Robotic Inference for Surgical Tools

Christian Rabus

Abstract—In this report, two image data sets are used to train for each of them a neural network using NVIDIA's DIGITS workflow. The trained models can later be used for robotics inference tasks on an NVIDIA Jetson TX2 board. The first data set (data-set A, supplied data set) contains 3 different classes: bottles, candy boxes and an empty conveyor belt. The purpose of the model is that on a conveyor belt a passing object (e.g. a bottle) of a certain class is detected in real time and then if necessary automatically sorted out. The second data set (data set B, data set for robotics inference project) contains 7 different classes of surgical instruments and are subset of tools that are used by a surgeon during spinal surgery. The purpose of the corresponding trained model is that later a "robotic scrub nurse" recognizes the different instruments on a sterile table (depending on the clinical work flow or requested by the a physician), grabs the right instrument and then passes it to the surgeon.

Index Terms—robotics, surgical robotics, service robotics, robotics in neurosurgery, robotic scrub nurse, scrub nurse robot, conveyor belt, deep learning.

1 Introduction

1.1 Data set A - Conveyor Belt

There are many interesting situations in the industry where sorting or classification of objects transported on a conveyor belt is performed. In the past, this process was carried out manually by humans. For economic reasons and due to automation, these tasks are increasingly performed by machines. This automated process usually consists of generally considering the objects distributed on the conveyor belt, locating each individual object, classifying it according to features that make it acceptable or unacceptable, and, if necessary, grasping, removing or separating to make the necessary separation.

Selver et al. [1] describes the design of an automated conveyor system where marble slabs are classified according to their quality (e.g. homogeneity, texture, colour, distribution of limestone, cohesive material). For this purpose, images of the marble slabs are taken, features extracted and then classified using cluster methods. Pneumatic pistons then separate the slabs.

Mattone et al [2] deals with the localization and identification of objects on the conveyor belt by using the height profile of objects scanned by a 3D optical device (laser beam plus CCD camera). Using fuzzy techniques, the objects are geometrically described and then classified with a neural network.

The supplied data set (**Data set A**) also deals with a more and more important topic: the sorting of garbage on a conveyor belt. The world generates 2.01 billion tonnes of municipal solid waste annually. If no action is taken, global waste will increase 70% to 3.4 billion tonnes by 2050. Especially in low-income countries, over 90% waste is mismanaged and worldwide only 13.5% is recycled. [3]

1.2 Data set B - Surgical Tools

Both the appearance of surgical robots and the increase in the complexity of surgery have increased the difficulty of a scrub nurse's work. The nurse not only has to be able to handle a larger number of instruments, the handling of those is also becoming more and more demanding. This task requires a high level of skill and concentration.

It is often a stressful job: the scrub nurse must pay attention all the time to the needs of the surgeon and must give the demanded instrument safely to him, because he usually doesn't look away from the operating field when he receives a tool. At the same time, however, she also has to clean and reprocess the instruments contaminated with blood and tissue that she received back from the surgeon. In addition, cutting or pointed instruments are often used, so that the task also involves a risk of injury. For this reason, the surgical nurse must have the skills to prevent injuries to herself, the patient and the surgeon. In many cases, the duration of a surgery can take several hours.

Furthermore, the development and commercial use of surgical robots has increased significantly in recent years. Many companies have specialized in specific clinical fields. The oldest and best-known example is the da Vinci (Intuitive Surgical Inc., Sunnyvale, CA, USA), a telemanipulated robot used in visceral surgery [4]. Other examples are the robots Mako [5] (Stryker Inc., Kalamazoo, MI, USA) used for joint replacement, or Mazor X [6] (MAZOR Robotics Ltd., Caesarea, Israel) used for spine surgery.

Based on these facts, the idea now arises to develop a robotic assistance system that functions as a kind of "robotic scrub nurse" and supports and relieves the human scrub nurse in its work.

