Spatial-temporal Modelling for Surgical Tool Classification in Cholecystectomy Videos

Tamer Abdulbaki Alshirbaji[1], Nour Aldeen Jalal[1], Thomas Neumuth[2] and Knut Moeller[3]

[1]Institute of Technical Medicine (ITeM), Furtwangen University, and Innovation Centre Computer Assisted Surgery (ICCAS), University of Leipzig
{Nour.A.Jalal,abd}@hs-furtwangen.de
[2] Innovation Centre Computer Assisted Surgery (ICCAS). University of Leipzig
thomas.neumuth@uni-leipzig.de
[3]Institute of Technical Medicine (ITeM), Furtwangen University
Knut.Moeller@hs-furtwangen.de

Abstract. Surgical tool classification is an essential component to analyse the surgical workflow of laparoscopic intervention. It has many potential applications, for instance, developing decision-support systems, automatic indexing of laparoscopic videos, and assessing surgical skills. In this work, a framework for surgical tool presence detection is presented. The proposed approach consists of a CNN model and two LSTM units to model spatial and temporal information encoded in the laparoscopic video. The proposed approach achieved a mean average precision of 94.57%. Experimental results show the value of temporal modelling in improving the classification performance of surgical tools.

Keywords: Surgical tool classification; Laparoscopic video; CNN; LSTM.

1 Introduction

Current operating rooms are equipped with advanced surgical devices and instrumentations. Those devices enable performing the surgical intervention and provide the surgical team with the necessary information. However, it is challenging for the surgeon to process all available data from different surgical devices and keep focusing on surgical actions. Hence, active research has been conducted to analysis surgical workflow, in particular for laparoscopic interventions.

Laparoscopic interventions provide a wealth of data as this type of procedures are performed with special instrumentations like laparoscopic camera which enable monitoring the procedure [1]. Thus, the main focus of the conducted research has been on analyzing laparoscopic videos for several purposes. Identifying surgical tools in laparoscopic images is essential for recognizing surgical actions and phases. That kind of knowledge can serve other intelligent systems with a variety of potential applications. For instance, notifying the surgeon with possible complications, provide assistive guidance, automatic indexing of videos for training purpose and predicting required time to optimize schedule of operating room [1–3].

The revolution of computing power and availability of surgical data have empower applying deep learning approaches. Twinanda et al. released Cholec80 dataset and proposed a convolutional neural network (CNN) architecture called EndoNet to perform surgical phase and tool recognition [4]. However, the prediction of EndoNet model was based only on a single image. Nevertheless, many disturbances can occur in laparoscopic images, for example, emergence of smoke [5, 6], bleeding or light reflections. Hence, anatomical structures and surgical tools might be covered due to such disturbances, and thus impeding capability of image-based approaches. To alleviate the challenging nature of laparoscopic images, modelling temporal information along the video were addressed using different techniques. Hidden Markov model (HMM) [4, 7], graph convolutional network [3], nonlinear autoregressive network with exogenous inputs (NARX) [2], and long short-term memory (LSTM) [8–12] were used to detect surgical tools and/or phases.

In this work, a deep learning approach for classification surgical tools in laparoscopic images was proposed. A CNN model was used to encode spatial features from laparoscopic images. Two LSTM units utilised CNN features of labelled and some unlabelled frames to model temporal dependencies along short and complete sequence of procedure video.

2 Methods

2.1 Model overview

A pipeline consisting of a CNN and two LSTM units was implemented. **Fig. 1** depicts an overview of the methodology pipeline. The CNN was employed to capture visual features of a laparoscopic frame. To obtain high-level discriminative features, the CNN model was initially trained on cholecystectomy images to perform surgical tool detection and surgical phase recognition. To this ends, the architecture of the CNN was modified to perform those tasks similar to the EndoNet architecture [4]. Every single labelled frame and a number of surrounding unlabelled frames formed a short sequential data. Using the trained CNN model, visual features were extracted for every frame in the short sequences. The first LSTM unit exploited the CNN-features of every short data sequence. Thus, temporal dependencies across the short sequence were utilised. To enhance the performance further, a second LSTM unit was employed to consider temporal information along the entire video.

