Deep Neural Networks Predict Remaining Surgery Duration from Cholecystectomy Videos

Ivan Aksamentov[1](✉), Andru Putra Twinanda[1], Didier Mutter[2], Jacques Marescaux[2], and Nicolas Padoy[1]

[1] CNRS, IHU Strasbourg, ICube, University of Strasbourg, Strasbourg, France
ivan.aksamentov@etu.unistra.fr
[2] IRCAD, IHU Strasbourg, University Hospital of Strasbourg, Strasbourg, France

Abstract. For every hospital, it is desirable to fully utilize its operating room (OR) capacity. Inaccurate planning of OR occupancy impacts patient comfort, safety and financial turnover of the hospital. A source of suboptimal scheduling often lies in the incorrect estimation of the surgery duration, which may vary significantly due to the diversity of patient conditions, surgeon skills and intraoperative situations. We propose automatic methods to estimate the remaining surgery duration in real-time by using only the image feed from the endoscopic camera and no other sensor. These approaches are based on neural networks designed to learn the workflow of an endoscopic procedure. We train and evaluate our models on a large dataset of 120 endoscopic cholecystectomies. Results show the strong benefits of these approaches when surgeries last longer than usual and promise practical improvements in OR management.

Keywords: Remaining duration prediction · Surgical workflow analysis · Operating room management · Deep learning · Recurrent neural networks

1 Introduction

The surgery department is one of the busiest units in a hospital. This creates the need for an accurate surgery duration prediction. It plays an important role in optimizing the resources of the surgical facility, especially since high expenditure mainly comes from: (1) duration overestimation leading to underutilization of resources and (2) duration underestimation causing high patient waiting time [7]. Additionally, a more deterministic time table improves patient safety by reducing the duration of anesthesia, ventilation and time spent in the intensive care. However, it is difficult to optimally allocate the OR resources due to the uncertainty of interventional duration, caused by the diversity of patient conditions, surgeon skills and the variety of intraoperative situations. For instance, [12] reports that general surgeons underestimated the time required for the procedure by 31 min

I. Aksamentov and A.P. Twinanda—Equal contribution.

M. Descoteaux et al. (Eds.): MICCAI 2017, Part II, LNCS 10434, pp. 586–593, 2017.
DOI: 10.1007/978-3-319-66185-8_66

in average, while anaesthesiologists underestimated it by 35 min. A recent study over 157 cholecystectomy procedures [5] has also shown that there is a large variation of patient preparation and waiting time (47 ± 17 min), while it typically requires 25 min to prepare the patients.

Several approaches have been proposed to address the OR scheduling problem, aiming to preoperatively predict the surgery duration. For instance, the "Last 5 Case" estimate, proposed in [9], predicts the surgery duration based on the procedure-surgeon historical data. Other data, such as patient's age [2], operational (e.g., OR assignment and assigned surgical team) and temporal factors (e.g., the weekday, month, and year) [8] have also been investigated to predict the surgery duration. However, such preoperative approaches still face challenges due to the uniqueness and unpredictability of each surgical procedure.

One possible solution to these challenges is to dynamically adapt the schedule as the day progresses. Typically, verbal communication with the surgical staff can be used to obtain an estimate of the remaining surgery duration (RSD). However, this disrupts the smoothness of surgical workflow and may compromise the safety in the OR [14]. Thus, semi-automatic methods, such as the one presented in [3] which requires the input of anaesthesiologist during surgery, are not desirable to predict the RSD. Other signals, such as surgical tool usage [10,11], surgeon's right hand [10], and low-level task representations (i.e., tool, organ, and action) [4] have also been used to perform the RSD prediction.

However, in these studies, the signals are obtained through manual annotation, which currently renders the methods impractical for intraoperative applications. In [5], the activation of the electrosurgical device was utilized to answer the question: "should the next patient be called now?" The pipeline is constrained to start the detection after the procedure has progressed for 15 min and assumes that the next patient should be prepared 25 min before the surgery ends. These constraints render the method less broadly applicable than general RSD estimation.

