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

A comparative evaluation of spatial pooling methods for surgical tool detection

https://doi.org/10.1515/cdbme-2023-1054

Abstract: Surgical tool detection is an important aspect for recognising surgical activities and understanding surgical workflow. Laparoscopic videos represent an information source that can be used for recognising surgical tools. However, manual labelling of tool incidence and location in such data is extremely time intensive. Therefore, weakly-supervised approaches have been developed to perform tool localisation. In this study, three types of spatial pooling methods were implemented to evaluate the influence of each method on the performance of weakly-supervised model. The best achieved performance was a mean average precision (mAP) of 94% for tool classification and a f1-score of 70% for tool localisation. Experimental results showed the importance of selecting an appropriate pooling function to enhance model performance.

Keywords: Convolutional neural network (CNN), surgical tool localisation, laparoscopic videos, weakly-supervised learning.

***Corresponding author: Tamer Abdulbaki Alshirbaji:** Institute of Technical Medicine (ITeM), Furtwangen University, Jakob-Kienzle-Strasse 17, 78054 Villingen-Schwenningen, Germany, and Innovation Centre Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany, e-mail: abd@hs-furtwangen.de
Nour Aldeen Jalal: Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany, and Innovation Centre Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany.
Paul D. Docherty: Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand, and Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany.
Thomas Neumuth: Innovation Centre Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany.
Knut Moeller: Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany.

1 Introduction

Identifying surgical tools appearing in surgical images is a key component to recognising surgical activities and understanding the current situation of surgery. Conceiving surgical workflow enables the development of context-awareness systems (CAS) [1–3]. Potential applications for CAS include predicting the course of upcoming surgical activities [4], supporting surgical team in making decisions [1] and alerting them about possible hazards. Additionally, the resources of the surgical department could be optimised by predicting the time remaining for the surgery [5,6].

Laparoscopic videos offer a great source of visual information that can potentially be used to detect the surgical tools. Image-based approaches have been proposed for classifying surgical tools in laparoscopic images. In [7], a convolutional neural network (CNN) was proposed for recognising surgical phases and tools in cholecystectomy procedures. A CNN model trained using loss-sensitive learning was introduced in [8] for surgical tool classification. In other studies, temporal dependencies were incorporated with spatial information to improve classification performance of surgical tools. In [9], long short-term memory (LSTM) model was used to learn temporal information in short video clips. In a subsequent work, temporal clues across video sequences and along the complete laparoscopic video were leveraged [10].

Other studies proposed approaches capable of generating localisation bounding boxes and/or segmentation masks for the surgical tools presented in the image [11–15]. Most of the proposed approaches were based on deep learning models that required a huge amount of labelled data. However, in [11,14,15], weakly-supervised learning approaches were introduced. These approaches used solely binary signals of surgical tool presence. Nwoye et al. [11] proposed a network consisting of a CNN and a convolutional LSTM to perform classification, localisation and tracking of surgical tools. Jalal et al. adapted the ResNet-50 model to perform tool localisation in weakly supervised manner by adding a multi-map

convolutional layer and attention modules [14]. A temporal model was added to the previous approach to perform additionally tool classification and surgical phase recognition in [15].

In the aforementioned tool localisation approaches [11,14,15], a spatial pooling layer was required to get the confidence of tool presence using the localisation map of the corresponding tool. To date, no studies have investigated the influence of the spatial pooling function on the classification performance. In this work, the weakly-supervised learning approach introduced in [14] was used to evaluate the performance of different spatial pooling methods on tool detection. Three types of spatial pooling were investigated, namely global average pooling (GAP), MinMax pooling (MMP), and evidence ratio pooling (ERP).

2 Methodology

A CNN model was built to perform surgical tool localisation and classification following the architecture introduced in [14]. The architecture consists of a base CNN model (ResNet-50 [16]), attention modules, multi-map convolutional layer and pooling layers. Figure 1 depicts the model architecture.

2.1 Architecture

The core component of the tool detection approach was ResNet-50 model, since this architecture achieved high performance for classifying surgical tools as presented in [10]. The architecture of this model is composed of five convolutional blocks. Four attention modules were added after the second, third, fourth and fifth blocks to improve feature representation (see Figure 1). The attention modules were of type Squeeze-and-Excitation (SE) modules [17]. Similar to [11,14], the stride of convolutional layers in the last two blocks was decreased to enhance spatial resolution of generated feature maps.

