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

Assessing Generalisation Capabilities of CNN Models for Surgical Tool Classification

Abstract: Accurate recognition of surgical tools is a crucial component in the development of robust, context-aware systems. Recently, deep learning methods have been increasingly adopted to analyse laparoscopic videos. Existing work mainly leverages the ability of convolutional neural networks (CNNs) to model visual information of laparoscopic images. However, the performance was evaluated only on data belonging to the same dataset used for training. A more comprehensive evaluation of CNN performance on data from other datasets can provide a more rigorous assessment of the approaches. In this work, we investigate the generalisation capability of different CNN architectures to classify surgical tools in laparoscopic images recorded at different institutions. This research highlights the need to determine the effect of using data from different surgical sites on CNN generalisability. Experimental results imply that training a CNN model using data from multiple sites improves generalisability to new surgical locations.

Keywords: CNN generalisability, surgical tool detection, laparoscopic images.

https://doi.org/10.1515/cdbme-2021-2121

***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

In recent decades, artificial intelligence (AI) has been progressing to develop more powerful and effective solutions in many disciplines. AI, particularly deep learning, has achieved impressive success on image and video processing applications such as object detection and action recognition [1]. Therefore, research has been initiated to develop and utilise AI solutions in the medical field to enhance and ease medical diagnosis. Moreover, AI in the surgical field has also gained interest, due to the promising potential applications that might weave the next era of operating rooms (OR) [1, 2].

Minimally invasive surgery may profit from developing AI systems. This type of surgery is performed using a laparoscopic camera that captures the view of the surgical site. Thus, the video signal is rich in information. Future OR systems could potentially use similar video signals to analyse surgical workflow. Such systems have a variety of intra- and post-operative applications. They can support the surgeon in decision-making, provide relevant information about the executed surgical task and notify the surgical team about possible complications. This may improve surgical practice and surgical outcomes. Additionally, such AI systems can enhance training novice surgeons and improve their skills [1, 3].

Analysing the surgical workflow involves many aspects. An essential aspect is surgical tool classification. Various systems and approaches have been investigated for detecting surgical tools in laparoscopic interventions. Recently, deep learning approaches have been wieldy applied in many research works. Various architectures of convolutional neural networks (CNNs) have been employed to model spatial information of laparoscopic images [4, 5]. Other approaches have leveraged temporal information encoded across complete laparoscopic videos or short clips containing unlabelled frames around a labelled one [6-9]. Long short-term memory (LSTM) [7, 8, 10], GRU [11] or Graph Convolutional Networks [9] have been used for modelling sequential data.

CNN-based approaches have demonstrated positive performance for identifying surgical tools in laparoscopic

images [5, 9]. However, the proposed CNN models were trained and evaluated using a single source of data. In other words, the data were recorded in one surgical site and for a particular type of procedure. Therefore, likely performance of CNN models across differing procedures should be evaluated on different sources of data.

Robustness of deep learning-based approaches has been addressed in some studies that handled processing laparoscopic images. In the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge 2019, the results of surgical tool segmentation were obtained on images of different procedure types [12]. In other works, the robustness of CNN models was studied for tool segmentation [13], tool detection [14] and surgical phase recognition [15, 16].

In this work, the generalisability of four CNN architectures of different depths and sizes was studied. The architectures were VGG-16, ResNet-50, DenseNet-121 and EfficientNet-B0. Two datasets containing data recorded at one hospital (single-site dataset) and at multiple hospitals (multi-site dataset) were used to conduct the study. The CNN models were trained on each dataset and their generalisation capabilities were evaluated on the other dataset. The effect of using multi-site data for training on CNN generalisability was investigated.

2 Methods

2.1 Datasets

Two datasets were used to study CNN generalisability. One of the datasets is the publicly available Cholec80 [4]. It contains laparoscopic videos for 80 procedures performed at the University Hospital of Strasbourg. The other dataset (Cholec20) was recorded at two different hospitals. It contains 20 videos. In both datasets, procedures are cholecystectomy.