Recently, it has been shown that visual information contains more discriminative features than tool binary signals to perform surgical phase recognition [13]. In this paper, we argue that the visual information also contains discriminative characteristics for RSD prediction. To the best of our knowledge, this is the first work to address such a problem by relying solely on visual information from videos. Here, we propose and evaluate two approaches to perform RSD prediction during cholecystectomy procedures. In the first approach, we use the prediction of the current surgical phase and phase statistics to estimate the RSD. In the second approach, we directly perform RSD prediction via regression from the video information available up to the current time. For both approaches, we propose a pipeline consisting of a convolutional neural network (CNN) and a long short-term memory (LSTM) network. These approaches are compared to two baselines: the first baseline estimates the RSD by relying on simple statistics of surgery durations, while the second relies on an expert observer to manually indicate surgical phase transitions.

The evaluation is performed on a large cholecystectomy video dataset, containing 120 cholecystectomy videos.

In summary, the contributions of this paper are threefold: (1) we propose a deep learning approach, solely relying on the visual information, to predict the remaining surgery duration; (2) we perform a wide range of comparisons on RSD prediction on a large cholecystectomy video dataset; and (3) we show that the proposed approaches yield promising practical results, especially when the surgeries are shorter or longer than usual.

2 Methodology

2.1 Methods for RSD Prediction

Naïve Approach. The most straight-forward approach to perform RSD prediction is to use the historical data of the surgeries: at time t during a surgery, the RSD t_{rsd} is obtained by computing $max(0, t_{ref} - t)$, where t_{ref} is a referential duration derived from the dataset (e.g., mean or median). The $max(\cdot, \cdot)$ operator is used to ensure that t_{rsd} is always positive. However, this method does not take into account any intraoperative information and only relies on the statistics of the historical data. One way to incorporate intraoperative information into the model is by using key time points related to the progression of the surgery.

Phase-Inferred Estimation. The execution of a surgery is guided by a surgical workflow representing the sequence of the tasks to be performed during the procedure. Here, we argue that these tasks could be used as intraoperative information to estimate the RSD since some tasks are performed uniquely at certain times of the procedure. For example, gallbladder packaging during cholecystectomy procedure indicates that the surgery is ending soon. Specifically, we use surgical phases as key information [13]. We perform RSD prediction by computing $max(0, t^p_{ref} - t^p) + \sum_{m=p+1}^{N} t^m_{ref}$ where t^m_{ref} is the referential duration of phase m, t^p is the elapsed time in current phase p, and $N = 7$ is the number of defined phases. This approach requires the phase information p at each time step. The best possible phase information could be obtained from an expert observer, i.e. a clinician who informs the system about phase transitions during a surgery. To remove the requirement for human intervention, which is expensive and may introduce disruptions in the workflow, we automatically obtain the phase information by using a deep learning pipeline, consisting of a CNN and an LSTM network. For CNN, we adopt the residual network architecture (ResNet) [6]. This network is chosen since it is the state-of-the-art network in the computer vision community and has outperformed other networks, such as AlexNet for surgical phase recognition in our early experiments (around 10% difference in accuracy). We connect the visual features coming from the dense layer of ResNet to the LSTM network. The output of the LSTM network is then passed to a dense layer, consisting of $N = 7$ nodes, the value of each node represents the confidence being in a certain phase.

Note that this approach is similar to the idea proposed in [11], where the current phase is inferred by a linear HMM and the expected RSD is then computed from the linear HMM model. The main differences are that we use a stronger temporal model (LSTM) and rely only on visual input.

Time Regression for RSD. The aforementioned phase-inferred approaches are strongly driven by the phase information, which only models the surgery progress in a coarse manner. Therefore, we propose to address the RSD prediction using direct estimation via regression. The RSD regression is carried out using a deep pipeline similar to the one used for phase recognition. The difference is that here the LSTM network is trained to perform regression. The output of LSTM is connected to a dense layer of one node containing the predicted RSD, which is ultimately smoothed using a 15-second window.