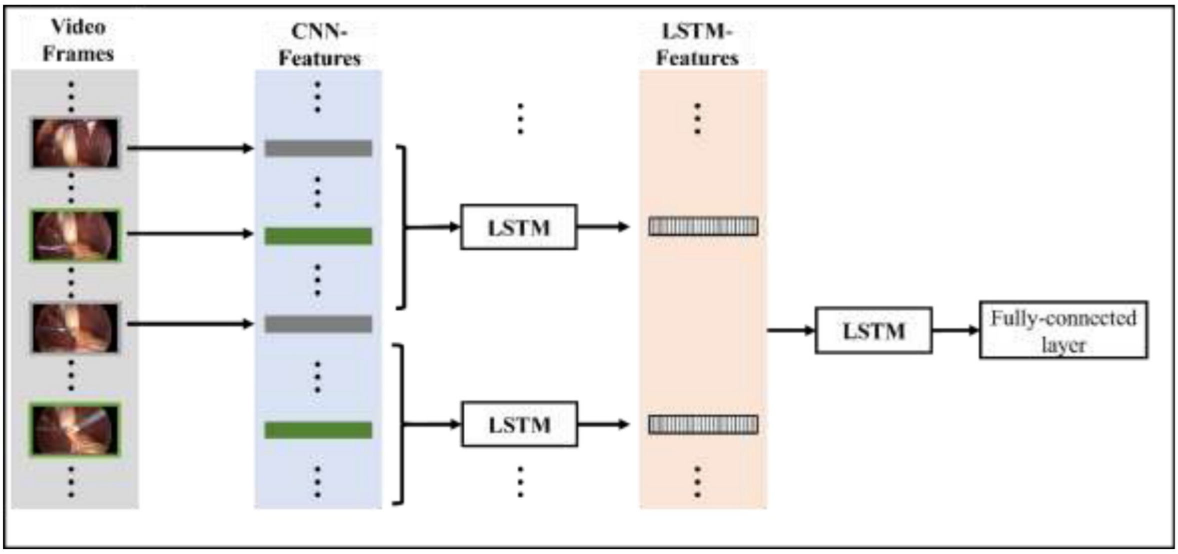

Fig. 1. Pipeline architecture of the proposed approach. Labelled and unlabelled frames are in green and grey rectangles, respectively.

2.2 Dataset description

Eighty cholecystectomy videos of Cholec80 dataset [4] were used in this work. The videos were recorded at University Hospital of Strasbourg at 25 Hz. The dataset contains labels for surgical phases at 25 Hz and labels for surgical tools at 1 Hz. The surgical tools defined in Cholec80 dataset are grasper, bipolar, hook, scissors, clipper, irrigator, specimen bag. Forty videos were used for training, and the remaining videos were used for performance evaluation. The distribution of training and testing data are shown in **Fig. 2**.

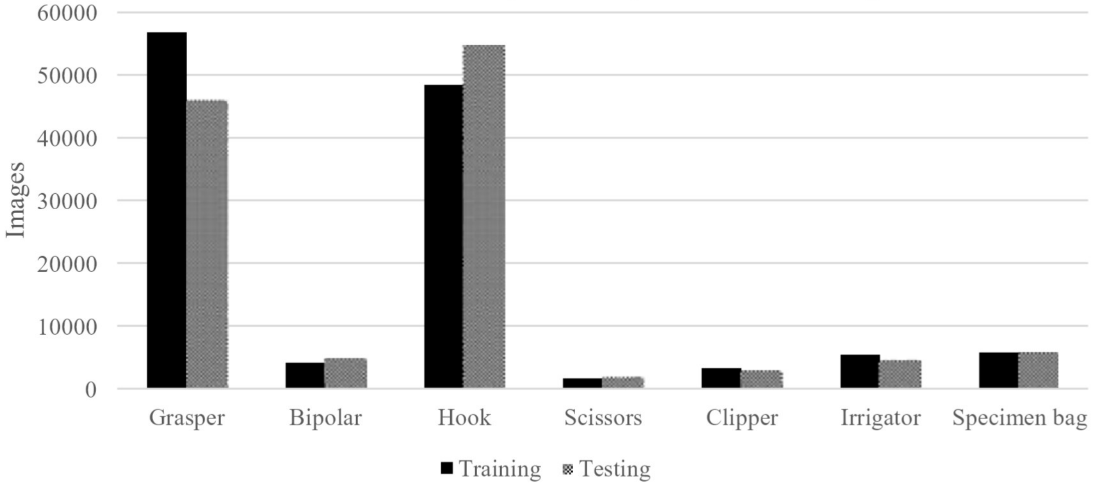

Fig. 2. Distribution of surgical tools in the training and testing data.