Carpintero et al. [7] and Perez-Vidal et al. [8] describe in their papers the development of a robotic scrub nurse, which is equipped with a speech recognition module to recognize commands from the surgeon. The robotic system itself then locates the requested instrument on a storage tray by a pattern recognition algorithm, grasps the instrument with an electromagnetic gripper and places it on an interchange tray where the surgeon can pick it. The system can identify upt to 27 surgical tools and more than 82 spoken instructions.

Zhou et al. [9] describes in his paper the development of a robotic scrub nurse (RSN). The instruments are first segmented (patch-based segmentation) and pose is estimated.

Then the RS system picks up the unknown instruments to present them to an optical sensor in a determined pose. Finally, the instrument is recognized and delivered to the surgeon. The recognition accuracy is up to 95.6%.
The aim of this robotic inference project is to train a deep neural network with a **data set B** which contains the information of 6 (7) different surgical instruments. Those surgical instruments are subset of tools that are used by a surgeon during an image-guided spinal surgery (especially for pedicle screw placement) and are listed and presented below:

- **Pointer**: the Pointer is a pointed instrument to register the preoperative CT data of the patient to the patient itself. [Fig. 9]
- **ICM**: the ICM (instrument calibration matrix) is used during setup to register and calibrate the Drill Guide with the image-guided navigation system [Fig. 8]
- **Drill Guide**: the Drill Guide is held and placed by the surgeon on the vertebrae of the patient, to drill a hole into pedicle. [Fig. 6]
- **Trocar**: the Trocar is attached to the Drill Guide by the scrub nurse and aids during minimally invasive surgery. It protects the sleeve of the drill guide from tissue while preparation. [Fig. 10]
- **Depth Control**: the Depth Control is attached to the Drill Guide by the surgeon and is used to adjust the drilling depth. [Fig. 5]
- **Holder**: the Holder is attached to a mechatronic holding arm. It has an interface to the Drill Guide so that the surgeon can lock the position of the Drill Guide during surgery once he has aligned it [Fig. 7]
- **No Item**: no item, just blank images of the empty table.

2 Background / Formulation

2.1 Data set A - Conveyor Belt

The first step of building a trained neural network with the data set A was, creating a new Image Classification Dataset in DIGITS. The predefined default settings have been selected for this purpose:

- **Image Type**: Color
- **Image Size**: 256x256px
- **Resize Transformation**: Squash
- **Minimum samples per class**: 2
- **Maximum samples per class**: -
- **% for validation**: 25
- **% for testing**: 0
- **DB backend**: LMDB
- **Image Encoding**: PNG

The next step was to select a suitable NN to be trained with the image classification data set. In DIGITS, there are 3 known standard networks available:

- **LeNet**: requires 28x28 gray images
- **AlexNet**: requires 256x256 color images
- **GoogLeNet**: requires 256x256 color images

Considering the available data (256x256 color images) only AlexNet and GoogLeNet are appropriate. As already explained in detail in section 1.1, the trained model will later be used to sort garbage. The minimum requirement is to achieve an accuracy of 75% or better and an inference time of 10ms or less. According to the figure in class lesson 7.5, GoogLeNet is generally more accurate than AlexNet. However, AlexNet is faster than GoogLeNet. So there is a tradeoff between accuracy and speed. Considering the task ”sorting huge amounts of garbage”, it is important that the interference runs as fast as possible in order to reduce the amount of unsorted garbage. The small amount of incorrectly sorted waste can be corrected manually in a subsequent process. Therefore the standard network AlexNet is selected. AlexNet [10] is an 8-layer CNN and won 2012 the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by reducing the top-5 error from 26% to 15.3%. It consists 11x11, 5x5,3x3, convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum. It attachs ReLU activations after every convolutional and fully-connected layer.
For training the model within the DIGITS workflow, the standard settings have been chosen for the solver. The model was trained with 8 epochs [Fig. 2] and also with 5 epochs [Fig. 1]. The results are discussed in part 4.1 of this paper.