The features learnt by the last three attention modules were combined and passed to a convolutional layer. This convolutional layer was employed to generate localisation maps for every surgical tool. The layer has a kernel size of 3×3 and a filter size of 4×7, and thus, four localisation maps were yielded for each of the seven surgical tools.

2.2 Pooling

2.2.1 Class-wise pooling

A class-wise pooling was implemented to combine the localisation maps generated by the multi-map convolutional layer into a single map for each surgical tool. The localisation map of a surgical tool was computed by using the max pooling operator. The class-wise pooling layer produces seven localisation maps, each for a surgical tool.

2.2.2 Spatial Pooling

A spatial pooling layer was applied to aggregate the localisation maps into features. The features were provided into a fully-connected layer (FC) to perform surgical tool classification. The FC layer had seven nodes and a sigmoid activation function.

Three pooling functions were implemented which are global average pooling (GAP), MinMax pooling (MMP) [18–20], and evidence ratio pooling (ERP). The performance of surgical tool classification and localisation was evaluated for each pooling function. GAP computes the average of every localisation map yielded from the class-wise pooling layer, and thus, a vector of seven features was the output of GAP.

MMP was applied according to Equation 1. This pooling method considered a number of the highest and lowest elements of each localisation map, denoted n_{top} and n_{low}, respectively. n_{top} and n_{low}were set to 50. $\mathbf{h}_{\text{top}}$ and $\mathbf{h}_{\text{low}}$ are the localisation map elements with the highest and lowest scores, respectively. The lowest score elements were weighted by a factor α. Based on the ablation study conducted in [19] α was set to 0.6.

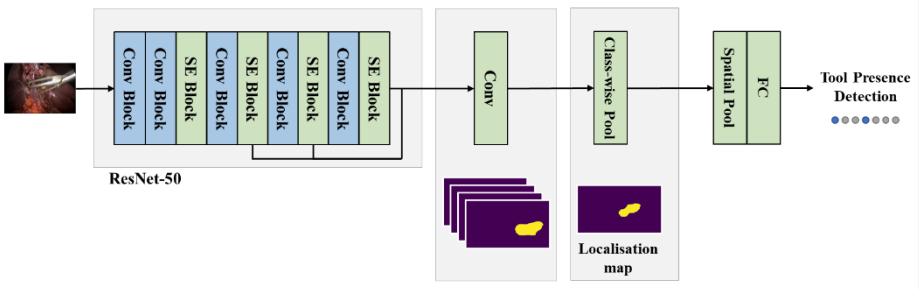

Figure 1: Architecture of the implemented model.

$$MMP = \frac{1}{n_{top}} \sum \mathbf{h_{top}} + \frac{\alpha}{n_{low}} \sum \mathbf{h_{low}} \quad (1)$$

The third pooling method (ERP) transfers every localisation map into two features. The features are the top evidence ratio (TER) and low evidence ratio (LER). *TER* and *LER* are calculated as shown in Equation 2 and 3.

$$PER = \frac{\sum_{i,j} x_{i,j}}{h.w}, \quad \begin{matrix} x_{i,j} = 1 \text{ for } x > 0.9 \\ x_{i,j} = 0, \text{otherwise} \end{matrix} \quad (2)$$

$$NER = \frac{\sum_{i,j} x_{i,j}}{h.w}, \quad \begin{matrix} x_{i,j} = 1 \text{ for } x < 0.1 \\ x_{i,j} = 0, \text{otherwise} \end{matrix} \quad (3)$$

where h and w are height and width of the tool localisation map, respectively.

2.3 Dataset

The cholec80 dataset introduced by Twinanda et al. [7] was used. The dataset contains laparoscopic videos of 80 cholecystectomy procedures. The videos were labelled for surgical tools and surgical phases. Seven surgical tools were used in the Cholec80 procedures. The model was trained using the first 40 videos of Cholec80 dataset and their tool binary labels. The remaining videos were used to evaluate the performance for surgical tool classification. The model learnt tool localisation in weakly-supervised manner. The surgical tools in five videos were labelled with bounding boxes to evaluate the model performance for tool localisation.

