Exploring the Benefits of Multi-Task Learning in Surgical Phase Recognition and Tool Detection

Drishti Sharma Author[a]

[a]Queen's University, Kingston, Canada

1. ABSTRACT

Surgical video analysis involves two critical and complex tasks: detecting the presence of surgical tools and recognizing the surgical phase. These tasks are vital for many modern operating room applications. Although they are interrelated in clinical practice because of the well-defined surgical process, previous methods have tackled them separately, without leveraging their connection. In this project, I have implemented an approach that introduces a multi-task recurrent convolutional network with correlation loss (MTRCNet-CL) that utilizes their relatedness to enhance the performance of both tasks simultaneously.

2. INTRODUCTION

Recognising the surgical phase and tools can help clinicians during and after the operation. During the surgery, it helps by generating real-time warnings by detecting some unexpected anomalies. It can also help estimate the time to complete the surgery and thus in better scheduling of the resources. On the other hand, after the surgery, the data can be used for documentation and indexing purpose which can further be used for training and skill assessment. But recognising the surgical phase and tools is a very challenging task. Various features have been used for phase recognition tasks but so far they were either manually annotated or handcrafted which lead to the loss of other significant characteristics during feature extraction.

Convolutional Neural Network(CNN) is used to learn visual features in surgical videos for phase recognition and overcome the above-mentioned challenges. A lot of work has been done in the past for phase detection using Neural networks. It includes the work of DiPietro, Lea, Malpani, Ahmidi, Vedula, Lee, Lee, Hager, 2016,[1]Sahu, Mukhopadhyay, Szengel, Zachow, 2017.[2] However, all these works consider phase recognition and tool detection as two independent tasks. The above-mentioned works ignore the high correlation between the surgical phase and tool usage.

In the realm of surgery, there is a strong relationship between the different phases of a surgical procedure and the tools that are used during each phase. Surgeons are expected to utilize specific instruments for each stage of the procedure to ensure that the operation is carried out effectively and safely. An example of this is the use of hooks during the dissection phase and clippers and scissors during the cutting and clipping stage of a cholecystectomy procedure.

Thus, tool detection can be further used to generate better features for phase recognition tasks.

I found the paper that Yeung Jin et al.[3] wrote on these concepts very interesting. After reading the current literature on this topic, I was inspired by the results of the paper written by Yueming Jin et al. I believed that implementing this methodology would allow me to gain a deeper understanding of the concepts involved and provide valuable insights into the application of the approach in a practical context.

In this report, I present the details of my implementation of the Multi-task recurrent convolutional network with correlation loss for surgical video analysis. I also present the results of my implementation and discuss the insights that I gained from this project. In future work, I have mentioned a new project that I would like to work on based on my current work.

In conclusion, this project has created a solid foundation for me so I am able to continue exploring the possibilities in surgical video analysis.

Further author information: (Send correspondence to A.A.A.)
A.A.A.: E-mail: 21ds128@queensu.ca

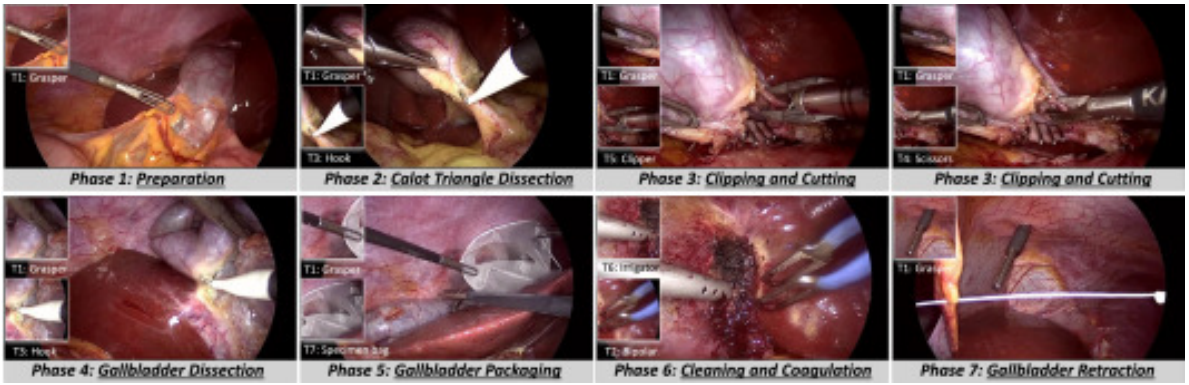

Figure 1. Figure Showing how phase recognition and tool detection are correlated, taking cholecystectomy procedure as an example[3]

3. RELATED WORK

3.1 Surgical Video Analysis

Until now, surgical video analysis has typically treated phase recognition and tool detection as separate tasks. In earlier methods, some researchers approached this problem as a multilabel classification task, leveraging the correlations among the various tools employed during the surgery.Wang et al. (2017)[4] Sahu et al. (2017)[2]Addressed it as a multi-label imbalance problem of surgical tool detection using CNN. Al Hajj et al. (2017)[5] used multi-image fusion inside a convolutional neural for surgical tool detection in a cataract surgery video network. Wang et al. (2019)[6] used Graph convolutional nets for tool presence detection in surgical videos.

