Ranking robot-assisted surgery skills using kinematic sensors

Burçin Buket Oğul[1,2], Matthias Felix Gilgien[2] and Pınar Duygulu Şahin[1]

[1] Department of Computer Engineering, Hacettepe University, Ankara, Turkey
[2] Department of Physical Performance, Norwegian School of Sport Sciences, Oslo, Norway
bb.ural@gmail.com, matthias.gilgien@nih.no,
pinar@cs.hacettepe.edu.tr

Abstract. Assessing surgical skills is an essential part of medical performance evaluation and expert training. Since it is typically conducted as a subjective task by individuals, it may lead to misinterpretations of the skill performance and hence lead to suboptimal training and organization of the surgical activities. Therefore, objective assessment of surgical skills using computational intelligence techniques via sensory data has received attention from researchers in recent years. So far, the problem has been approached by employing a classification model where a query action for surgery is assigned to a predefined category that determines the level of expertise. In this study, we consider the skill assessment problem as a pairwise ranking task where we compare two input actions to identify better surgical performance. To this end, we propose a hybrid Siamese network that takes two kinematic motion data acquired from robot-assisted surgery sensors and report the probability of the first sample having a better skill than the second one. Experiments on annotated real surgery data reveals that the proposed framework has high accuracy and seems sufficiently accurate for use in practice. This approach may overcome the limitations of having consistent annotations to define skill levels and provide a more interpretable means for objective skill assessment.

Keywords: Skill assessment, ambient intelligence in education, ambient intelligence in health, robot-assisted surgery, Siamese networks, LSTM.

1 Introduction

A major task in medical training is the assessment of surgical actions to grade current performance of the candidate and monitor the development of skills during training activities. These activities are usually performed manually in an operation room under supervision of expert surgeons [7]. Manual assessment, even being performed by experts, has several limitations, including subjectivity, lack of consistency and reliability.

Recent developments in computer-assisted surgery provide new opportunities to employ ambient intelligence techniques for objective skill assessment [12]. Data col-

lected during surgery activity, either from sensory or multimedia interfaces, can serve as platform for offline analysis of the action of surgeon post operation. Recently the machine learning community has made an effort to realize such analysis, which includes the development of computational methods that can automatically identify the surgery skill level as “expert”, “intermediate” or “novice” from surgical action data [3, 4, 5, 13, 14, 15]. The major short coming of these systems is their limited ability to predict a fixed number of predefined, possibly inconsistent, categories for expertise levels. They are unable to model skill levels between these category labels. Furthermore, they can model only overall expertise level of surgeons although it seems obvious that a surgeon’s performance may vary between different surgery action.

In two recent studies, the authors considered the problem as a task of learning to rank video recordings [2; 10] instead of assigning them into predefined labels. The study aimed to build models with wide applicability of skill determination in any domain, but algorithms were also tested for surgical skill assessment in addition to other tasks. In [2], they introduced a two-stream Temporal Segment Network to capture both the type and quality of actions. As an alternative to that [10] integrated an attention pooling and temporal aggregation mechanism to a two-stream CNN model. Skill assessment through video recordings has two main limitations. First, video data processing is time and resource inefficient, which makes it difficult to run the algorithms in conventional personal computers. Second, video can record the actions in two dimensions. This is unfortunate since tracking of trajectories and velocities can only be measured in two dimensions and important information of surgery skills is lost, if the third dimension is lacking.

In this study, we use kinematic data collected from robot-assisted surgery environment to develop a method for rank-based assessment of surgery skills. The study is considered as an emerging application of data-driven ambient intelligence in education of healthcare professionals. We introduce a novel deep learning framework based on Siamese of recurrent neural networks for pairwise ranking of motion kinematics in the form of multi-variate time-series data. We present the results of experimental evaluation of our method on real life data collected from human-controlled robot arms for surgical skill assessment. We argue that our approach provides a more interpretable and reliable view of objective skill assessment while it overcomes the limitations caused by inconsistencies in subjective skill grading scales.

