Finn Behrendt*, Nils Gessert and Alexander Schlaefer

Generalization of spatio-temporal deep learning for vision-based force estimation

https://doi.org/10.1515/cdbme-2020-0024

Abstract: Robot-assisted minimally-invasive surgery is increasingly used in clinical practice. Force feedback offers potential to develop haptic feedback for surgery systems. Forces can be estimated in a vision-based way by capturing deformation observed in 2D-image sequences with deep learning models. Variations in tissue appearance and mechanical properties likely influence force estimation methods' generalization. In this work, we study the generalization capabilities of different spatial and spatio-temporal deep learning methods across different tissue samples. We acquire several data-sets using a clinical laparoscope and use both purely spatial and also spatio-temporal deep learning models. The results of this work show that generalization across different tissues is challenging. Nevertheless, we demonstrate that using spatio-temporal data instead of individual frames is valuable for force estimation. In particular, processing spatial and temporal data separately by a combination of a ResNet and GRU architecture shows promising results with a mean absolute error of 15.450 compared to 19.744 mN of a purely spatial CNN.

Keywords: deep learning; laparoscopic imaging; spatio-temporal data; vision-based force estimation.

Introduction

Modern robot-assisted surgical systems for minimally invasive surgery offer high dexterity to surgeons and allow delicate operations. Features like tremor filtering or motion scaling enabling precise handling of the tools and can reduce physical trauma and hospital stay [1, 2]. By implementing force feedback, tissue interaction forces could be presented to the surgeon and thus facilitate the interventions especially for less experienced surgeons [3, 4]. In order to implement force feedback, interaction forces between instruments and tissues have to be measured. One approach is to measure the forces by external electro-mechanical force sensors, assembled at the tip of the instruments [5]. However, implementing sensors in existing systems while preserving full dexterity of the instruments and enable sterilization entails challenging problems [6]. To overcome disadvantages of electro-mechanical force sensors, vision-based force estimation (VBFE) is a promising approach. With VBFE, interaction forces between tissue and instruments can be estimated from images of the deformed tissue.

Recently, deep learning approaches that deal with image sequences and thus temporal information instead of single 2D-images have proven to be beneficial for force estimation:

In 2017 Aviles et al. [7] manually extracted the deformed structure of the tissue from stereo-image sequences and processed the features with a recurrent neural network (RNN).

In the work of Marban et al. [8], a convolutional neural network (CNN) is used as feature extractor instead of extracting the features by hand.

Recent approaches often use external cameras to acquire images. To implement VBFE into clinical application, it is beneficial to use established imaging systems for minimally invasive surgery: Laparoscopes. Furthermore, a systematic comparison of spatio-temporal approaches is missing and there is no evaluation of robustness across different ex-vivo tissues.

In this work, we compare different approaches of processing the temporal information systematically. Furthermore, we make use of different *ex-vivo* chicken heart tissues to evaluate robustness and generalization across tissues.

Materials and methods

Data-set generation

The visual information of the deformation of *ex-vivo* tissues is acquired along with the corresponding forces, measured by physical force-sensing, to train the models in a supervised fashion. The models learn to extract features from the input images and to map the deformation

***Corresponding author: Finn Behrendt,** Institute of Medical Technology, Hamburg University of Technology, Am Schwarzenberg-Campus 3, Hamburg, Germany, E-mail: finn.behrendt@tuhh.de
Nils Gessert and Alexander Schlaefer, Institute of Medical Technology, Hamburg University of Technology, Hamburg, Germany

of the tissues to a corresponding force. For inference, images of the tissues are provided to the models and the forces are estimated.

