Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-15 (2021) 334–339

A Deep Learning Framework for Recognising Surgical Phases in Laparoscopic Videos

Nour Aldeen Jalal*, Tamer Abdulbaki Alshirbaji, Paul D. Docherty***, Thomas Neumuth**** and Knut Moeller†**

**Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany; and Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany (Tel: +49 7720 3074638; e-mail: Nour.A.Jalal@hs-furtwangen.de).*

*** Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany; and Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany, (e-mail: abd@hs-furtwangen.de)*

**** Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand; and Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany, (e-mail: paul.docherty@canterbury.ac.nz)*

***** Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany, (e-mail: thomas.neumuth@uni-leipzig.de)*

† Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany, (e-mail: moe@hs-furtwangen.de)

Abstract: Image-based surgical phase recognition is a fundamental component for developing context-aware systems in future operating rooms (ORs) and thus enhance patient outcomes. To date, phase recognition in laparoscopic videos has been investigated, and spatio-temporal deep learning-based approaches have been introduced. However, phase recognition in laparoscopic videos is still a challenging task and requires ongoing research. In this work, a spatio-temporal deep learning approach for recognising surgical phases is proposed. The proposed framework consists of a convolutional neural network (CNN) and a cascade of three long short-term memory (LSTM) networks. The first and second LSTM networks were trained to learn temporal information from short video clips and the complete video sequence to perform tool detection. The last LSTM was employed to enforce temporal constraints of surgical phases. The proposed approach was thoroughly evaluated on the Cholec80 dataset, and the experimental results demonstrate the high recognition performance of this method.

Keywords: Surgical phase recognition, CNN, Laparoscopic video, LSTM, Context-aware system.

1. INTRODUCTION

The accelerating development of technologies for the operating room (OR) environment has increased the complexity of the surgical workflow where increasing rates of information need to be processed and interpreted by the surgical team. Therefore, a persistent need to develop context-aware systems (CASs) that are able to comprehensively analyse available data and communicate relevant information to human operators during surgeries has arisen (Lalys and Jannin 2014, Maier-Hein et al. 2017). Recognising surgical activities is an essential component of CASs in future ORs because it provides a consistent support to the medical staff and improves, therefore, patient safety (Lalys and Jannin 2014, Maier-Hein, et al. 2017, Padoy et al. 2012). In this context, different granularity levels for describing the surgical activities have been proposed in the literature under surgical process modelling terminology (Lalys and Jannin 2014). Surgical phases are the highest level and represent the main tasks performed during surgery. Indeed, automatic recognition of surgical phases as the surgeon is performing them is important for developing intelligent operating rooms that interact with the surgical team and promote a better awareness (Maier-Hein, et al. 2017). Moreover, by recognising surgical phases, the medical staff outside the OR is informed about the progress of the undergoing procedure and the schedule of the surgical department can be, therefore, optimised.

Surgical phase recognition is a two-fold problem where an adequate data source is firstly required, and then an effective methodology for performing the recognition must be developed. Initial studies have deployed sensor-based signals such as surgical tool usage signals obtained using radiofrequency identification (RFID) tags (Neumuth and Meißner 2012) to carry out the recognition (Stauder et al. 2014). For instance, Padoy et al. employed Hidden Markov Models (HMMs) to capture temporal information from tool binary usage signals (Padoy, et al. 2012). In addition, image-based approaches that made use of intraoperative videos such