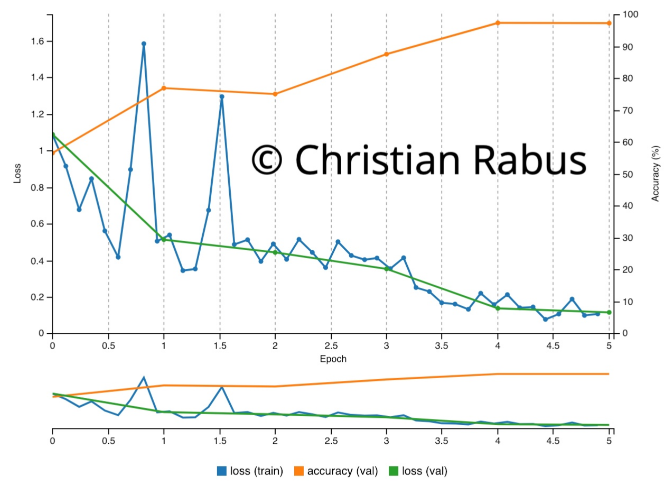

Fig. 1. Training curve with 5 epochs

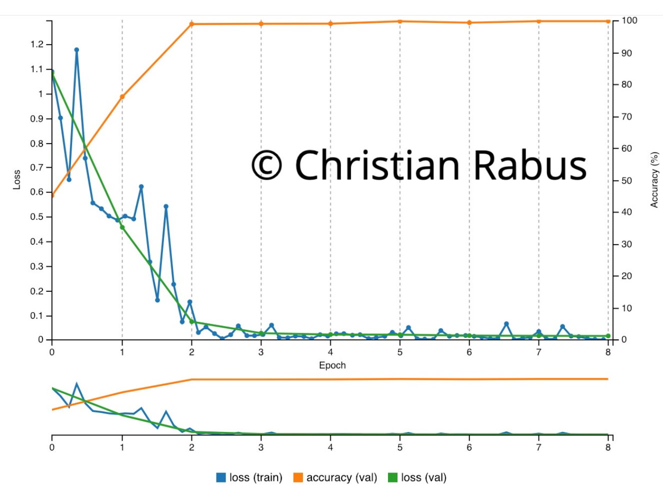

Fig. 2. Training curve with 8 epochs

2.2 Data set B - Surgical Tools

The approach of training a suitable model with data set B is similar to the approach as for data set A. After data acquisition (see section 3.2 of this report), the data set B has been uploaded to the DIGITS workspace. Then, a new image classification data set within DIGITS was created. The predefined default settings have been selected for this purpose:

- **Image Type**: Color
- **Image Size**: 256x256px
- **Resize Transformation**: Squash
- **Minimum samples per class**: 2
- **Maximum samples per class**: -
- **% for validation**: 25
- **% for testing**: 0
- **DB backend**: LMDB
- **Image Encoding**: PNG

The next step was to select a suitable NN to be trained with the created image classification data set. As already described in 2.1, there are 3 known standard networks available. Data set B images are colored and of size of 256x256px, so same as for data set A, therefore only AlexNet and GoogLeNet make sense. As already described in 1.2, a ”robotic scrub nurse” shall identify the correct instrument out of a large number of surgical instruments on a sterile table and then hand it to the surgeon. This process usually occurs only a few times per minute during a surgery. Since, in this scenario the required inference time does not have to be within milliseconds, but it is of crucial importance that the right instrument is recognized correctly. Hence, accuracy has precedence and the standard network GoogLeNet is selected. The CNN GoogLeNet [11] has 22 layers and is the winner of the ILSVRC 2014 competition. It achieved a top-5 error rate of 6.67%.

For training the model within the DIGITS workflow, the standard settings have been chosen for the solver. The model was first trained with 10 epochs [Fig. 3] and then set as pre-trained model (named: **10 epochs GooLeNet**). This pre-trained model was trained 5 epochs more with the same image classification data set and stored again as a pre-trained model (named: **10 epochs GooLeNet +5**). That step was repeated for a third trained model (named: **10 epochs GooLeNet +5+5**), so finally, the model was trained with a total of 20 epochs. On a second approach, the untrained GoogLeNet model was trained with 15 epochs [Fig. 4] ((named: **15 epochs GooLeNet**), whereas the default settings for the solver have been selected. This trained model was set to pre-trained and used for a second training with 5 epochs more (named: **15 epochs GooLeNet +5**). After each training job, the trained models were evaluated with new image data of data set B (in total 40 images of all 7 classes). The results are discussed in part 4.2 of this paper.