To obtain data to train and evaluate the algorithms, we use the following experimental setup. The main parts of the setup are a laparoscope (Olympus, Endoeye Flex 3D), a force-torque sensor (ATI Nano45), different *ex-vivo* tissues, a needle as instrument head phantom and a hexapod-robot (Physik Instrumente GmbH). A Schematic drawing of the setup is depicted in Figure 1. The force-torque sensor (2) is fixed to an adapter plate that is attached to the hexapod-robot (1). A drill chuck screwed on the force-torque sensor fixes the needle (3). The laparoscope (5) is attached to a carrier system. Three different chicken hearts (4) are used to evaluate the generalization ability across different chicken hearts of the applied algorithms. For measurement, the needle is driven by the hexapod and manipulates the tissue while the laparoscope is used to acquire images. The needle is driven with a smooth random movement pattern along it's shaft-axis. The force-torque sensor measures the interaction-forces induced by the needle. The acquisition is performed in 12 different trials per tissue. For each trial, needle position and maximal deformation depth is altered to evaluate the robustness of the algorithms. The acquisition procedure results in a data-set including chicken heart 1, chicken heart 2 and chicken heart 3 (CH1, CH2 and CH3), each consisting of 12 trials. In total, the tissue data-set consists of 9,707 + 9,835 + 9,624 = 29,166 images. To achieve equal frame rates, the force stream is down-sampled to the image acquisition frequency of 25 Hz. Images and the corresponding force values are synchronized by adding a delay between force and video stream manually. Figure 1 shows an exemplary image, fed to the network. The image has an original resolution of 865 px × 487 px and is cropped to a 400 px × 400 px region of interest. For evaluation across different chicken heart tissues, we apply a 3-fold cross validation where in each fold two different tissues are used as training data and the remaining tissue as test data. For tuning hyperparameters, we exclude four trials per fold to use them as validation-set.

Architectures and training

We investigate three network architectures (See Figure 2). For processing spatial data, i.e., RGB-images $x \in \mathbb{R}^{h \times w \times 3}$ a 2D-Resnet architecture (2DRN) is used. For processing spatio-temporal data, i.e. RGB-image-sequences $x \in \mathbb{R}^{h \times w \times 3 \times f}$, where f is the number of frames in the input-sequence, two approaches are provided: A 3D-Resnet architecture (3DRN) that processes spatial and temporal information simultaneously and a combination of a 2D-Resnet and a GRU-RNN architecture (CNNGRU) that processes the information separately. For all approaches, the residual architecture ResNet-34 [9] serves as backbone. For spatio-temporal models, the sequence length is set to f=4.

2DRN (Figure 2A) is the plain ResNet-34 architecture from [9] and serves as a baseline for comparison. Only the last layer is replaced by a fully connected layer with one output.

3DRN (Figure 2B) extends the 2DRN architecture to 3D convolutions. Therefore the 2D-Kernels are replaced by 3D-Kernels. For temporal dimension a stride of $s = 1$ is used.

CNNGRU (Figure 2C) decouples the processing of spatial and temporal data. First, the 2DRN is used to extract a feature representation of the images in the input-sequence. The feature vector $x_{out} \in \mathbb{R}^{2048 \times f}$ is then fed to the GRU-RNN, consisting of two hidden layers with 1024 neurons. Two fully connected layers with a dropout-layer and a ReLU activation in between produce the scalar output y_{out}.

The root mean square error (RMSE) serves as loss function. We train for 300 epochs with the Adam algorithm. 2DRN and 3DRN are trained from scratch and no pre-training is applied. For the CNNGRU, 2DRN pretrained on the same data-set is loaded and further trained along with the GRU-RNN. The mean average error (MAE) is reported as absolute metric. Relative metrics are provided by the relative MAE (rMAE) and pearson's correlation coefficient (PCC). The rMAE is determined by dividing MAE by the standard deviation of the targets. Additionally, the interquartile range is provided. To test for significant differences, the wilcoxon signed rank test is used with a significance level of α=0.01. To visualize learned features from input-images, a technique called Guided-backpropagation [10] is used for generating saliency maps.

Results

Table 1 shows that models that process temporal and spatial information separately, outperform other models by a margin. CNNGRU performs best and also 3DRN outperforms its 2D counterpart. The performance differences are statistically significant for all models. All differences of the MAE between the proposed models are significant with a p-value of p<0.01.