10.1016/j.ifacol.2021.10.278

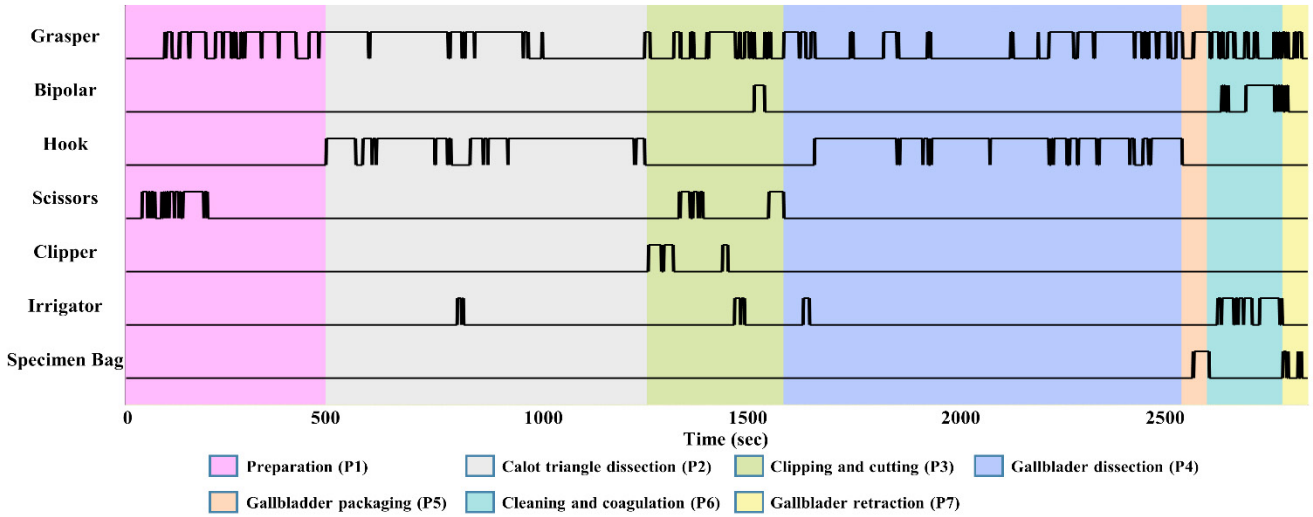

Figure 1. Visualisation of surgical tool usage in each surgical phase along the surgical procedure in the Chole80 dataset.

as laparoscopic video have also been introduced (Abdulbaki Alshirbaji et al. 2020, Dergachyova et al. 2016, Jin et al. 2017, Jin et al. 2020, Twinanda 2017, Twinanda et al. 2016). Early image-based approaches focused on extracting visual features and then employed a classifier to determine the surgical phase (Dergachyova, et al. 2016). With the emergence of deep learning in image classification tasks, most recent approaches relied on utilising convolutional neural networks (CNNs) to learn visual features and recurrent neural networks (RNNs) to model temporal dependencies along the laparoscopic video (Abdulbaki Alshirbaji, et al. 2020, Alshirbaji et al. 2021, Czempiel et al. 2020, Jalal et al. 2018, Jin, et al. 2017, Jin, et al. 2020, Twinanda 2017, Twinanda, et al. 2016).

Since each surgical phase is accomplished using corresponding surgical tools (see Fig. 1), many studies exploited the tool-phase relation to develop more consistent phase recognition approaches. Jalal et al. proposed a deep learning pipeline consisting of a CNN and a nonlinear autoregressive network with exogenous inputs (NARX) (Jalal et al. 2019). The CNN model was utilised to perform tool classification. The tool binary classifications were then provided to the NARX to perform phase prediction. Twinanda et al. introduced a multi-task model, called EndoNet, that jointly performs surgical phase recognition and tool presence detection (Twinanda, et al. 2016). To enforce temporal constraints, they employed a Hierarchical HMM (HHMM) to learn temporal information and refine the classification obtained by EndoNet. Additionally, Twinanda et al. replaced the HHMM by a long short-term memory (LSTM) network to overcome drawbacks imposed by the HMM and enhance temporal modelling (Twinanda 2017). Similarly, Jin et al. applied a CNN-LSTM pipeline (SV-RCNet) in an end-to-end manner to carry out phase recognition, and they proposed a prior knowledge inference scheme (Jin, et al. 2017). Jin et al. proposed also a multi-task framework (MTRCNet) consisting of a CNN and a LSTM network (Jin, et al. 2020). They also introduced a correlation loss to identify tool-phase relatedness. Czempiel et al. proposed TeCNO framework that relies on temporal convolutional networks (TCNs) to learn temporal information from pre-extracted visual features (Czempiel, et al. 2020).

In this paper, a spatio-temporal deep learning approach for recognising surgical phases in laparoscopic videos was proposed. The proposed approach relied on learning fine-level temporal information in short video clips including unlabelled frames and the tool-phase relation. Initially, a CNN model (ResNet-50) was trained in a multi-task manner to perform tool detection and phase recognition, and it was utilised to extract visual features. Then, two LSTM models, termed as LSTM-clip and LSTM-video, were employed (similar to (Abdulbaki Alshirbaji, et al. 2020)) to carry out tool presence detection. On top of the previous LSTM cascade, another LSTM model, termed as LSTM-phase, was utilised to perform phase recognition. Finally, the proposed approach was evaluated on the large dataset Cholec80.

