

Hook

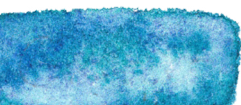
The applicability of deep learning to detect the progress of laparoscopic surgery using video recordings

Master Thesis
S.E.P. Meij



Grasper

Final check



Clipping & cutting



Master Thesis

The applicability of deep learning to detect the progress of laparoscopic surgery using video recordings

by

Senna Eleonora Petrus Meij
4213564

Department of Biomedical Engineering
Faculty of Mechanical, Maritime and Material Engineering
Delft University of Technology

12 December 2019

Thesis Committee:

Dr. J.J. van den Dobbelsteen

Delft University of Technology (supervisor)

Dr. F.M. Vos

Delft University of Technology

T.S. Vijfvinkel, MD

Delft University of Technology

Dr. ir. A.C.P. Guédon

Spaarne Gasthuis (supervisor)

Spaarne  Gasthuis

 **TU**Delft

 **COSMONIO**

Abstract

Hook

Grasper

Final check

Clipping & cutting

The operating room is one of the most complex and expensive environments in the hospital. Research has been focusing on improving the efficient use of the OR time, for instance by using intraoperative data to update the planning of the OR during the day. This thesis used a deep learning network to automatically recognize surgical tools and pre-defined surgical phases present in the recordings, to ultimately track the progress of the procedure. The aim of this thesis is to assess the performance and applicability of this deep learning method for the use of image recognition in a medical environment. To ultimately predict the remaining surgery duration and improve the efficiency of the OR planning.

Two datasets of laparoscopic recordings were used, one containing laparoscopic cholecystectomies and one containing total laparoscopic hysterectomies. The surgical tools and the pre-defined phases of the procedure were annotated in every recording, after which the deep learning network was trained with this data. The performance of the network was tested in multiple experiments.

The results showed that the performance of the deep learning network was promising and in line with published literature, but that the results varied between recordings. An experiment using three different sized datasets showed that a larger dataset corresponded with the best results and results that varied the least between recordings. Testing the generalizability of the network showed that a network trained on one type of surgery can also be used to recognize similar tools in a different type of surgery. Important is that the tools have the same design. It was found that the most important resources for a project like this are a dedicated hardware with image recognition software and time.

This thesis showed the applicability of a deep learning network to automatically recognize the progress of a surgery and provided insight into the steps that need to be taken to use it on a larger scale.

Preface

Hook

Grasper

Final check

Clipping & cutting

This master thesis is part of an ongoing research of dr. ir. Annetje Guédon at the Spaarne Gasthuis in Hoofddorp. The goal of this research is to predict the remaining surgery duration in real time in the OR using a deep learning method, to be able to adjust the OR-planning of the day to the unexpected events in the OR. I had the honour to work alongside Annetje and be part of the start of this research project. Being part of the development of this project was a great learning experience, but is not necessarily covered in the main part of this thesis. Therefore, I wrote an additional appendix about this part of the project (appendix A)

For this research we started a collaboration with COSMONiO. This is a company specialized in image recognition using machine learning located in Groningen. They provided the hardware and programming of the software for this research. Together with COSMONiO we developed a method to use image recognition on video recordings.

This thesis describes what I have done during my graduation project. The materials and methods are described first, then the results chapter is divided into six subchapters. Because of the many different experiments each of them contains a small recap of the method, the results and a short discussion for the specific experiment. At the end of this thesis, there is the overall discussion containing a more in-depth analysis about the complete thesis.

Table of contents

Abstract	4
Preface	6
Abbreviations	10
1. Introduction	12
2. Method	16
2.1. Datasets	17
2.2. Deep learning method	20
2.3. Annotating	21
2.4. Training the ANN	21
2.5. Evaluation of the performance	22
2.5.1. Recall, precision and accuracy	22
2.5.2. Visualisation phase results	25
2.5.3. Tool - phase relation	25
2.5.4. Different sizes of training subsets	26
2.5.5. Cross validation	26
2.5.6. Resources	26
3. Results	28
3.1. Recall, precision and accuracy	29
3.2. Visualisation phase results	31
3.3. Tool - phase relation	35
3.4. Different sizes of training subsets	36
3.5. Cross validation	39
3.6. Resources	40
4. Discussion	42
4.1. Results	43
4.2. Deep learning in healthcare	44
4.3. Future research	45
References	46
Appendix A	50
Appendix B	52
References appendix	60

Abbreviations

#	Number
AI	Artificial Intelligence
ANN	Artificial Neural Network
CPU	Central Processing Unit
csv	comma separated values
exp	experiment
FN	False Negatives
FP	False Positives
fps	frame per second
GPU	Graphics Processing Unit
Lap Chol	Laparoscopic Cholecystectomy
NaN	Not a number
OR	Operating Room
RAM	Random-Access Memory
ResNet50	Residual Network 50
RFID	Radio-Frequency Identification
RSD	Remaining Surgery Duration
TB	Terabyte
TLH	Total Laparoscopic Hysterectomy
TN	True Negatives
TP	True Positives
UI	User Interface

1. Introduction

One of the most complex and expensive environments inside the hospital is the operating room (OR) [1]–[3]. With increasing healthcare costs hospitals are looking for ways to contain their expenses [4]. A lot of research has therefore focussed on ways to use the OR more efficiently.

Veen-Berkx et al. show there is room for improvement in the planning of OR procedures to reduce the time the OR is unused [5]. Part of the solution they suggested is to better determine the procedure duration. Travis et al. conducted a study on how to determine the procedure time more accurately, finding that surgeons underestimate the procedure time by 28.7% [6]. Different approaches to estimate the procedure time using preoperative data are discussed in literature. Historical surgery data about the surgeon and the procedure, or patient specific characteristics can be used for more accurate estimations [7], [8]. Additionally, research has been done into predicting the remaining surgery duration (RSD) intraoperatively, using information about the devices and tools usage during procedures in the OR [9]. Intraoperative information about the progress of a specific surgery can be used to update the planning of the OR during the day, to minimize the amount of underused OR time and decrease patient waiting time. This thesis is part of ongoing research which focuses on the use of intraoperative information, and predicting the RSD to improve the efficiency of the OR.

The prediction of the RSD is determined by information acquired during procedures in the OR. Video recordings from these procedures can provide this information. Specifically, laparoscopic recordings, made during minimally invasive surgeries, can give insight in the progress of the surgical process. Currently, these recordings are primarily used to guide the surgical team, however, the information they contain can also be used on a wider range of applications. For instance, with the use of data analysis, specific information can be extracted to predict the RSD.

Previous research into the progress of the surgical process shows that the use of the tools can provide this information, as some tools are related to specific steps of the surgery process. By knowing the surgical process, the use of these tools can be used to predict the RSD: Guédon et al. monitored the use of the electrosurgical device to predict the RSD [10]. Another way to predict the RSD is to divide the surgical process into several phases. Each phase contains a predefined step of the procedure. The RSD can be predicted by knowing in which phase the surgery is and knowing the sequence of the phases. Aksamentov et al. used a phase-inferred approach to predict the RSD with the use of laparoscopic recordings [9]. This phase-inferred

approach is often used together with information about the tool usage. Meeuwsen et al. used RFID tags to track tools during the procedure as an input to recognize the phases [11]. Twindanda et al. used the visual information of laparoscopic recordings to extract information about the tools and phases to predict the RSD [12].

This thesis focuses on the use of laparoscopic recordings to detect the progress of the surgical process, using tool and phase information. The combination of tools and phases provides overall more information to eventually detect the progress of a surgical procedure. This information needs to be extracted from the recordings to be able to use it, this will be done using a deep learning method. Deep learning is a specific machine learning method where an artificial intelligence (AI) network uses input provided by the user to train itself to perform future tasks. Machine learning methods are often used to automatically extract information from images [9]. By providing the network with the information about the tool usage and the passing phases it can train itself and can learn on its own. Afterwards, it can give an output in terms of tool and phase recognition on new recordings. In this thesis, such a method is used to automatically recognize tools and phases in laparoscopic recordings. The aim of this research is to assess the performance and applicability of this method for the use of image recognition in a medical environment.

2. Method

The setup of this research consisted of five aspects, which are presented in this section. This chapter provides information about the composition of the datasets, the use of the deep learning method, the annotating process, how the network was trained and lastly how the performance was evaluated. Furthermore will this chapter explain which recordings that were used, the AI method that was applied and how the recordings were annotated, which means how the recordings were labelled with the tools that were used and surgical phases of each procedure. These annotations were used to train the AI and evaluate the performance.

2.1. Datasets

Two datasets with laparoscopic recordings were obtained for this research. The first dataset contained 36 recordings of a laparoscopic cholecystectomy (Lap Chol), the second dataset contained two recordings of a total laparoscopic hysterectomy (TLH). These two surgery types were used because of their high variety in length, their level of standardization and because they are frequently performed [13], [14]. The surgeries were provided by the Spaarne Hospital and acquired from their medical video archive, DAX. They were recorded between 2016 and 2019, the Lap Chol procedures were performed by five different surgical teams and the TLH were performed by two different surgical teams. To prepare the datasets the recordings were trimmed to ensure that they did not contain footage of outside the body for longer than five minutes at the beginning or end of the recording.

Each frame of all the recordings, of both datasets, were labelled with the tools that were used and the surgical phases of the procedures. These labels were defined in collaboration with corresponding surgeons and are shown in table 1-3. The phases were defined in the chronological order they almost always occur in during surgery. With the exception of the bleeding phase, none of the phases could occur simultaneously. The bleeding phase could occur at any given moment during each phase and was therefore categorized as an additional phase. It was important to annotate the bleeding phase, as it could be the cause of prolonging the surgery, which could provide valuable information about the length of the surgery.

Table 1. Tools used during a Lap Chol procedure and a TLH procedure.

Laparoscopic cholecystectomy	Total laparoscopic hysterectomy
Grasper	Grasper
Scissors	Scissors
Monopolar hook (hook)	Monopolar hook (hook)
Irrigator & suction device (irrigator)	Irrigator & suction device (irrigator)
Clipper	Ligasure
Bag	Uterus mobiliser
Drain	Morcellator
	Needle feeder
	Needle & thread

Table 2. Surgical phases of the Lap Chol.

Phase	Start cue	End cue
1. Trocar & tools insertion	First frame with a view of the inside of the body	First frame with a tool in view
2. Preparation & dissection	Frame after first tool in view	Frame before the clipper is in view
3. Clipping & cutting	First frame with the clipper in view	Last frame with the scissors in view
4. Gallbladder dissection	Frame after the last scissors in view	Frame before the bag is in view
5. Gallbladder packaging & retrieval	First frame with the bag in view	Last frame with the bag in view
6. Liver bed coagulation	First frame where the gasper is used to coagulate	Last frame the grasper is used to coagulate
7. Final check & irrigation	Frame after the last frame of coagulating or the last frame with the bag in view	Last frame before removing the trocars or leaving the body
8. Closing & desufflation	First frame of removing the trocars or leaving the body	First frame outside the body or the end of video
Additional phase	Start cue	End cue
Bleeding	First frame with blood and the irrigator or a gauze	Last frame with the irrigator or a gauze

Table 3. Surgical phases of the TLH.