The results show an improved performance for CNNGRU. It is of interest how the averaged MAE is

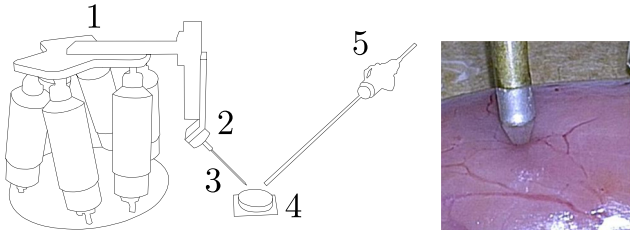

Figure 1: Experimental setup for acquiring the data-set. Left: Setup with all its components. Right: Exemplary acquired image.

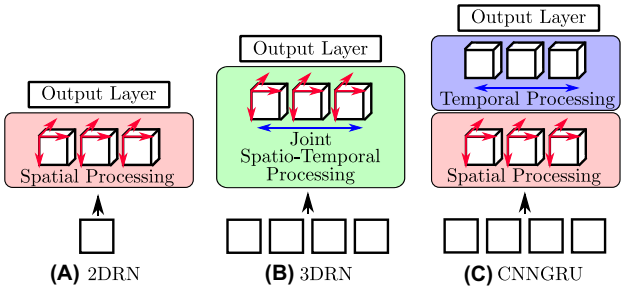

Figure 2: Proposed network architectures. 2DRN refers to 2D-Resnet, 3DRN refers to 3D-Resnet and CNNGRU refers to a combination of 2D-Resnet and a GRU-RNN.

Table 1: Comparison of the performance of all approaches. To compare the models, the performance is averaged across the three test- and training-set combinations of the chicken hearts (CH1, CH2, CH3).

Network	MAE [mN]	rMAE	PCC
2DRN	19.744(22.050)	0.423(0.423)	0.917(0.035)
3DRN	18.805(18.248)	0.403(0.391)	*0.922(0.032)*
CNNGRU	*15.450(16.126)*	*0.331(0.346)*	0.910(0.038)

composed regarding the different test-sets (CH1, CH2 and CH3) and their corresponding training-sets. Therefore, Figure 3 shows a boxplot, where CNNGRU and 2DRN are compared across different test- and training-set compositions. It shows large differences between the test-sets CH1, CH2 and CH3. Regarding the interquartile range, for every test-set a wide spread of the errors can be observed. It is noticeable that for 2DRN the absolute error is the lowest for CH3 as test-set and the highest for CH1 as test-set. All differences between the test- and training-set combinations regarding the absolute error are significant with a p-value of $p<0.01$. It is shown that 2DRN performs similar for test-set CH2 and CH3 but considerably worse for test-set CH1 compared to CNNGRU. To further investigate the different performance of CNNGRU and 2DRN across the chicken hearts (CH1 and CH3) saliency maps are presented in Figure 4. The shown saliency maps provide a possibility to gain insight to what pattern and regions the network relies on to solve a given task. Since the performance of 2DRN and CNNGRU differs for the different test-sets, a saliency map is provided for test-set CH1 and CH3 for both models in Figure 4. The blue regions of the images correspond to regions the network is assumed to not rely on for force estimation. The red regions indicate regions where information about the applied force is collected. For 2DRN, using test-set CH3 shows a cloud-like region around the tip of the needle where the network focuses on. For test-set CH1, the red region is fissured and not solely located around the tip of the needle. It can be observed that the region implicitly includes bulging tissue behind the needle and fat-tissue in general. For CNNGRU similar behavior can be observed for test-set CH3. However, for test-set CH1 the saliency map shows a smooth region around the needle.

