Surgical Endoscopy (2022) 36:679–688
https://doi.org/10.1007/s00464-021-08336-x

A contextual detector of surgical tools in laparoscopic videos using deep learning

Babak Namazi[1] · Ganesh Sankaranarayanan[2] · Venkat Devarajan[3]

Received: 23 August 2020 / Accepted: 13 January 2021 / Published online: 8 February 2021

Abstract

Background The complexity of laparoscopy requires special training and assessment. Analyzing the streaming videos during the surgery can potentially improve surgical education. The tedium and cost of such an analysis can be dramatically reduced using an automated tool detection system, among other things. We propose a new multilabel classifier, called LapTool-Net to detect the presence of surgical tools in each frame of a laparoscopic video.

Methods The novelty of LapTool-Net is the exploitation of the correlations among the usage of different tools and, the tools and tasks—i.e., the context of the tools' usage. Towards this goal, the pattern in the co-occurrence of the tools is utilized for designing a decision policy for the multilabel classifier based on a Recurrent Convolutional Neural Network (RCNN), which is trained in an end-to-end manner. In the post-processing step, the predictions are corrected by modeling the long-term tasks' order with an RNN.

Results LapTool-Net was trained using publicly available datasets of laparoscopic cholecystectomy, viz., M2CAI16 and Cholec80. For M2CAI16, our exact match accuracies (when all the tools in one frame are predicted correctly) in online and offline modes were 80.95% and 81.84% with per-class F1-score of 88.29% and 90.53%. For Cholec80, the accuracies were 85.77% and 91.92% with F1-scores if 93.10% and 96.11% for online and offline, respectively.

Conclusions The results show LapTool-Net outperformed state-of-the-art methods significantly, even while using fewer training samples and a shallower architecture. Our context-aware model does not require expert's domain-specific knowledge, and the simple architecture can potentially improve all existing methods.

Keywords Convolutional neural networks · Recurrent neural networks · Tool detection · Laparoscopic surgery · Label power-set

Numerous advantages of minimally invasive surgery such as shorter recovery time, less pain and blood loss, and better cosmetic results, make it the preferred choice over conventional open surgeries [1]. In laparoscopy, the surgical instruments are inserted through small incisions in the abdominal wall and the procedure is monitored using a laparoscope. The special way of manipulating the surgical instruments and the indirect observation of the surgical scene introduce more challenges in performing laparoscopic procedures [2]. The complexity of laparoscopy requires special training and assessment for the surgery residents to gain the required bimanual dexterity. Analyzing the streaming videos during the surgery and the recorded videos from previously accomplished procedures can potentially improve the outcomes. The tedium and cost of such an analysis can be dramatically reduced using an automated tool detection system, among other things and is, therefore, the focus of this paper.

Tracking surgical tools is essential in understanding the workflow of a procedure and in the assessment and rating of the videos. For example, it has been shown that experts have a better economy of motion compared to novice or less experienced surgeons [3, 4]. Also, by detecting the tools, we can check for wrong tool usage, monitor activation time of electro-surgical tools, and the use of proper technique (how

Presented as poster at SAGES 2017.

✉ Ganesh Sankaranarayanan
ganesh.sankaranarayanan@bswhealth.org

[1] Baylor Scott & White Research Institute, Dallas, TX, USA

[2] Department of Surgery, Baylor University Medical Center, 3500 Gaston Ave, Dallas, TX 75246, USA

[3] Electrical Engineering Department, University of Texas at Arlington, Arlington, TX, USA

a needle is positioned and moved with a needle driver during suturing), etc.

Manual annotation of long videos from surgeries is a time-consuming and expensive task. A vision-based algorithm for automated detection of the presence, location, or movement of surgical tools is indispensable in designing a fast and objective surgical evaluation system. A well-annotated database of surgical videos can also be used in information retrieval and is a reliable source for education and training of the future surgeons.

During surgery, monitoring the usage of surgical tools can provide real-time feedback to the surgeons and operating room staff. Furthermore, in computer-aided intervention, the surgical tools are controlled by a surgeon with the aid of a specially designed robot [5], which requires a real-time understanding of the current task. Therefore, detecting the presence, location, or pose of the surgical instruments is useful in robotic surgeries as well [6–8]. Finally, an automated tool usage detector can help to generate an operative summary.

To track surgical instruments, several approaches have been introduced, which use the collected signals during the procedure. For instance, in vision-based methods, the instruments can be localized using the videos captured during the operation. These methods are generally reliable and inexpensive. Traditional vision-based methods rely on extracted features such as shape, color, the histogram of oriented gradients, etc., along with a classification or regression method to estimate the presence, location, or pose of the instrument in the captured images or videos. However, these methods are dependent on pre-defined and painstakingly extracted hand-crafted features. Just logically defining and extracting such features alone is a major part of the detection process. Thus, these hand-crafted features and designs are not suitable for real-time applications.

Compared with the other surgical video tasks, detecting the presence and usage of surgical instruments in laparoscopic videos has certain challenges that need to be considered.