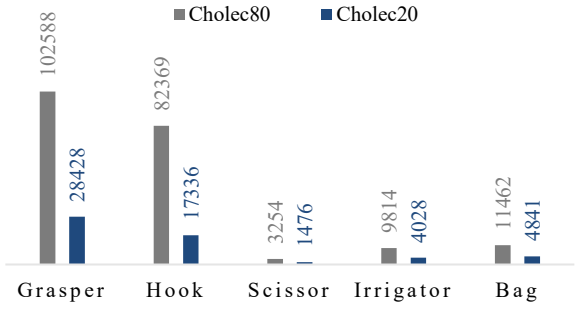

Figure 1: Distribution of surgical tools in Cholec80 and Cholec20.

Seven surgical tools were used in Cholec80 procedures. The labels for surgical tools are provided at 1 Hz. Cholec20 was labelled for surgical tools at 1 Hz also. Although the same type of procedures was executed in both datasets, some of the tools are not similar (see Table 1). Hospitals may use tools from different manufacturers; therefore, tools of the same functionality might have a different visual appearance. Thus, only similar tools presented in all videos were considered in this study. Those tools are grasper, hook, scissors, irrigator and bag.

2.2 CNN Training

Four CNN models of different architectures were used. The models are VGG-16, ResNet-50, DenseNet-121 and EfficientNet-B0. The models differ in depth and number of learnable parameters.

The architectures of the CNN models were modulated for the tool classification task. The last layer in each model was replaced by a fully-connected layer with five nodes, each for a surgical tool. In this layer, the sigmoid activation function was used since the task is a binary classification.

In both datasets, some surgical tools appear more often than others. For instance, grasper is present in most images, because it is used alone or with other tools more frequently for holding anatomical structures or executing surgical actions. Figure 1 shows the distribution of surgical tools in Cholec80 and Cholec20 datasets. To reduce the biasing effect of the imbalanced distribution of training samples, loss-sensitive learning was adapted [5]. The loss was computed for every tool using the binary cross-entropy function [6].

Table 1: Presented surgical tools in Cholec80 and Cholec20 datasets. (✓) indicates tool presence and (-) indicates that tool is not present or has a different visual appearance.

Surgical Tool	Cholec80	Cholec20
Grasper	✓	✓
Bipolar	✓	-
Hook	✓	✓
Scissors	✓	✓
Clipper	✓	-
Irrigator	✓	✓
Bag	✓	✓

Two experiments were performed using the modulated CNN models. In the first experiment, denoted as single-site training, models were trained on data from a single surgical location. While in the second experiment, denoted as multi-

