Tamer Abdulbaki Alshirbaji*, Nour Aldeen Jalal and Knut Möller

A convolutional neural network with a two-stage LSTM model for tool presence detection in laparoscopic videos

https://doi.org/10.1515/cdbme-2020-0002

Abstract: Surgical tool presence detection in laparoscopic videos is a challenging problem that plays a critical role in developing context-aware systems in operating rooms (ORs). In this work, we propose a deep learning-based approach for detecting surgical tools in laparoscopic images using a convolutional neural network (CNN) in combination with two long short-term memory (LSTM) models. A pre-trained CNN model was trained to learn visual features from images. Then, LSTM was employed to include temporal information through a video clip of neighbour frames. Finally, the second LSTM was utilized to model temporal dependencies across the whole surgical video. Experimental evaluation has been conducted with the Cholec80 dataset to validate our approach. Results show that the most notable improvement is achieved after employing the two-stage LSTM model, and the proposed approach achieved better or similar performance compared with state-of-the-art methods.

Keywords: convolutional neural network (CNN); endoscopic video; spatiotemporal feature; surgical tool detection.

Introduction

Analysing surgical workflow is a key factor in establishing intelligent technologies that aim to support surgical teams and optimize patient treatment inside the operating room (OR). For instance, recognizing surgical workflow is a fundamental component to develop context-aware systems [1]. These systems can effectively monitor the workflow and communicate relevant information to human operators with different perspectives, i.e. surgeon and anaesthesiologist. In laparoscopic surgeries, the complete intervention can be captured using a laparoscope, giving rise to analyse surgical videos, particularly, for recognizing surgical activities [2–4] and detecting surgical tools [5–9].

Several studies have investigated surgical tool presence detection in laparoscopic images. The methods proposed so far are based on either learning a classifier on top of handcrafted features [5] or using convolutional neural networks (CNNs) [6–9]. However, laparoscopic images differ from other types of images in terms of characteristics and qualities. Additional challenges arise due to various reasons: multi-tool classification task, image blur resulting from rapid movement of camera and tools masked with blood or tissues or obscured by smoke from electro-surgical cutting and coagulation. Therefore, temporal information along neighbouring frames may be considered to correct misclassified images. Recently, Wang et al. demonstrated the feasibility of considering information along continuous video frames [6]. With their approach, a graph convolutional network (GCN) was applied to capture temporal information from a video clip. Chen et al. explored using a 3D convolutional network to extract spatiotemporal features from a video clip to perform tool classification [8].

The aim of this study was to investigate whether tool presence classification in laparoscopic videos can be improved by incorporating temporal features from a short clip and also from the whole video sequence. First, a CNN model, namely VGG-16 [10], was fine-tuned to learn and extract spatial features from laparoscopic images. Then, since the data is sparsely annotated at 1 fps, we employed a long short-term memory (LSTM) to learn temporal coherence across video clip, termed as LSTM-clip. On top of the LSTM-clip, we added another LSTM, termed LSTM-video, to learn temporal features across the entire surgical video frames. Finally, the proposed approach was evaluated on the large public dataset Cholec80 [2] by determining the area under the precision-recall curve, i.e. average precision (AP).

***Corresponding author: Tamer Abdulbaki Alshirbaji,** Institute of Technical Medicine, Furtwangen University, Jakob-Kienzle-Straße 17, Villingen-Schwenningen, Germany, E-mail: abd@hs-furtwangen.de
Nour Aldeen Jalal and Knut Möller, Institute of Technical Medicine, Furtwangen University, Villingen-Schwenningen, Germany, E-mail: ja@hs-furtwangen.de (N.A. Jalal), moe@hs-furtwangen.de (K. Möller)

Methods

Dataset

Cholec80 dataset [2] was used in this work. It contains 80 cholecystectomy videos captured at a frame rate of 25 fps and annotations for surgical tools (1 fps) and surgical phases. Processed frames were downsized to 224 × 224 and colour channels were rescaled to the range [0, 1]. The first 40 videos were used for training, and the 40 remaining were used for validation and test.

Overview

The proposed method consists of three main components: CNN model for extracting visual features and two LSTMs for incorporating temporal information (see Figure 1). The CNN model, namely VGG-16, was fine-tuned using the concept of transfer learning for detecting surgical tools and phases in laparoscopic images [2]. The first LSTM, termed LSTM-clip, predicts tools presence in a frame using a sequence of previous neighbouring frames. The second LSTM considers the whole video as a sequence and takes features from LSTM-clip as input for every time step.