2 Methods

2.1 Siamese framework for ranking

Given two surgical actions, *m* and *n*, with their kinematic data of $\boldsymbol{x^m}$ and $\boldsymbol{x^n}$, which are in the form of multi-variate time series, the task is to determine which surgical action is performed with better quality. We denote this output by p_{mn} where;

$$p_{mn} = \begin{cases} 1 & m \text{ performs better than } n \\ 0.5 & m \text{ and } n \text{ show equal performance} \\ 0 & n \text{ performs better than } m \end{cases} \quad (1)$$

We interpret this as the probability of the first surgical action being performed better than the second. Our goal is then to train a model that minimizes the probabilistic loss in human-annotated samples for surgical skills. To this end, we introduce a novel framework based on a Siamese network of recurrent neural networks integrated with a probabilistic ranking layer, which can take the case of skill equivalence into consideration (Fig.1).

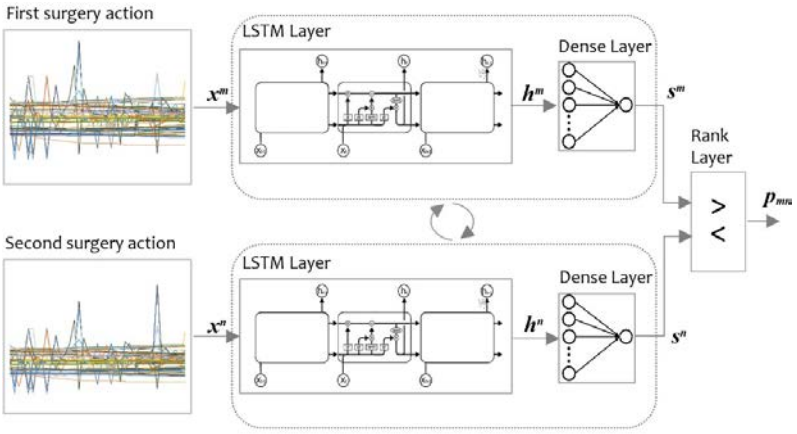

Fig.1 Siamese network for pairwise ranking of surgery actions

LSTM Layer: The kinematic data at one input of the Siamese network is given in the form of multi-variate time-series and fed into a long short-term memory (LSTM) network [9]:

$$h_t = LSTM(h_{t-1}, x_t) \qquad (2)$$

where x_t and h_t are the inputs at time t. The LSTM is parameterized by output, input and forget gates, controlling the information flow within the recursive operation. At every time step t, LSTM outputs a hidden vector h_t that reflects the skill representation of the kinematic motion at position t.

Dense Layer: A fully-connected layer takes the vector of skill representation at the output of an LSTM layer, h^m for any of the input m, and transforms it into a scalar, s^m, which is directly comparable with the output, s^n, at the other end of the Siamese network.

Rank Layer: This layer adapts a probabilistic loss function, which was originally introduced to learn how to rank text objects using gradient descent [1]. The pairwise rank between two surgery action inputs is required to be represented by p_{mn}, which is interpreted as the probability of m performing a better action than n. We denote the posterior probability distribution $P_{ij}=P(i \succ j)$, where $\succ$ refers to the skill superiority of i to j and let $\acute{P}_{ij}$ be desired target values for those posteriors, such that $\acute{P}_{ij} \in \{1,0.5,0\}$. The goal is then to minimize the distance between these two entities.

2.2 Implementation

We used a bidirectional LSTM [8] to allow the modelling of two-way temporal dependencies in actions. The rank layer was implemented by a sigmoid activation followed by a binary cross-entropy loss function. We used the following hyper-parameters: a learning rate of 0.001, a batch size of 2 and a unit size of 64 with single hidden layer. The framework was implemented in Keras using TensorFlow back-end.

