Computer Assisted Instruction in Laparoscopic Surgery using Deep Learning

Jaafari Jaafar[1,*], *Pr. Douzi Samira*[2], *Pr. Douzi Khadija*[1], *and Pr. Hssina Badr*[1]

[1]Faculty of Sciences and Technologies of Mohammedia, Hassan II University of Casablanca, Morocco,

[2]Faculty Of Medicine And Pharmacy, Mohammed V University, Rabat, Morocco,

Abstract. Minimally invasive surgery (MIS) is famous to cause less harm to the skin compared with regular operation, due to the tiny surgical instruments and the small incisions used. It provides many advantages to the patients like a shorter hospital stays, reduced pain and faster recovery. In addition, MIS offers the possibility of video record the surgery. These videos are used for teaching purposes, evaluating surgeons and also they are treated as evidence in case of lawsuits from patients. On the other hand, these types of surgeries are difficult to learn and teach. That's why surgeons tend to check MIS videos for a possible technical error. Since MIS medias are commonly very long, this manual surgical quality assessment (SQA) process, without any support of video search, take so much time and effort. To surmount this issue, we present a neural network based solution, to identify surgical instruments and index these videos, using three fine-tuned Convolutional Neural Network VGG19, Inception v-4 and NASNet-A. Finally, we present the benefits of the proposed approach on the Cholec80 dataset.

1- Introduction

In endoscopic surgery, surgeons use video technology to perform operations with a fewer incisions instead of the large cut of the open surgery. It offers plentiful benefits compared to open surgery [1] in different types of operations. MIS leaves smaller and less visible scars, it has a reduced risk of infections [2] and it is characterized by a short and less painful recovery time.

There are two majors' types of MIS: Laparoscopic surgery and Robotic surgery.

- In **Laparoscopy**, small incisions are made in order to pump carbon dioxide gas to inflate the abdomen, which allows the surgeon to see the organs more clearly, then a small tube called laparoscope that contains a high-resolution camera and a light source is used.
- On the other side, **Robotic surgery** allows doctors to execute numerous forms of difficult operations more precisely, flexibility and control.

The videos recorded during the surgery offers limitless possibilities. In fact, in some countries, it is mandatory to store the surgery as evidence; because it is the only proof that surgeons didn't make a mistake.

On the other hand, those videos are widely used in the teaching field; because junior surgeons are not allowed to operate on real patients, then, surgery videos are a precious tool to learn procedures [3]. Particularly, there are plenty of online platforms that offer learning surgery procedures, including YouTube.

Human error and negative events can go unnoticed during the surgery, that's why Surgical Quality Assessment (SQA) , using the video recorded surgery is very important. SQA is an internal verification approach, to evaluate the surgeons via the surgery videos and give a review based on some standardized rating checklist.

Nevertheless, surgery videotapes could reach few hours easily. Therefore, searching a specific segment and navigating over these medias is time consuming.

To solve this problem, we present a solution based on CNN (convolutional neural networks), to detect and classify surgical instruments, in surgery videos, then, store the outcomes in a database, in order to perform particular requests, and enable junior surgeons and post-operative controllers, to manage and access to requested video segment.

The paper is organized as follow: Sec. II defines the computer vision concepts. Sec. III presents the related works, while Sec. IV describes the methodology, whereas Sec. V contains the experimental results and the discussion. Finally, the sec. VI contains the conclusion and future works.

2- Computer vision

Machine learning algorithms are trained to improve automatically from the learned experience over time, with the aim of achieving better results. The objective is to imitate the human and acquire the ability to learn through experience.

Deep Learning algorithms are a subset of machine learning, it is a mechanism that uses deep-layered neural network architectures. It had seen an important rise in

*Corresponding author: jaafar.jaafari@etu.fstm.ac.ma

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

the last decade, impelled by the publication of massive datasets and the expansion of machine power, because a deep artificial neural network model needs more time and CPU power compared to other machine learning models.
Computer vision (CV) is the most benefited field from deep learning. The aim of computer vision is to grant machines with humanoid awareness skills, in order to feel the environment, figure out the felled data, take suitable actions and learn from this experience in order to improve future acts.
Convolutional neural networks (CNNs) [4] are a branch of deep neural networks, most frequently used for image recognition and classification. CNNs have several layers of neurones that extract data from pictures to determine their classes. There are constituted of three majors types of layers: Convolutional layer, Pooling Layer, and Fully connected layer.
Convolutional Neural Network (CNN) has made state-of-art results in the field of medicine.