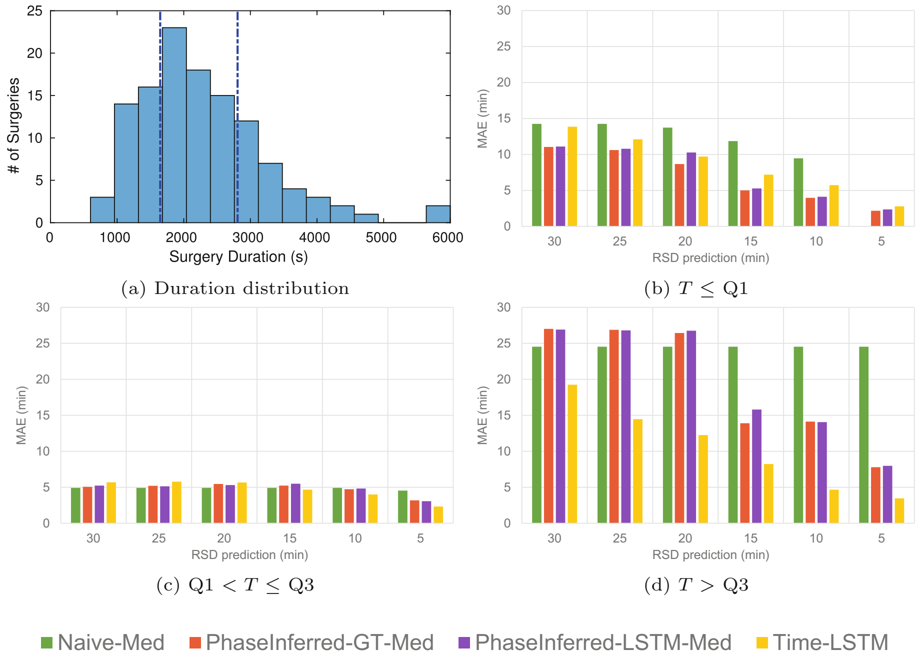
Note that RSD regression is a harder problem than phase classification: frames with the same RSD label but from videos of different length may belong to different surgical phases and thus differ significantly in visual appearance.

2.2 Training Strategy

Two-Step Optimization. Since ResNet is a large network and the cholecystectomy videos are of long durations, it is difficult to train the complete pipeline in an end-to-end manner due to memory constraints. To alleviate this problem, we use a two-step optimization process. First, we train the CNN by finetuning a pre-trained ResNet model to perform surgical phase recognition so that the network extracts discriminative and semantically meaningful visual features from the videos [13]. It is beneficial to train the CNN on surgical phase recognition because, thanks to the surgical workflow, the visual features extracted by the CNN will then contain information about the progression of surgical procedure, which is correlated to the RSD. Once finetuned, the CNN is then used to extract the visual features, which will later be utilized to train the LSTM networks (for either phase recognition or RSD prediction) on complete sequences.

RSD Normalization for Regression. The regression function has a wide range of possible values. Since the sigmoid function $\sigma(x)$, which is used in the sigmoid cross-entropy loss function, plateaus after $x = 6$, we normalize the RSD value by dividing it with the highest duration. At test time, the estimated RSD is denormalized to obtain the final RSD.

Dataset Balancing. To perform the evaluation, the dataset has to be divided into subsets. If done improperly, this could lead to an imbalance in the resulting subsets, e.g., all long videos grouped into the same subset, which is undesirable. Here, we balance dataset subsets using a genetic algorithm to minimize the following functions: (1) difference between minimal and maximal mean durations (to encourage similar mean durations among subsets), and (2) negative sum of standard deviations of durations in subsets (to encourage diversity of durations within subsets). The objective functions are chosen to be competing with each other, so that both inter-subset similarity and intra-subset diversity are developed.



(a) Duration distribution

(b) $T \leq$ Q1

(c) Q1 $< T \leq$ Q3

(d) $T >$ Q3

Fig. 1. (a) Distribution of the surgery duration T in the dataset with dashed blue lines indicating the boundaries of Q1 and Q3 (first and third quartiles) of the dataset. (b, c, d) MAE against RSD prediction for lower, middle, and upper duration ranges.

3 Experimental Setup

Dataset. The dataset contains 120 videos, which is generated by combining the Cholec80 dataset [13] with additional 40 cholecystectomy videos. All videos are annotated with the phases defined in [13]. They are recorded at 25 fps and accumulate over 75 h of recordings. Since the distribution of the surgery durations is asymmetric (as shown in Fig. 1-a), we compare the mean and median of durations for the referential durations t_{ref} and t^{p}_{ref}.