Phase	Start cue	End cue
1. Trocar & tools insertion	First frame with a view of the inside of the body	First frame with a tool in view
2. Uterus dissection	Frame after first tool in view	Frame before the hook is used on the vaginal cuff
3. Uterus separation from the vagina	First frame the hook is used on the vaginal cuff	First frame the uterus is fully separated from the vagina
4. Uterus retrieval: transvaginal	Frame after the first frame the uterus is fully separated from the vagina	Last frame the uterus is pushed through the vaginal canal
5. Uterus retrieval: morcellation	Frame after First frame the uterus is fully separated from the vagina	Last frame with the bag in view
6. Vaginal cuff closure	First frame the uterus and/or bag is not in view	Last frame the needle feeder and/or needle and thread are in view
7. Final check & irrigation	First frame after the needle feeder and/or needle and thread are in view	Last frame before removing the trocars or leaving the body
8. Closing & desufflation	First frame of removing the trocars or leaving the body	First frame outside the body or the end of video
Additional phase	Start cue	End cue
Bleeding	First frame with blood and the irrigator or a gauze	Last frame with the irrigator or a gauze

The Lap Chol dataset was divided into three subsets. Firstly, a training subset containing 26 recordings, secondly a testing subset containing one recording, and thirdly a validation subset containing nine recordings. The distribution of the recordings to the different subsets was done randomly, while preserving that every different tool and phase appeared in each subset. The TLH dataset was not divided into subsets, this dataset only functioned as a validating set, as the ANN did not train on the TLH dataset.

2.2. Deep learning method

Deep learning is a method of machine learning which contains artificial neural networks (ANN) and is often used for visual recognition tasks [12]. An ANN is a system that can learn by giving it examples. Such a deep learning network consists of several progressive layers; each layer recognizes more specific features of the recognition task. The differences between deep learning and other machine learning techniques is that a deep learning network can train itself and learn how to improve certain layers to provide a better outcome. Because the deep learning network can train itself it is not known how the network improves these layers and trains itself, therefore it is often called a black box. For this research, the presence of a specific tool and which phase was taking place needed to be extracted from each frame of the recordings. This is a so-called classification task; this is different from a detection task where the specific location of the specific tool is also recognized.

The specific primary network that was used in this research was an InceptionV3. This network has an inception architecture as proposed by Szegedy et al [15], which is the state-of-the-art architecture for a classification and detection network [16]. It was designed to improve the performance in terms of speed and accuracy compared to previous architectures [17]. This network uses 42 layers and Szegedy et al. show the layout of the network [15]. Beside this primary network a residual network was used, called the Residual Network 50 (ResNet 50). This network helps to build deeper layers in the network and to propagate information through the layers. These ANNs were implemented by COSMONiO in the programming interface Keras and modified to fit the needs of the application. It ran on a dedicated server, the NOUS learner, which was developed by COSMONiO and was located at the Spaarne Hospital. Table 4 shows the specifications of the server.

Table 4. Specification of the NOUS Learner.

GPU	NVIDIA Titan RTX - 24GB memory
CPU	Intel Core i7-9700K 8 Cores Processor Base Frequency: 3.60 GHz Max Turbo Frequency: 4.90 GHz Cache: 12 MB SmartCache
RAM	32GB DDR4-3200
SSD	1TB
HDD	6TB
Power supply	850W

2.3. Annotating

In order to train the ANN, the recordings of each dataset needed to be annotated. Every frame needed to be labelled with the correct tools and the phase present in that frame for the ANN to be able to train itself. This annotating process was done using the user interface (UI) NOUS, also developed by COSMONiO. This software ran on a Microsoft Surface Pro 6, which was connected to the NOUS Learner via a WIFI connection. First all the recordings were uploaded and stored onto the NOUS Learner. During this process all the frames were extracted and the amount of frames was reduced from 25 fps to 1 fps. The annotating process was done in two steps, first the tools were annotated and afterwards the phases. The entire datasets were fully annotated in NOUS.

The tool classification was a binary problem, which meant that every tool was either labelled as present or not present in the frame. This was done by chronologically propagating through all the frames of each recording and labelling them with the correct tools. Four students were hired to annotate the datasets under the supervision of the primary researcher. They received a face to face introduction on the first day and a document with some reminders and examples of all the tools they had to annotate, which is shown in Appendix B.

The phase classification was a multiclass problem, which meant that each frame was labelled with one of the eight phases. Except the additional bleeding phase, this was a binary problem just like the tools. The phases were annotated by selecting a range of multiple frames at once and labelling them with the specific phase. This was done by the primary researcher, because this step required a better understanding of the surgeries.

2.4. Training the ANN

After the annotating process the ANN was trained on the training subset of the Lap Chol dataset, which contained 26 recordings. The settings for this training can be found in Table 5.

Table 5. Training settings of the ANN.

Image size	299 x 299 pixels
Epochs	50 epochs
Batch size	32 frames
Learning rate	0.0001 with Adam optimizer
Data augmentation	Scaling and rotation
Power supply	850W

One Epoch is when all the data, meaning every frame, has gone through the ANN once. For instance, if the training of the ANN ran for 50 epochs the ANN had seen every frame of every recording 50 times. To pass all the data through the ANN it was divided into batches. The batch size is the amount of frames in each batch. The learning rate indicated how quickly the ANN can adapt to the classification problem. The lower the learning rate the more epochs are needed. With a low learning rate, the ANN only makes a small adjustment after each epoch. Therefore, a lot of epochs are needed to optimize the ANN [18]. However, if the learning rate is very high, the ANN makes larger adjustments which could cause the ANN to make less optimal changes and not reach the optimal results. An Adam optimizer was used to optimize the learning rate for the different features of the ANN to get the optimal results. The data augmentation was used to make the ANN more generalizable; it ensures that the ANN also recognizes the visuals in each image when they have a different size or when they are rotated [19].

Initially the training ran for 50 epochs, however there were two rules which could stop the training prematurely.

- When the ANN reached a 100% accuracy
- When there was no improvement of accuracy for 20 epochs

The ANN used the testing subset, which contained one recording, to check its performance during the training and to eventually stop prematurely if needed.

2.5. Evaluation of the performance

2.5.1. Recall, precision and accuracy

The trained ANN was evaluated on the validation subset of the Lap Chol, which contained nine recordings. The system provided two csv files for every recording, one contained the annotated labels and one contained the recognized labels for each frame. From this a confusion matrix for each tool and for all the phases was produced. The confusion matrix consists of the amount of true positives (TP), false positives (FP), the true negatives (TN) and the false negatives (FN) (table 6 and 7).

Table 6. Example confusion matrix for a binary problem.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Table 7. Example confusion matrix for a multiclass problem.

	Predicted phase 1	Predicted phase 2	Predicted phase 3	Predicted phase 4
Actual phase 1	TN	FP	TN	TN
Actual phase 2	FN	TP	FN	FN
Actual phase 3	TN	FP	TN	TN
Actual phase 4	TN	FP	TN	TN

The performance of the ANN will be presented in the form of recall, precision, and accuracy. The data that was used is very imbalanced, some tools appear more often than other tools and some phases are much longer than other phases. Because of this the accuracy results can give a distorted image of the performance, however recall and precision can be used to effectively evaluate the performance of imbalanced data [20]. The recall, precision and accuracy will be calculated using the following equations:

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \quad (2.1)$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \quad (2.2)$$

$$\text{Accuracy tools} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \quad (2.3)$$

$$\text{Accuracy phases} = \text{TP}/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \quad (2.4)$$

The accuracy is the percentage of frames which the system recognized correctly. The recall is the percentage of all the frames on which a phase or tool is present (TP+FN) where the system actually recognized a tool or phase correctly (TP). The precision is the percentage of all the frames on which the system recognized a phase or tool (TP+FP) where the system recognized a tool or phase correctly (TP). Figure 1 provides a visual representation of the recall and precision. All the black dots represent frames containing a specific tool or phase. All the white dots represent frames not containing this specific tool or phase. The dots within the circle represent the frames on which the system recognized this tool or phase and the dots outside the circle represent the frames in which the system did not recognize this. The recall is the black dots inside the circle divided by all the black dots. The precision is shown by the black dots inside the circle divided by all the dots inside the circle.

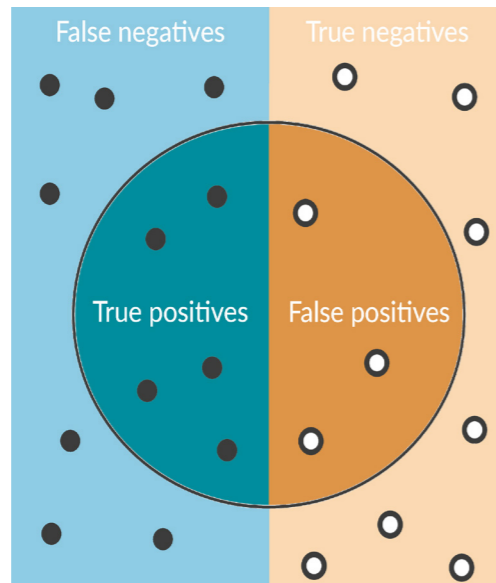


Figure 1. Visual representation of the recall and precision, adapted from [21].

The performance was evaluated per tool and phase individually for each recording. In addition, also the weighted accuracy, recall and precision were calculated for each tool and the weighted recall and precision for each phase. This was done by multiplying the accuracy, recall and precision by the number of frames the tool or phase was actually present and dividing it by the total number of frames which contained a tool or a phase. This can be seen in the following equations:

$$\text{Weighted recall} = \frac{(\text{Recall}) (\# \text{ of frames the tool or phase is present})}{\text{Total \# of frames in which tools or phases are present in}} \quad (2.5)$$

$$\text{Weighted precision} = \frac{(\text{Precision}) (\# \text{ of frames the tool or phase is present})}{\text{Total \# of frames in which tools or phases are present in}} \quad (2.6)$$

$$\text{Weighted accuracy} = \frac{(\text{Accuracy}) (\# \text{ of frames the tool or phase is present})}{\text{Total \# of frames in which tools or phases are present in}} \quad (2.7)$$

Afterwards the not-weighted average and weighted average per tool and phase were calculated as well as the overall average over all nine recordings of the validating subset. When the not-weighted average was used, the tool or phase present in only a few frames had the same contribution to the average as a tool or phase which was present in a large amount of frames. With the weighted average the amount of frames a tool or phase was present in was taken into account, which provides a more accurate representation of the performances for this type of imbalanced data. In some cases a specific tool or phase was not present in the recording. When calculating the averages for this tool or phase, the calculation could give an error, because it was divided by zero. In those cases the errors were not taken into account for both averages. Lastly, the standard deviation was calculated to show the variation of the results.