3- Related works

Different approaches have been issued on surgical tool detection. Kranzfelder [5] used radio frequency identification (RFID) tags to track surgical instruments in laparoscopic surgery. Twinanda [6] introduced Endonet which is a deep convolutional neural networks (CNN) for phase recognition and tool presence detection tasks in a multi-task manner on laparoscopic videos. Kletz [7] presented a R-CNN, in order to identify the surgical instrument in a personalized laparoscopic gynecological dataset. Amy. J [8] proposed a tool detection and operative skill assessment in surgical videos using R-CNN network and VGG-16 on M2CAI dataset. Wang [9] associates two SOTA networks namely VGGNet and GoogLeNet, then used this combination to classify surgical instruments. Kanakatte [10] used a segmentation algorithm, which segments and localizes the surgical tool using spatiotemporal deep network. Colleoni [11] presented an encoder–decoder architecture to detect and localize surgical instrument joint using three-dimensional convolutional layers to exploit spatiotemporal features from laparoscopic videos. Rocha [12] proposed a self-supervised approach in a robot-assisted context based on kinematic model of the robot in order to generate training labels.

4- Methodology

In this section, we present the main architecture of our algorithm to detect surgical tools.
The appearance of surgical tools in laparoscopic surgery videos is defined by the visual features. We approach the task of identifying these tools by using three variants of CNN. Moreover, we used myriad data augmentation methods and measure the achievement of the proposed models on Cholec80 dataset. Then, we implement the classification outcomes in a framework intended for teaching purposes.

a. Network architectures

In this research, three deep architectures were used because they are the state-of-the-art in the computer vision community:
- VGG-19 [13]: Visual Geometry Group (VGG) is a convolutional neural network that is 19 layers deep. It is a variant of the VGG model and consist of 16 convolution layers, 3 Fully connected layer, 5 MaxPool layers and 1 SoftMax layer. The default input RGB image size of this network is 224 x 224 pixels. It is pre-trained on the well-known ImageNet [16] dataset, for that reason, the final connected layer consists of 1,000 channels, representing the number of ImageNet classes.
- Inception v-4 [14] : is a convolutional neural network that introduce the inception layer. It allows the internal layers to pick and choose which filter size will be relevant to learn the required information. Thus, there is no need to think of which filter size should be used at each layer. The default input RGB image size of this network is 299 x 299.
- NASNet-A [15]: in Search Architecture (NAS) Network, the blocks or cells are not predefined by authors. Instead, they are searched by reinforcement learning search method. It obtains state of the art performance on CIFAR-10 and ImageNet competition. The default input RGB image size of this network is 331 x 331.
Transfer learning is utilizing a network trained on a dataset, and uses it on new image/object categories. Fundamentally, we can employ the powerness of filters learned by SOTA architectures on huge sets of data such as ImageNet, to classify new data. It has manifold advantages, but the essential profits are reducing the training period, a better results and the possibility to train models using little data. For these reasons, we fine-tuned these models, pre-trained earlier on ImageNet dataset.

b. Dataset and preprocessing

The Cholec80 [6] dataset is used for performance evaluation. Cholec80 is a minimally invasive surgery video dataset including 80 videos of cholecystectomy surgeries, with over 75 hours of recordings. The medias are recorded at 25 fps(frame per second) and annotated in image-level surgical tool labels for 1 fps. The different tools (Fig.1) present in this dataset are grasper, bipolar, specimen bag, clipper, hook, scissors and irrigator.

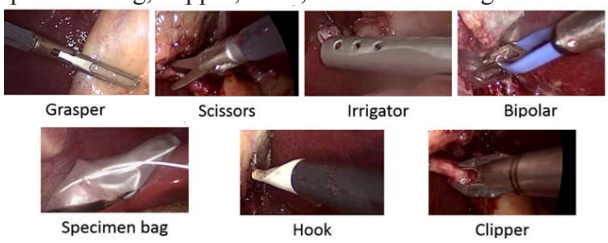

Fig.1. Surgical tools annotated in videos

To split the Cholec80 videos, we used FFmpeg v3.0 in one frame per second images, then we resize them from

1920 x 1080 pixels to 224 x 224 pixels, 299 x 299 and 331 x 331, in order to fit our models.

The distribution of image per surgical tools is shown in fig.2. As we can see, some tools are used more frequently than the others, leading to an unbalanced data issue.

Imbalanced data refers to classification problem where data inputs are not displayed equally. Considering the majority of machine learning algorithms get a better accuracy in balanced data, using the unbalanced dataset Cholec80 could lead to an unsatisfactory predictive performance, especially for the minority class.

To overcome this problem, we over-sample images from the under-represented classes, by adding more copies of these instances. Image transformation techniques like rotation, mirroring and padding are used.