Firstly, since multiple instruments might be present at the same time, detecting the presence of these tools in a video frame is a multilabel (ML) classification problem. In general, ML classification is more challenging compared to the well-studied multiclass (MC) problem, where every instance is related to only one output. These challenges include but are not limited to using correlation and co-existence of different objects/concepts with each other and the background/context and the variations in the occurrence of different objects.

Second, as opposed to other surgical videos, such as cataract surgery, robot-assisted surgery, or videos from a simulation, where the camera is stationary or moving smoothly, in laparoscopic videos, the camera is constantly shaking. Due to the rapid movement and changes in the field of view of the camera, most of the images suffer from motion blur, and the objects can be seen in various sizes and locations. Also, the camera view might be blocked by the smoke caused by burning tissue during cutting or cauterizing to arrest bleeding. Therefore, using still images is not sufficient for detecting the instruments.

Third, surgical operations follow a specific order of tasks. Although the usage of the tools does not strictly adhere to that order, it is nevertheless highly correlated with the task being performed. The performance of the tool detection can be improved with the information about the task and the relative position of the frame with regard to the entire video.

At last, since the performance of a deep classifier in a supervised learning method is highly dependent on the size and the quality of the labeled dataset, collecting and annotating a large dataset is a crucial task.

Recent years have witnessed great advances in deep-learning techniques in various computer vision areas such as image classification, object detection, and segmentation etc., and in medical imaging [9]. Therefore, there is a trend towards using these methods in analyzing the videos taken from laparoscopic operations.

Endonet [10] was the first deep-learning model designed for detecting the presence of surgical instruments in laparoscopic videos, wherein Alexnet [11] was used as a Convolutional Neural network (CNN), for feature extraction and was trained for the simultaneous detection of surgical phases and instruments. Inspired by this work, other researchers used different CNN architectures [12, 13] to classify the frames based on the visual features. For example, in [14], three CNN architectures were used and [15] proposed an ensemble of two deep CNNs.

Sahu et al. [16] were the first to address the imbalance in the classes in a MultiLabel (ML) classification of video frames. They balanced the training set according to the combinations of the instruments. The data were re-sampled to have a uniform distribution in label-set space and, class re-weighting was used to balance the data. Despite the improvement gained by considering the co-occurrence in balancing the training set, the correlation of the tools' usage was not considered directly in the classifier and the decision was made solely based on the presence of single tools. Alshirbaji et al. [17] used class weights and re-sampling together to deal with the imbalance issue.

In order to consider the temporal features of the videos, Twinanda et al. employed a hidden Markov model (HMM) in [10] and Recurrent Neural Network (RNN) in [18]. Sahu et.al utilized a Gaussian distribution fitting method in [12] and a temporal smoothing method based on a moving average in [16] to improve the classification results, after the CNN was trained. Mishra et al. [19] were the first to apply a Long Short-Term Memory model (LSTM) [20], as an RNN

to a short sequence of frames, to simultaneously extract both spatial and temporal features for detecting the presence of the tools by end-to-end training.

A variety of different approaches were as following. Hu et al. [21] proposed an attention-guided method using two deep CNNs to extract local and global spatial features. In [22], a boosting mechanism was employed to combine different CNNs and RNNs. In [23], the tools were localized, after labeling the dataset with bounding boxes containing the surgical tools.

It should be noted that none of the previous methods takes advantage of any knowledge regarding the order of the tasks and, the correlations of the tools are not directly utilized in identifying different surgical instruments. In this paper, we propose a novel context-aware model called LapTool-Net to detect the presence of surgical instruments in laparoscopic videos. The uniqueness of our approach is based on the following three original ideas:

- A novel ML classifier is proposed as a part of LapTool-Net, to take advantage of the co-occurrence of different tools in each frame—in other words, the context is taken into account in the detection process.
- The ML classifier and the decision model are trained in an end-to-end fashion.
- The model's prediction for each video is sent to another RNN to consider the order of the usage of different tools/tool combinations and long-term temporal dependencies; yet another consideration for the context.

The pre-print version of this paper with more results and detailed discussions can be found in [24]. The preliminary results were presented at the SAGES 2017 Annual Meeting.

Materials and methods

The overview of the proposed model is illustrated in Fig. 1. The goal is to design a classifier that maps the frames of surgical videos, to the tools in the observed scene. The overall system is described based on the dataset from M2CAI16[1] tool detection challenge, which is a subset of Cholec80 dataset [10]. We chose the smaller dataset to highlight the improvements caused by the main contributions of this paper. The dataset contains 15 videos from cholecystectomy procedure, which is the surgery for removing the gallbladder. All the videos are labeled with seven tools for every 25 frames. The tools are Bipolar, Clipper, Grasper, Hook, Irrigator, Scissors, and Specimen bags. There are ten videos for training and five videos for validation. The type and shape of all seven tools remain the same for the training and validation sets.

Since the publicly available Cholec80 dataset was used in this study to train and test our deep-learning model, an Institutional Review Board (IRB) approval is not required for this study.

Spatio-temporal features

To detect the presence of surgical instruments in laparoscopic videos, the visual features (intra-frame spatial and inter-frame temporal features) need to be extracted. We use CNN to extract spatial features. CNN is a type of artificial neural network that is capable of processing still images and has been successfully applied to many computer vision tasks that involve image classification or object recognition. As shown in Fig. 1, the input frame x_{ij} is sent through the trained CNN and the output of the last convolutional layer (after pooling) forms a fixed size spatial feature vector v_{ij}.