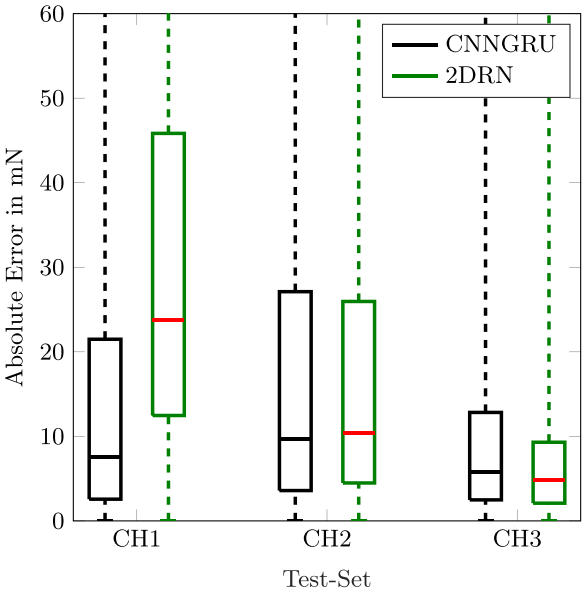

Figure 3: Absolute error for Resnet-34 (2DRN) in green and the combined Resnet-34 and GRU-RNN (CNNGRU) in black evaluated on different chicken heart data-sets (CH1, CH2 and CH3).

Discussion

With a reduced MAE of 3DRN and CNNGRU compared to 2DRN in Table 1 we show that including temporal information is beneficial for the task. The MAE of our proposed CNNGRU outperforms other models by a margin. Thus, we show that utilizing temporal and spatial information separately further improves the generalization ability compared to a simultaneously processing. Figure 3 indicates that generalization across different tissues is a challenging task. Differences across tissues regarding the stiffness, applied forces, general surface appearance and inhomogeneities within each tissue impede the force estimation across the tissues. This is reflected in varying performances between different test-sets and large interquartile ranges. Therefore, with respect to the clinical application, sufficient reliability of the force estimation for quantitative measurement appears to be difficult with the proposed surface-based method. The saliency maps in Figure 4 show that temporal information can guide the network to learn meaningful features. With respect to the large variability of the tissues, extending the data-set with more tissues is assumed to improve the overall performance. Also providing additional information of the underlying tissue structure and bio-mechanical properties is assumed to be beneficial for the task.

Conclusion

In this work, we investigate robustness and generalization of deep learning models utilizing temporal data for VBFE with videos acquired by a clinical laparoscope.