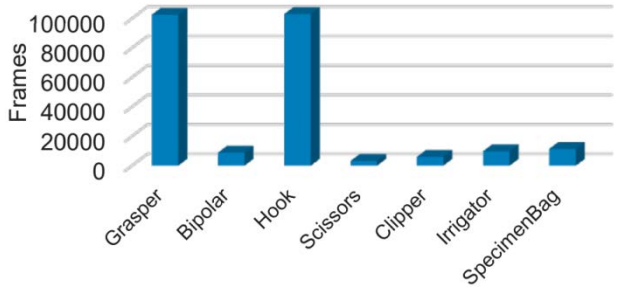

Fig. 2. Distribution of surgical tools

5- Results and discussion

a. Implementation parameters

EfficientNet model was applied using TensorFlow, with weights pre-trained on ImageNet. Our experiments are performed using an online implementation of this architecture. The original size of the last layer was 1,000 categories, corresponding to the number of ImageNet classes. It was reduced to seven neurons, which represent the number of surgical tools in Cholec80 dataset. As we mention earlier, Cholec80 dataset contains 80 videos of gallbladder removal surgery. We divide the dataset into 60 videos (75%) and 20 videos (25%) (for training and testing respectively).
The performance of the proposed approach on the classification task is calculated based on the average precision (AP), which is a measure that combines recall and precision for a particular surgical tool. The AP is calculated as:

$$AP = \frac{1}{k} \sum_{Recall_i} Precision(Recall_i) \quad (1)$$

where

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

and

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The achievement of all the proposed architectures is measured by the mean Average Precision (mAP). It is represented as:

$$mAP = \frac{1}{7} \sum_{i=1}^{7} AP_i \quad (4)$$

Where 7 is the total surgical instruments.
Since multiple instruments can be present simultaneously, the tool detection is Multilabel and Multiclass task, therefore, final activation function is a Sigmoid function, on the other hand, the binary Cross-Entropy function is utilized as Loss function.
The network is trained on Intel® Core™ i7-9700K Processor, 16GB, NVIDIA Geforce GTX 2080.

b. Results and discussion

Surgical tool appearance for image-level detection is investigating the presence or not of each surgical instruments.

The results are summarized in Table 1. It presents the AP per instrument, in addition to the mAP for each approach. A high average precision means that the models classify correctly the surgical tool.

As illustrated by Table 1, the NASNet-A architecture (96.45%) outperforms the Inception v-4 (93.01%) and VGG19 (95.85%) architecture in the classification task.

The Hook achieves the highest average precision among all the surgical tools, followed by the clipper in all the models. One possible explanation is that they have a high number of sampling (Fig.2). Another potential reason is their specific tip shape, making it easily detectable. Clipper, grasper, and specimen bag have also a encouraging results, whereas bipolar and irrigator are often misclassified.

Furthermore, in EndoNet [6] and Amy.J [8], the authors didn't use augmentation data to overcome the unbalanced data issue, that explains the low average precision of the Scissors (58.60% and 70.80% respectively) , which have

Table 1. Average precision of the frame-level presence per class and mAP

Tool	EndoNet[6]	Amy.J[8]	VGG19	Inception v-4	NasNet-A
Grasper	84.8	87.2	97.89	96.54	97.32
Bipolar	86.9	75.1	96.72	94.33	97.11
Hook	95.6	95.3	99.83	99.70	99.89
Scissors	58.6	70.8	87.59	80.84	90.06
Clipper	80.1	88.4	97.65	93.67	98.54
Irrigator	74.4	73.5	96.10	92.08	95.91
SpecimenBag	86.8	82.1	95.21	93.94	96.35
Average (mAP)	81.0	81.8	95.85	93.01	96.45

the minor representation in the dataset. In our case, NASNet-A, Inception v-4 and VGG19 made an average precision of 87.59%, 80.84% and 90.06% respectively.

We can also notice remarkable differences between the mean average precision of the proposed and the old approaches, with an improvement of more than 15%.

6- Conclusion and future works

Automatic surgical tools classification from the laparoscopic surgery video can help the surgeons community to easily learn difficult procedures and also as a surgical quality assessment tool. An example of the utility of this task is shown in fig.3. It represents a framework based on the result of this study, allowing surgeons to navigate easily in the minimally invasive surgery video, by searching specific tools.

In this work, we presented a deep learning classifier, based on three different Convolution Neural Network to detect and classify surgical tools appearing in the Cholec80 dataset, which is minimally invasive surgery video collection, including 80 videos of cholecystectomy surgeries. The proposed models perform better than the other studies, with a mean average precision of 96.45%, 93.01% and 95.85% for the NASNet-A, Inception v-4 and VGG19 architectures.

The unbalanced data problem of Cholec80 dataset is resolved by some data augmentation techniques. The average precision of the minority class (Scissors) increased significantly, from 58.60% and 70.80% in the compared approaches, to 87.59%, 80.84% and 90.06% in the proposed models.