Since there is a high correlation among video frames, it can be exploited by an RNN to improve the performance of the tool detection algorithm. An RNN uses its internal memory (states) to process a sequence of inputs for time series and videos-processing tasks [25]. This helps the model to identify the tools even when they are occluded or not clear due to motion blur. For this purpose, short sequences of frames (say 5 frames) are selected. We called the model consisting of a CNN and an RNN, a Recurrent Convolutional Neural Network (RCNN).

For each frame x_{ij}, the sequence of the spatial features is the input for the RNN. The total length of the input is no longer than one second, which ensures that the tools remain visible during that time interval. We selected Gated Recurrent Unit (GRU) [26] as our RNN for its simplicity. The final hidden state h_{ij} is the output of the GRU and is the input to a fully connected neural network FC1.

Tool combination

In a laparoscopic cholecystectomy surgery, not all the 2^K combinations are possible as the total number of incisions are typically 3 or 4. Figure 2 shows the percentage of the most likely combinations in the M2CAI dataset. The first 15 classes out of a possible maximum of 128 span more than 99.5% of the frames in both the training and the validation sets, and the tools combinations have almost the same distribution in both cases. Extracting the pattern in the surgical tool combination can potentially improve the performance of an automated tool detection algorithm. Furthermore, modeling the tools' co-occurrence is beneficial for assessing the performance by monitoring the wrong combinations.

To consider the tool combinations, in the well-known Label Power-set (LP) method, multiple tools are combined

[1] http://camma.u-strasbg.fr/m2cai2016/index.php/program-challenge.

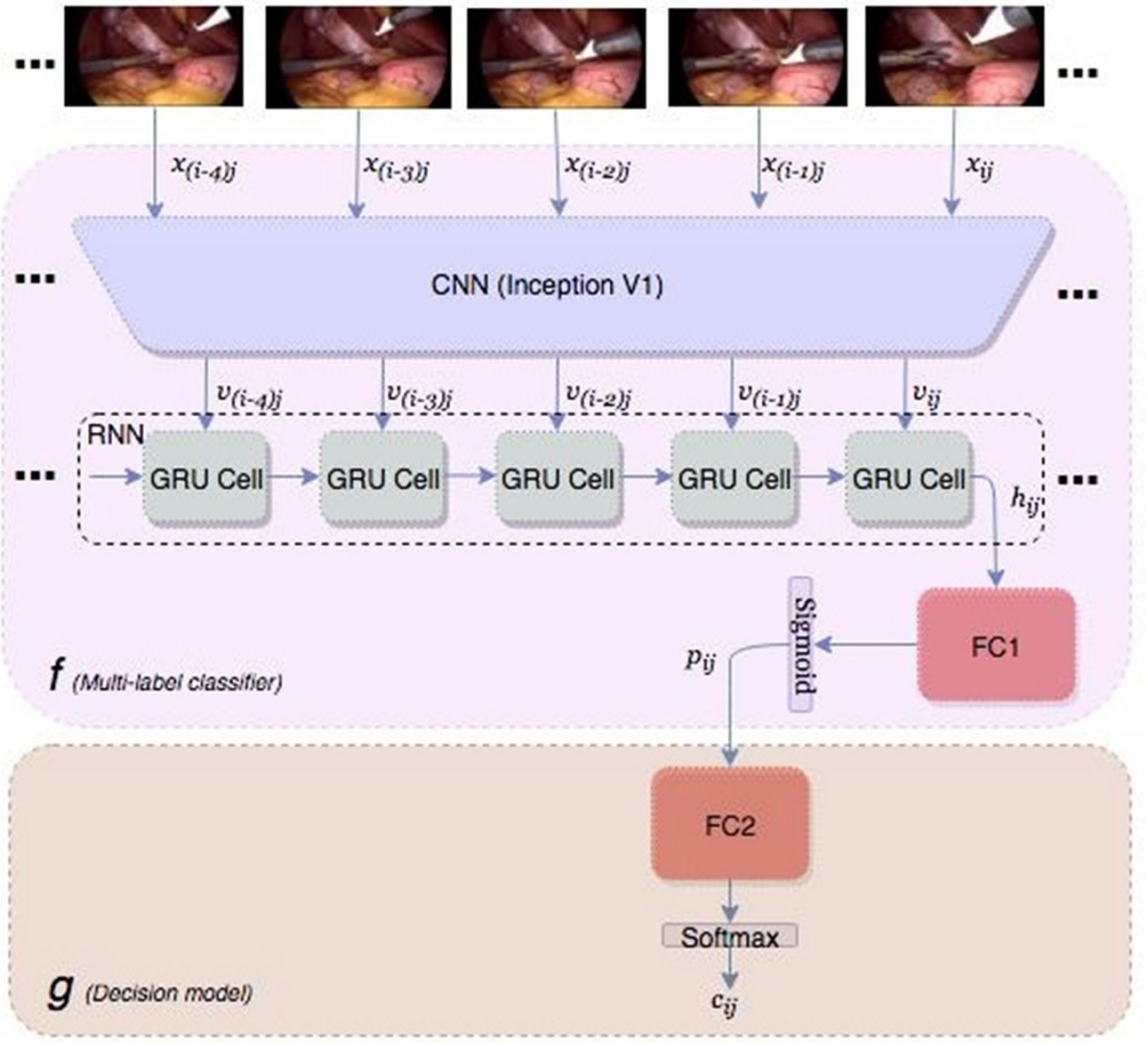

Fig. 1 Block diagram of the proposed classifier for detecting the presence of surgical tools in each frames of a laparoscopic video

into one superclass (combination) and the problem is transformed into a multiclass classification. The advantage of LP is that the class dependencies are automatically considered. Also, by eliminating uncommon combinations from the outputs, the classifier's attention is directed towards the more possible combinations.