2. METHODOLOGY

2.1 Architecture

The proposed approach consists of a CNN and a cascade of three LSTM models. The CNN was used to extract visual features from laparoscopic images. The CNN features were passed to the first LSTM (LSTM-clip) to model temporal information within short video clips. The last two LSTM models, which are LSTM-video and LSTM-phase, capture temporal dependencies along the complete video. The overall framework is illustrated in Fig. 2.

2.1.1 CNN

The first stage of our approach is to encode visual information of laparoscopic images into a vector of features. Therefore, a CNN model was initially trained on laparoscopic images to learn discriminative features. Then, the trained model was used to extract features from images.

The ResNet-50 (He et al. 2016) was fine-tuned in a multi-task manner to classify surgical tools and recognise surgical phases. The last layer of ResNet-50 was substituted by a fully-connected layer with seven nodes, termed fc-tool, to classify surgical tools. Another fully-connected layer, termed fc-phase, was added to identify surgical phases. The outputs of fc-tool and the global average pooling layer form the input for fc-phase.

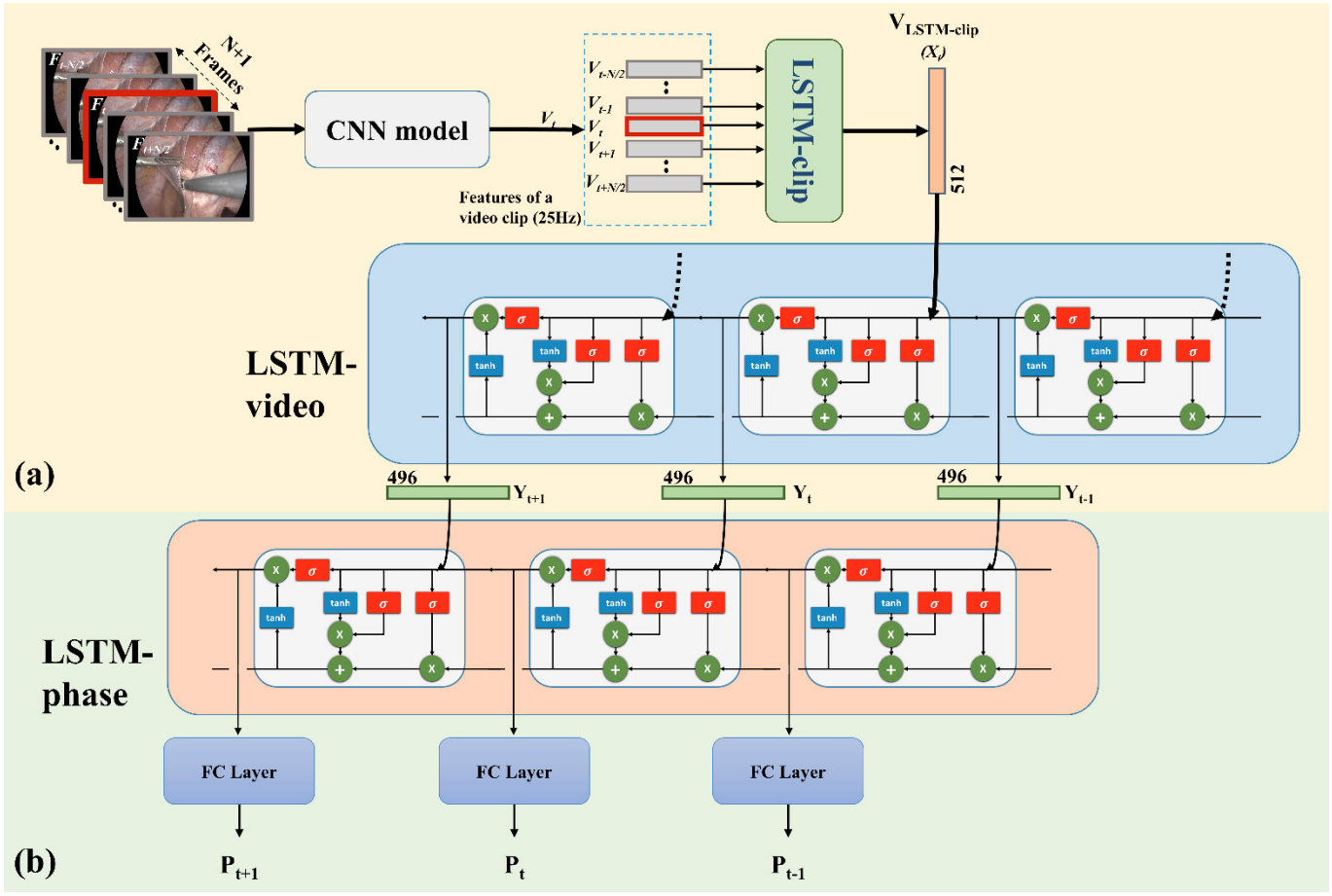

Figure 2. The proposed framework for surgical phase recognition. (a) Overview of the full pipeline for detecting surgical tools, (b) the LSTM-phase model. Red and grey rectangles indicate labelled and unlabeled frames, respectively. *V* vectors represents visual features obtained using the CNN model.

2.1.2 Cascade-LSTM

In the second stage, a cascade of LSTM models was employed to incorporate the temporal information. The LSTM is an effective type of recurrent neural network (RNN) for modelling sequential data. Unlike traditional RNN, LSTM prevents vanishing and exploding gradients problems. Hence, LSTM is more able to capture long and short-term sequential dependencies (Hochreiter and Schmidhuber 1997).

The cascade-LSTM consists of three LSTM models which are LSTM-clip, LSTM-video and LSTM-phase. Each LSTM model performs a single task. The LSTM-clip and LSTM-video classify surgical tools, whilst the LSTM-phase recognise surgical phases. Since both tasks are closely related, the learned knowledge by the first two LSTM is highly beneficial to surgical phase recognition in the LSTM-phase.

2.1.2.1 LSTM-clip

This model was engaged to leverage sequential information at a fine-level by incorporating a fixed number of unlabelled frames surrounding a labelled one. LSTM-clip has many-to-one configuration to capture temporal relations across adjacent frames of a short video clip. The laparoscopic video of every intervention is segmented into short clips. Each video clip contains one tool-labelled frame and N unlabelled frames around it. A feature vector is extracted for every frame in the clip and that results in a sequence of feature vectors, as in (1).