2.5.2. Visualisation phase results

The recall and precision provide information about how often the ANN makes a wrong recognition. In this thesis, it is interesting to provide information about where in the surgery process these wrong recognitions occur. For example, wrong recognitions could occur during a certain surgical phase, or aggregated around the phase transitions. The recognitions were made visual in a line graph for each recording of the validating subset.

2.5.3. Tool - phase relation

The surgical phases were defined in relation to the tools that were used during the surgery. These relations between the tools and the phases were not explicitly provided to the ANN. However, they were implicitly provided by annotating the phases according to the tool usage. Table 8 shows the relations between the specific tools and phases of the Lap Chol.

Table 8. Tools and their corresponding phase of Lap Chol.

Tools	Phases
Grasper	2. Preparation & dissection
Clipper	3. Clipping & cutting
Monopolar hook	4. Gallbladder dissection
Bag	5. Gallbladder packaging & retrieval
Irrigator & suction device	Bleeding

The strength of the relation between the tool and phase was calculated in order to provide an indication of which information was used by the ANN to recognize phases. The tool-phase relation provided the percentage of frames where both the tool and the corresponding phase were recognized, divided by the total amount of frames the tool was recognized. The higher the tool-phase relation the stronger the connection between the tool and the phase in the results of the ANN. This relation was calculated on the recognitions of the ANN with the following equation:

$$\text{Tool - phase relation (\%)} = \frac{\# \text{ of frames containing the tool and corresponding phase}}{\text{Total \# of frames containing the tool}} \quad (2.8)$$

The higher the reference rate the stronger the connection the ANN had made between the tool and the phase.

2.5.4. Different sizes of training subsets

For an ANN network more training data often results in a better performance. However, there is a point that the increase in performance start to decrement. An indication of the increase in performance can be provided by training multiple ANNs with training subsets of different sizes while testing them on the same validation subset. For this purpose, two additional training subsets of the Lap Chol dataset were made, beside the primary training subset of 26 recordings (experiment 1). A training subset containing 18 recordings (experiment 2), and a training subset containing nine recordings (experiment 3). These two additional training subsets contained a selection of the recordings of the subset of exp 1. Two new ANNs were trained on these two additional subsets, the performance of these were evaluated by calculating the recall, precision and accuracy as discussed in the section 2.5.1.

2.5.5. Cross validation

The TLH dataset was used for a cross validation test in order to test the generalizability of the ANN. A generalizable ANN is preferred because it can be expanded onto different surgery types, as the training of the ANN and the preparation of the training data is very time consuming. When the ANN could use its training of one surgery type on another surgery type this would decrease the amount of training data and time needed to train for the other surgery type.

The TLH dataset contained two recordings and was submitted to the ANN that was trained on exp 1 of the Lap Chol dataset. The performance of the ANN was only evaluated on the recognition of the tools that are used in both types of surgeries. This showed if the ANN could recognize information on which it was trained under different circumstances and could be expanded to different surgery types. The remaining tools and phases were not evaluated as the ANN did not contain the labels of these tools and phases of the TLH.

2.5.6. Resources

The resources that were used during this research have been tracked, to give insight in what was needed to use this system on a larger scale in medical practice. One of these resources was time needed to annotate all the data. Comparing the time it took to execute the annotations with the length of the recordings provided insight in the speed of the annotating process. This gave an estimation of the time needed when more annotations are needed in the future. Another resource was equipment, in terms of hardware and software but also data and software to prepare the data.

3. Results

This chapter shows the results of the six separate parts of the performance evaluation as discussed in subchapter 2.5, with the results of each test presented in an individual subchapter. For the sake of clarity, each subchapter starts with a short summary of the method and ends with a short concluding part including a short discussion about the results of the specific test. The overall discussion of all the results are covered in chapter four.

3.1. Recall, precision and accuracy

The accuracy of the tool recognition, and the recall and precision of the tool and phase recognition were calculated with the results of experiment 1. These three metrics were defined as follows:

- Accuracy is the percentage of frames in which the tool is correctly recognized.
- Recall is the percentage of frames the ANN misses a tool or phase that is present in the frame.
- Precision is the percentage of frames where the recognition that a tool or phase is present in the frame is correct.

Table 9 shows the results of the tool recognition: the accuracy, recall and precision (using equation 2.1-2.3) per tool over all recordings of the validating subset. It also shows the average and weighted average over all the tools (using equation 2.5-2.7). The results for the (weighted) average for recall and precision were lower than for accuracy, and the standard deviation was larger, meaning there was a large variation.

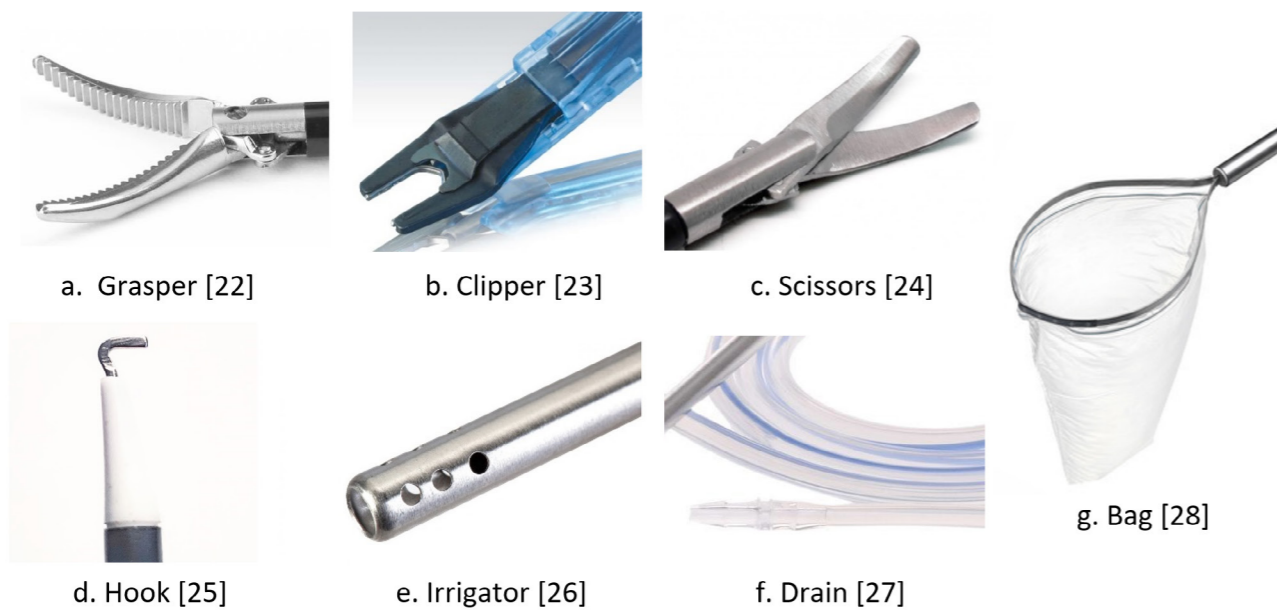
Looking specifically at this variation in the recall results, higher values were reached for the tools that appeared in many frames. The grasper appeared the most often in frames and showed the highest recall. The drain appeared the least and showed the lowest recall.

Looking specifically at the precision results, a connection could be seen with the design of the tools. Figure 2 shows pictures of the different used tools [22]–[28]. The tools with a more characteristic design, such as the clipper and the bag, gave a relative higher precision despite the lower amount of frames. A higher precision was reached for the clipper, which had a more characteristic shape, than for the scissors, even though the scissors and clipper were used about as often.

Table 9 The results of the tool detection of validation subset 1.

	Accuracy (%)	Recall (%)	Precision (%)	# of frames
Grasper	89.2	83.0	84.5	14019
Clipper	99.0	71.9	83.2	651
Monopolar hook	95.7	79.7	66.3	4492
Scissors	98.2	51.5	52.3	600
Irrigator & suction device	97.5	65.3	91.2	1780
Bag	98.7	66.2	100.0	741
Drain	99.7	1.6	84.5	182
Overall average	96.9	69.1	77.8	
Standard deviation	3.6	27.7	16.2	
Weighted overall average	92.0	78.2	88.0	

Figure 2 Tools that are used during Lap Chol surgeries [22]-[28].



The results of the phase recognition are shown in table 10. It can be seen that there was a large difference between the average and the weighted average. The phase with the most frames (phase 2) had the highest recall and precision, the three phases with the lowest amount of frames (phase 6, 7 and 8) had the lowest recall and precision. This had a large impact on the difference between the average and the weighted average. Phase 6 was only performed

in two recordings causing a recall of 0.0% and not having a value for precision, because the phase was not recognized by the ANN. Bleeding was mentioned separately because it was an additional phase. The recall and precision were lower for bleeding than the average results of the phases, but also lower than the phases with a similar amount of frames.

Table 10. The results of the phase detection of validation subset 1. Abbreviations: NaN = Not a number.

	Recall (%)	Precision (%)	# of frames
1. Trocar & tools insertion	66.0	71.2	2235
2. Preparation & dissection	82.9	79.8	15694
3. Clipping & cutting	49.5	67.2	1931
4. Gallbladder dissection	70.5	60.7	6133
5. Gallbladder packaging & retrieval	67.4	69.0	1181
6. Liver bed coagulation	0.0	NaN	16
7. Final check & irrigation	30.3	19.3	811
8. Closing & desufflation	35.8	30.3	230
Overall average	57.4	56.8	
Standard deviation	27.1	22.8	
Weighted overall average	71.1	76.2	
Bleeding	35.2	46.5	2231

The results presented in table 8 and 9 are in line with what can be found in literature. Comparing the results of Twinanda et al. [12] the average precision of their tool recognition was slightly higher, 81.0%. However, the precision of individual tools was similar, only the precision of the hook was much higher, 95.6%. The average recall of the phases was 66.0% ($\pm 12.0\%$) and the average precision was 70.0% ($\pm 8.4\%$). These average results of Twindanda et al. were lower than the results shown in table 11. However, the variation of the phase results of Twinanda et al. was smaller. This could be explained by the different sized training subset; with a larger training subset the ANN has more frames to train on and could perform more consistently.