CNN model: VGG-16 model [10] was fine-tuned using 40 cholecystectomy videos. The model architecture was modified, similarly to Twinanda work in Ref. [2], to recognise surgical phases and tools. The last layer was replaced with seven fully-connected layers, each for classifying surgical tool, and outputs from tool layers were concatenated with a vector of visual features obtained from the previous fully-connected layer.

LSTM-clip: The CNN model might not be able to recognise surgical tools using the spatial information of a single frame alone, especially if the tool is masked or obscured. Therefore, it might be beneficial to consider some preceding frames to perform tool classification. In our dataset, videos were captured at 25 fps, whereas surgical tools were annotated at 1 fps. Thus, we propose to incorporate unlabelled frames using the LSTM model. The spatial features for all frames were extracted using VGG-16, and they were arranged in sequences, each containing features of an annotated frame and n preceding frames. LSTM-clip was trained with the aim to predict tool presence in the last frame of every sequence $S_t=[V(F_{t-n}),\ldots, V(F_t)]$, where $V(F_t)$ denotes a vector of spatial features of the t frame. The clip-sequences in the training set were shuffled before every training epoch.

LSTM-video: The LSTM-clip is trained on short video clips and it does not profit from temporal cues along the entire video. Therefore, an LSTM layer, termed LSTM-video, was built on the top of LSTM-clip to model temporal dependencies across consecutive video clips of a laparoscopic video. For every clip S_t in a video, a spatial-temporal feature vector $H(S_t)$ was obtained from LSTM-clip in order to form a sequence of feature vectors $S^V=[H(S_1),\ldots., H(S_t),\ldots., H(S_T)]$, where T is the length of the video in seconds and V denotes video number. The LSTM-video was trained on sequences of training videos to identify surgical tools in every clip of the video.

Experimental setup

The VGG-16 model was trained using Adam optimizer with initial learning rate=10^{-4} for all layers except the tool and phase layers, that were added with random initialization and trained at a higher learning rate of 0.002. A drop-out layer of rate 0.6 was used after each fully-connected layer. Training was carried out for 10 epochs with a batch size of 50 images and a weight decay of 9×10^{-4}.

The LSTM-clip and LSTM-video have one LSTM layer with 512 and 4,096 cells respectively. For both models, the initial learning rate was set to 10^{-4} with weight decay of 10^{-3}, and they were trained for 30 epochs. LSTM-clip was trained with batch size of 50 clip-sequences, each contains spatial features of 11 consecutive frames. The training for the LSTM-video was conducted using one video-sequence S^V every training iteration. Otherwise padding of these sequences is required to guarantee same length.

Since the tool classification is a binary multi-label task, a sigmoid activation function was used for the tool classification output layers and the loss was computed using cross-entropy function. All experiments were implemented in Keras using NVIDIA GeForce RTX 2080 TI GPUs.

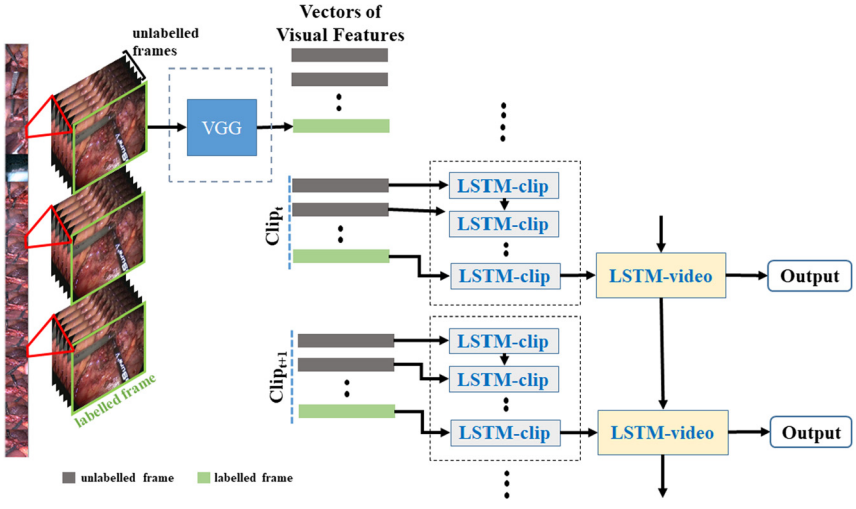

Figure 1: An overview of the proposed method for surgical tool presence detection in laparoscopic videos.

Table 1: Tool presence detection average precision (AP) of evaluated models, and a comparison with baseline methods (bold values indicate best performance for each tool).