2.3 Experimental setup

The base model ResNet-50, pretrained on ImageNet dataset, was employed. The model was trained for ten epochs with an initial learning rate of 2.10^{-3}. The batch size was set to 50 images. The first LSTM contained 512 cells, whereas the second LSTM had 4096 cells. Each of the LSTMs was trained for 30 epochs with an initial learning rate of 1.10^{-4}. The first LSTM had a batch size of 50 short sequences. The length of every short sequence was set to 21 frames, where a labelled frame was in the middle of the sequence. The second LSTM was trained with a batch size of one video.

A fully-connected layer with seven nodes was used in each model to perform tool classification. This layer had a sigmoid activation function since the task was a binary classification. The cross-entropy function was used to compute loss of each tool. Tool losses were weighted according to the number of images belonging to each tool in the training data, as in [13]. The proposed approach was implemented using Keras framework. The implementation was conducted on a PC with NVIDIA GeForce RTX 2080Ti GPU.

3 Results

Each component of the proposed approach was evaluated on the same testing set. **Fig. 3** presents the average precision (AP) of each surgical tool and the mean AP over all tool for CNN and LSTMs.

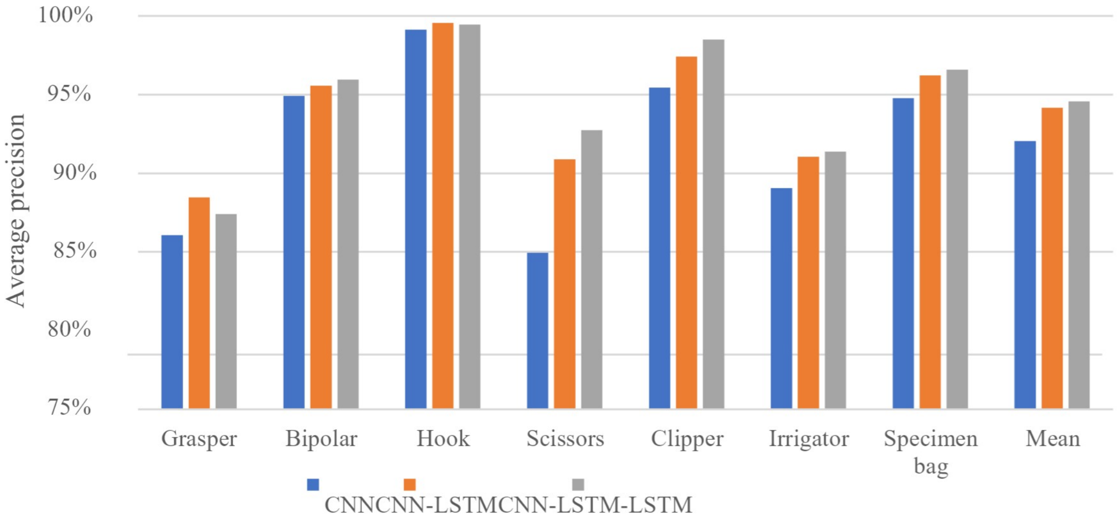

Fig. 3. Average precicion of tool classification yielded from each component of the proposed pipeline.

4 Discussion

This work presents a framework for classifying surgical tools in laparoscopic videos. The framework is based on utilisation of spatial and temporal information encoded in videos. To this end, ResNet-50 and two LSTM units were employed.

The CNN model had a high capability to identify surgical tools in a laparoscopic image. However, the CNN model failed to detect surgical tools when they were partially appeared in the scene or were covered by smoke, blood or a tissue. Therefore, using some unlabelled frames before and after the target frame (labelled frame) helps to recognise surgical tools. Hence, applying the first LSTM unit improved the classification results for all surgical tools (see **Fig. 3**).

The laparoscopic procedure can be segmented into some surgical phases which are executed in some specific order. Since particular surgical tools are used in each surgical phase, there is a correlation between surgical phases and tool usage. Thus, modelling sequential dependencies along entire video, conducted by the second LSTM unit, enhanced the average precision for all tools, except for grasper (see **Fig. 3**). Grasper is used frequently during the entire procedure and appear in all surgical phases, and hence, applying the second LSTM had a marginal effect on enhancing performance for this surgical tool.