3.2 MultiTask Learning

As previously stated, knowing how tools are used can be beneficial for determining the input signal's phase. Consequently, by integrating tool presence detection with phase recognition, information about tool usage can be employed to indirectly improve phase recognition through shared features.

Andru P. Twinanda employed EndoNet, To carry out both phase recognition and tool presence detection tasks concurrently,a CNN architecture was employed. This architecture comprised two branches that shared initial layers, which were utilized to extract visual features. In order to improve the accuracy of phase recognition by imposing temporal constraints, a Hierarchical HMM was implemented.[7] However, despite its impressive performance, the unified framework did not fully integrate the temporal dependencies needed for phase analysis.

4. METHODS AND MATERIALS

The project has four training classes. mtrcnet, "mtrcnet_cl", singlenet_tool and singlenet_phase. Thus, in the project, I have compared the performance of four algorithms. Two single-task algorithms are used, one to recognise phase(train_singlenet_phase) and the other to detect tools(train_singlenet_tool). In the Multi-task recurrent convolutional network algorithm, I have used multitask network to train for both phase recognition and tool detection. And in Multi-task recurrent convolutional network with Correlation Loss, other than multi-task I have also used correlation loss to train for both phase recognition and tool detection.

In MTRCNet, there are two branches of the network that share early feature encoders, meaning that they both receive the same initial set of low-level features as input. However, each branch has its own set of higher-level layers that are specifically designed for a different task, allowing the network to perform multiple tasks simultaneously while leveraging the same initial feature extraction process.

The phase recognition task requires the network to analyze a sequence of input data over time (e.g. video frames or sensor data) and make predictions based on the temporal evolution of the data. This is in contrast to the tool presence detection task, which can be performed on individual frames or snapshots of the surgical scene. Thus, LSTM is embedded in the phase branch.

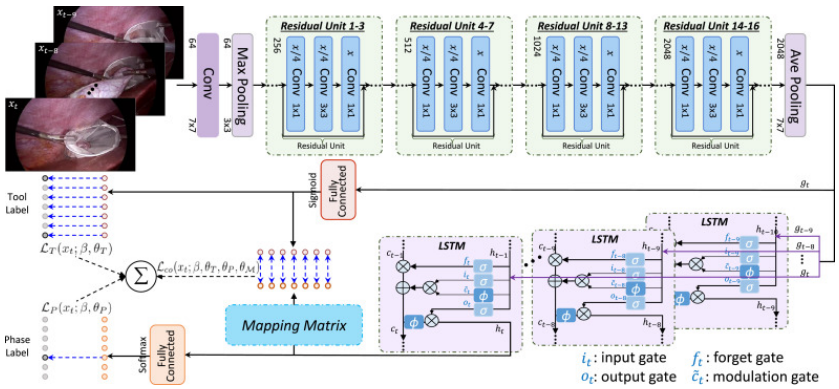

Figure 2. Figure showing The overview of MTRNET-CL[3]

And finally, The correlation loss is introduced to encourage the network to produce output probabilities that are consistent with each other, thus leveraging this natural correlation between the two tasks. Specifically, the loss penalizes the divergence between the distributions of predicted probabilities for the two tasks, encouraging the network to produce outputs that are more aligned with each other.

To address the challenges of identifying surgical videos, I utilized a deep learning approach known as a 50-layer residual convolutional network to extract meaningful and high-level features from the videos. Each convolutional layer applies a set of filters to the input image to extract various features. The network architecture comprises multiple residual units, with each unit consisting of three convolutional layers, a batch normalization layer, and a ReLU activation function layer. By stacking 16 of these units, I constructed a deep residual network that can better identify intricate details in the surgical videos. The network produced a 2048-dimensional feature vector by applying an average pooling layer to the backbone part of the network. This 50-layer network serves as the backbone layer for the two branches.

A fully-connected layer is directly connected to the backbone network with a sigmoid layer followed to produce predictions for the tools. As mentioned earlier, for phase recognition which relies on temporal information, I have connected the shared backbone layers with LSTM units in this branch. This forms a recurrent convolutional network to process multi-tasks on surgical videos (MTRCNet).