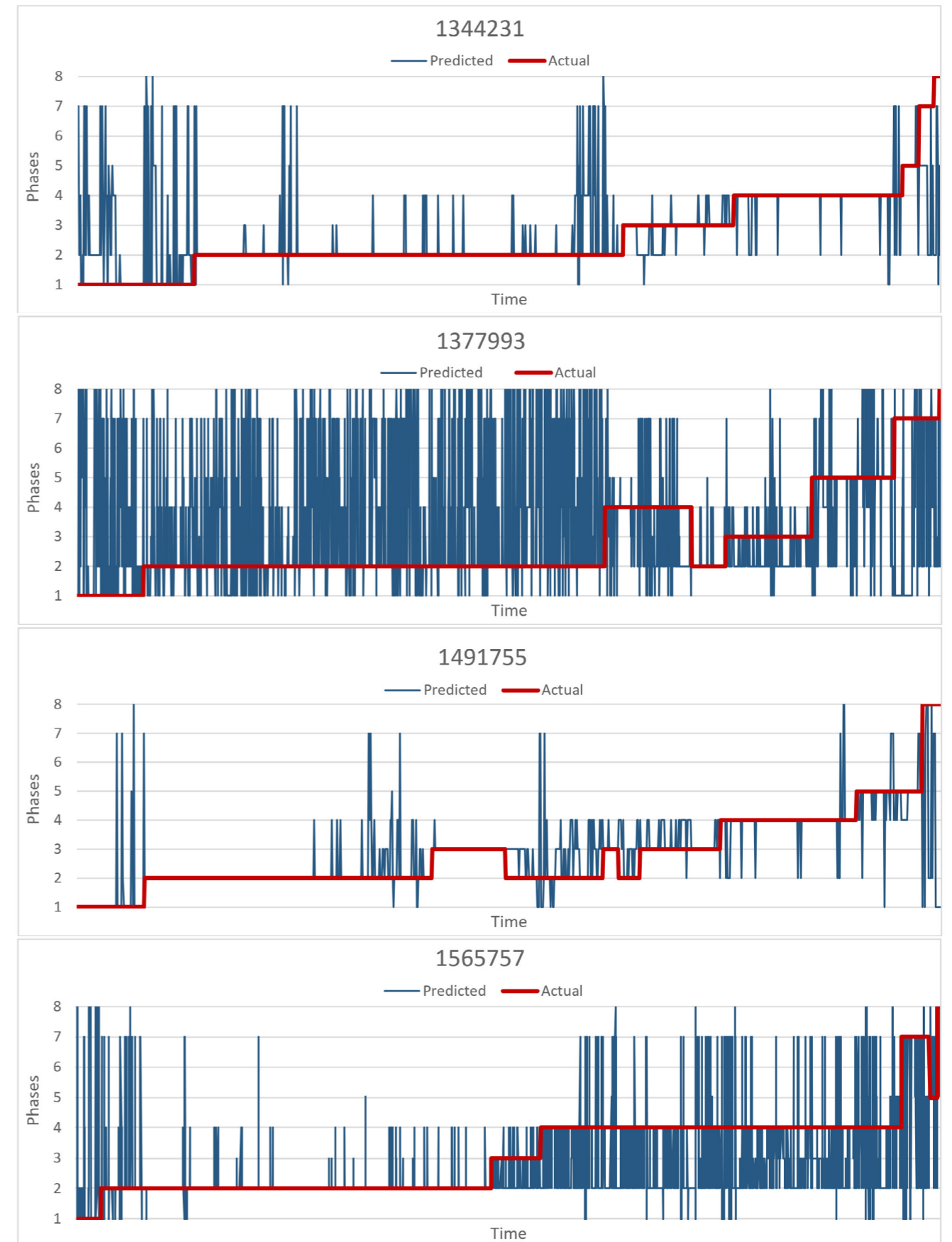
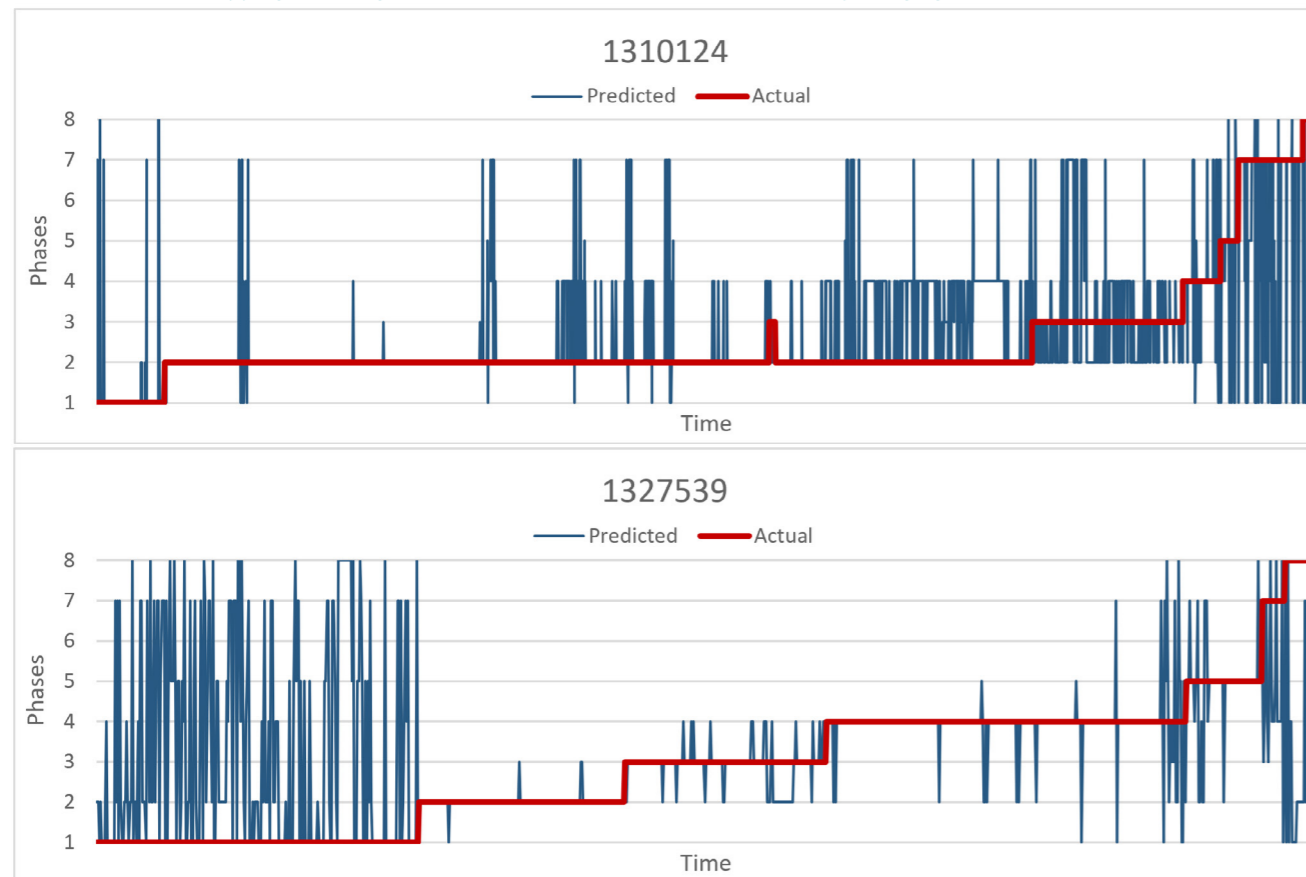
3.2. Visualisation phase results

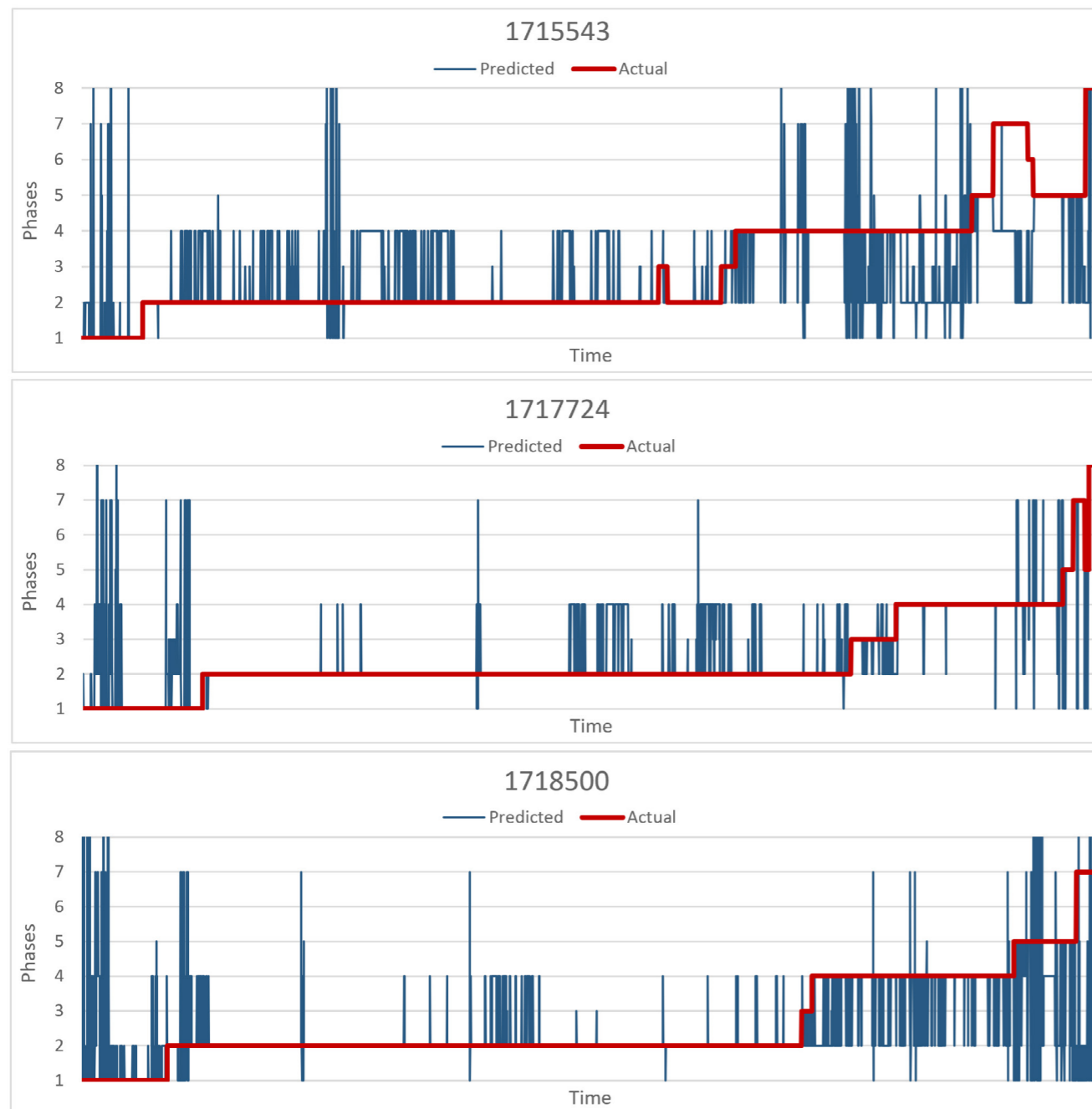
Besides the performance, in terms of recall and precision for the phase recognition, it was important to know which points during the surgical procedure were difficult to recognize by the ANN, because this could provide valuable information for the RSD prediction.