$$S_t^{clip} = [V_{CNN}(F^U_{t-\frac{N}{2}}), \dots, V_{CNN}(F^L_t), \dots, V_{CNN}(F^L_{t+\frac{N}{2}})] \quad (1)$$

where S_t^{clip} is a sequence of feature vectors for a tool-labelled frame F_t^L at time *t*, V_{CNN} is a vector of CNN features and F^U denotes unlabeled frame.

A sequence S^{clip} is arranged for every tool-labelled frame to be taken by LSTM-clip as input. The output of this LSTM model is passed to a fully-connected layer with seven nodes for classifying surgical tools. During training, the LSTM-Clip model learns a vector of features (V_{clip}) which is used by LSTM-video model.

2.1.2.2 LSTM-video

This model exploits the temporal information along the complete video. The features of LSTM-clip for every clip in the video were arranged in a sequence S^{video}, as in (2).

$$S^{video} = [V_{clip}(1), \dots, V_{clip}(t), \dots, V_{clip}(T)] \quad (2)$$

where S^{video} is a sequence of LSTM-clip features along a video of length T seconds, i.e. T tool-labelled frames. This sequence forms the input for the LSTM-video. A fully-connected layer was added atop the LSTM model to perform

tool classification. The configuration of LSTM-video is many-to-many, thus it gives a vector of features V_{video} for every tool-labelled frame in the video. The features obtained using the LSTM-video were then passed to the LSTM-phase.

2.1.2.3 LSTM-phase

The LSTM-phase profits from the accumulated knowledge through the prior models to identify the surgical phases. Similar to the preceding model, LSTM-phase models the temporal dependencies along the entire laparoscopic video. It utilises the features extracted from LSTM-video as input. Equation (3) shows a sequence of LSTM-video features S^{phase} for a video of length T seconds. To perform the phase recognition task, the output of this LSTM model was connected to a fully-connected layer.

$$S^{phase} = [V_{video}(1), \dots, V_{video}(t), \dots, V_{video}(T)] \quad (3)$$

2.2 Experimental Setup

2.2.1 Dataset

The dataset used in this work is the Cholec80 (Twinanda, et al. 2016). Cholec80 dataset was collected at the University Hospital of Strasbourg. It contains laparoscopic videos for 80 cholecystectomy procedures. The videos were recorded at a frame rate of 25 frames per second (fps). The resolution of the videos is either 854×480 or 1920×1080. Frames were resized into 224×224×3 since the CNN model accepts this input size.