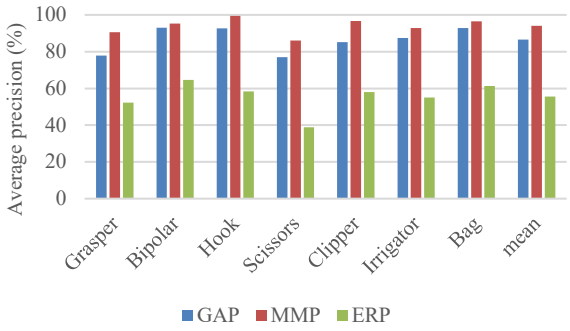

Figure 2: Average precision of surgical tool presence detection using spatial pooling methods (GAP, MMP and ERP).

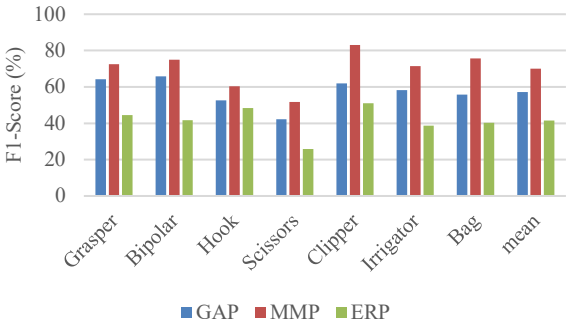

Figure 3: F1-score of surgical tool localisation using spatial pooling methods (GAP, MMP and ERP).

3 Results

Surgical tool classification and localisation were evaluated for different spatial pooling methods. Average precision (AP) was used to evaluate the performance of classifying surgical tools. Figure 2 shows the average precision achieved by the three spatial pooling methods.

The f1-score metric was used to evaluate localisation performance. F1-score was computed based on the tool presence confidence and the intersection over union (IOU) as in [14]. F1-scores for different spatial pooling (GAP, MMP and ERP) are presented in Figure 3.

4 Discussion

This work presents a comparative evaluation for various spatial pooling methods. Each of the pooling methods was integrated into a deep learning pipeline that was weakly-supervised trained to perform surgical tool classification and localisation.

MMP method achieved the best performance for surgical tools classification and localisation with a mean AP of 94.1% and a f1-score of 70%, respectively. Conventional pooling methods compute the average of a feature map (i.e., GAP) or consider solely the region with the maximum score (i.e., max pool). On the contrary, MMP does not use only the maximum scoring element in the map since it can be an outlier and cause misclassification. Hence, MMP consider a number of elements that have the highest score. Additionally, some minimum scoring elements are considered as a supportive information for the absence of a surgical tool. Thus, MMP achieved better performance than conventional pooling operators such as GAP which resulted in 86.6% mean AP and 57% f1-score.

On the other hand, ERP uses the percentage of high scoring (higher than 0.9) and low scoring (lower than 0.1) elements in the localisation map as an evidence of tool presence and tool absence, respectively. However, results showed that the ERP method had a lower classification and localisation performance than GAP and MMP. Due to the fixed upper and lower threshold in ERP method, no element was selected when the maximum score in the map was less than the upper threshold or when the minimum score was higher than the lower threshold. Consequently, the model failed to detect surgical tools in many frames.

GAP and MMP methods showed high classification performance for all surgical tools. Scissors were the only surgical tool that was classified using MMP with an AP less than 90%. Both MMP and GAP had also the lowest localisation performance for scissors. On the other hand, MMP method achieved the best classification performance for hook with an AP of 99%. However, localisation performance for

hook (f1-score of 60%) was lower than some tools such as clipper and bipolar. This lower localisation performance for the hook was due to the deviation between ground truth and detected bounding boxes. Bounding boxes of the ground truth were around the tool tip. As the model was trained solely on binary tool presence labels, the generated bounding boxes contained in many instances, in addition to the hook tip, the bottom part of the shaft. This deviation decreased IOU, and thus, lower f1-score was obtained for hook.