Since an LP classifier is MC, training a deep-learning model with Softmax loss requires the classes to be mutually exclusive. In other words, each superclass is treated as a separate class, i.e., separate features activate a superclass. This causes performance degradation in the classifier and therefore, more data are required for training. We address this issue by a novel use of LP as the decision model g, which we apply to the ML classifier f. The decision model is a fully connected neural network (FC2), which takes the confidence scores of f and maps them to the corresponding superclass (Fig. 1). Our method helps the classifier to consider our superclasses as the combinations of classes rather than separate mutually exclusive classes.

Class imbalance

In a laparoscopic surgery, some tools are used more often than the others. For instance, in our dataset, Grasper is present in almost 80% of the procedure, whereas the Scissors are visible in less than five seconds in each video. It is

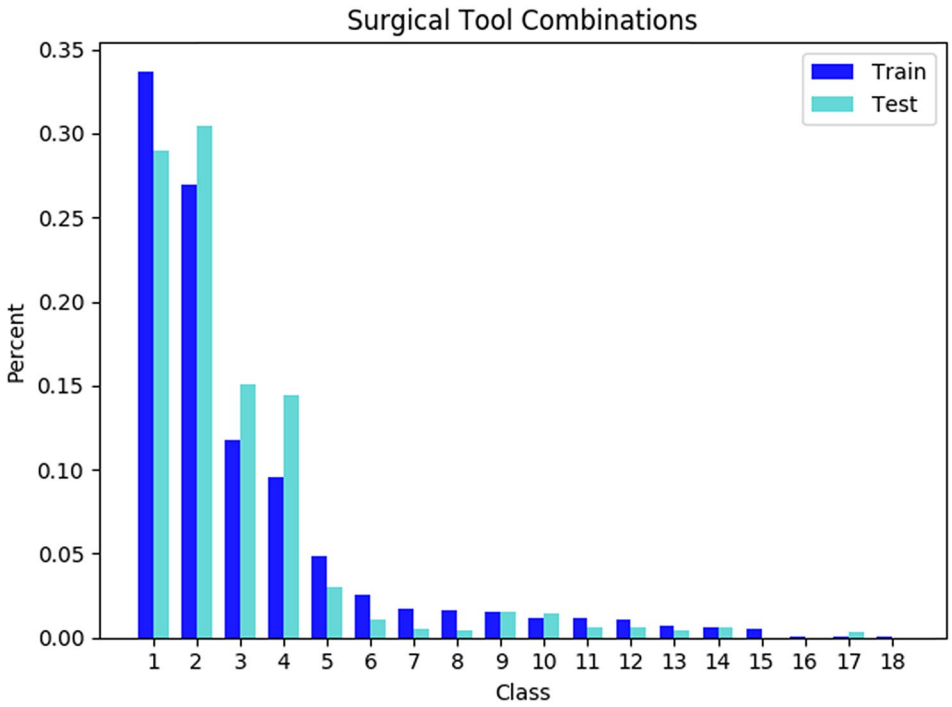

Fig. 2 The distribution for the combination of the tools in M2CAI dataset

known that in skewed datasets, the classifier's decision is inclined towards the majority classes. Therefore, it is always beneficial to have a uniform distribution for the classes during training. This can be accomplished using over-sampling for the minority classes and under-sampling for the majority classes. However, in ML classification, finding a balancing criterion for re-sampling is challenging [27].

To overcome imbalance, we perform under-sampling to have a uniform distribution of the combination of the classes. The main advantage of under-sampling over other re-sampling methods is that it can also be applied to avoid overfitting caused by the high correlation between the neighboring frames of a laparoscopic video. Therefore, we try different under-sampling rates to find the smallest training set without sacrificing the performance.

Figure 3 shows the relationship among the tools after re-sampling. It can be seen that the LP-based balancing method not only tends to a uniform distribution in the superclass space, it also improves the balance of the dataset in the single class space (with the exception of Grasper, which can be used with all the tools).

Training

We train the model to simultaneously identify the presence of each tool and the tools combinations. Having the vector of the confidence scores P, the ML loss L_f is the sigmoid cross-entropy (CE) and the Softmax CE loss function L_g is used for training the decision model. We use the joint training paradigm for optimizing the ML, and MC losses as a multitask-learning approach.