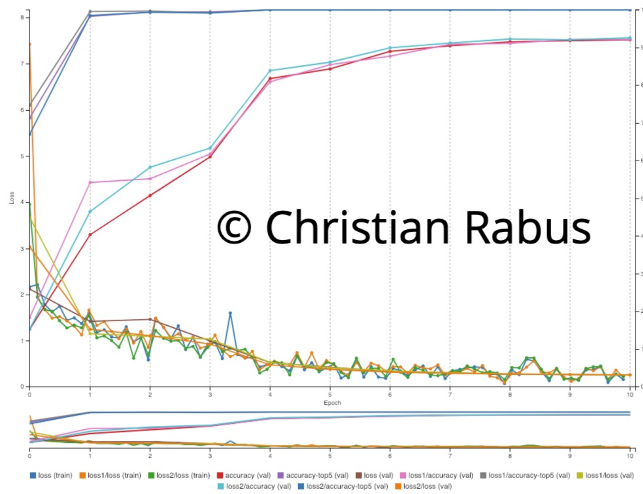

Fig. 3. Training of GoogLeNet with 10 epochs

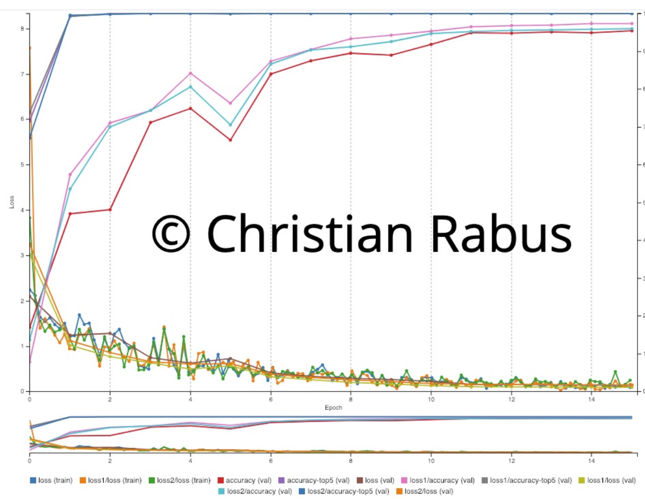

Fig. 4. Training of GoogLeNet with 15 epochs

3 Data Acquisition

3.1 Data set A - Conveyor Belt

The images of the supplied data set are divided in 3 categories: ”Bottle”, ”Nothing”, ”Candy Box”. The total number of images is 10,094 whereas ”Bottle” has 4,568, ”Nothing” has 3,031 and ”Candy Box” has 2,495. The original size of each image is 500x500 pixels, color (RGB) and PNG encoding. All images have been re-sized to 256x256, but no changes to image type and encoding. 75% of the images (7,570) are used for training the model and 25% of the images (2,524) are used for validation.

3.2 Data set B - Surgical Tools

The images of the robotic inference project are divided in 7 categories: ”Depth Control”, ”Drill Guide”, ”Holder”, ”ICM”, ”Pointer”, ”Trocar” and ”No item”. The total number of images is 3,780. The following table shows more details on the images.

