Manish Sahu*, Angelika Szengel, Anirban Mukhopadhyay and Stefan Zachow

Surgical phase recognition by learning phase transitions

<https://doi.org/10.1515/cdbme-2020-0037>

Abstract: Automatic recognition of surgical phases is an important component for developing an intra-operative context-aware system. Prior work in this area focuses on recognizing short-term tool usage patterns within surgical phases. However, the difference between intra- and inter-phase tool usage patterns has not been investigated for automatic phase recognition. We developed a Recurrent Neural Network (RNN), in particular a state-preserving Long Short Term Memory (LSTM) architecture to utilize the long-term evolution of tool usage within complete surgical procedures. For fully automatic tool presence detection from surgical video frames, a Convolutional Neural Network (CNN) based architecture namely ZIBNet is employed. Our proposed approach outperformed EndoNet by 8.1% on overall precision for phase detection tasks and 12.5% on meanAP for tool recognition tasks.

Keywords: CNN; endoscopic videos; RNN; state-preserving LSTM; surgical workflow; surgical phase recognition.

Introduction

Decomposing a surgical process into a sequence of abstract tasks independent of the physical factors was introduced by MacKenzie et al. [1] under the term Surgical Process Modeling (SPM). SPM centers around a concept of multi-scale temporal abstraction – termed as *granularity* [2]. In SPM, the highest *granularity* level is called surgical phase. A fully automatic recognition of such phases can enable multiple applications [3, 4]. Understanding the temporal *evolution* of tool usage patterns and discovering their relation to respective surgical phases can provide important cues for an automatic recognition of surgical phases [5]. Henceforth, by the term *evolution*, we will refer to the sequential process of surgical tool usage (including

co-occurrence) over a complete surgical procedure. The term *pattern* refers to the visibility of tools within endoscopic images.

The general trend in literature on fully automatic surgical phase recognition is to explicitly model a low level SPM abstraction followed by a global temporal alignment, using some variant of Hidden Markov Models (HMM). For example, Padoy et al. [5] and Twinanda et al. [6] modeled tool presence information using Dynamic Time Warping (DTW) and CNN respectively, followed by a hierarchical HMM. Dergachyova et al. [7] used image and instrument signals for modeling low level information, followed by a Hidden semi Markov Model for global alignment. A spatio-temporal CNN was introduced by Lea et al. [8] for modeling tool motion followed by a semi Markov model or DTW. LSTM has been introduced by DiPietro et al. [9] for surgical phase recognition from kinematics data and more recently utilized by Jin et al. [10] in combination with low-level image features (obtained from a CNN) and heuristic post processing.

Key observations of the present work are the inter- and intra-phase differences in tool usage evolution, as shown in Figure 1. Unlike prior methods, which solely learn *within-phase* information, we learn key changes for sequences of tool usage that uniquely identify phase transitions in an endoscopic video sequence. In its essence, the proposed method utilizes tool presence prediction of ZIBNet [11] in a state-preserving LSTM [12] framework for encoding complete evolution of tool usage during a surgical procedure.

The major contribution of the proposed work is a novel framework for learning evolution of surgical tool usage patterns in relation to inter- and intra-phases, where many-to-one LSTM sequence learning is utilized for fully automatic learning of long-term tool usage evolution.

Method

A set of video sequences, each composed of individual frames $\{I_t\}_{t=1}^T$ in combination with ground truth annotations of corresponding phases $\{y_t\}_{t=1}^T$ are given as input for training. The goal of fully automatic phase recognition is to learn a mapping scheme from I_t to y_t . However, instead of

*Corresponding author: Manish Sahu, Zuse Institute Berlin, Berlin, Germany, E-mail: sahu@zib.de

Angelika Szengel, 1000shapes GmbH, Berlin, Germany

Anirban Mukhopadhyay, TU Darmstadt, Darmstadt, Germany

Stefan Zachow, Zuse Institute Berlin, Berlin, Germany

Figure 1: Corrplot visualization of the evolution of surgical tool usage (including tool co-occurrences) at different phases of a surgery as provided with the *Cholec80* dataset.

predicting phase $\hat{y}_t$ directly from I_t , we use *ZIBNet* [11] to automatically detect tools associated with I_t and train an LSTM on top of this tool presence information, as shown in Figure 2. To keep this article self-contained, we will briefly introduce *ZIBNet* in the next section, before describing the state-preserving LSTM methodology.

ZIBNet

Sahu et al. introduced *ZIBNet* as a specialized transfer learning framework developed on top of generic CNN feature learning architectures for surgical tool presence detection [13, 11]. *ZIBNet* considers surgical tool presence detection as a multi-label classification task and tool co-occurrences (second and third order) are treated as separate classes. A detrimental effect of imbalance in tool usage on the performance of the CNN was analyzed. A stratification technique to counter the imbalance was employed

during CNN training. Moreover, online post-processing using temporal smoothing was introduced to enhance run time prediction. In contrast to [11] which only considered AlexNet [14] as the base CNN architecture, we also investigate Residual Neural Networks (ResNet [15]) trained for ImageNet classification as the base CNN architecture of *ZIBNet*.