The trainable weights for the ML optimizer are all the weights in the CNN, the weights in the RNN, and FC1. On the other hand, for the MC optimizer, the CNN, RNN, and FC2 are trainable. Note that the shared weights between the two optimizers are the RCNN weights. By keeping the FC1 layer untouched by the MC optimizer, the spatio-temporal features are extracted by the RCNN, considering both the presence of each tool and the combination of them, and FC2 is solely trained as a decision model.

Post-processing

To smooth the RCNN prediction and consider the long-term ordering of the tools, we model the order in the usage of the tools with an RNN over all the frames of each video [28]. Due to memory constraints, the final predictions of the RCNN (j) are selected as the input for the post-processing RNN.

In the online mode, only the past frames are available for classifying the current frame. In the offline mode, future frames can also be used along with past frames to improve the classification results of the current frame. To accomplish this, a bi-directional RNN is employed. The post-processing RNN is a two-layer GRU with 128 and 32 units in each layer.

The post-processing method described in this section is similar to [22] in extracting the long-term temporal features using RNNs. However, in contrast to these researchers, we used the final predictions of the RCNN model instead of the

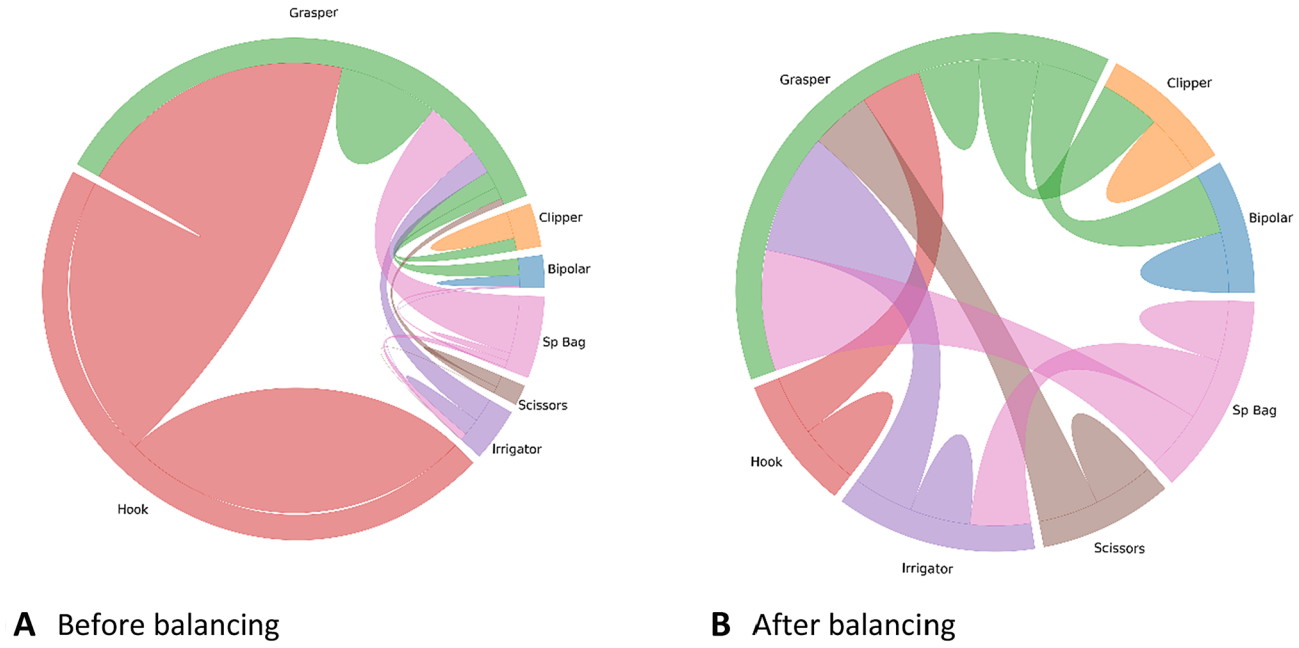

Fig. 3 The chord diagram for the relationship between the tools before and after balancing based on the tools' co-occurrences

vector of confidence scores of the tools. Besides containing the information about the co-occurrences, training RNNs can be accomplished easier with a single scalar versus the vector of the size of the total number of tools or the tools' combinations. With the aid of the shorter size input, we were able to train larger sequences, even after performing the temporal data augmentation (to be explained later).

Results

In this section, the performance of the different parts of the proposed tool detection model is validated through numerous experiments using the appropriate metrics. We selected Tensorflow [29] for all of the experiments. The CNN in all the experiments was Inception-V1 [30]. To have better generalization, extensive data augmentation, such as random cropping, horizontal and vertical flipping, rotation and a random change in brightness, contrast, saturation, and hue were performed during training. The initial learning rate was 0.001 with a decay rate of 0.7 after 5 epochs, and the results were taken after 100 epochs. The batch size was 32 for training the CNN models and 40 for the RNN-based models. All the experiments were conducted using an Nvidia TITAN XP GPU.

LapTool-Net results on M2CAI dataset

Since the dataset was labeled only for one frame per second (out of 25 frames/sec), there was a possibility of using the unlabeled frames for training, as long as the tools remain the same between two consecutive labeled frames. We used this unlabeled data to balance the training set, according to the LPs.