The images have been gathered with the standard Camera app of an iPhone 6 (iOS 12.1.2, Apple Inc., Cupertino), by pressing and holding the shutter release button. This allows a series of photos to be taken in a short time. The respective instrument was placed on a table covered with a green sheet (sterile tables are usually covered with blue or green disposable sheet). The lighting conditions were adjusted to those of an operating theatre. The standard image size of the raw data is 2448x3264px, color (RGB) and jpeg-format. Then, the series of photos have been copied to

TABLE 1
Surgical Tools

Category	Img Training	Img Validation	Total
Depth Control	431	144	575
Drill Guide	499	166	665
Holder	444	148	592
ICM	386	128	514
Pointer	383	127	510
Trocar	376	125	501
No Item	311	104	415

a MacBook Pro (2.8GHz, IntelCore i7, 16GB DDR3, macOS Mojave 10.14.1, Apple Inc., Cupertino) and there re-sized with the standard Preview app. The new size of each image was 360x480px by keeping image type, resolution and color. Uploaded to DIGITS Workspace, a new image Classification Data Set was created with all 7 categories. The selected image type was color, the image size changed to 256x256px using the Squash re-size transformation. There, the data set was divided into 75% for training and 25% for validation. The following figures show the different tools (expect ”No Item”) after creation of the image classification data set.

Fig. 5. Depth Control

Fig. 6. Drill Guide

Fig. 7. Holder

Fig. 8. ICM

Fig. 9. Holder

Fig. 10. ICM

4 Results

4.1 Dataset A - Conveyor Belt

The requirements to meet are an **inference time <10ms** and **model accuracy >75%**. The results of the evaluation for a trained AlexNet model with 5 epochs [Fig. 11] and 8 epochs [Fig. 12] are shown here [Table 2].

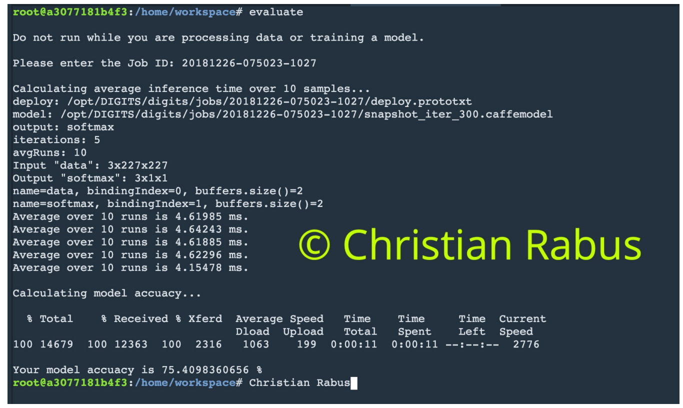

Fig. 11. Evaluation of trained AlexNet with 5 epochs

Fig. 12. Evaluation of trained AlexNet with 8 epochs

TABLE 2
Evaluation of AlexNet for supplied Data

Epochs	Inference Time	Accuracy	Meets Reqs
5	4.53	75.41%	yes
8	4.28	73.77%	no

Table 2 shows, that a trained AlexNet Model with 5 epochs fulfills the requirements.

4.2 Dataset B - Surgical Tools

In section 2.2 it was explained that the standard GoogLeNet network was used to create and evaluate 5 differently trained models for the inference project:

- **10 epochs GoogLeNet**: untrained GoogLeNet model, trained with 10 epochs
- **10 epochs GoogLeNet +5**: 10 epochs trained GoogLeNet model, trained with 5 more epochs

- **10 epochs GoogLeNet +5+5**: 10 +5 epochs trained GoogLeNet model, trained with 5 more epochs
- **15 epochs GoogLeNet**: untrained GoogLeNet model, trained with 15 epochs
- **15 epochs GoogLeNet +5**: 15 epochs trained GoogLeNet model, trained with 5 more epochs

Each trained model was evaluated with 40 new images. These images are not included in the training data set B but are of the same classes. The 40 images are distributed among the classes as follows:

- **Depth Control**: 5 images
- **Drill Guide**: 6 images
- **Holder**: 7 images
- **ICM**: 6 images
- **Pointer**: 6 images
- **Trocar**: 5 images
- **No Item**: 5 images

Table 3 shows, the results of the 5 trained GoogLeNet models. The best result was achieved by the trained network **10 epochs GoogLeNet +5** with a total accuracy of **90.17%**, where 37 of 40 instruments were correctly recognized. In contrast, the worst result was achieved by the trained network **15 epochs GoogLeNet +5** with a total accuracy of **60.73%**, where only 30 of 40 instruments were correctly identified.
The instruments (classes) with the best results, independent of the trained network, are **ICM** with **96.27%**, **Trocar** with **98.53%** and **No Item** with **99.50%**. These instruments (classes) were always identified correctly. On the other hand, the instruments (classes) with the worst results are **Drill Guide** with **59.95%** and **Holder** with **48.39%**. Even the network with the best results could not correctly identify all instruments in these two classes.

