. The accuracy for phase midpoint was 63.69% for SSM-LSTM vs. 56.23% for LSTM. The performance benefits from the SSM module for both phase and transition inference, likely due to a better inference of phase start/end points and phase duration.

V. CONCLUSIONS

In this paper we propose a novel SSM-LSTM model that aggregates temporal information to augment LSTM inference. The proposed model is validated on two large surgery datasets, Cholec80 and MGH100, surpassing state-of-the-art performance on a more clinically relevant and harder taxonomy. We demonstrate the advantage of the proposed model over the existing methods in several clinically important scenarios. Using different SSM features, the model benefits from complementary approaches for temporal analysis, and improves understanding long-range, clinically relevant temporal interactions in surgical workflows.

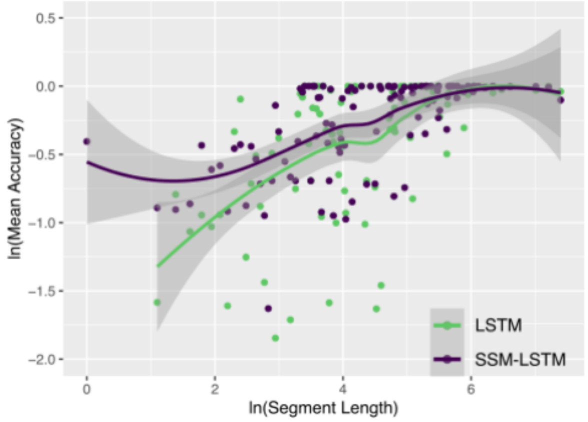


Fig. 4. Accuracy of phase segments in different length in MGH100 dataset. (plot in log space) Green: LSTM; Purple, SSM-LSTM. Our approach consistently achieves better results compared to LSTM approach, especially for short and alternating annotation segments.

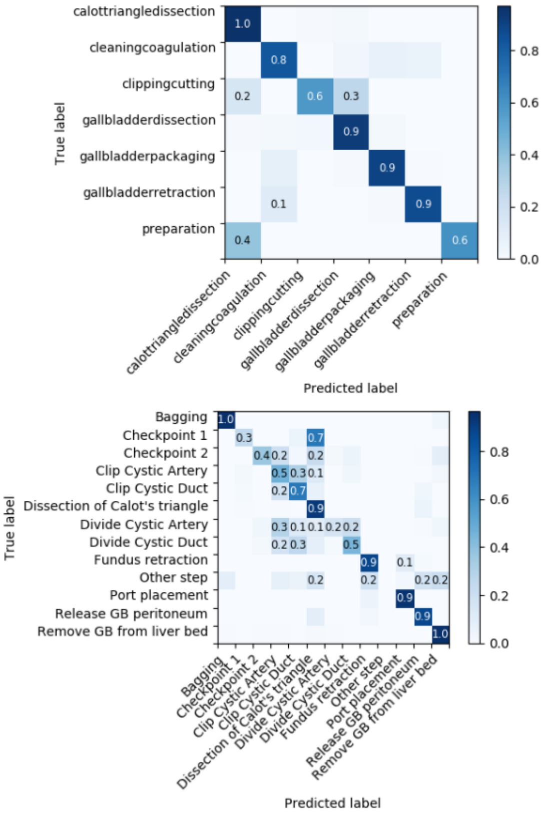


Fig. 5. Confusion matrix for each phase on Cholec80 dataset (up), MGH100 dataset (down). The proposed method has good performance on Cholec80 phases. On MGH100, the model performs well on long phases (e.g. Release GB Peritoneum), but for the phases which are short and cannot be assisted by temporal cues, the performance are not as good. e.g. Checkpoint 1.

References

[1] J. Lin, C. Gan, K. Wang, and S. Han, "Tsm: Temporal shift module for efficient and scalable video understanding on edge devices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[2] X. Long, G. De Melo, D. He, F. Li, Z. Chi, S. Wen, and C. Gan, "Purely attention based local feature integration for video classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[3] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *CVPR*, 2015, pp. 2568–2577.

[4] Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horaud, "Tracking multiple persons based on a variational bayesian model," in *European Conference on Computer Vision*. Springer, 2016, pp. 52–67.

[5] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational bayesian inference for audio-visual tracking of multiple speakers," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[6] Y. Xu, A. Osep, Y. Ban, R. Horaud, L. Leal-Taixé, and X. Alameda-Pineda, "How to train your deep multi-object tracker," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6787–6796.

[7] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[8] D. A. Hashimoto, C. G. Axelsson, C. B. Jones, R. Phitayakorn, E. Petrusa, S. K. McKinley, D. Gee, and C. Pugh, "Surgical procedural map scoring for decision-making in laparoscopic cholecystectomy," *The American Journal of Surgery*, vol. 217, no. 2, pp. 356–361, 2019.

[9] D. Kitaguchi, N. Takeshita, H. Matsuzaki, H. Takano, Y. Owada, T. Enomoto, T. Oda, H. Miura, T. Yamanashi, M. Watanabe, D. Sato, Y. Sugomori, S. Hara, and M. Ito, "Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach," *Surgical Endoscopy*, Dec. 2019.