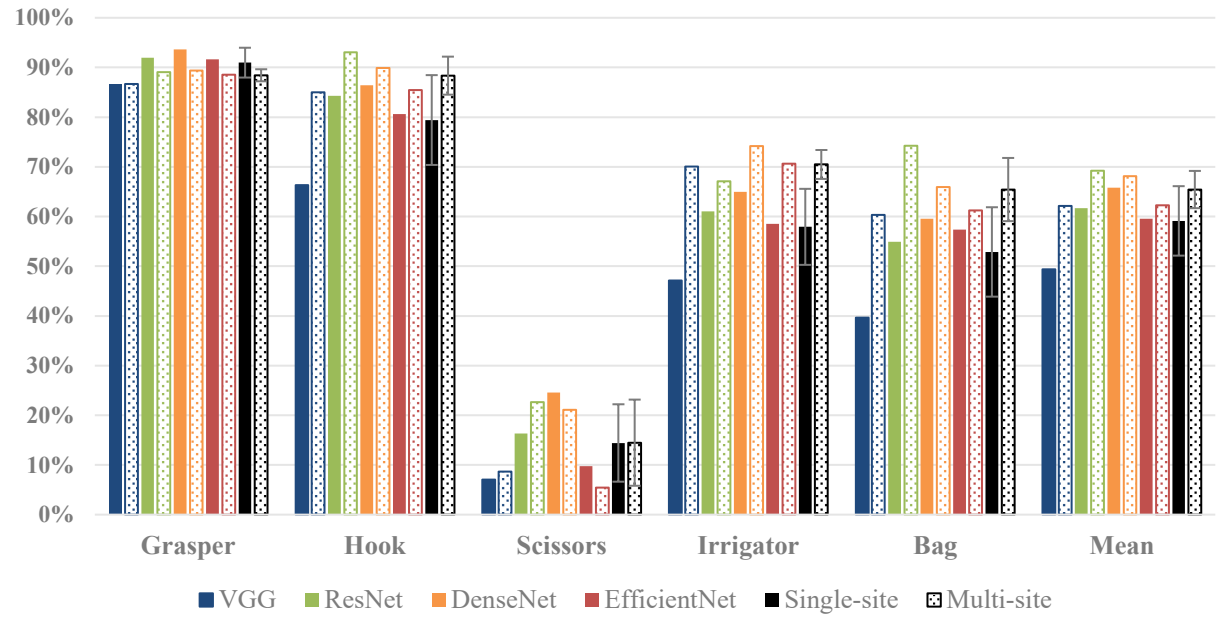

Figure 2: Average precision of surgical tools for VGG-16, ResNet-52, DenseNet-121 and EfficientNet-B0 using single- and multi-site training. Solid bars and Dotted bars are the results of CNN model trained on single-site and multi-site data, respectively. The bars in black colour are the mean and standard deviation over all models trained on single-site and multi-site data.

site training, the models were trained on data recorded at multiple surgical locations. The trained models, from each experiment, were tested on the other dataset, which did not contribute to the training. Table 2 presents a description of the training and testing data used in each experiment. For a fair comparison between single- and multi-site training, roughly the same number of training images was used in both experiments.

Models were trained using transfer learning approach. Thus, they were initialised with weights learned on ImageNet[17]. The initial learning rate was set to 1×10^{-4} and weight decay was set to 9×10^{-4}. Implementation was performed using Keras framework on a computer with an NVIDIA GeForce RTX 2080Ti GPU.

3 Results

The average precision metric was used to evaluate the classification performance of CNN models. Figure 2 shows results of testing CNN models on unseen data from other surgical locations. The abbreviation "-S" and "-M" is added to the model's name and refers to the results of single-site training and multi-site training experiments, respectively. Additionally, Figure 2 presents mean and standard deviation over the different CNN models and over the surgical tools.

Table 2: Description of training and testing set used in the conducted experiments.

Experiment	Training source	Number of training images	Testing set
Single-site Training	Cholec80	46078	Cholec20
Multi-site Training	Cholec20	46227	Cholec80

4 Discussion

In this study, generalisation of four CNN models to unseen data from different surgical sites was evaluated. Moreover, the improvement of using data from multiple surgical sites on CNN generalisability was assessed. The aim of this work is to support the deployment of deep learning solutions in real settings across different hospitals.

The CNN models show high generalisation capability for grasper in both experiments. On the contrary, the models were poorly generalised to identify scissors in across surgical settings. These results are due to characteristics of surgical data. They are not representative and not equally distributed over the classes. Additionally, capturing and labelling surgical data is challenging and therefore, available labelled datasets are limited in size. Those characteristics impact the training process of deep learning models. In our case, the distribution of data (shown in Figure 1) has biasing effect towards the over-

presented tool and thereby better learning and generalisation for grasper and lower performance for under-presented tools (scissors).

Results imply enhancement in classification performance of multi-site training over single-site training. Data acquired from multiple surgical sites are more diverse and thus improves generalisation capabilities of deep learning models. The studied models showed better generalisation for most tools when the models were trained on multi-site data. VGG-16 model has the most notable improvement (~23%) for detecting irrigator. The mean average precision of hook, irrigator and bag over the four CNN models were improved by 9%, 13% and 13%, respectively.

