Literature Review: "Learning Invariant Representation of Tasks for Robust Surgical State Estimation."

Hao Ding
hding15@jh.edu

1. Project Summary

The quality of a robot-mediated surgery is highly related to the skill of the surgeon. Thus, improving technical proficiency is always worth for researchers to look into. Our goal for this project is to develop a robust and interpretable system to empower the novice surgeons. Recent deep learning algorithms dominate large amounts of benchmarks for some areas including action recognition and prediction. We may treat the assistance as a kinematic prediction task using a deep neural network with the provided context and surgery task as input. However, deep learning methods have generalization issue and low interpretability. These drawbacks make the assistance system not reliable enough to put into practice. Causal inference mechanisms, with causal relation provided, are recently being highly focused on and explored by researchers to deal with the generalization issues and low interpretability of the deep learning algorithms. We will follow some recent works in incorporating causal inference mechanisms into deep learning architectures and try to improve the robustness of the assistance system through counterfactual inquiry.

2. Paper Selection

This paper is selected because it has several similarities with our project. First, we both work on Robot-Assisted Surgeries (RAS) data. Second, we both aim at improving robustness of the algorithm. Third, their invariant representation shares some ideas with causal factors. So, we may benefit from how it designs the architecture and training pipeline to learn the invariant feature representation.

3. Background and Contributions

a. Background

Autonomy is the trend, and real-time estimation of the current surgical state is a key prerequisite. The incorporation of multiple types of data (robot kinematics, endoscopic vision, and system events) can improve surgical state estimation accuracy, but not widely used. Nowadays, prior surgical state estimators relied

All figures and tables in this review are from the original paper: Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, andMahdi Azizian. Learning invariant representation of tasks forrobust surgical state estimation, 2021

heavily on RAS (Robot-Assisted Surgeries) datasets for model fitting/training, which leads to overfitting. Factors like endoscope lighting and viewing angles, surgical backgrounds are considered as potential nuisance factors that increase the training difficulty of a robust surgical state estimator. So, this paper tries to make surgical state estimation invariant to irrelevant nuisances and surgeon techniques if latent representations of the input data contain minimal information about those factors.

b. Contributions

The main contribution of this paper is introducing the StiseNet - adversarial model design that promotes invariance to nuisance and surgical technique factors in RAS data, as well as process to learn invariant latent representations of real-world RAS data streams, minimizing the effect of factors such as patient condition and surgeon technique.

4. Authors' work

The main part of the methods is the architecture and the learning pipeline of the Stisenet, it can be divided into following sections:

a. Feature extraction

The Stisenet simultaneously extract multimodal feature from **visual, kinematic,** and **system event** to form a feature representation H**.** It uses convolutional layers and LSTM encoder to extract features from RGB image concatenated with a segmentation mask predicted by a pre-trained segmentation network as H^{vis}. It uses LSTM encoder along with attention to extract kinematic and system event feature, H^{kin} and H^{evt} respectively. Then, it concatenates H^{vis}, H^{kin}, and H^{evt} as a 1-d vector and forms H.

b. Invariant representation learning

Feature vector H contains both nuisance feature and valid feature. It uses an encoder network E to divide H into $e_1$and e_2, where e_1 will be feed into a LSTM decoder M to predict the surgical state s. This means v is the valid feature while e_2 should be trained as nuisance feature. To make e_2 contains nuisance feature instead of meaningless noise, it adds a reconstructor R to reconstruct H from e_2 and a randomly dropouted e_1. To make e_1 and e_2 mutually exclusive, it adds two disentanglers f_1 and f_2 to try to infer e_1 from e_2 and infer e_2 from e_1

respectively. To make e_1 invariant to the surgical technique it adds a discriminator D for e_1 to infer a dataset specific style identifier l.

To learn these components and make them work properly, this paper follows a two-player game pipeline. In this two-player game L_M is minimized w.r.t M and E to make better prediction for surgical states. L_R is minimized w.r.t R and E to make e_2 contains necessary information to reconstruct feature. L_{f_1}, L_{f_2} are maximized w.r.t f_1 and f_2 to make the e_1 and e_2 mutually exclusive. L_D is maximized w.r.t D to make e_1 can't be discriminable.

The experiment part of this paper is also mainly two sections – a quantitative results and qualitative visualizations.

a. Quantitative results:

In quantitative results, some notations are introduced here: StiseNet-NO separates useful information and nuisance factors but excludes the invariance to surgical techniques. StiseNet-NA omits the adversarial component P2 entirely and uses H for estimation with Estimator. Non-causal setting: information from future time frames. Causal setting: current and preceding time frames.

From both table 1 and table 2, we see that the improvement is significant on some datasets but not stable. Especially from StiseNet-NA to StiseNet-NO, there is also significant decrease case which makes us think about the effectiveness of their proposed methods.