Through end-to-end training of the entire framework, it becomes feasible to simultaneously and interactively identify both the tool and phase. The initial layers extract common visual features, while the two branches are optimized jointly, resulting in mutual learning benefits for both tasks.

For correlation loss, I have constructed a correlation cell, i.e. a mapping matrix with 128×7 dimensions to linearly cast the high-dimensional spatial-temporal features to a compact semantic space with the meanings of surgical tools. Furthermore, the divergence of the probability distributions of the tool usage is minimized via a derived correlation loss, penalizing the inconsistency between the tool branch and the phase branch. The correlation loss for the multi-tasking network is based on KL-Divergence.

the correlation loss forces the phase branch to encode tool presence information into feature vectors, and meanwhile, it constrains the tool branch to take into account phase representation by enforcing the tool branch to learn from the perspective of phase.[3]

For data processing, I have converted the videos to frames using FFmpeg, then downsampled to 1fps from 25fps and then Resized the original frame to the resolution of 250 * 250. With an aim to enlarge the dataset, I performed the data augmentations with cropping and mirroring.

4.1 Dataset

I utilized the Cholec80 dataset in my study, which is a compilation of 80 endoscopic videos depicting cholecystectomy surgeries performed by 13 surgeons. The videos were recorded at a capture rate of 25 frames per second

but were downsampled to 1 frame per second for processing purposes. The dataset includes annotations for both the surgical phase and the presence of surgical tools.

All frames in the dataset are labelled with one of the 7 predefined surgical phases by experienced surgeons. Additionally, tool annotations are conducted at a 1 fps resampling rate and consist of 7 categories. Recognizing tools visually can be challenging as they may not always be clearly visible in the images. Therefore, the presence of a tool is defined as when at least half of its tooltip is visible in the image

Out of the 80 videos, 40 were planned to be used as training videos, and the rest 40 videos as testing. however, due to computational resource limitations, the current stated results are based on the limited dataset of 10 videos.

Out of the 10 videos, 5 are for training and 5 for testing.

5. RESULTS

To compare the performance of a single-task network, a multi-task network and a multi-task network with correlation loss, we have the accuracies of four different configurations. 1. Independent single-task tool detection network - singleNet_tool class 2. Independent single task phase recognition network - singleNet_phase class 3. end-to-end multi-task network without correlation loss - mtrcNet class 4. end-to-end multi-task network with correlation loss. - mtrcNet_cl class

Method	Phase	Tool
SingleNet_phase	72.97	-
SingleNet_tool	-	94.08
MTRCNet	77.20	94.42
MTRCNet_CL	75.54	94.33

When we compare the phase recognition task using singleNet architecture with our multi-task architecture-MTRCNet, we can see a drastic improvement in accuracy from 72.97 to 77.20. Thus, MRTCNet performed better in phase recognition. It shows that MTRCNet, even with the absence of correlation loss, can achieve consistent improvements. Similarly, for tool detection tasks, MTRCNets achieve better performance compared with the single-trained network. However, there is a very small improvement from 94.08 to 94.42.

More importantly, when we add correlation loss to enforce the prediction consistency, the results of our MTRCNet-CL for both phase and tool tasks are better than the single-task network but the performance decreases as compared to the multi-task network. As we can see for MTRCNet with Correlation loss has a phase recognition accuracy of 75.54, which is still better than singleNet architecture, which has a value of 72.97 but when we compare it with the MTRCNet algorithm, its accuracy is dropped by 2 per cent. In all the cases, the performance of tool detection isn't much different ranging from 94.08 for singleNet, 94.2 for MTRCNet and 94.33 for MultiTask Network with Correlation loss.

Figure 3, 4 and 5 show the confusion matrix for Phase recognition. To keep it neat, I kept the labels for phases from 0 to 6 in the confusion matrix. Following are the corresponding values for the respective phases. 0 = Preparation 1 = CalotTriangleDissection 2 = ClippingCutting 3 = GallbladderDissection 4 = GallbladderPackaging 5 = CleaningCoagulation 6 = GallbladderRetraction

6. DISCUSSION

The above results show that the multi-task network MTRCNet achieves improvement in both tasks. This demonstrates that multi-task learning is better for both phase recognition and tool detection. However, unlike our expectations, the introduction of correlation loss doesn't give much improvement as compared to a single-task network and performs worse than a multi-task network. Though the current literature mentions and has proved enough that a correlation loss between the two tasks improves performance, I believe my current results could be because of the small dataset that I opted to work upon.