Graph 1 shows the phase recognition of the ANN in blue and the actual course of the phases in red. There was a large variation between the recordings, where recording #1491755 was recognized the best and recording #1377993 was by far the worst. Overall, phase two was best recognized. When this phase was recognized incorrectly, it was mostly confused with phase four. Moreover, was phase four often confused with phase two, the other phases were not particularly confused with one other phase. Phase six was never recognized, as can also be seen in the graph.

Furthermore, not all procedures followed the phases chronologically from phase 1 until phase 8, they sometimes jumped back and forth between phases, as can be seen by the red lines. However, the procedures that did follow the phases in chronological order, did not show a considerable better phase recognition. Lastly, there was no visual link between wrong recognitions and the phase transitions, or a specific part of the phase. There was not only a lot of variation between the phases but also between the recordings.

Graph 1. The course of the phases of each recording. The blue lines are the predictions and the red lines the actual course of the phases. The horizontal axes represent the time, the vertical axes the phases. Each number corresponds with the number of each phase. 1 = Trocar & tools insertion, 2 = Preparation & dissection, 3 = Clipping & cutting, 4 = Gallbladder dissection, 5 = Gallbladder packaging & retrieval, 6 = Liver bed





Although the weighted average of the recall and precision was quite high, the visual representation of the results showed that there was still a large amount of phase recognitions that were incorrect. The big variation in performance between the phases, but also between the recordings, could again be explained by the training data. With a larger amount of training data the ANN has more information to learn from. This is also the case for more diverse training data, both should improve the performance of the ANN. However, phase six was represented in the training data and the ANN still did not recognize this phase. This shows

that the ANN probably did not have enough data to recognize phase six. This could either mean that phase six is not commonly executed during a Lap Chol and therefore needs to be removed as phase, or the ANN needs more data to properly train.

Furthermore, this graph shows where the ANN made mistakes, during a phase or a phase transition. The phase transitions provide information on when a phase ended and a new phase started. This information could be used for the prediction of the RSD, and therefore needs to be the most accurate. However, the graph does not indicate that the ANN performed different during the phase or at a phase transition.

3.3. Tool - phase relation

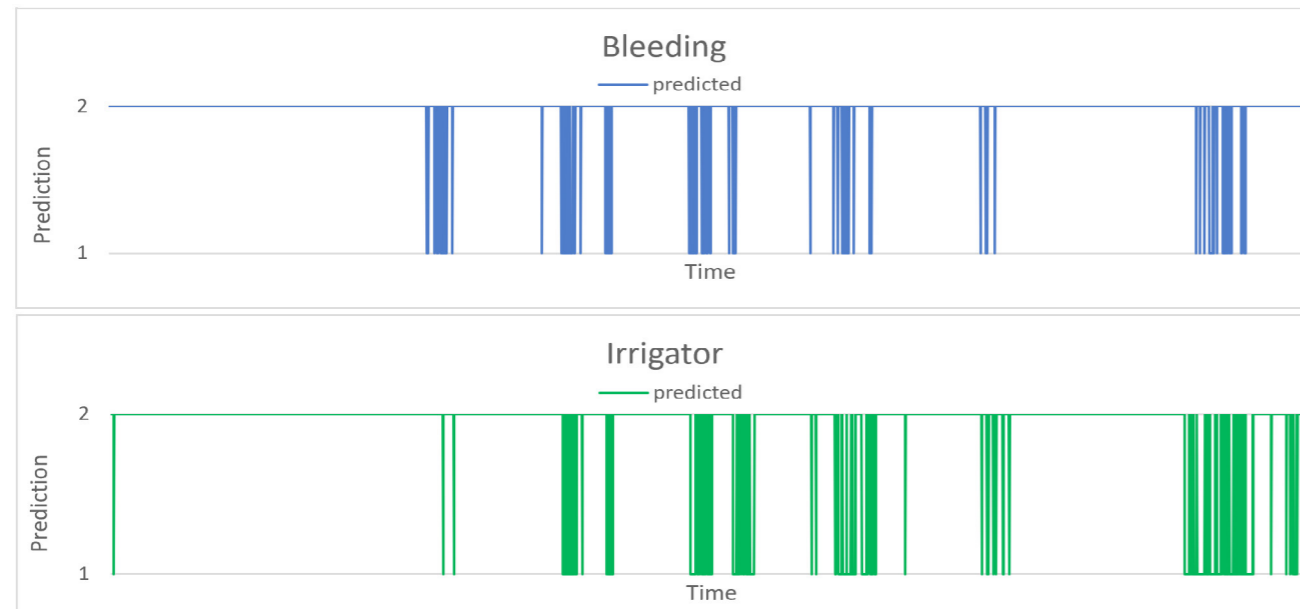
The phases were determined in relation to the tool usage, as shown in table 8. To have an idea how the ANN works, the tool-phase relation was calculated using equation 2.8. This showed the percentage of frames where a tool together with the corresponding phase were recognized, from the total amount of frames the tool was recognized.

The tool-phase relation was calculated for every recording individually, after which the average over all recordings was calculated together with the standard deviation, as seen in table 11. The clipper and the bag have the strongest relation with their corresponding phases. In 98.3% of the frames that the ANN recognized a bag, the ANN also recognized phase five. The hook and irrigator had the weakest relation to their corresponding phases and had the highest standard deviation.

Table 11. The results of the tool-phase relations.

Video numbers	Grasper - phase 2 (%)	Clipper - phase 3 (%)	Hook - phase 4 (%)	Bag - phase 5 (%)	Irrigator- Bleeding (%)
1310124	84.9	85.1	35.9	100.0	38.5
1327539	61.8	91.1	97.1	100.0	-
1344231	80.4	100.0	63.5	100.0	-
1377993	76.7	89.4	60.0	97.5	36.2
1491755	83.8	98.9	96.8	100.0	50.0
1565757	79.9	71.2	42.8	100.0	64.1
1715543	81.6	91.9	68.6	97.7	81.5
1717724	79.1	97.0	61.8	100.0	34.9
1718500	83.2	100.0	80.6	89.2	80.0
Average	79.1	91.6	67.4	98.3	55.0
Standard deviation	6.9	9.3	21.3	3.5	20.3

Graph 2. The recognition of the bleeding phase and irrigator of recoding 1717724. 1=present, 2=not present.



Graph 2 visualizes the tool-phase relation, the recognition of the bleeding and the irrigator are plotted over the length of the recording. On the vertical axis, a 1 represents that a bleeding or irrigator is recognized to be present and a 2 represents that a bleeding or irrigator is recognized to be not present. The graph shows that often a bleeding and irrigator were recognized at the same time.

The results of the tool-phase relation suggest that the ANN made a connection between the tools and phases that was similar to the one shown in table 8. However, some relations were stronger than others, this difference could be explained. The clipper and the bag had a strong relation with their corresponding phases, these tools were very specific for particular steps of the surgical process and were never used outside their corresponding phases. The grasper and hook were tools that were mainly used during their corresponding phases, although they could also be used during other phases, which explains the weaker relation.

The bleeding phase was different because it was an additional phase and occurred simultaneously with the other phases. The use of the irrigator was an indication of bleeding, but only in combination with visual blood. Like the grasper and the hook, the irrigator could also be used when there was no bleeding, which explains the lower relation.

3.4. Different sizes of training subsets

The size of the training subset is important for the performance of the ANN. In general, a larger training subset is preferred because it provides more information for the ANN to learn

from. However, there is a point where the larger subset does not outweigh the improvement of the performance. A larger training subset also increases the amount of work, as every additional recording needs to be screened, prepared and annotated manually. Therefore, within the training subset of exp 1, two smaller subsets were made to train two additional ANNs and to evaluate their performances. These smaller subsets are; exp 2, containing 18 recordings, and exp 3, containing nine recordings.

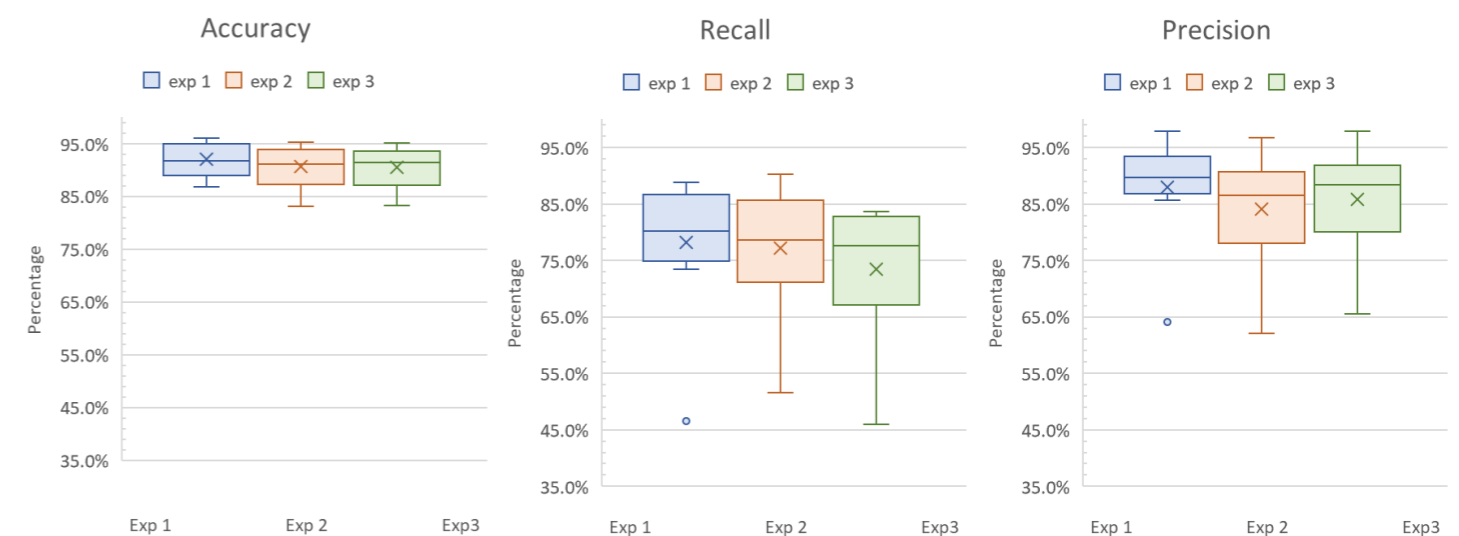
Table 12. The results of the tool detection of the three different validating subsets.

	Accuracy (%)	Recall (%)	Precision (%)
Exp 1	92.0	78.2	88.0
Exp 2	90.6	77.1	84.1
Exp 3	90.6	73.4	85.9

Table 12 shows the weighted overall average accuracy, recall and precision per experiment for the tool recognition. There was a slight decrease in performance the smaller the training subset got.

The results are plotted in boxplots in graph 3. The accuracy of exp 2 and exp 3 were comparable to exp 1, there was not a big difference between the experiments.

Graph 3. Three boxplots from the weighted average accuracy, recall and precision of the tool recognition of the three validating subsets.