To balance the datasets, 15 superclasses were selected and the original frames were re-sampled to have a uniform distribution. The numbers of frames for each superclass were randomly selected to be 400, forming a training set of 6000 frames. In other words, under-sampling was performed based on the tool combinations.

We tested the model before and after adding the decision model. For training the RCNN model, we used 5 frames at a time (current frame and 4 previous frames) with an inter-frame interval of 5, which resulted in a total distance of 20 frames between the first and last frames. The RCNN model was trained with a Stochastic Gradient Descent (SGD) optimizer. The data augmentation for the post-processing model includes adding random noise to the input and randomly dropping frames to change the duration of the sequences; the final predictions of the RCNN model are saved every 20 frames, and the frames are dropped with the probability of 10–30%. Table 1 shows the results of the proposed RCNN and LapTool-Net.

In the table, CNN represents the model that uses only still images, CNN-LP is the results after considering the tool combinations in still images, RCNN considers spatio-temporal features from several successive frames, and LapTool-Net represents the performance of the mode after considering the long-term ordering of the tools usages.

Table 1 Final results for the proposed model on M2CAI dataset

	Acc (%)	F1-macro (%)	F1-micro (%)
CNN	74.36	74.43	87.70
CNN-LP	76.31	78.32	88.53
RCNN	77.51	81.95	89.54
RCNN-LP	78.58	84.89	89.79
Laptool-net(online)	80.95	88.29	91.24
Laptool-net(offline)	81.84	90.53	91.77

Table 2 The precision, recall, and F1-score of each tool for the ML classifier in RCNN-LP after removing the decision model

Tool	Precision (%)	Recall (%)	F1 (%)
Bipolar	77.62	83.57	80.49
Clipper	83.22	81.90	82.56
Grasper	69.99	90.28	78.85
Hook	95.33	93.43	94.37
Irrigator	77.27	83.60	80.31
Scissors	82.91	82.91	82.91
Specimen bag	76.96	94.91	85.00
Mean	80.55	87.22	83.50

It can be seen that by considering the temporal features through the RCNN model, the exact match accuracy and F1-macro were improved by 3.15% and 7.52%, respectively. Also, the F1-macro improves by 2.94% after adding the LP decision model.

The higher performance of the LapTool-Net, shown in Table 1, is due to consideration of the long-term order of the usage of the tools. In the offline mode, the utilization of the frames from both the past and the future of the current frame causes the improvements over the online model in accuracy and F1-scores.

To check the effectiveness of the multitask approach used for the end-to-end training of the RCNN-LP model, we took the output of the ML classifier, after removing the decision model from the trained RCNN-LP. In other words, we replaced the LP-based decision layer of the trained model with the threshold-based decision method. The results are shown in Table 2. It is worth mentioning that this results show that the RCNN model without the LP decision can be taken for making prediction for all the combinations including the rare combinations that were originally excluded during training.

In order to localize the predicted tools, the attention maps were visualized using grad-CAM method [31]. The results for some of the frames are shown in Fig. 4. In order to avoid confusion with frames that multiple tools, only the class activation map of a single tool is shown based on the prediction of the model. The results show that the visualization of the attention of the proposed model can also be used in reliably identifying the location of each tool without any additional annotations for the location and shape of the tools.

Comparison with current work

To validate the proposed model, we compared it with previously published research on the M2CAI dataset. The result is shown in Table 3. We show that our model outperformed previous methods by a significant margin even when choosing a relatively shallower model (Inception-V1) and while using less than 25% of the labeled images.