The dataset was manually labelled for surgical phases at 25 fps and for surgical tools at 1 fps. The defined surgical phases and used tools are presented in Fig. 1. The first 40 videos of cholec80 were used for training. The remaining videos were kept to enable evaluation of the performance of each model in our framework.

2.2.2 Training process

Each model was trained separately. The training parameters for each model are presented in Table 1. The CNN model was initialised with weights learned from training on ImageNet (Deng et al. 2009). The CNN training started at a learning rate of 10^{-4}, except the fully-connected layers on top of the model had a higher rate of 20×10^{-4}. The LSTM-clip was trained using 20 unlabelled frames around every tool-labelled one. The LSTM-video and LSTM-phase were trained every training iteration with one S^{video} and one S^{phase}, respectively. Accordingly, applying zero-padding on sequences of different lengths, that is required when multiple sequences are processed per batch, was avoided.

Sigmoid and softmax activation functions were used for tool classification and phase recognition, respectively. The loss was computed for tool classification task using binary cross-entropy function, and for phase recognition using softmax multinomial logistic loss function. An Adam optimiser was employed to minimise the loss. The implementation was carried out in Keras using NVIDIA GeForce RTX 2080 TI GPUs.

Table 1. Training parameters for all models

Model	CNN	LSTM-clip	LSTM-video	LSTM-phase
Batch size	50 images	50 S^{clip}	1 S^{video}	1 S^{phase}
Epochs	10	30	30	20
Initial learning rate	20×10^{-3}	10^{-4}	10^{-4}	10^{-4}
Weight decay	9×10^{-4}	10^{-3}	10^{-3}	10^{-3}
Memory cells	-	512	4096	4096

3. RESULTS

The accuracy, precision (PR) and recall (RE) were utilised to evaluate the performance of the of the proposed framework. The PR and RE were first calculated for each phase and the mean PR and mean RE were then calculated. The accuracy was calculated for the entire data. A comprehensive analysis about the improvement achieved using the proposed framework over the ResNet-50 model are presented in Table 2. The predictions of all phases enhanced using the proposed approach. Furthermore, to highlight the ability of the proposed framework to refine intra- and inter-phase predictions, the confusion matrices are further visualised in Fig. 4.

Table 3 presents comparison results of phase recognition with the reference methods. To achieve consistency with prior papers, the same data split was used to evaluate the proposed approach.

Table 2. Precision and recall of phase recognition results for all phases using the proposed framework.

Phase	Precision		Recall	
	ResNet-50	Our approach	ResNet-50	Our approach
P1	71.8	97.8	54.3	80.6
P2	84.0	97.9	85.1	98.3
P3	73.7	89.8	69.7	82.2
P4	85.4	92.7	84.7	98.8
P5	62.5	87.8	79	82.2
P6	68.7	92.5	71.2	66.9
P7	58.2	72.6	58.4	86.3
Mean	72.0	90.1	71.8	85.1

Table 3. Baseline comparisons with other methods. Bold values indicate best performance.

Method	Accuracy	Precision	Recall
PhaseNet	78.8	71.3	76.6
EndoNet + HHMM	81.7	73.7	79.6
EndoNet + LSTM	88.6	84.4	84.7
SV-RCNet	85.3	80.7	83.5
MTRCNet-CL	89.2	86.9	**88.0**
TeCNO	88.6	86.5	87.6
Our approach	**92.9**	**90.1**	85.1

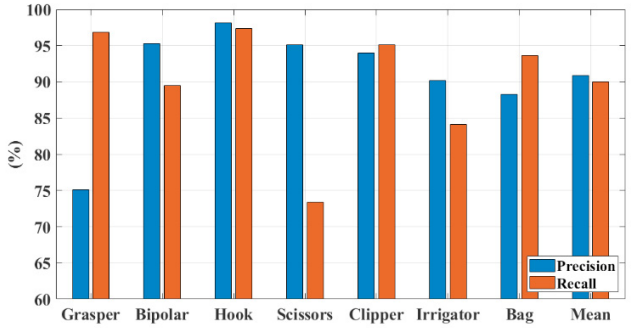

Figure 3. Tool presence detection results using LSTM-video. Note the truncated scale of the y-axis