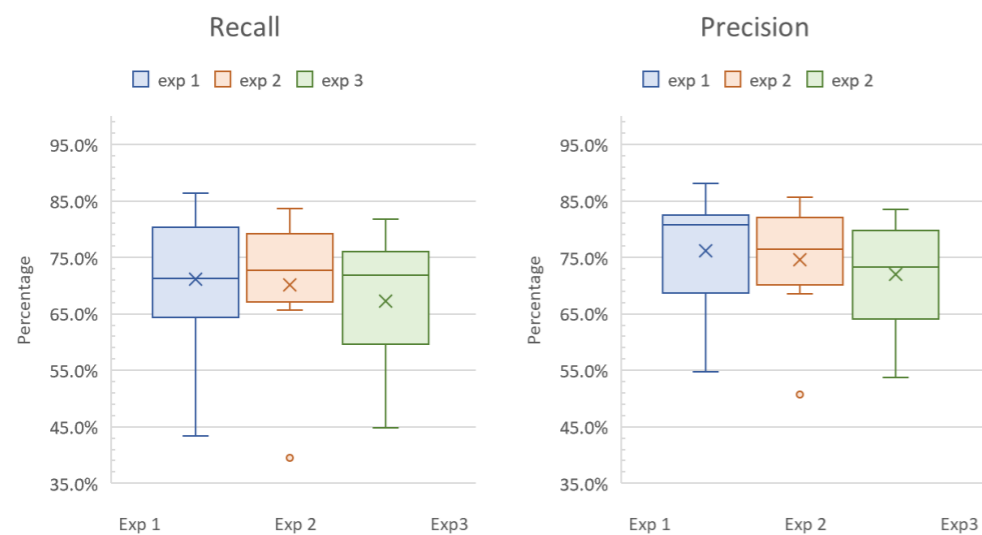
For the recall and precision this was not the case. For both, the overall weighted average of exp 2 and 3 were lower than exp 1 and the maximum and minimum of both exp 2 and 3 had a wider range than exp 1.

The results for the phases can be found in table 13, the recall and precision were used to evaluate the performance of the phases. Similarly as for the tools, there was a decrease in performance for both the recall and precision as the training subset got smaller. Graph 4 shows the results plotted in boxplots. These show that the weighted overall average of exp 2 and exp 3 were lower than exp 1 for both recall and precision. The range of the minimum and maximum of exp2 and exp 3 were smaller than the range of exp 1. Especially exp 2, although exp 2 had one outlier lower than the minima of exp 1 and exp 3.

Table 13. The results of the phase detection of the three different validating subsets.

	Recall (%)	Precision (%)
Exp 1	71.1	76.2
Exp 2	70.1	74.6
Exp 3	67.2	72.0

Graph 4. Two boxplots from the weighted average accuracy recall and precision of the phase recognition of the three validating subsets.



Looking at the results of the tool recognition, the larger training subset of exp 1 showed better and more consistent results. The lower range between the maximum and minimum of exp 1 showed that the results were more reliable considering all the recordings. Looking at the results of the phase recognition, a similar conclusion could not be drawn. These results were more comparable between the experiments and did not show a substantial difference.

3.5. Cross validation

To test the generalizability of the ANN a cross validation experiment was done. The ANN which was trained on the Lap Chol dataset was tested on two TLH recordings. The experiment only included the tools that occur in both types of surgeries, because the ANN could only use the labels of the Lap Chol experiment. The phases were not included because they were too surgery specific. The results will show if the ANN can also recognize the tools under different circumstances and can be extended to different surgery types.

Table 14 shows the average results of the two TLH recording for each tool used in the TLH and the Lap Chol procedure and the overall average results, next to the corresponding results of experiment 1 of the Lap Chol dataset. For each metric the overall average of the TLH was lower than the Lap Chol and the standard deviation was much higher for the TLH. The recall of the grasper and precision of the irrigator of the TLH came closest to the values of the Lap Chol. The recall of the grasper showed the smallest difference of 1.9%. The hook and the scissors showed the lowest accuracy, recall and precision and showed the biggest difference between the two surgery types. The recall of the hook showed the largest difference of 78% lower for the TLH than for the Lap Chol.

Table 14. The results of the tool recognition of the Lap Chol ANN on the TLH dataset.

	Accuracy (%)		Recall (%)		Precision (%)	
	TLH	Lap Chol	TLH	Lap Chol	TLH	Lap Chol
Grasper	75.5	89.2	80.1	83.0	70.3	87.5
Monopolar hook	76.4	95.7	1.7	79.7	23.7	83.2
Scissors	96.9	98.2	12.5	51.5	3.8	66.3
Irrigator & suction device	87.3	97.5	41.8	65.3	47.6	52.3
Overall average	84.0	95.2	34.0	69.9	36.4	72.3
Standard deviation	10.1	4.1	35.1	14.5	28.9	16.2

These preliminary results do not represent a definite answer if the ANN is truly generalizable, as the dataset of the TLH only contained two surgeries. The results of the grasper, as well as the irrigator, were promising when compared to the results of the Lap Chol. The low recall and precision of the scissors could be explained by the fact that the scissors were only used in less than 10 frames during each TLH recording. The hook on the other hand was one of the most used tools in the TLH, however there are different brands of hooks with different designs used in the hospital. The surgeons performing the Lap Chol procedures all used the same hook, but the surgeons performing the TLH used a different brand of hook.

3.6. Resources

Time is an imported resource in a project using machine learning. Preparing the dataset and training a deep learning network was a very time consuming process. The rule is that the quality of input is the quality of output. Therefore, it is preferred to screen, prepare and carefully annotate the data before providing it to the ANN. Especially the annotating part was very time consuming, as every frame had to be manually labelled with the correct tools and phase. To illustrate the time needed to do this, the time to annotate the tools was tracked during this project.

Table 15 shows the time needed to annotate 36 recordings of a Lap Chol procedure, together with the total time of all recordings added together. In total it took 99:15 hour to annotate all recordings, this was 3.5 times the total length of all recordings.

Table 15. Recorded time for the tool annotations of 36 Lap Chol recordings.

Total time recordings ([hh:mm:ss])	28:23:17
Total time annotating ([hh:mm:ss])	99:15:00
Number of recordings	36
Average time recording ([hh:mm:ss])	00:47:19
Average time annotating p. rec. ([hh:mm:ss])	02:45:27
Factor time annotating	3.5x

Annotating the tools was only a part of the annotating process, the phases also needed to be annotated separately. Although the tool annotations were the most time consuming part it was only one of several parts of the preparation that needed to be completed. The data also needed to be downloaded, checked if the recordings contained the correct and fully executed surgical procedure and the beginnings and ends needed to be trimmed before the annotating process could start. This showed a lot of time is needed to start a project like this, even when only one type of surgery was used. To use this method on a larger scale and more recordings the time needed to execute it would increase relatively with the increase of recordings. If twice as many recordings would be used, twice as much time would be needed. That is why it is desirable to explore ways to optimize this part of the project and produce an ANN that can be widely used.

4. Discussion

4.1. Results

The overall verdict about the results is that they are in line with literature on this topic. As explained in the previous chapter, the tool and phase recognition are similar to the results of Twinanda et al. [12], but are also similar to Alshirbaji et al. [29]. Alshirbaji et al. focused only on the tool detection of 80 Lap Chols and show the accuracy, recall and precision of their tool detection. The average results are respectively 93.95%, 78.62% and 77.57% but they do not provide the variation of their results.

Although the results of this thesis look promising, there are still some improvements that can be made to reach a more accurate tool and phase recognition. There is a large variation especially in the results of the phase recognition as visualised in section 3.2, which makes the results less reliable and not suited for practice yet. Twinanda et al. used a dataset of 80 recordings and trained their ANN on 40 recordings (almost twice as many than the training subset of this thesis) and showed a lower variation in their results. This indicates that increasing the size of the training subset should make the performance of the ANN used in this thesis more consistent.

However, the ANN could also be optimised besides increasing the size of the dataset. First, the cutoff threshold could be optimised. This is the threshold on which the ANN decides if a tool or phase is present or not [30]. The default threshold is 0.5 in a range from zero to one. For instance, if the ANN finds a probability for the presence of a tool that is lower than 0.5, it will label the frame with no tool. If the value is higher than 0.5, the frame will be labelled with the tool. Tweaking this threshold for every individual tool and phase will probably improve the results. Secondly, temporal information could be implemented in the ANN [31]. Most phases proceed in chronological order. Providing a timestamp of each frame to the ANN would increase the amount of information to train on and probably improve the performance of the phase recognition.

Applying this method on a larger scale would be demanding at this moment. Preparing the dataset and training an ANN is a time-consuming project. The annotating of 36 recordings took nearly 100 hours, which is only a small part of preparing the dataset. This needs to be improved in order to apply this method on a larger scale. There is room for improvement in the annotating process itself, in the method of annotating, the user-friendliness of the software, as well as the speed of the software and the server.

The generalizability of the ANN to other types of surgery would be a way to efficiently broaden this research and make this method relatively less time-consuming. The cross validation test was performed to assess the generalizability. This is only a preliminary test, as the dataset for the TLH was very small. However, the results for the recognition of the grasper and irrigator show that the ANN is not strictly dependant on one type of surgery. It also shows that the type or brand of the instrument can influence the design and this can have a large impact on the recognition of the ANN, as with the hook of the TLH. Nonetheless this indicates that it would be possible to use the ANN, that is trained on a certain surgery type, as a base to train for a different surgery type. The further training of the ANN only needs to be done on the additional tools and phases that are specific to the different surgery type. This could save a considerable amount of time.

4.2. Deep learning in healthcare

The use of deep learning in healthcare is a sensitive topic because deep learning is considered as a black box. It is unknown how the ANN actually executes the recognition task. The healthcare sector is hesitant to use deep learning in practice because the results of the ANN can not be validated while they could influence the choices for diagnosis or therapy of patients [32]. For the application of this thesis, this is less a concern as the ANN is not directly involved in the care of the patient. The focus of the research project this thesis is part of is merely on the optimisation of the planning of the OR. It is key that the ANN performs well to be able to optimize the planning of the OR, but a false recognition will not put lives of patients at risk. However, having insights in how the ANN works can still be beneficial, for optimizing the performance.

One way to get an idea of how the ANN works is by reverse engineering the results. During this thesis this was done by calculating the tool-phase relation to see if the ANN uses the implicit relationships between certain tools and phases. This test was not found in literature of this field of research before. The results suggest that the ANN did pick up on these relationships. However, it is not certain to say that the ANN also made or used these relationships. There is no causal relationship, because it is unknown how the ANN performs the recognitions due to how the ANN is structured.

Another way is producing a “heatmap” of the pixels of a frame the ANN uses to recognize the tools and phases [33]. This does not show the decisions the ANN makes, but it does show which parts of the image the ANN uses. This was not done during this thesis, but it could be valuable to perform this in the future to get a better understanding and discover how to further improve the ANN.