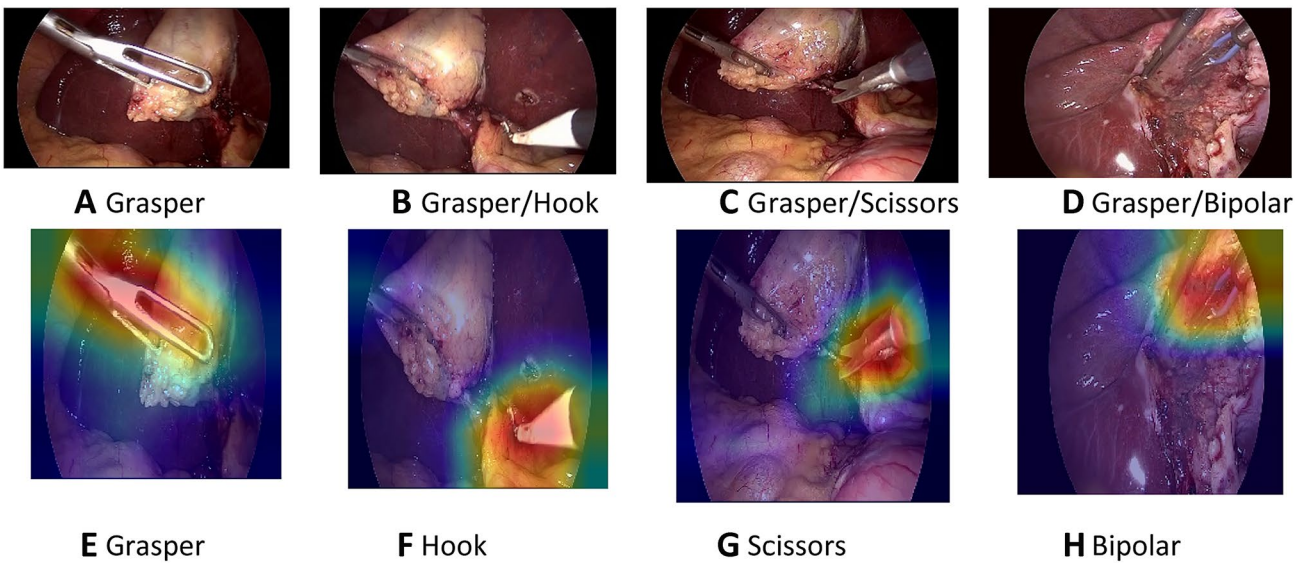

Fig. 4 The visualization of the class activation maps for some examples, based on the prediction of the model

Table 3 Comparison of tool presence detection methods on M2CAI

Method	CNN	Map (%)	F1-macro (%)
Laptool-net(offline)	Inception-V1	–	90.53
Laptool-net(online)	Inception-V1	–	88.29
RCNN(ours)	Inception-V1	89.88	81.95
[21]	Resnet-101 [34]	86.9	–
[23]	VGG	81.8	–
[16]	Alexnet [11]	65	–
[15]	Inception-v3 [35]	63.8	–
[12]	Alexnet	61.5	–
[36]	Alexnet	52.5	–

It is worth mentioning that a fair comparison with previous work on the same dataset is not feasible, since the evaluation metrics might not be the same. Nevertheless, we compared our ML classifier f, which is the RCNN model, along with the final models to show the superiority of our balancing and temporal consideration methods. Regardless of the choice of the CNN architecture, which is the most dominant component that can affect the results, the superiority of our model over the works in Table 3 is due to the end-to-end temporal consideration and the inclusion of the context such as the co-occurrence and tasks ordering, which are the main contributions of this paper.

LapTool-Net results on Cholec80 dataset

In this section, the performance of our model is evaluated on a larger dataset of laparoscopic cholecystectomy videos called Cholec80. We used the first 40 videos for training and the remaining 40 videos for testing our model.

The total number of tool combinations in Cholec80 dataset is 32, out of which 20 combinations are present in over 99.5% of the duration of videos. Compared with M2CAI dataset, the higher number of tool combinations is due to the more diversity in the larger dataset. Nonetheless, the extra five superclasses in Cholec80 dataset contain less than 0.4% of all frames. For each of the 20 tool combinations, 1500 samples were selected, forming a uniform class distribution on 30 K frames.

We used the same model as for M2CAI dataset for extracting the spatio-temporal features, the decision policy, and the post-processing step, as well as the training strategy. The results for the different parts of the model are shown in Table 4. Compared with the M2CAI results in Table 1, we can see significant improvement in accuracy and F1-scores. For example, the F1-macro of the CNN on the balanced Cholec80 is 9.19% higher than M2CAI dataset.

As was to be expected, the accuracy and F1-scores increase after adding the LP-based decision layer. However, the improvements are relatively smaller compared with the M2CAI results. For instance, the F1-macro of the RCNN-LP is less than one percent higher than RCNN. Similarly, the increase in the F1-macro for the CNN and RCNN is less compared with M2CAI dataset (less than 5% versus over 10% in M2CAI). The reason behind this observation is likely due to the fact that while the end-to-end training of the CNN, RNN, and LP layer results in the richer discriminating features, considering the co-occurrence and temporal coherence, the performance is dominated and bounded by the capacity of the CNN.

Table 4 Final results for the proposed model on Cholec80 dataset

	Acc (%)	F1-macro (%)	F1-micro (%)
CNN	75.41	83.62	89.05
CNN-LP	76.30	86.16	89.56
RCNN	77.77	88.39	90.41
RCNN-LP	79.95	89.17	91.21
Laptool-net(online)	85.77	93.10	93.71
Laptool-net(offline)	91.92	96.11	96.40

Discussion

In this paper, we proposed a novel system called LapTool-Net, for automatically detecting the presence of tools in every frame of a laparoscopic video. The main feature of the proposed RCNN model is the context awareness, i.e., the model learns the short-term and long-term patterns of the usage of the tools by utilizing the correlation between the usage of the tools with each other and, with the surgical steps. Our method outperformed all previously published results on M2CAI dataset, while using less than 1% of the total frames in the training set.

While our model is designed based on the previous knowledge of the cholecystectomy procedure, it does not require any domain-specific knowledge from experts and can be effectively applied to any video captured from laparoscopic or even other forms of surgeries. Also, the relatively small training set after under-sampling suggests that the labeling process can be accomplished faster by using fewer frames (e.g., one frame every 5 s). Moreover, the simple architecture of the proposed LP-based classifier makes it easy to use it with other proposed models such as [22] and [21], or with weakly supervised models [32, 33] to localize the tools in the frames. To accomplish that, the threshold mechanism of the ML classifier in all these papers can be simply replaced by our combination-aware decision model.

Acknowledgements The authors would like to thank NVIDIA Inc. for donating the TITAN XP GPU through the GPU grant program.

Funding This work was supported by Joseph Seeger Surgical Foundation award from the Baylor University Medical Center at Dallas.