Table 1. Non-causal performance

	Non-causal			
	Input data	JIGSAWS	RIOUS+	HERNIA-20
TCN [11]	kin	79.6	82.0	72.1
TCN [11]	vis	81.4	62.7	61.5
Bidir. LSTM [12]	kin	83.3	80.3	73.8
LC-SC-CRF [14]	vis+kin	83.5	-	-
3D-CNN [13]	vis	84.3	-	-
Fusion-KVE [6]	vis+kin+evt	86.3	**93.8**	78.0
StiseNet-NA	vis+kin+evt	86.5	93.1	80.0
StiseNet-NO	vis+kin+evt	87.9	90.3	83.2
StiseNet	vis+kin+evt	**90.2**	92.5	**84.1**

All figures and tables in this review are from the original paper: Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, andMahdi Azizian. Learning invariant representation of tasks forrobust surgical state estimation, 2021

Table 2. causal performance

	Causal			
	Input data	JIGSAWS	RIOUS+	HERNIA-20
TCN [11]	vis	76.8	54.8	58.3
TCN [11]	kin	72.4	78.4	68.1
Forward LSTM [12]	kin	80.5	72.2	69.8
3D-CNN [13]	vis	81.8	-	-
Fusion-KVE [6]	vis+kin+evt	82.7	89.4	75.7
StiseNet-NA	vis+kin+evt	83.4	88.9	77.3
StiseNet-NO	vis+kin+evt	84.1	88.9	81.0
StiseNet	vis+kin+evt	**85.6**	**89.5**	**82.7**

b. Qualitative visualizations:

The paper shows a 2D UMAP for the feature space of e_1 and e_2 and shows that they are well separated and e_1 are meaningful for state recognition while e_2 is not as informative.

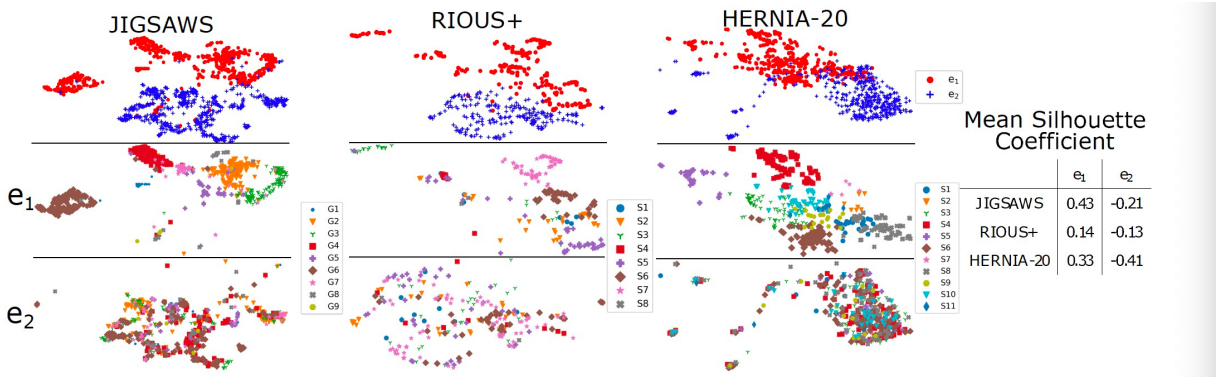

Fig. 5: 2D UMAP plots of information enclosed in e_1 and e_2 at each state instance. **Top row:** e_1 and e_2 segregates into distinguishable clusters, which indicates little overlap in information. **Middle row:** Information in e_1 color-coded by surgical states clusters relatively neatly. **Bottom row:** Information in e_2 is more intertwined and non-distinguishable by state. The mean silhouette coefficient $\overline{d}$ of each graph is shown, with a larger $\overline{d}$ indicating better clustering quality.

Figure 1. 2D UMAP

5. Assessment

This paper Provides a practical method for multi model feature extraction and a good two-player game thoughts to learn invariant features. The reconstructor idea to get avoid of trivial solution is also informative.

All figures and tables in this review are from the original paper: Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, andMahdi Azizian. Learning invariant representation of tasks forrobust surgical state estimation, 2021

However, it might have made some mistakes in the optimization design. As it says, the L_{f_1}, L_{f_2} are maximized w.r.t f_1 and f_2 to make the e_1 and e_2 mutually exclusive. L_D is maximized w.r.t D to make e_1 can't be discriminable. From my own point of view, the L_{f_1}, L_{f_2} should be minimized w.r.t f_1 and f_2 and maximized w.r.t E. By doing this the f_1 and f_2 are optimizing towards inferring e_1 and e_2 from each other and E is optimized towards making them mutually exclusive which is what it is supposed to do. Optimization on L_D is same. So, I guess this is the reason why the improvement in the quantitative results is not stable.

6. Relevance and Next Steps

This paper provides a thought for improving robustness of a RAS related neural network, which may be adapted to our architecture. Also, its idea of using two-player game to achieve invariance could also be explored to achieve causal relation for the feature extraction of our assistance system.

Once we successfully implemented a vanilla transformer network for kinematic prediction, we will try to incorporate some idea or method in this paper to test whether it is able to improve the robustness under our scenario.

7. Conclusions

From this paper we get some inspiration from the two-player game to learn invariant feature representation. This method might also be adapted to learn some causal relations in the network, which is always perceived as a black box. Also, learning an invariant feature representation itself is an effective approach to make the assistance system more robust.

All figures and tables in this review are from the original paper: Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, andMahdi Azizian. Learning invariant representation of tasks forrobust surgical state estimation, 2021

Reference

1. Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, andMahdi Azizian. Learning invariant representation of tasks forrobust surgical state estimation, 2021

All figures and tables in this review are from the original paper: Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, andMahdi Azizian. Learning invariant representation of tasks forrobust surgical state estimation, 2021