However, the weakness of the proposed approaches is that the temporal information of the videos are ignored, since we split the video to images.

In future works, we plan to use other neural network architectures, with image enhancement techniques, to raise the average precision of the model. On the other hand, we project to use the temporal information.

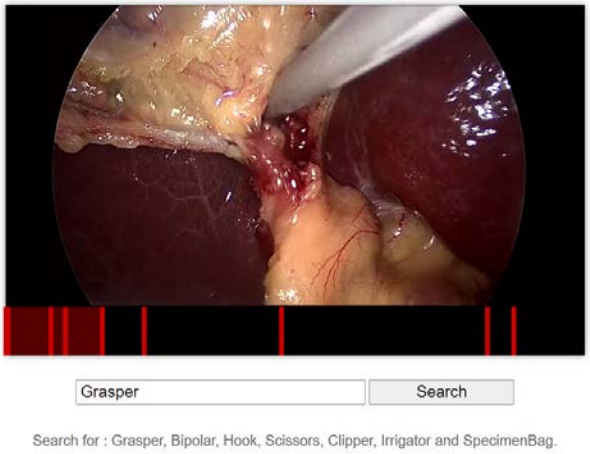

Fig. 3. Prototype of the proposed web-based system for surgical tools navigation in MIS videos

References

1. Chen, Q., Merath, K., Bagante, F., Akgul, O., Dillhoff, M., Cloyd, J., &Pawlik, T. M. (2018). A Comparison of Open and Minimally Invasive Surgery for Hepatic and Pancreatic Resections Among the Medicare Population. Journal of Gastrointestinal Surgery. doi:10.1007/s11605-018-3883-x

2. Ee, W. W. G., Lau, W. L. J., Yeo, W., Von Bing, Y.,& Yue, W. M. (2013). Does Minimally InvasiveSurgery Have a Lower Risk of Surgical Site Infections Compared With Open Spinal Surgery?

3. Mota, P., Carvalho, N., Carvalho-Dias, E., Jo˜ao Costa, M., Correia-Pinto, J., & Lima, E. (2018). Video-Based Surgical Learning: Improving Trainee Education and Preparation for Surgery. Journal of Surgica Education, 75(3), 828–835. doi:10.1016/j.jsurg.2017.09.027

4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015)

5. Kranzfelder. M, Schneider. A, Fiolka. A, Schwan. E, Gillen. S, Wilhelm. D and Feussner. H, Real-time instrument detection in minimally invasive surgery ising radiofrequency identification technology, in J.Surg. Res., vol. 185, 2013, pp.704-710.

6. Twinanda A.P, Shehata S., Mutter. D, Marescaux. J, De Mathelin. M and Padoy. N, Endonet: A deep architecture for recognition tasks on laparoscopic videos, IEEE Trans. Med. Imag, vol. 36, 2016, pp. 86-97.

7. Kletz, S., Schoeffmann, K., Benois-Pineau, J., &Husslein, H. (2019). Identifying Surgical Instruments in Laparoscopy Using Deep Learning Instance Segmentation. 2019 International Conference on Content-Based Multimedia Indexing (CBMI). doi:10.1109/cbmi.2019.8877379

8. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks - Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, ArnoldMilstein, Li Fei-Fei. (2018) IEEE Winter Conference on Applications of Computer Vision (WACV)

9. Wang, S., Raju, A., & Huang, J. (2017). Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos.
10. Kanakatte, Aparna; Ramaswamy, Akshaya; Gubbi, Jayavardhana; Ghose, Avik; Purushothaman, Balamuralidhar (2020). 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) - Surgical tool segmentation and localization using spatio-temporal deep network. , 1658–1661. doi:10.1109/EMBC44109.2020.9176676
11. Colleoni, Emanuele; Moccia, Sara; Du, Xiaofei; De Momi, Elena; Stoyanov, Danail (2019). Deep Learning Based Robotic Tool Detection and Articulation Estimation With Spatio-Temporal Layers. IEEE Robotics and Automation Letters, 4(3), 2714–2721. doi:10.1109/LRA.2019.2917163
12. Cristian da Costa Rocha; Nicolas Padoy; and Benoit Rosa (2019). Self-Supervised Surgical Tool Segmentation using Kinematic Information. International Conference on Robotics and Automation (ICRA) Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019
13. Simonyan, K., Zisserman, A., May 2015. Very deep convolutional networks for large-scale image recognition. In: Proc ICLR. San Diego, CA, USA.
14. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., Feb. 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proc AAAI. San Francisco, CA, USA, pp. 4278–4284.
15. Zoph, B., Vasudevan, V., Shlens, J., Le, Q. V., Jul. 2017. Learning transferable architectures for scalable image recognition. arXiv:1707.07012 [cs, stat].
16. Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2009.5206848