[10] O. Zisimopoulos, E. Flouty, I. Luengo, P. Giataganas, J. Nehme, A. Chow, and D. Stoyanov, "DeepPhase: Surgical Phase Recognition in CATARACTS Videos," *arXiv:1807.10565 [cs, stat]*, July 2018.

[11] F. Lalys, L. Riffaud, X. Morandi, and P. Jannin, "Surgical phases detection from microscope videos by combining SVM and HMM," in *International MICCAI Workshop on Medical Computer Vision*. Springer, 2010, pp. 54–62.

[12] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C.-W. Fu, and P.-A. Heng, "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Medical image analysis*, vol. 59, p. 101572, 2020.

[13] Y. Qin, S. A. Pedram, S. Feyzabadi, M. Allan, A. J. McLeod, J. W. Burdick, and M. Azizian, "Temporal segmentation of surgical subtasks through deep learning with multiple data sources," *arXiv preprint arXiv:2002.02921*, 2020.

[14] T. M. Ward, P. Mascagni, Y. Ban, G. Rosman, N. Padoy, O. Meireles, and D. A. Hashimoto, "Computer vision in surgery," *Surgery*, 2020.

[15] M. Volkov, D. A. Hashimoto, G. Rosman, O. R. Meireles, and D. Rus, "Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 754–759.

[16] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] D. A. Hashimoto, G. Rosman, E. R. Witkowski, C. Stafford, A. J. Navarette-Welton, D. W. Rattner, K. D. Lillemoe, D. L. Rus, and O. R. Meireles, "Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy," *Annals of surgery*, vol. 270, no. 3, pp. 414–421, 2019.

[19] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy, "Deep neural networks predict remaining surgery duration from cholecystectomy videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 586–593.

[20] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 1069–1078, 2018.

[21] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] X. Shi, Y. Jin, Q. Dou, and P.-A. Heng, "LRTD: Long-range temporal dependency based active learning for surgical workflow recognition," *arXiv preprint arXiv:2004.09845*, 2020.

[24] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.-W. Fu, and P.-A. Heng, "SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network," *IEEE transactions on medical imaging*, vol. 37, no. 5, pp. 1114–1126, 2017.

[25] Z. Zhao, T. Cai, F. Chang, and X. Cheng, "Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade," *Healthcare Technology Letters*, vol. 6, no. 6, pp. 275–279, 2019.

[26] Y. Qin, S. Feyzabadi, M. Allan, J. W. Burdick, and M. Azizian, "davincinet: Joint prediction of motion and surgical state in robot-assisted surgery," *arXiv preprint arXiv:2009.11937*, 2020.

[27] C.-A. Saint-Pierre, J. Boisvert, G. Grimard, and F. Cheriet, "Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images," *Machine Vision and Applications*, vol. 22, no. 1, pp. 171–180, 2011.

[28] L. C. Garcia-Peraza-Herrera, W. Li, L. Fidon, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren, *et al.*, "Toolnet: holistically-nested real-time segmentation of robotic surgical tools," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5717–5722.

[29] Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 440–448.

[30] Z. Zhao, Y. Jin, X. Gao, Q. Dou, and P.-A. Heng, "Learning motion flows for semi-supervised instrument segmentation from robotic surgical video," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 679–689.

[31] D. Rivoir, S. Bodenstedt, I. Funke, F. von Bechtolsheim, M. Distler, J. Weitz, and S. Speidel, "Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 752–762.

[32] Y.-Y. Tsai, B. Huang, Y. Guo, and G.-Z. Yang, "Transfer learning for surgical task segmentation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9166–9172.

[33] S. Kannan, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Future-state predicting LSTM for early surgery type recognition," *IEEE Transactions on Medical Imaging*, vol. 39, no. 3, pp. 556–566, 2019.

[34] O. Zisimopoulos, E. Flouty, I. Luengo, P. Giataganas, J. Nehme, A. Chow, and D. Stoyanov, "Deepphase: surgical phase recognition in cataracts videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 265–272.

[35] T. M. Ward, D. A. Hashimoto, Y. Ban, D. W. Rattner, H. Inoue, K. D. Lillemoe, D. L. Rus, G. Rosman, and O. R. Meireles, "Automated operative phase identification in peroral endoscopic myotomy," *Surgical Endoscopy*, pp. 1–8, 2020.

[36] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[37] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, 2013, pp. 1310–1318.

[38] D. Liu and T. Jiang, "Deep reinforcement learning for surgical gesture segmentation and classification," in *International conference on med-*

ical image computing and computer-assisted intervention. Springer, 2018, pp. 247–255.

[39] X. Gao, Y. Jin, Q. Dou, and P.-A. Heng, "Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search," *arXiv preprint arXiv:2002.08718*, 2020.

[40] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim, and N. Navab, "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks," *arXiv preprint arXiv:2003.10751*, 2020.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.07122

[43] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.

[44] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Medical image analysis*, vol. 16, no. 3, pp. 632–641, 2012.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *preprint arXiv:1412.6980*, 2014.

[46] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.