5 Discussion

5.1 Dataset A - Conveyor Belt

The required values were achieved with the trained AlexNet model - even if narrowly in terms of accuracy. The maximum allowed inference time, on the other hand, was clearly surpassed. In order to achieve a higher accuracy, it would be reasonable to train a GoogLeNet model with the data, even if the inference time would increase.

5.2 Dataset B - Surgical Tools

The evaluation from section 4.2 shows that very good results (accuracy of **90.17%**) can already be achieved with the trained network **10 epochs GoogLeNet +5** - at least with the available test data set of 40 images. Nevertheless, when a more detailed look is taken at the results, there are also some poor values. So why are there still some poor results? Looking into the details, for the two instruments **Drill Guide** and **Holder** not all test images were correctly identified. For the Drill Guide 1 of 6 image was not correctly assigned and for the Holder 2 of 7 images were incorrectly identified. Table 4 shows how the other trained networks have interpreted these test images. It is easy to see that the other networks also misinterpreted these test images. Only the trained model **15 epochs GoogLeNet** has correctly identified the drill_guide_03.JPG test image but only with a prediction of 51.76%. It is also interesting to note in Table 4 that if the instrument was detected incorrectly, the misinterpretation is the same for all networks. This means that the trained networks all wanted to have the wrong instrument identified.

Looking at these images, a possible explanation could be found. Unfortunately, the 3 test images show the instruments very well. A misinterpretation by human would hardly be imaginable. The selection of poor test images (e.g. blurred or strongly distorted) can therefore be excluded. Only for the test image holder_05.JPG [Fig. 14] it could be explained why the trained models might have recognized another instrument (Depth Control).
Nevertheless, the result of **90.17%** accuracy is already good, but does not reach the value by Zhou et al. [9], who has gained a recognition accuracy up to 95.6%.

Fig. 13. drill_guide_03.JPG

Fig. 14. holder_05.JPG

Fig. 15. holder_07.JPG

TABLE 3
Evaluation of GoogLeNet for inference project

Instrument	10 epochs GoogLeNet	10 epochs GoogLeNet +5	10 epochs GoogLeNet +5+5	15 epochs GoogLeNet	15 epochs GoogLeNet +5	Accuracy of Instrument
Depth Control (5)	82.15% (5)	99.17% (5)	96.66% (5)	57.55% (3)	52.29% (4)	77.56% (22/25)
Drill Guide (6)	53.09% (4)	76.38% (5)	70.02% (5)	81.66% (6)	18.61% (2)	59.95% (22/30)
Holder (7)	45.54% (4)	61.45% (5)	63.42% (5)	50.25% (4)	21.29% (3)	**48.39% (21/35)**
ICM (6)	97.80% (6)	98.62% (6)	98.80% (6)	99.51% (6)	86.61% (6)	96.27% (30/30)
Pointer (6)	92.57% (6)	95.65% (6)	95.96% (6)	93.29% (6)	53.78% (5)	86.25% (29/30)
Trocar (5)	98.95% (5)	99.97% (5)	99.93% (5)	98.90% (5)	94.88% (5)	98.53% (25/25)
No Item (5)	99.99% (5)	99.93% (5)	99.96% (5)	99.97% (5)	97.64% (5)	99.50% (25/25)
Overall Accuracy	81.44% (35/40)	**90.17% (37/40)**	89.25% (37/40)	83.02% (35/40)	60.73% (30/40)	-