Dataset Split. To train and test the approach, the dataset is split into 4 parts: T1 (40 videos), T2 (40 videos), V (10 videos), and E (30 videos). Subset T1 is used to train the CNN, while the combination of T1 and T2 is used to train the LSTM. The CNN is only trained on T1 to avoid overfitting of the LSTM.

Subset V is used as validation during both CNN and LSTM training. Ultimately, subset E is used to evaluate the trained CNN-LSTM pipeline. We perform the evaluation on the dataset using a four-fold cross validation. We obtain the folds by employing the aforementioned dataset balancing method.

Training Setup and Parameters. The pipeline is trained and tested at 1 fps. A 152-layer ResNet model, pretrained on the ImageNet dataset, is finetuned with batch size 48 on our dataset; while the LSTM is trained on complete sequences (the longest is 5987 s). In order to mitigate the exploding gradient problem, we employ gradient clipping. To obtain the best models, we perform an extensive hyperparameter search, including the LSTM hidden size and dropout rate, using the training and validation subsets. The training process is considered finished when there is no improvement observed on the validation subset for 20 epochs. Models are trained using TensorFlow [1] and NVIDIA Titan X GPUs. At test time, each model runs at 1 fps on a conventional laptop's CPU.

Evaluation Metrics. We use mean absolute error (MAE) as evaluation metric, which is obtained by averaging the absolute difference of the ground truth and the estimated RSD in second. This is the natural metric for the task, as it is easily interpretable by clinicians, showing the under- and overestimation of RSD.

4 Experimental Results

In Table 1, we show the RSD prediction results. It can be seen that the Naïve approach yields the highest MAE. This is expected since this model does not consider any intraoperative information. When we incorporate the surgical phase information, significant improvements can be observed. When we compare our proposed automatic method (PhaseInferred-LSTM) to the semi-automatic method (PhaseInferred-GT) which requires an expert observer to provide extra information during the procedure, there is no significant difference observed in the results. In other words, we could remove the expert observer in the RSD prediction process without sacrificing the performance of the system. This is thanks to the high performance for online phase recognition given by PhaseInferred-LSTM, i.e., 89% accuracy on this dataset.

It can also be seen that the Time-LSTM approach outperforms other methods, yielding an MAE of 460 s, despite the challenges of predicting RSD via regression (e.g., high variation on visual appearance for frames with same RSD labels).

The results in Table 1 also show that the proposed approaches do not significantly improve RSD prediction on videos from the middle range ($\mathrm{Q1} < T \leq \mathrm{Q3}$). This is however expected since the surgery durations in this range are close to the median of the duration distribution. However, the proposed approaches significantly outperform the Naïve approach on surgeries which deviate from the "average" surgery, i.e., surgeries in lower and upper ranges ($T \leq \mathrm{Q1}$ and $T > \mathrm{Q3}$, respectively).

Table 1. RSD prediction results. The MAEs are shown for the complete dataset and the lower, middle, and upper ranges. Q1 and Q3 are shown in Fig. 1-a.

Method		Mean absolute error (MAE in second)			
		Complete	$T \leq$ Q1	Q1 $< T \leq$ Q3	$T >$ Q3
Naive	Mean	668 ± 481	1036 ± 235	300 ± 177	1035 ± 523
	Median	640 ± 478	855 ± 229	281 ± 152	1146 ± 507
PhaseInferred-GT	Mean	487 ± 345	668 ± 231	252 ± 117	775 ± 411
	Median	479 ± 388	426 ± 195	256 ± 153	978 ± 409
PhaseInferred-LSTM	Mean	498 ± 350	611 ± 299	354 ± 266	642 ± 422
	Median	487 ± 390	454 ± 282	354 ± 301	749 ± 483
Time-LSTM		460 ± 310	591 ± 234	288 ± 130	672 ± 422