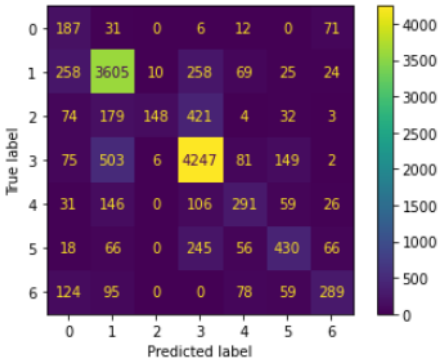

Figure 3. Confusion Matrix for single-task phase recognition

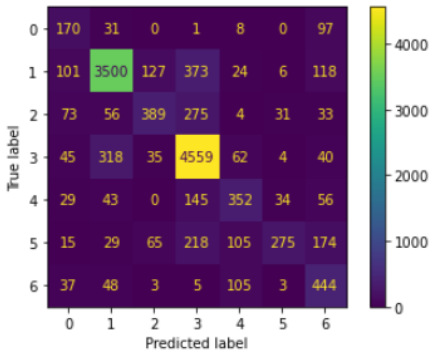

Figure 4. Confusion matrix for multi-task without correlation loss for phase recognition

The other reason could be the difference in hyperparameter values. I changed a few hyperparameters such as batch size. I reduced the batch size from 800 to 100, so I could train the model faster. I believe this may have affected the performance of the algorithms

Even in the confusion matrix, as we can see, the general trend is that phase 1 i.e. Calot Triangle Dissection and phase 3 i.e. gall bladder dissection are the easiest recognizable phases for all three algorithms.

Single-task network works slightly better than other algorithms for phase 1. On the other hand, the multi-task algorithm works better for phase 3. However, unfortunately, none of the algorithms shows good results for other phases in the confusion matrix.

The GitHub link for my project can be found here: https://github.com/Drishti2996/20353192_drishti_sharma.git

7. FUTURE WORK

Currently, I have used 10 videos as my dataset. However, I would like to use the larger dataset to see the difference between my current results and the new results. I believe by doing so, I will be able to get better results for correlation loss class too.

In future, I would like to compare the performance of the MTRCNet-CL with other similar state-of-the-art methods. these methods include "ToolNet" - which is a single-task network solely meant to perform tool detection; and "EndoNet" - which is a multi-layer network that leverages both tool and phase annotations.

For phase recognition, I will use (1) binary tool usage information generated from the manual annotations; (2) PhaseNet- which soleley uses the phase annotations and again (3) EndoNet.

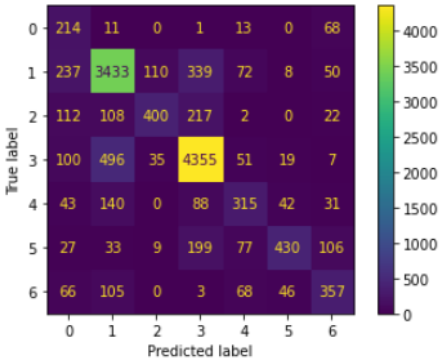

Figure 5. Confusion matrix for multi-task correlation loss for phase recognition

I would like to expand this project further. The biggest challenge I faced was of getting access to the good quality dataset. Therefore, I would like to work on Transfer learning for MultiTask Learning in surgical video analysis to see if we can train our model ona relatively easily available dataset and use them to classify our medical data.

REFERENCES

[1] DiPietro, R., Lea, C., Malpani, A., Ahmidi, N., Vedula, S. S., Lee, G. I., Lee, M. R., and Hager, G. D., "Recognizing surgical activities with recurrent neural networks," in [*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19*], 551–558, Springer (2016).

[2] Sahu, M., Mukhopadhyay, A., Szengel, A., and Zachow, S., "Addressing multi-label imbalance problem of surgical tool detection using cnn," *International journal of computer assisted radiology and surgery* **12**, 1013–1020 (2017).

[3] Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., and Heng, P.-A., "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Medical image analysis* **59**, 101572 (2020).

[4] Wang, S., Raju, A., and Huang, J., "Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos," in [*2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*], 620–623, IEEE (2017).

[5] Al Hajj, H., Lamard, M., Charriere, K., Cochener, B., and Quellec, G., "Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network," in [*2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*], 2002–2005, IEEE (2017).

[6] Wang, S., Xu, Z., Yan, C., and Huang, J., "Graph convolutional nets for tool presence detection in surgical videos," in [*Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*], 467–478, Springer (2019).

[7] 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 transactions on medical imaging* **36**(1), 86–97 (2016).