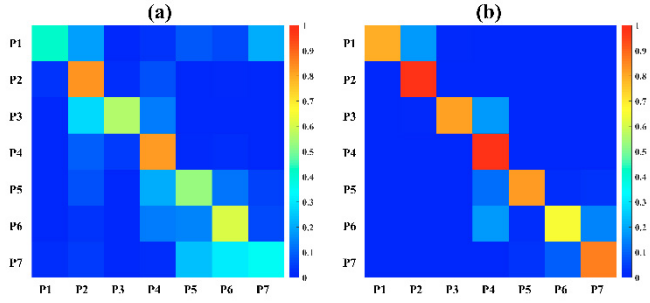

Figure 4. Confusion matrices for (a) the ResNet-50 model and (b) the proposed approach.

4. DISCUSSION

This work presents a deep learning framework for recognising surgical phases in cholecystectomy videos. The proposed approach consists of two main stages. Initially, a CNN model followed by a cascade of two LSTM networks were trained to perform tool presence detection. Then, spatiotemporal features extracted using the aforementioned approach were provided into a LSTM to perform phase recognition.

The proposed framework achieved mean precision and mean recall of 89.0% and 84.5%, respectively. These values improved on the established ResNet-50 model mean precision and mean recall values of 72.0% and 71.8%, respectively (Table 2). The LSTM-clip and LSTM-video approaches contributed effectively to extract more discriminative spatiotemporal features not only for detecting surgical tools but also for recognising surgical phases. Additionally, using the LSTM-phase on top of the LSTM-video to enforce transitions between surgical phases achieved notable improvement of all phases (see Table 2).

Fig. 1 shows the tool-phase relation in a laparoscopic procedure of the Cholec80 dataset. The surgical tool usage signals are not only discriminative for intra-phases (i.e. within the surgical phase) but also for inter-phases (i.e. transitions between phases). Therefore, the proposed approach relied on employing the ResNet, LSTM-clip and LSTM-video cascade to extract input features for the LSTM-phase. In this context, the recognition results are highly correlated to the tool detection results and can be interpreted accordingly.

The best recognition results were obtained for P2 (calot triangle dissection) and P4 (gallbladder dissection) (see Table 2). Indeed, these two phases were accomplished using the grasper and hook tools (see Fig. 3). The sensitivity of the grasper and the hook were 96.8% and 97.4%, respectively. Consequently, the proposed framework achieved sensitivity of 98.3% and 98.8% for P2 and P4, respectively. In contrast, the recognition results for P3 (clipping and cutting) were much lower. The sensitivity of the scissors and the clipper, that were utilised to perform this phase, were 73.4% and 95.1%, respectively. Moreover, the scissors usually appear at the middle and end of P3. Therefore, some frames belonging to the third phase were misclassified as P4 (see Fig. 4).

The recognition results of the last three phases, especially P6, are significantly lower than the results of P2 and P4, which is due to the non-linear transition between these three phases. Moreover, the specimen bag, that typically appears in P5 and P7, appeared in P6.

Table 3 shows the reference methods and their corresponding recognition precision and recall. Like this work, MTRCNet-CL, EndoNetLSTM and SV-RCNet utilised LSTM for temporal refinement. In MTRCNet-CL and EndoNetLSTM, the CNN model was trained similar to this study to perform both tool detection and phase recognition tasks. Moreover, MTRCNet-CL was trained in end-to-end fashion, so the visual and temporal features are jointly learned. However, the MTRCNet-CL was trained using short video sequences in the patch because of hardware constraints. In contrast, the LSTM networks used in the proposed framework were trained using complete video sequences, where the input features of each network were extracted prior to the training process. While MTRCNet-CL reported an average recall of 88.0% (which exceeds the average recall of 85.1% obtained using the presented framework) the average accuracy and average precision achieved in this study are higher (Table 3). A more recent study Czempiel et al. proposed using a multi-stage TCN (TeCNO) to temporally refine phase predictions, and they reported an average precision and average recall of 86.5% and 87.6%, respectively. In general, the proposed approach achieved comparable average recall compared to the reference methods, and better average precision and accuracy. However, more detailed comparison of the precison and recall of each phase was not possible since the results of each phase were not reported in most of the published papers.