Considering a number of L tools are necessary to perform a surgical procedure, *ZIBNet* learns the mapping between I_t and the ground truth tool labels. During testing, given a video frame I_t , *ZIBNet* predicts the probability of a tool $x_t \in \mathbb{R}^L$ present in the frame.

State-preserving LSTM

The goal of an LSTM is to learn the mapping between tool labels x_t generated by *ZIBNet* and the corresponding phase label y_t . The main reason behind considering an LSTM unit is their built in memory states, for which the LSTM learns to store, update and output information [12].

Typical LSTM architectures map an input with fixed dimensionality to an output vector. However, this setup offers a serious limitation for online phase detection since the length of a surgical procedure is not known a-priori. One key design choice of our particular learning strategy is to focus on learning phase *transitions* rather than the overall description of phases. In this work, we formulate the online phase detection as a *many-to-one LSTM sequence learning task*. Figure 2 visualizes the pipeline of the proposed method. In particular, the input sequence of tool predictions is fed into the LSTM, one mini-sequence at a time, and the respective state of the LSTM is preserved for the following mini-sequence in order to retain long term dependency.

Figure 2: Overall design of the proposed surgical phase recognition framework.

Experiments and results

The performance of the proposed method was quantitatively evaluated in view of a fully automatic surgical phase

Table 1: Comparison of phase recognition results with other approaches on *Cholec80* dataset (PN $\rightarrow$ PhaseNet [6], EN $\rightarrow$ EndoNet [6] and EH $\rightarrow$ EndoNet + HHMM [6])

	PN	EN	EH	Proposed
Avg. precision	67.0	70.0	73.7	81.8
Avg. recall	63.4	66.0	79.6	80.9

recognition. In the following, the used data set, the experimental strategy, the quantitative analysis and the comparison with related state-of-the-art methods are described.

Data preparation and experimental settings

All our experiments were conducted with the *Cholec80* dataset [6] in an *online* setting (i.e. no future information was used for training). The *Cholec80* dataset comprises 80 videos of cholecystectomy procedures performed by 13 surgeons. The frame rate of 25 fps was down-sampled to 1 fps for a ground truth annotation of tool labels as well as for further processing. For SPM, a cholecystectomy procedure is divided into seven surgical phases, where seven tools are commonly used to perform the procedure (see Figure 1). We follow the evaluation strategy of Twinanda et al. [6] for both tool presence detection and phase recognition in order to provide direct comparison with their method.

The training data was created by converting the video into mini-sequences of five frames while maintaining the original sequential order. We utilized eight LSTM units, categorical cross entropy as loss function and Adam as optimization algorithm with a learning rate of 0.001.

Table 2: Comparison of tool recognition performance between EndoNet [6] (EN) and two flavors of ZIBNet: ZIBNet-AlexNet (ZA) and ZIBNet-ResNet (ZR) on the *Cholec80* dataset.

Tools	EN	ZA	ZR
Grasper	84.8	86.6	90.5
Bipolar	86.9	94.8	95.6
Hook	95.6	98.8	99.3
Scissors	58.6	68.8	86.8
Clipper	80.1	85.0	96.5
Irrigator	74.4	77.2	89.7
Specimen bag	86.8	88.3	96.1
MeanAP	81.0	85.7	93.5

Quantitative comparison

Our proposed framework outperforms EndoNet [6] on phase recognition evaluation metrics as developed by Padoy et al. [5], namely average precision and average recall. Results of top performing methods on the *Cholec80* dataset from Twinanda et al. [6] are listed together with our results in Table 1. In particular, the performance of our proposed method is compared to those of PhaseNet, EndoNet and EndoNet followed by HHMM (EndoNet + HHMM) based global alignment. Our proposed method leads to approx. 8% improvement in terms of average precision, compared to the second best result (EndoNet + HHMM).

ZIBNet performance for tool detection

Due to the dependence of our proposed method on tool presence, it is important to employ an accurate and reliable method for the detection of tool presence right at the outset. To this end, we compared the performance of two state-of-the-art tool detection methods on the *Cholec80* dataset using Average Precision (AP) as a metric. We considered two flavors of ZIBNet [11] with base CNN as either AlexNet [14] or ResNet [15], reported as ZIBNet-AlexNet and ZIBNet-ResNet respectively. It becomes evident from Table 2 that both versions of ZIBNet outperformed EndoNet [6] in terms of Mean Average Precision (MeanAP) for all tools. However, the ResNet flavor beats the two other methods in each tool category and achieved an overall MeanAP of 93.5. An interesting observation is that the less frequent tools like scissors, irrigators, specimen bags etc. are more related to phase transitions. The general superior performance of ZIBNet over EndoNet for these tools (for example, an increase by approx. 30% for scissors), as shown in Table 2, prompted us to choose ZIBNet