The proposed approach achieved a mean average precision of 94.57% higher performance than state-of-the-art methods. Twinanda et al. reported a mean average precision of 81% with EndoNet model [4]. Jin et al. used a loss function which models correlation between surgical phases and tools, and an average precision of 89% for tool presence detection was reported [14]. Similar to our approach, Wang et al. proposed a CNN and graph convolutional networks (GCN) to model spatial and temporal information, respectively, across short video clips. The CNN-GCN approach achieved a mean average precision of 90.13% [3].

The CNN and LSTM models were trained separately. It would be interesting to train the complete framework and investigate the effectiveness of end-to-end training. Moreover, the robustness of the proposed approach to data from different surgical sources could be evaluated.

5 Conclusion

Experimental results demonstrate the value of exploiting temporal information for surgical tool classification. Moreover, this study highlights feasibility of using unlabelled data to improve the classification performance.

References

1. Anteby, R., Horesh, N., Soffer, S., Zager, Y., Barash, Y., Amiel, I., Rosin, D., Gutman, M., Klang, E.: Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surgical Endoscopy. 35, 1521–1533 (2021). https://doi.org/10.1007/s00464-020-08168-1.
2. Jalal, N.A., Alshirbaji, T.A., Möller, K.: Predicting surgical phases using CNN-NARX neural network. Current Directions in Biomedical Engineering. 5, 405–407 (2019). https://doi.org/10.1515/cdbme-2019-0102.
3. Wang, S., Xu, Z., Yan, C., Huang, J.: Graph convolutional nets for tool presence detection in surgical videos. In: International Conference on Information Processing in Medical Imaging. pp. 467–478. Springer (2019). https://doi.org/10.1007/978-3-030-20351-1_36.
4. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging. 36, 86–97 (2016). https://doi.org/10.1109/TMI.2016.2593957.
5. Abdulbaki Alshirbaji, T., Jalal, N.A., Mündermann, L., Möller, K.: Classifying smoke in laparoscopic videos using SVM. https://doi.org/10.1515/cdbme-2017-0040.
6. Jalal, N.A., Alshirbaji, T.A., Mündermann, L., Möller, K.: Features for detecting smoke in laparoscopic videos. Current Directions in Biomedical Engineering. 3, 521–524 (2017). https://doi.org/10.1515/cdbme2017-0110.
7. Jalal, N.A., Alshirbaji, T.A., Möller, K.: Evaluating convolutional neural network and hidden markov model for recognising surgical phases in sigmoid resection. Current Directions in Biomedical Engineering. 4, 415–418 (2018). https://doi.org/10.1515/cdbme-2018-0099.
8. Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks. arXiv preprint arXiv:1805.08569. (2018).
9. Jalal, N.A., Abdulbaki Alshirbaji, T., Docherty, P.D., Neumuth, T., Möller, K.: Surgical Tool Detection in Laparoscopic Videos by Modeling Temporal Dependencies Between Adjacent Frames. In: European Medical and Biological Engineering Conference. pp. 1045–1052. Springer (2020).
10. Alshirbaji, T.A., Jalal, N.A., Docherty, P.D., Neumuth, T., Möller, K.: A deep learning spatial-temporal framework for detecting surgical tools in laparoscopic videos. Biomedical Signal Processing and Control. 68, 102801 (2021).
11. Alshirbaji, T.A., Jalal, N.A., Möller, K.: A convolutional neural network with a two-stage LSTM model for tool presence detection in laparoscopic videos. Current Directions in Biomedical Engineering. 6, (2020).
12. Jalal, N.A., Alshirbaji, T.A., Docherty, P.D., Neumuth, T., Moeller, K.: A Deep Learning Framework for Recognising Surgical Phases in Laparoscopic Videos. IFAC-PapersOnLine. 54, 334–339 (2021). https://doi.org/10.1016/j.ifacol.2021.10.278.
13. Alshirbaji, T.A., Jalal, N.A., Möller, K.: Surgical tool classification in laparoscopic videos using convolutional neural network. Current Directions in Biomedical Engineering. 4, 407–410 (2018). https://doi.org/10.1515/cdbme-2018-0097.
14. Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., Heng, P.-A.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Medical image analysis. 59, 101572 (2020). https://doi.org/10.1016/j.media.2019.101572.