TABLE 4
Results (excerpt) of the test image data set

Image Name	10 epochs GoogLeNet	10 epochs GoogLeNet +5	10 epochs GoogLeNet +5+5	15 epochs GoogLeNet	15 epochs GoogLeNet +5	Accuracy of Instrument
drill_guide_03.JPG	84.49% (Holder)	68.59% (Holder)	72.96% (Holder)	51.76% (Drill Guide)	64.32% (Holder)	10.35% (1 of 5)
holder_05.JPG	86.24% (Depth Control)	59.96% (Depth Control)	95.49% (Depth Control)	97.69% (Depth Control)	72.77% (Depth Control)	0.00% (0 of 5)
holder_07.JPG	96.05% (Drill Guide)	97.60% (Drill Guide)	96.97% (Drill Guide)	97.29% (Drill Guide)	69.81% (Drill Guide)	0.00% (0 of 5)

Considering the requirements of the application area (inference accuracy before inference time), it is desirable to achieve even better results.

6 CONCLUSION / FUTURE WORK

6.1 Dataset A - Conveyor Belt

The required goals where achieved with the trained AlexNet model. In order to present a realistic scenario for ”sorting of garbage”, it would be indispensable to have a much larger number of data and classes available. Only then it is possible to critically assess which neural network is suitable for this purpose.

6.2 Dataset B - Surgical Tools

In the present report it could be shown that with the help of the DIGITS workflow a known neural network can be trained and delivers acceptable results for the beginning. The idea of developing a Robotic Scrub Nurse and commercial use is given by the current ”robotization” [4], [5] and [6] in the OR. However, there are still many steps to be taken for an initial clinical test:

- Improvement of inference accuracy, e.g. by gathering further training data
- Gather data for other new surgical instruments
- Realization of the hardware, e.g. an industrial robot from Universal Robot, could be used cost-effectively as a basis
- Setup of a first complete system (Hard- and Software)
- Usability tests together with surgeons

REFERENCES

[1] M. A. Selver, O. Akay, F. Alim, S. Bardakçı, and M. Ölmez, “An automated industrial conveyor belt system using image processing and hierarchical clustering for classifying marble slabs,” *Robotics and Computer-Integrated Manufacturing*, vol. 27, no. 1, pp. 164–176, 2011.

[2] R. Mattone, G. Campagiorni, and F. Galati, “Sorting of items on a moving conveyor belt. part 1: a technique for detecting and classifying objects,” *Robotics and Computer-Integrated Manufacturing*, vol. 16, no. 2-3, pp. 73–80, 2000.

[3] S. Kaza, L. C. Yao, P. Bhada-Tata, and F. Van Woerden, *What a Waste 2.0 : A Global Snapshot of Solid Waste Management to 2050*. Washington, DC: World Bank, 2018. https://openknowledge.worldbank.org/handle/10986/30317 License: CC BY 3.0 IGO.

[4] da Vinci, Intuitive Surgical Inc., https://www.intuitive.com/en/products-and-services/da-vinci.

[5] Mako, Stryker Inc., https://www.stryker.com/us/en/portfolios/orthopaedics/joint-replacement/mako-robotic-arm-assisted-surgery.html.

[6] Mazor X, MAZOR Robotics Ltd., https://www.mazorrobotics.com/en-us/product-portfolio/mazor-x/mazorx-how-it-works.

[7] E. Carpintero, C. Perez, R. Morales, N. Garcia, A. Candela, and J. Azorin, “Development of a robotic scrub nurse for the operating theatre,” in *Biomedical Robotics and Biomechatronics (BioRob), 2010 3rd IEEE RAS and EMBS International Conference on*, pp. 504–509, IEEE, 2010.

[8] C. Pérez-Vidal, E. Carpintero, N. Garcia-Aracil, J. Sabater-Navarro, J. M. Azorín, A. Candela, and E. Fernandez, “Steps in the development of a robotic scrub nurse,” *Robotics and Autonomous Systems*, vol. 60, no. 6, pp. 901–911, 2012.

[9] T. Zhou and J. P. Wachs, “Needle in a haystack: Interactive surgical instrument recognition through perception and manipulation,” *Robotics and Autonomous Systems*, vol. 97, pp. 182–192, 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.