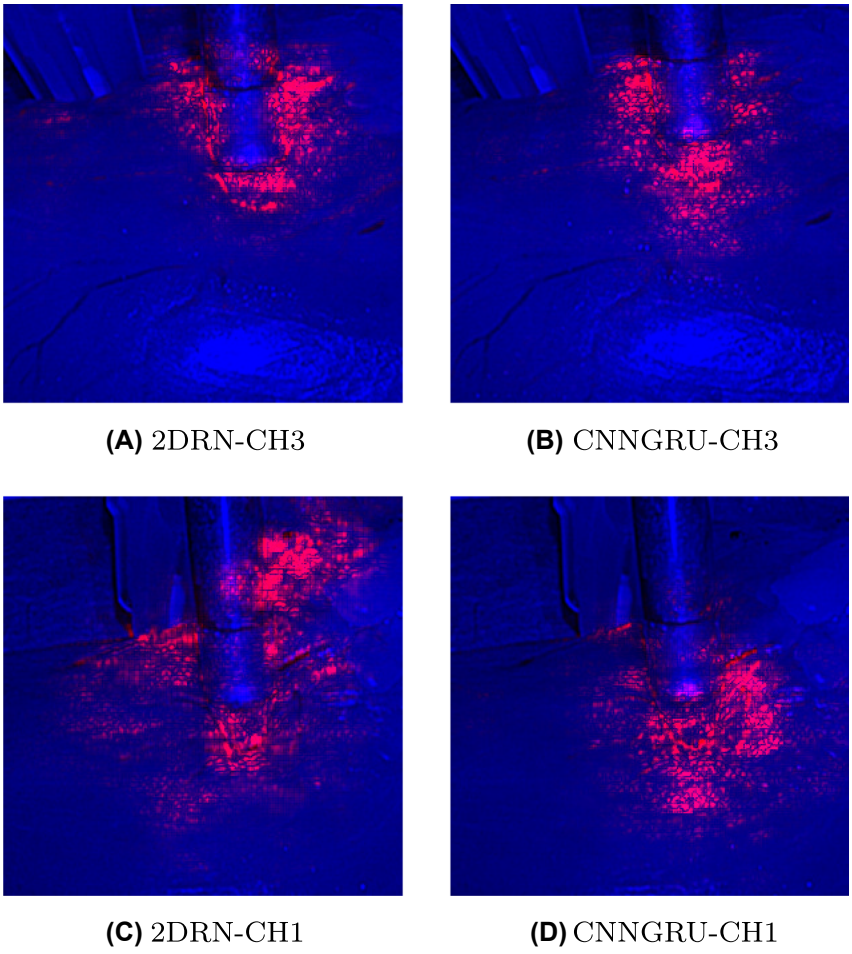

Figure 4: Saliency-map for 2DRN and CNNGRU. (A) 2DRN, trained with trials of CH1 and CH2, evaluated with CH3. (B) CNNGRU, trained and evaluated as in (A). (C) 2DRN, trained with trials of CH2 and CH3, evaluated with CH1. (D) CNNGRU, trained and evaluated as in (C).

We show that using videos with spatio-temporal deep learning methods rather than single images improves the robustness and generalization of force estimation. Furthermore, we show that processing temporal and spatial information separately is superior to simultaneous processing. Large variability across the tissues compared to the data-set size and missing information of underlying tissue structures or biomechanical properties aggravate the force estimation. In future work, tissue identification could be used to improve the generalization ability of the proposed deep learning models. Also, force prediction could be performed by estimating future force values based on past values as shown in [11].

Research funding: The laparoscopic system was provided by Olympus.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interest: Authors state no conflict of interest.

References

1. Moorthy K, Munz Y, Dosis A, Hernandez J, Martin S, Bello F, et al.. Dexterity enhancement with robotic surgery. Surg Endosc 2004; 18: 790–5.
2. Kim VB, Chapman WHH, Albrecht RJ, Bailey BM, Young JA, Nifong LW, et al. Early experience with telemanipulative robot-assisted laparoscopic cholecystectomy using da vinci. Surg Laparosc Endosc Percutaneous Tech 2002; 12: 33–40.
3. Reiley CE, Akinbiyi T, Burschka D, Chang DC, Okamura AM, Yuh D. Effects of visual force feedback on robot-assisted surgical task performance. J Thorac Cardiovasc Surg 2008; 135: 196–202.
4. Wagner CR, Howe RD. Force feedback benefit depends on experience in multiple degree of freedom robotic surgery task. IEEE Trans Robot 2007; 23: 1235–40.
5. Haidegger T, Benyó B, Kovács L, Benyó Z. Force sensing and force control for surgical robots. IFAC 2009; 42: 401–6.
6. Sokhanvar SS. Tactile sensing and displays: haptic feedback for minimally invasive surgery and robotics. Chichester, West Sussex: Wiley; 2013.
7. Aviles AI, Alsaleh SM, Hahn JK, Casals A. Towards retrieving force feedback in robotic-assisted surgery: a supervised neuro-recurrent-vision approach. IEEE Trans Haptics 2017; 10: 431–43.
8. Marban A, Srinivasan V, Samek W, Fernández J, Casals A. A recurrent convolutional neural network approach for sensorless

force estimation in robotic surgery. Biomed. Signal Proces. 2019; 50: 134–50.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: CVPR. Piscataway, New Jersey, US: IEEE; 2016: 770–8 p.
10. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. In: ICLR; 2015.
11. Gessert N, Bengs M, Schlüter M, Schlaefer A. Deep learning with 4d spatio-temporal data representations for oct-based force estimation. Med Image Anal 2020;64:101730.