Compliance with ethical standards

Disclosures Babak Namazi, Ganesh Sankaranarayanan, and Venkat Devarajan have no conflicts of interest or financial ties to disclose.

References

1. Velanovich V (2000) Laparoscopic vs open surgery. Surg Endosc 14(1):16–21
2. Ballantyne GH (2002) The pitfalls of laparoscopic surgery: challenges for robotics and telerobotic surgery. Surg Laparosc Endosc Percutan Tech 12(1):1–5
3. Sherman V, Feldman LS, Stanbridge D, Kazmi R, Fried GM (2005) Assessing the learning curve for the acquisition of laparoscopic skills on a virtual reality simulator. Surgical Endoscopy 19(5):678–682
4. Perrenot C, Perez M, Tran N, Jehl J-P, Felblinger J, Bresler L, Hubert J (2012) The virtual reality simulator dVTrainer® is a valid assessment tool for robotic surgical skills. Surgical Endoscopy 26(9):2587–2593
5. Antico M, Sasazawa F, Wu L, Jaiprakash A, Roberts J, Crawford R, Pandey AK, Fontanarosa D (2019) Ultrasound guidance in minimally invasive robotic procedures. Med Image Anal 54:149–167
6. Du X, Allan M, Dore A, Ourselin S, Hawkes D, Kelly JD, Stoyanov D (2016) Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery. Int J Comput Assist Radiol Surg 11(6):1109–1119
7. Allan M, Ourselin S, Thompson S, Hawkes DJ, Kelly J, Stoyanov D (2013) Toward detection and localization of instruments in minimally invasive surgery. IEEE Trans Biomed Eng 60(4):1050–1058
8. Allan M, Ourselin S, Hawkes DJ, Kelly JD, Stoyanov D (2018) 3-D pose estimation of articulated instruments in robotic minimally invasive surgery. IEEE Trans Med Imaging 37(5):1204–1213
9. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
10. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36(1):86–97
11. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on neural information processing systems 25:1097–1105
12. Sahu M, Mukhopadhyay A, Szengel A, Zachow S (2016) Tool and Phase recognition using contextual CNN features, arXiv :1610.08854
13. Prellberg J, Kramer O (2018) Multi-label classification of surgical tools with convolutional neural networks, 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1–8. https://doi.org/10.1109/IJCNN.2018.8489647
14. Zia A, Castro D, Essa I (2016) Fine-tuning Deep Architectures for Surgical Tool Detection. In: Workshop and challenges on modeling and monitoring of computer assisted interventions (M2CAI)
15. Wang S, Raju A, Huang J (2017) Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In: Proceedings—international symposium on biomedical imaging
16. Sahu M, Mukhopadhyay A, Szengel A, Zachow S (2017) Addressing multi-label imbalance problem of surgical tool detection using CNN. Int J Comput Assist Radiol Surg 12(6):1013–1020
17. Abdulbaki Alshirbaji T, Jalal NA, Möller K (2018) Surgical tool classification in laparoscopic videos using convolutional neural network. Curr Dir Biomed Eng 4(1):407–410
18. Twinanda AP, Padoy N, Troccaz MJ, Hager G (2017) Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos, Ph.D. thesis. https://tel.archives-ouvertes.fr/tel-01557522/document
19. Mishra K, Sathish R, Sheet D (2017) Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)
20. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
21. Hu X, Yu L, Chen H, Qin J, Heng P-A (2017) AGNet: attention-guided network for surgical tool presence detection. Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, pp 186–194
22. Al Hajj H, Lamard M, Conze P-H, Cochener B, Quellec G (2018) Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. Med Image Anal 47:203–218
23. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, Fei-Fei L (2018) 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)
24. Namazi B, Sankaranarayanan S, Devarajan V (2019) LapToolNet: a contextual detector of surgical tools in laparoscopic videos based on recurrent convolutional neural networks. Arxiv Preprints, https://arXiv.org/1905.08983
25. Jin Y, Dou Q, Chen H, Yu L, Qin J, Fu C-W, Heng P-A (2018) SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans Med Imaging 37(5):1114–1126
26. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)
27. Charte F, Rivera AJ, del Jesus MJ, Herrera F (2015) Addressing imbalance in multilabel classification: measures and random resampling algorithms. Neurocomputing 163:3–16
28. Namazi B, Sankaranarayanan G, Devarajan V (2018) Automatic detection of surgical phases in laparoscopic videos. In: Proceedings on the international conference in artificial intelligence (ICAI)
29. Abadi et al (2015) TensorFlow: large-scale machine learning on heterogeneous distributed systems. https://www.tensorflow.org/
30. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, 2015, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

31. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV)
32. Nwoye CI, Mutter D, Marescaux J, Padoy N (2019) Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. Int J Comput Assist Radiol Surg 14(6):1059–1067
33. Vardazaryan A, Mutter D, Marescaux J, Padoy N (2018) Weakly-supervised learning for tool localization in laparoscopic videos. Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis. Springer, Cham, pp 169–179
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR)
35. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR)
36. Twinanda AP, Mutter D, Marescaux J, de Mathelin M, Padoy N (2016) Single- and multi-task architectures for tool presence detection challenge at M2CAI 2016