Table 3: Analysis and comparison of recognition performance for each phase using Precision(Recall) on the *Cholec80* dataset (EH $\rightarrow$ EndoNet + HHMM [6])

Phases	EH	Proposed
Preparation	90.0 (85.5)	99.1 (87.6)
Calot triangle dissection	96.4 (81.1)	98.5 (85.3)
Clipping and cutting	69.8 (71.2)	73.6 (68.8)
Gall bladder dissection	82.8 (86.5)	81.6 (92.7)
Gall bladder packaging	55.5 (75.5)	83.0 (83.4)
Cleaning and coagulation	63.9 (68.7)	64.7 (77.4)
Gall bladder retraction	57.5 (88.9)	71.9 (68.1)
Average	73.7 (79.6)	81.8 (80.9)

with ResNet [15] for initial fully automatic tool presence detection. Note that in Tables 1 and 3 the performance of ResNet flavor of *ZIBNet* is reported in the columns entitled *Proposed*.

Quantitative analysis

We quantitatively analyzed the specific design choices of our proposed method and its performance on each of the seven surgical phases. Precision and recall are considered as the metric for measuring the performance of fully automatic phase recognition. We specified the combined performance of EndoNet and HHMM (reported as EndoNet + HHMM) for each surgical phase in Table 3. Our proposed method outperformed EndoNet in terms of precision and recall.

Our proposed framework is implemented in Keras-Theano, running at 6 fps on an NVIDIA GTX 1080 Ti when applied on a surgical video.

Discussions and conclusion

We presented a fully automatic technique for the recognition of surgical phases from endoscopic videos of cholecystectomy procedures. Unlike prior work, we focused on differentiating between evolution of tool usage at phase transitions and within phases. In particular, a many-to-one state-preserving LSTM was trained on the tool presence predictions of *ZIBNet* to learn evolution of tool transition patterns from endoscopic videos. The proposed method outperformed EndoNet methods on the *Cholec80* dataset by 8.1% on average precision. This study solely concentrated on tool presence as the initial source of information. However, future studies of surgical phase recognition from endoscopic videos may benefit from studying tool localization on the image context as the initial source. Finally, such fully automatic techniques are expected to be instrumental in advancing the computer assistance during surgical intervention from bench to bedside.

Research funding: This study was funded by the German Federal Ministry of Education and Research (BMBF) under the project COMPASS (grant no. - 16 SV 8019).

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Informed consent: This study contains patient data from a publicly available dataset.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. MacKenzie L, Ibbotson J, Cao C, Lomax A. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. *Minim Invasive Ther Allied Technol* 2001;10:121–7.
2. Lalys F, Jannin P. Surgical process modelling: a review. *IJCARS* 2014;9:495–511.
3. Blum T, Feußner H, Navab N. Modeling and segmentation of surgical workflow from laparoscopic video. In: *MICCAI* Springer; 2010, pp. 400–7.
4. Franke S, Meixensberger J, Neumuth T. Multi-perspective workflow modeling for online surgical situation models. *J Biomed Inf* 2015;54:158–66.
5. Padoy N, Blum T, Ahmadi SA, Feussner H, Berger MO, Navab N. Statistical modeling and recognition of surgical workflow. *Med Image Anal* 2012;16:632–41.
6. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. In: *IEEE TMI* 2017;36:86–97.
7. Dergachyova O, Bouget D, Huauilmé A, Morandi X, Jannin P. Automatic data-driven real-time segmentation and recognition of surgical workflow. *IJCARS* 2016;11:1081–9.
8. Lea C, Choi JH, Reiter A, Hager GD. Surgical phase recognition: from instrumented ORs to hospitals around the world. *M2CAI workshop, MICCAI*; 2016.
9. DiPietro R, Lea C, Malpani A, Ahmadi N, Vedula SS, Lee GI, et al. Recognizing surgical activities with recurrent neural networks. In: *MICCAI*, Springer; 2016, pp. 551–8.
10. Jin Y, Dou Q, Chen H, Yu L, Heng PA. SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. In: *IEEE TMI* 2018;37:1114–26.
11. Sahu M, Mukhopadhyay A, Szengel A, Zachow S. Addressing multi-label imbalance problem of surgical tool detection using CNN. *IJCARS* 2017;12:1013–20.
12. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9:1735–80.
13. Sahu M, Mukhopadhyay A, Szengel A, Zachow S. Tool and phase recognition using contextual CNN features. Tech report – M2CAI challenge. *MICCAI*; 2016.
14. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *NeurIPS*; 2012, pp. 1097–105.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *IEEE CVPR*; 2016, pp. 770–8.