4.3. Future research

There is more research to be done on this topic besides the above mentioned areas. The main goal of the research project, this thesis is part of, is the prediction of the RSD. Therefore, work needs to be done on how to translate the information of the ongoing phases to a prediction model for the RSD. Preoperative factors that have an influence on the total surgery time (such as BMI, start time and number of OR staff) [34] could be added in the prediction model. The RSD also needs to be predicted in real time in an OR and provide information in a way suitable for the OR schedulers.

Overall this thesis is the first step towards RSD predictions. It shows that a deep learning method can be used to recognize the progress of a surgery automatically. This thesis provides in-depth information about the performance of the ANN and its applicability in medical practice. It also provides insights into the steps that need to be taken to use this deep learning method on a larger scale to improve OR planning.

References

- [1] A. Peltokorpi, "How do strategic decisions and operative practices affect operating room productivity?," *Health Care Manag. Sci.*, vol. 14, no. 4, pp. 370–382, Nov. 2011.
- [2] S.A. Erdogan and B.T. Denton, "Surgery Planning and Scheduling," in *Wiley Encyclopedia of Operations Research and Management Science*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2011.
- [3] A. Macario, T. Vitez, B. Dunn, and T. McDonald, "Where Are the Costs in Perioperative Care?: Analysis of Hospital Costs and Charges for Inpatient Surgical Care," *Anesthesiology*, vol. 83, no. 6, pp. 1138–1144, Dec. 1995.
- [4] A. K. Gupta, "JIT in Healthcare: An Integrated Approach," *Int. J. Adv. Manag. Econ.*, vol. 1, no. 2, pp. 20–27, 2012.
- [5] E. Van Veen-Berkx, "Enhancement opportunities in operating room utilization; With a statistical appendix," *J. Surg. Res.*, vol. 194, no. 1, pp. 43-51.e2, Mar. 2015.
- [6] E. Travis, S. Woodhouse, R. Tan, and S. Patel, "Operating theatre time, where does it all go? A prospective observational study," 2014.
- [7] A. Macario and F. Dexter, "Estimating the duration of a case when the surgeon has not recently scheduled the procedure at the surgical suite," *Anesth. Analg.*, vol. 89, no. 5, pp. 1241–1245, 1999.
- [8] B. J. Ammori, M. Larvin, and M. J. McMahon, "Elective laparoscopic cholecystectomy: Preoperative prediction of duration of surgery," *Surg. Endosc.*, vol. 15, no. 3, pp. 297–300, 2001.
- [9] I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy, "Deep neural networks predict remaining surgery duration from cholecystectomy videos," vol. 10434 LNCS. Springer, Cham, 2017.
- [10] A. C. P. Guédon, "It is Time to Prepare the Next patient' Real-Time Prediction of Procedure Duration in Laparoscopic Cholecystectomies.," *J. Med. Syst.*, vol. 40, no. 12, p. 271, Dec. 2016.
- [11] F. C. Meeuwssen, "Intraoperative monitoring of surgical instrument use with Radio Frequency Identification – a pilot study," 2019.
- [12] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 86–97, Jan. 2017.
- [13] S. J. Connor, W. Perry, L. Nathanson, T. B. Hugh, and T. J. Hugh, "Using a standardized method for laparoscopic cholecystectomy to create a concept operation-specific checklist,"

HPB, vol. 16, no. 5, pp. 422–429, 2014.

- [14] J. I. Einarsson and Y. Suzuki, "Total laparoscopic hysterectomy: 10 steps toward a successful procedure.," *Rev. Obstet. Gynecol.*, vol. 2, no. 1, pp. 57–64, 2009.
- [15] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.
- [16] C. Szegedy et al., "Going deeper with convolutions," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.
- [17] B. Raj, "A Simple Guide to the Versions of the Inception Network," 2018. [Online]. Available: <https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202>. [Accessed: 24-Sep-2019].
- [18] J. Brownlee, "Understand the Impact of Learning Rate on Neural Network Performance," 2019. [Online]. Available: <https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/>. [Accessed: 14-Nov-2019].
- [19] J. Brownlee, "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning," 2017. [Online]. Available: <https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/>. [Accessed: 14-Nov-2019].
- [20] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
- [21] Walber, "[CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>)]," 2014. [Online]. Available: <https://commons.wikimedia.org/wiki/File:Precisionrecall.svg>. [Accessed: 23-Nov-2019].
- [22] "Maryland Dissectors | Laparoscopic Surgery Instruments," 2017. [Online]. Available: <https://www.gerati.com/product/maryland-dissector/>. [Accessed: 25-Nov-2019].
- [23] "Reflex ® ELC530 Disposable Laparoscopic Clip Applier," 2017.
- [24] "Laparoscopic Scissors – Rekhison Quality Products." [Online]. Available: <https://rekhison.com/product/laparoscopic-scissors/>. [Accessed: 25-Nov-2019].
- [25] "Laparoscopic L-Hook. Endoscopic Solutions." [Online]. Available: <https://endoscopic.net/laparoscopic-l-hook>. [Accessed: 25-Nov-2019].
- [26] "Laparoscopic Irrigation Suction - Liss," 2019.
- [27] "Sterile Silicone Round Wound Drains with Trocars | Medline Industries, Inc." [Online]. Available: <https://www.medline.com/product/Sterile-Silicone-Round-Wound-Drains-with-Trocars/Drains/Z05-PF06832>. [Accessed: 25-Nov-2019].
- [28] "Covidien #173050G - Endo Catch Gold 10 mm Specimen Pouch 6/BX - CIA Medical." [Online]. Available: <https://www.ciamedical.com/covidien-173050g-endo-catch-gold-10-mm-specimen-pouch-6-bx>. [Accessed: 25-Nov-2019].
- [29] T. A. Alshirbaji, N. A. Jalal, and K. Möller, "Surgical tool classification in laparoscopic videos using convolutional neural network," *Curr. Dir. Biomed. Eng.*, vol. 4, no. 1, pp. 407–410, 2018.
- [30] T. Fawcett, "An introduction to ROC analysis," pp. 861–874, 2005.

- [31] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, "Video activity recognition: State-of-the-art," *Sensors (Switzerland)*, vol. 19, no. 14. MDPI AG, 02-Jul-2019.
- [32] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Futur. Healthc. J.*, vol. 6, no. 2, pp. 94–98, Jun. 2019.
- [33] W. Samek, A. Binder, G. Montavon, S. Bach, and K.-R. Müller, "Evaluating the visualization of what a Deep Neural Network has learned."
- [34] P. Phan, S.-H. Lee, T. Dai, N. Moran, V. J. Mathuranayagam, and J. Stonemetz, "Exploratory Study of Factors Influencing Surgery Scheduled Length, Deviation from Scheduled Length, and Impact on Length of Stay," *J. Am. Coll. Surg.*, vol. 227, no. 4, p. e160, Oct. 2018.

Appendix A

Hook

Grasper

Final check

Clipping & cutting

With this thesis I had the privilege to be part of the project from almost the beginning and this was a great learning experience. I started around the time the project application was filed for funding within the hospital. At that point I was working on my literature review. From there it took about six months until the equipment was installed, and the annotating process and training of the ANN could start. During these months a lot of preparatory work was done and I learned a lot about what it takes to start a new (research) project.

The legal affairs were an important topic at the start of the project, because the data that was used contained anonymous but personal data of patients. Although the data was anonymous it was still important to handle it with care and think about what data was needed and how to make sure it was used safely.

When the project was approved by the finance and the legal department of the hospital, I could start with the preparation and construction of the dataset. Which contained the downloading, checking if the recordings contained the correct and fully executed surgical procedure and if needed trimming the beginnings and ends of the recordings. Additionally, I had to familiarize myself and learn to understand the surgical procedures of the Lap Chol and the TLH and define the tool usage and surgical phases with the corresponding surgeons.

Together with Annetje, we thought of how to reach the end goal of predicting the RSD. We clarified what we wanted to do and discussed with COSMONiO if it was achievable within the possibilities of the funded project. We developed a method for the first step of this research, which are described in this thesis, with the future steps in mind. Furthermore, our workplace had to be prepared for the equipment of COSMONiO, the server and the tablet, that would be used. Together with ICT a solution was found to assure the installation and safe use with no interference with the hospital ICT infrastructure.

The last preparations towards the installation of the equipment was to find employees to do the annotation of the dataset. Via het FlexHuis, an organisation that provides flex workers, four employees were hired to work for two months. During this process I hired, instructed, supervised these employees and kept track of the expenses.

Throughout the entire project there was a lot of communication with COSMONiO, to constantly evaluate how it was going, where improvements could be made and how to adjust the software to the needs of the project. During the annotating process it was my responsibility to communicate the progress with COSMONiO and address and solve problems that occurred.


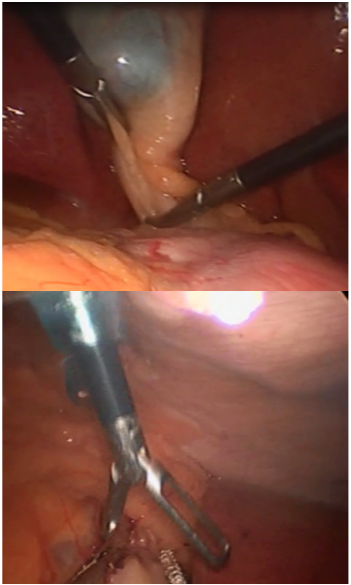

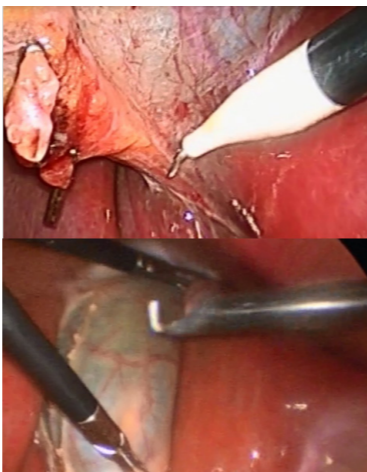

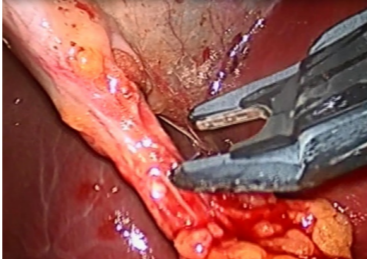
Over the course of the project I learned a lot about project management, all the different departments that are involved and all the details that need to be thought through when starting a (research) project. It gave me a better understanding about the health authorities. This part was not explicitly discussed in the main part of this thesis, nevertheless, it was a very valuable and enjoyable experience for me. I am thankful for the opportunity and the responsibilities Annetje entrusted me with.