Tool	EndoNet [2]	Endo3D [8]	GCN [6]	VGG-16	LSTM-clip	LSTM-video
Grasper	84.8	71.32	–	85.58	**86.3**	86.18
Bipolar	86.9	69.72	–	92.94	93.34	**94.72**
Hook	95.6	87.81	–	99.13	99.02	**99.23**
Scissors	58.6	**87.33**	–	72.32	74.5	80.89
Clipper	80.1	95.12	–	92.69	93.4	**96.13**
Irrigator	74.4	**96.43**	–	84.92	85.31	85.85
Specimen bag	86.8	**94.97**	–	93.58	93.18	94.26
Mean (mAP)	81.02	86.1	90.13	88.73	89.3	**91.04**

Results

Tool presence detection was evaluated using average precision (AP), defined as the area under the precision-recall curve. Table 1 shows results of the proposed model and a comparison with published methods EndoNet [2], GCN [6] and Endo3D [8]. The best classification performance was obtained after using both LSTM networks for learning temporal information.

Discussion

In this study, we present a deep learning pipeline for detecting surgical tool presence in laparoscopic videos. We propose to employ a VGG-16 network to learn discriminative features from frames and engage two cascaded LSTM networks to learn temporal dependencies among continuous frames. Experimental results demonstrate the advantage of combining spatial and temporal information to develop an effective and robust method for classifying surgical tools in laparoscopic videos. The most notable improvement is achieved after employing the second LSTM network, that is LSTM-video.

As can be seen from the aforementioned results (see Table 1), our method achieved higher performance for all tools than the baseline method EndoNet with a significant margin in terms of AP (81 vs. 91%). The gained improvement confirms that temporal features are helpful for detecting surgical tools in laparoscopic videos.

Wang et al. [6] and Chen et al. [8] tackled the problem in a similar way as we did and achieved a mean AP (mAP) of 90.13 and 86.1%, respectively. Our method, in contrast, models temporal dependencies not only across short video clips but also through the complete laparoscopic video. Moreover, Chen et al. [8] reported higher values of AP for the scissors and irrigator than our method, but for the other five tools, we achieved better results. Indeed, the model struggles to detect scissors and irrigator. This might be because of the imbalanced data problem that still affects training the CNN. Nwoye et al. [9] used a CNN with a Convolutional LSTM (ConvLSTM), and they reported a value of 92.9% for mAP which is better than our results.

The model performance depends on many aspects such as the temporal and spatial depth of the LSTM and CNN networks respectively. Therefore, a pre-evaluation was done with various pre-trained CNN architectures, where VGG-16 achieved best results. Regarding the LSTM, we used video clips with length of 11 frames in our implementation. Since the idea behind using the LSTM is to correct noise-related misclassified frames, it would be interesting to investigate whether increasing clip length could improve classification accuracy.

Overall, the proposed method provided strong classification results. However, the CNN and LSTM were trained separately and therefore additional work has to be carried out to form an end-to-end network that can produce more discriminative features.

Conclusions

This study evaluates a CNN with two-stage LSTM models for detecting surgical tool presence in laparoscopic videos. The first LSTM learns temporal features from short video clips, while the second LSTM models temporal dependencies across the entire video sequence. This approach was evaluated on the Cholec80 datasets and achieved better or similar performance compared with state-of-the-art methods.

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

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Informed consent: Informed consent is not applicable.
Ethical approval: The conducted research is not related to either human or animals use.

References

1. Forestier G, Riffaud L, Jannin P. Automatic phase prediction from low-level surgical activities. Int J Comput Ass Rad Surg 2015;10: 833–41.
2. Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imag 2016;36:86–97.
3. Funke I, Bodenstedt S, Oehme F, von Bechtolsheim F, Weitz J, Speidel S. Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2019. pp. 467–75.
4. Jin Y, Dou Q, Chen H, Yu L, Qin J, Fu CW, et al. SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans Med Imag 2017;37:1114–26.
5. Bouget D, Benenson R, Omran M, Riffaud L, Schiele B, Jannin P. Detecting surgical tools by modelling local appearance and global shape. IEEE Trans Med Imag 2015;34:2603–17.
6. 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. Cham: Springer; 2019. pp. 467–78.
7. Abdulbaki Alshirbaji T, Jalal NA, Möller K. Surgical tool classification in laparoscopic videos using convolutional neural network. Curr Dir Biomed Eng 2018;4:407–10.
8. Chen W, Feng J, Lu J, Zhou J. Endo3d: online workflow analysis for endoscopic surgeries based on 3d cnn and lstm. In: OR 2.0 Context-Aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis. Cham: Springer; 2018. pp. 97–107.
9. Nwoye CI, Mutter D, Marescaux J, Padoy N. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. Int J Comput Ass Rad Surg 2019;14:1059–67.
10. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA; 2015.