To better understand how accurate the RSD predictions are for practical applications, we investigate the reliability of the predictions by computing the MAEs with respect to several RSD predictions (from 5 to 30 min). In other words, this evaluation indicates how big the error is when the method predicts that the surgery will end in, for instance, 25 min. The MAEs are computed by using a two-minute window on the RSD predictions. We perform this experiment on all three ranges. As depicted in Table 1, all methods perform similarly on the middle range (Fig. 1-c). This is however not the case on lower and upper ranges (Fig. 1-b and d, respectively), where the proposed approaches PhaseInferred-LSTM-Median and Time-LSTM significantly outperform the Naïve approach. On the lower range, PhaseInferred-LSTM-Median performs better than the Time-LSTM approaches, however the difference of performance is not significant (i.e., 1.4 min). Note that the Naïve approach is never able to predict an RSD of 5 min because the surgeries in this range are much shorter than the median duration. On the other hand, the Time-LSTM approach significantly outperforms PhaseInferred-LSTM-Median on the upper range, yielding more than 9 min improvements in average. Compared to the Naïve approach, the Time-LSTM approach yields significantly better results, i.e., improvements by 14 min in average. This shows that our proposed approaches are more robust to the variation in surgery duration.

5 Conclusions

In this paper, we have presented two real-time approaches which only rely on the visual information coming from the videos to predict the remaining surgery duration (RSD) on cholecystectomy procedures. We have shown that the deep learning pipeline, performing RSD regression, outperformed both Naïve and semi-automatic methods, which solely rely on statistics and/or manually provided phase labels. The proposed automated RSD prediction approaches are particularly beneficial when surgery durations deviate from the average. Over a large number of surgeries, these methods have the potential to improve patient safety as well as to significantly reduce the clinical operative costs.

Acknowledgment. This work was supported by French state funds managed by the ANR within the Investissements d'Avenir program under references ANR-11-LABX-0004 (Labex CAMI) and ANR-10-IAHU-02 (IHU Strasbourg). The authors would also like to acknowledge the support of NVIDIA with the donation of the GPU used in this research.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
2. Ammori, B.J., Larvin, M., McMahon, M.J.: Elective laparoscopic cholecystectomy. Surg. Endosc. **15**(3), 297–300 (2001)
3. Dexter, F., Epstein, R.H., Lee, J.D., Ledolter, J.: Automatic updating of times remaining in surgical cases using bayesian analysis of historical case duration data and instant messaging updates from anesthesia providers. Anesth. Analg. **108**(3), 929–940 (2009)
4. Franke, S., Meixensberger, J., Neumuth, T.: Intervention time prediction from surgical low-level tasks. J. Biomed. Inform. **46**(1), 152–159 (2013)
5. Guédon, A.C.P., Paalvast, M., Meeuwsen, F.C., Tax, D.M.J., van Dijke, A.P., Wauben, L., van der Elst, M., Dankelman, J., van den Dobbelsteen, J.: Real-time estimation of surgical procedure duration. In: International Conference on E-health Networking, Application & Services, pp. 6–10 (2015)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
7. Kayış, E., Khaniyev, T.T., Suermondt, J., Sylvester, K.: A robust estimation model for surgery durations with temporal, operational, and surgery team effects. Health Care Manag. Sci. **18**(3), 222–233 (2015)
8. Kayis, E., Wang, H., Patel, M., Gonzalez, T., Jain, S., Ramamurthi, R.J., Santos, C.A., Singhal, S., Suermondt, J., Sylvester, K.: Improving prediction of surgery duration using operational and temporal factors. In: AMIA (2012)
9. Macario, A., Dexter, F.: Estimating the duration of a case when the surgeon has not recently scheduled the procedure at the surgical suite. Anesth. Analg. **89**, 1241–1245 (1999)
10. Maktabi, M., Neumuth, T.: Online time and resource management based on surgical workflow time series analysis. IJCARS **12**(2), 325–338 (2017)
11. Padoy, N., Blum, T., Feussner, H., Berger, M.O., Navab, N.: On-line recognition of surgical activity for monitoring in the operating room. In: IAAI, pp. 1718–1724 (2008)
12. Travis, E., Woodhouse, S., Tan, R., Patel, S., Donovan, J., Brogan, K.: Operating theatre time, where does it all go? A prospective observational study. BMJ **349**, g7182 (2014)
13. 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 (2017)
14. Wiegmann, D.A., ElBardissi, A.W., Dearani, J.A., Daly III, R.C., Sundt, T.M.: Disruptions in surgical flow and their relationship to surgical errors: an exploratory investigation. Surgery **142**(5), 658–665 (2007)