This study has two limitations. The first limitation is the small size of training data. Cholec20 has about 55.5k labelled images. 80% of Cholec20 was used in the multi-site training experiment, and the same amount of data was used for single-site training. Secondly, this study was conducted using one type of procedure which is cholecystectomy. Thus, CNN generalisation to other procedures is yet to be investigated.

5 Conclusion

Experimental results imply that training on data from different surgical sites improves generalisability of CNN models. Nevertheless, this study shows that acceptable generalisation is possible for surgical tools which are well presented in the training data. In future work, generalisation of CNN to different types of procedures will be studied.

Author Statement

Research funding: This work was supported by the German Federal Ministry of Research and Education (BMBF under grant CoHMed/IntelliMed grant no. 13FH5I01IA and 13FH5I05IA). Conflict of interest: Authors state no conflict of interest. Informed consent: A written informed consent was collected from each participant. Ethical approval: The ethical approval was obtained from the ethics commission of the Furtwangen University (application Nr. 19 -0306LEKHFU).

References

[1] Egert, M., Steward J. E., Sundaram C. P.: Machine learning and artificial intelligence in surgical fields. Indian journal of surgical oncology, 1-5 (2020).

[2] Bodenstedt, S., Wagner M., Müller-Stich B. P., Weitz J., Speidel S.: Artificial Intelligence-Assisted Surgery: Potential and Challenges. Visceral Medicine 36 (6), 450-455 (2020).

[3] Zhou, X.-Y., Guo Y., Shen M., Yang G.-Z.: Artificial intelligence in surgery. arXiv preprint arXiv:200100627, (2019).

[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 Trans Med Imaging 36 (1), 86-97 (2016).

[5] Alshirbaji, T. A., Jalal N. A., Möller K.: Surgical tool classification in laparoscopic videos using convolutional neural network. Current Directions in Biomedical Engineering 4 (1), 407-410 (2018).

[6] 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).

[7] Abdulbaki Alshirbaji, T., 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 (1), (2020).

[8] Jalal, N. A., Alshirbaji T. A., 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).

[9] 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).

[10] Wittenstein, J., Huhle R., Scharffenberg M., Kiss T., Herold J., Vivona L., Bergamaschi A., Schultz M. J., Pelosi P., Gama de Abreu M., Bluth T.: Effects of two stepwise lung recruitment strategies on respiratory function and haemodynamics in anaesthetised pigs: A randomised crossover study. Eur J Anaesthesiol 38 (6), 634-643 (2021).

[11] Namazi, B., Sankaranarayanan G., Devarajan V.: A contextual detector of surgical tools in laparoscopic videos using deep learning. Surg Endosc, 1-10 (2021).

[12] Ross, T., Reinke A., Full P. M., Wagner M., Kenngott H., Apitz M., Hempe H., Filimon D. M., Scholz P., Tran T. N.: Robust medical instrument segmentation challenge 2019. arXiv preprint arXiv:200310299, (2020).

[13] Ross, T., Zimmerer D., Vemuri A., Isensee F., Wiesenfarth M., Bodenstedt S., Both F., Kessler P., Wagner M., Müller B.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery 13 (6), 925-933 (2018).

[14] Abdulbaki Alshirbaji, T., Jalal N. A., Docherty P. D., Neumuth T., Möller K.: Cross-dataset evaluation of a CNN-based approach for surgical tool detection. In: The 15th Interdisciplinary Symposium Automation in Medical Engineering (AUTOMED 2021), Zenodo Basel, Switzerland (2021).

[15] 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 (1), 415-418 (2018).

[16] Bar, O., Neimark D., Zohar M., Hager G. D., Girshick R., Fried G. M., Wolf T., Asselmann D.: Impact of data on generalization of AI for surgical intelligence applications. Scientific reports 10 (1), 1-12 (2020).

[17] Krizhevsky, A., Sutskever I., Hinton G. E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097-1105 (2012).