5 Conclusion

This study demonstrates the effect of the pooling method on the performance of a deep model developed to detect the type and location of surgical tools in laparoscopic images. Experimental results show the role of spatial pooling type on model performance. GAP and MMP can obtain more effective features than ERP. This work was conducted using a single dataset, and therefore, in future work, spatial pooling methods will be evaluated on data obtained from different sources.

Author Statement

Research funding: This work was supported by the German Federal Ministry of Research and Education (BMBF under project CoHMed/DigiMedOP grant no. 13FH5I05IA and project PersonaMed-A grant no. 13FH5I06IA). Conflict of interest: 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 animal use.

References

[1] Maier-Hein L, Vedula SS, Speidel S, Navab N, et al. Surgical data science for next-generation interventions. Nature Biomedical Engineering. 2017;1(9):691–6.

[2] Jalal NA, Abdulbaki Alshirbaji T, Docherty PD, Neumuth T, Moeller K. A deep learning framework for recognising surgical phases in laparoscopic videos. IFAC-PapersOnLine. 2021;54(15):334–9.

[3] Jalal NA, Abdulbaki Alshirbaji T, Laufer B, Docherty PD, Neumuth T, Moeller K. Analysing multi-perspective patient-related data during laparoscopic gynaecology procedures. Scientific reports. 2023;13(1):1604.

[4] Neumuth T, Rockstroh M, Franke S. Context-aware medical technologies-relief or burden for clinical users? Current Directions in Biomedical Engineering. 2018;4(1):119–22.

[5] Jalal NA, Abdulnaki Alshirbaji T, Möller K. Evaluating convolutional neural network and hidden markov model for recognising surgical phases in sigmoid resection. Current Directions in Biomedical Engineering. 2018;4(1):415–8.

[6] Jalal NA, Abdulbaki Alshirbaji T, Möller K. Predicting surgical phases using CNN-NARX neural network. Current Directions in Biomedical Engineering. 2019;5(1):405–7.

[7] Twinanda AP, 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. 2016;36(1):86–97.

[8] Abdulbaki Alshirbaji T, Jalal NA, Möller K. Surgical tool classification in laparoscopic videos using convolutional neural network. Current Directions in Biomedical Engineering. 2018;4(1):407–10.

[9] Jalal NA, Abdulbaki Alshirbaji T, Docherty PD, Neumuth T, Möller K. Surgical Tool Detection in Laparoscopic Videos by Modeling Temporal Dependencies Between Adjacent Frames. In: 8th European Medical and Biological Engineering Conference: Proceedings of the EMBEC 2020, November 29–December 3, 2020 Portorož, Slovenia. Springer; 2021. p. 1045–52.

[10] Abdulbaki Alshirbaji T, Jalal NA, Docherty PD, Neumuth T, Möller K. A deep learning spatial-temporal framework for detecting surgical tools in laparoscopic videos. Biomedical Signal Processing and Control. 2021;68:102801.

[11] Innocent Nwoye C, Mutter D, Marescaux J, Padoy N. Weakly Supervised Convolutional LSTM Approach for Tool Tracking in Laparoscopic Videos. arXiv e-prints. 2018;arXiv: 1812.01366.

[12] Sznitman R, Becker C, Fua P. Fast part-based classification for instrument detection in minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17. Springer; 2014. p. 692–9.

[13] Jin A, Yeung S, Jopling J, Krause J, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE; 2018. p. 691–9.

[14] Jalal NA, Abdulbaki Alshirbaji T, Docherty PD, Arabian H, Neumuth T, Moeller K. Surgical tool classification & localisation using attention and multi-feature fusion deep learning approach. IFAC-PapersOnLine. 2023;

[15] Jalal NA, Abdulbaki Alshirbaji T, Docherty PD, Arabian H, et al. Laparoscopic Video Analysis Using Temporal, Attention, and Multi-Feature Fusion Based-Approaches. Sensors. 2023;23(4):1958.

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7132–41.

[18] Durand T, Thome N, Cord M. Weldon: Weakly supervised learning of deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 4743–52.

[19] Durand T, Mordan T, Thome N, Cord M. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 642–51.

[20] Vardazaryan A, Mutter D, Marescaux J, Padoy N. Weakly-supervised learning for tool localization in laparoscopic videos. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3. Springer; 2018. p. 169–79.