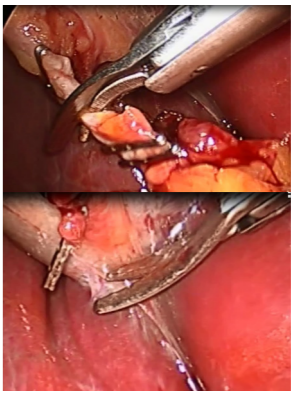





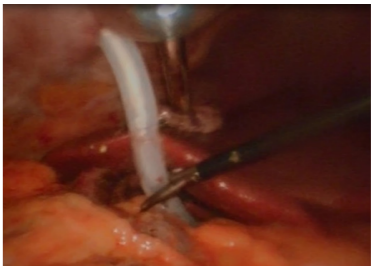
Appendix B

Instructions Laparoscopic Cholecystectomy (Lap Chol)

- The tools that are used can be found below.
- Annotate every tool that comes, or is, in view.
- About the tools that are marked red: when they come into view, annotate them only when at least 50% of the tip is in view.
 - The tip is the end of the tool until the hinge.
- When a tool is inside the border of the image and a part of the tool is hidden behind some of the anatomy or behind a different tool, but you know which tool it is (because of previous frames) you do annotate the tool.
 - This does not apply when the tip of the tool is more than 50% outside the border of the image.
- When the tool is hard to recognize because the image is blurry either because of movement or because of bad image quality, but you but you know which tool it is (because of previous frames) you do annotate the tool.
 - The image quality may never play a role in the annotation of the tools.
- Make sure you check yourself frequently, to see if the right labels are selected. Especially when the tools that are used have not changed for a long time.
- If you do not recognize a tool or if you are not sure, do not hesitate to ask. Better too ask to often than to annotate the wrong tool.

Tools:


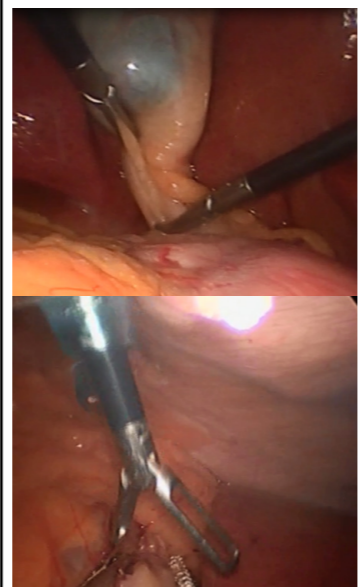
<p>Grasper [1], [2] (Multiple graspers could be used at once)</p>		
<p>Monopolar Hook [3]</p>		
<p>Clipper [4]</p>		


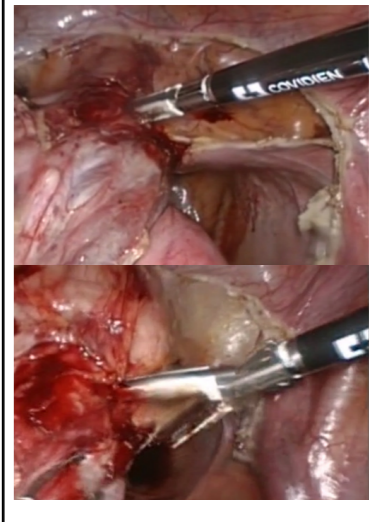

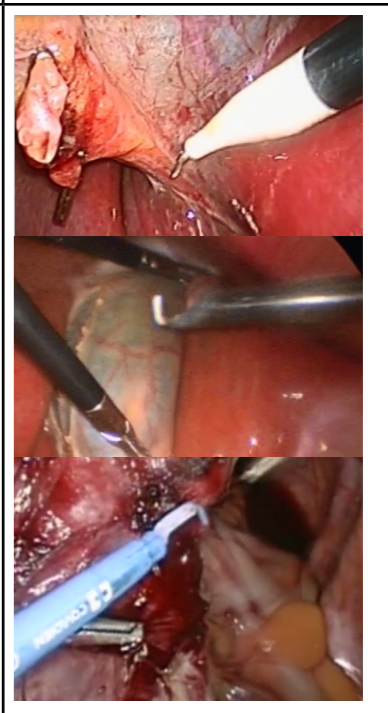

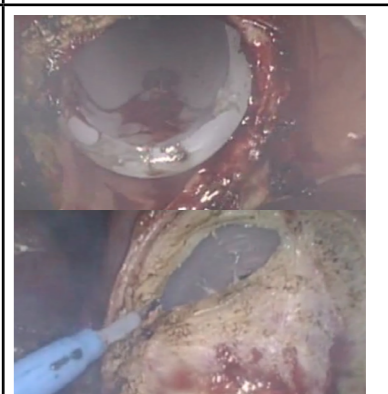
<p>Scissors [5]</p>		
<p>Bag [6]</p>		
<p>Irrigator [7]</p>		
<p>Drain [8]</p>		


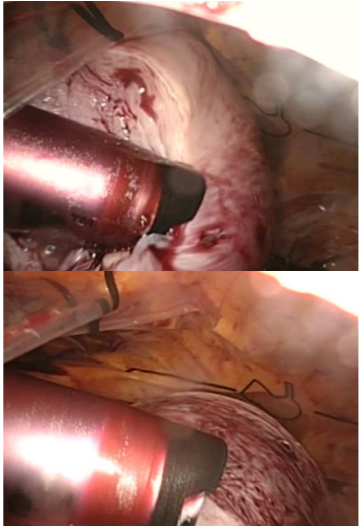

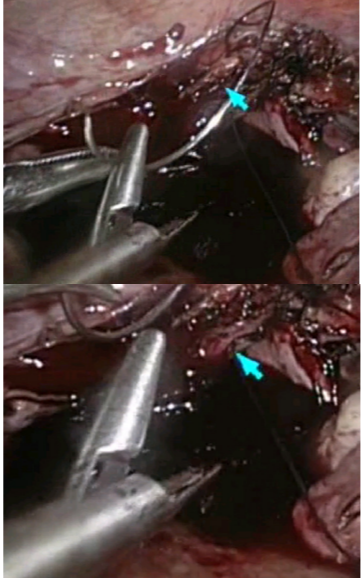

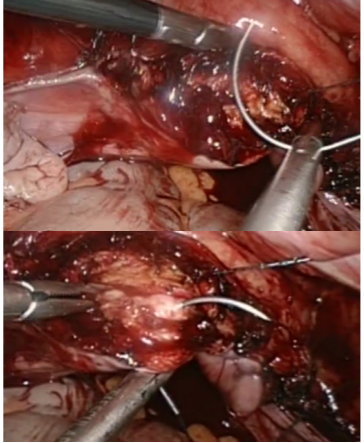
Instructions Total Laparoscopic Hysterectomy (TLH)


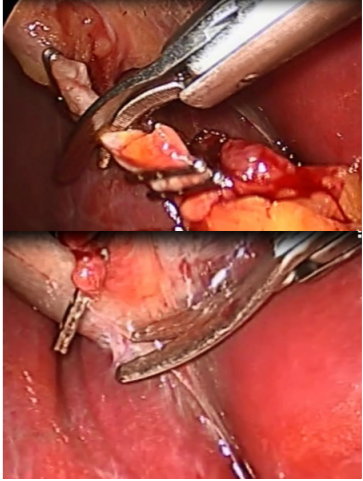


- The tools that are used can be found below.
- Annotate every tool that comes, or is, in view.
- About the tools that are marked red: when they come into view, annotate them only when at least 50% of the tip is in view.
 - The tip is the end of the tool until the hinge.
- When a tool is inside the border of the image and a part of the tool is hidden behind some of the anatomy or behind a different tool, but you know which tool it is (because of previous frames) you do annotate the tool.
 - This does not apply when the tip of the tool is more than 50% outside the border of the image.
- When the tool is hard to recognize because the image is blurry either because of movement or because of bad image quality, but you but you know which tool it is (because of previous frames) you do annotate the tool.
 - The image quality may never play a role in the annotation of the tools.
- Make sure you check yourself frequently, to see if the right labels are selected. Especially when the tools that are used have not changed for a long time.
- If you do not recognize a tool or if you are not sure, do not hesitate to ask. Better too ask to often than to annotate the wrong tool.

Tools:

<p>Grasper [1], [2] (Multiple graspers could be used at once)</p>		
---	---	--

<p>Ligasure [9]</p>		
<p>Monopolar Hook [3]</p>		
<p>Tube [10]</p>		

Morcellator [11]		
Needle feeder [12]		
Thread and needle [13]		

Scissors [5]		
Irrigator [7]		

References appendix

- [1] "Maryland Dissectors | Laparoscopic Surgery Instruments," 2017. [Online]. Available: <https://www.gerati.com/product/maryland-dissector/>. [Accessed: 25-Nov-2019].
- [2] "Laparoscopic Grasper - Glowcell instruments ." [Online]. Available: <https://www.glowcellinstruments.com/laparoscopic-grasper.html>. [Accessed: 28-Nov-2019].
- [3] "Laparoscopic L-Hook. Endoscopic Solutions." [Online]. Available: <https://endoscopic.net/laparoscopic-l-hook>. [Accessed: 25-Nov-2019].
- [4] "Reflex ® ELC530 Disposable Laparoscopic Clip Applier," 2017.
- [5] "Laparoscopic Scissors – Rekhison Quality Products." [Online]. Available: <https://rekhison.com/product/laparoscopic-scissors/>. [Accessed: 25-Nov-2019].
- [6] "Covidien #173050G - Endo Catch Gold 10 mm Specimen Pouch 6/BX - CIA Medical." [Online]. Available: <https://www.ciamedical.com/covidien-173050g-endo-catch-gold-10-mm-specimen-pouch-6-bx>. [Accessed: 25-Nov-2019].
- [7] "Laparoscopic Irrigation Suction - Liss ," 2019.
- [8] "Sterile Silicone Round Wound Drains with Trocars | Medline Industries, Inc." [Online]. Available: <https://www.medline.com/product/Sterile-Silicone-Round-Wound-Drains-with-Trocars/Drains/Z05-PF06832>. [Accessed: 25-Nov-2019].
- [9] "Covidien Valleylab Nano-Coated LigaSure Blunt Tip 44cm LF1844 Case of 6 - Global Medical." [Online]. Available: https://www.global-medical-solutions.com/Covidien-Valleylab-Nano-Coated-LigaSure-Blunt-Tip-44cm-LF1844-Case-of-6_p_9096.html. [Accessed: 28-Nov-2019].
- [10] "McCartney Tube - The O.R. Company." [Online]. Available: <https://theorcompany.com/products/mccartney-tube/>. [Accessed: 28-Nov-2019].
- [11] "LiNA Xcise Morcellator - Kebomed." [Online]. Available: <https://www.kebomed.co.uk/products/lina-xcise-morcellator-10/>. [Accessed: 28-Nov-2019].
- [12] "Laparoscopic Needle Holders - Endovision." [Online]. Available: <https://www.endovision.com.au/product/laparoscopic-needle-holders/>. [Accessed: 28-Nov-2019].
- [13] "Surgical Suture Needles with Thread - Lisen International, Inc." [Online]. Available: <http://www.liseninc.com/medical-instruments/tube-and-surgery/surgical-suture-needles-with-thread.html>. [Accessed: 28-Nov-2019].