Despite the high phase recognition performance achieved by the proposed framework, the study has limitations. The presented approach was evaluated using singular data split for training and testing. Therefore, cross-validation experiments should be carried out to ensure that the high performance achieved in this study was not uniquely optimised to the specific training and testing data described. Furthermore, the proposed framework was only evaluated on the Cholec80 dataset where only seven phases and seven tools are defined. Therefore, the proposed method should be evaluated on other datasets of the same surgery type and on other complex surgeries such as sigmoid resection where more surgical phases can be defined.

5. CONCLUSIONS

This study proposed a deep learning framework to recognise surgical phases in laparoscopic videos. The proposed approach depends on initially extracting spatiotemporal features using a cascade of a CNN and two LSTM networks that was trained mainly to perform tool presence detection. The extracted features are then provided into a LSTM model to perform phase recognition. Experimental results of this approach showed strong phase recognition performance indicating the promising potential to integrate this approach into smart systems in the operating theatre. Future work includes enhancing this method for inter-phase recognition performance especially for the last few phases.

ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry of Research and Education (BMBF under grant CoHMed/IntelliMed grant no. 13FH5I01IA and 13FH5I05IA).

REFERENCES

Abdulbaki Alshirbaji, T., Jalal, N. A., and Möller, K. (2020). A convolutional neural network with a two-stage LSTM model for tool presence detection in laparoscopic videos. *Current Directions in Biomedical Engineering,* vol. 6 (1).

Alshirbaji, T. A., Jalal, N. A., Docherty, P. D., Neumuth, T., and Möller, K. (2021). A deep learning spatial-temporal framework for detecting surgical tools in *laparoscopic* videos. *Biomedical Signal Processing and Control,* vol. 68, p. 102801.

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S. T., and Navab, N. (2020). TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Springer, 2020), pp. 343-52.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, (Ieee, 2009), pp. 248-55.

Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., and Jannin, P. (2016). Automatic data-driven real-time segmentation and recognition of surgical workflow. *International journal of computer assisted radiology and surgery,* vol. 11 (6), pp. 1081-9.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual *learning* for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 770-8.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term *memory*. *Neural computation,* vol. 9 (8), pp. 1735-80.

Jalal, N. A., Alshirbaji, T. A., and Möller, K. (2018). Evaluating convolutional neural network and hidden *markov* model for recognising surgical phases in sigmoid resection. *Current Directions in Biomedical Engineering,* vol. 4 (1), pp. 415-8.

Jalal, N. A., Alshirbaji, T. A., and Möller, K. (2019). Predicting surgical phases using CNN-NARX neural *network*. *Current Directions in Biomedical Engineering,* vol. 5 (1), pp. 405-7.

Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.-W., and Heng, P.-A. (2017). SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging,* vol. 37 (5), pp. 1114-26.

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

Lalys, F. and Jannin, P. (2014). Surgical process modelling: a review. *International journal of computer assisted radiology and surgery,* vol. 9 (3), pp. 495-511.

Maier-Hein, L., Vedula, S. S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., *et al.* (2017). Surgical data science for next-generation interventions. *Nature Biomedical Engineering,* vol. 1 (9), pp. 691-6.

Neumuth, T. and Meißner, C. (2012). Online recognition of surgical instruments by information fusion. *International journal of computer assisted radiology and surgery,* vol. 7 (2), pp. 297-304.

Padoy, N., Blum, T., Ahmadi, S.-A., Feussner, H., Berger, M.-O., and *Navab*, N. (2012). Statistical modeling and recognition of surgical workflow. *Medical image analysis,* vol. 16 (3), pp. 632-41.

Stauder, R., Okur, A., Peter, L., Schneider, A., Kranzfelder, M., Feussner, H., and Navab, N. (2014). Random forests for phase detection in surgical workflow analysis. In *International Conference on Information Processing in Computer-Assisted Interventions*, (Springer, 2014), pp. 148-57.

Twinanda, A. P. (2017). Vision-based approaches for surgical activity *recognition* using laparoscopic and RBGD videos. Strasbourg,

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., and Padoy, N. (2016). Endonet: a deep *architecture* for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging,* vol. 36 (1), pp. 86-97.