3 Results

3.1 Data

We evaluated our framework on a publicly available real surgery data set called JIGSAW [6]. The JIGSAWS dataset contains of surgical data collected from eight subjects with different skill levels performing three different surgical tasks using the da Vinci surgical system. The tasks are 4-throw suturing (39 trials), needle passing (26 trials), and knot tying (36 trials) performed on benchtop training phantoms. The dataset consists of 76 motion variables collected at 30 Hz, including tooltip positions and orientation, linear and rotational velocities, and gripper angle. Therefore, the kinematic data that we use in the study refers to a multi-variate time-series data set captured from the da Vinci robot. A trial is a part of the data set that corresponds to one subject performing one instance of a specific task. Each subject is categorized by a fixed expertise level but each trial may have a different score. This score is annotated using OSATS as a grading system [11]. OSATS consists of different grading criteria like respect for tissue, time and motion, flow of operation, overall performance and the quality of the final product. The system reports a final grade to represent general surgical skill.

3.2 Evaluation Setup

We performed four-fold cross validation to evaluate the prediction performance. In this setup, the pairs between ¾ of the actions were used for training and the remaining pairs were used for testing. Therefore, test samples include both the pairs where neither video has been used in a pair for training and the pairs where the other video was used for training in a different pairing. We used pairwise ranking accuracy, which is the percentage of correctly ordered pairs, produced by each fold. We reported two accuracy results for the cases where the skill equivalence is counted and not counted.

3.3 Empirical Results

We applied our model for each surgery task separately to rank surgery actions by their skills. Table 1 discerns the accuracy for each task where the pairs with equal skill scores are taken into account.

Table 1. Results of pairwise ranking including skill equivalence

Surgery type	Accuracy (%)
Knot tying	75.1
Needle passing	74.4
Suturing	60.3

To our knowledge, this is the first study that applies a method for pairwise ranking on kinematic data of surgery skills. Therefore, there is no previous study to benchmark our method. [2; 10] are the most relevant studies, which used video data for skill ranking and tested their methods in the same dataset. Another difference is the fact that they work for only binary ranking cases, where the equivalent skills were found to be inconsistent. To make a comparison with these methods we ran our model with complementary kinematic data from which the equally-rated pairs were removed. The results are shown in Table 2. [2; 10] did not give accuracies separately for each task, but rather reported overall performance in surgery dataset. [2] tested two different versions of their model, where only spatial information was used, and temporal data was combined with spatial data in two-streams. Our model can achieve a competitive accuracy with video-based models. Moreover, the present model is built upon only kinematic data, which reduces the computational resources compared to approaches which use videos. [2] reported that average running time to train a single fold is 18 hours with NVIDIA TITANX GPU, whereas learning a fold in our model is conducted in less than an hour with a conventional CPU.

Table 2. Results of pairwise ranking excluding skill equivalence

Method	Action data	Surgery type	Accuracy (%)
Present method	Kinematic	Knot tying	79.6
		Needle passing	77.5
		Suturing	63.5
		Average	73.5
Doughty et al. 2018 (spatial)	Video	-	66.5
Doughty et al. 2018 (two-stream)	Video	-	74.4
Li et al. 2019	Video	-	73.1

4 Conclusion

We introduce a novel framework for objective skill assessment for robot-assisted surgery. The contribution of the study is twofold. First, kinematic-based surgical skill assessment problem is approached for the first time as a pairwise ranking task instead of the direct assignment of samples into predefined skill categories. This approach provides a more interpretable and reliable skill assessment while it overcomes the limitations caused by inconsistencies in subjective skill grading scales. Compared to video-based solutions, the use of kinematic data reduces the demands on computational power and is therefore a more applicable alternative for the practical implemen-

tation in a hospital setting. Second, a novel deep learning framework based on Siamese of recurrent neural networks is introduced for pairwise ranking of multi-variate time-series data. Experimental results on surgical skill assessment data have justified the applicability of the proposed models for this task. Since the system does not rely on learning from any problem-specific features, the framework can be easily adopted for other problems in data-driven ambient intelligence with sensory interfaces.

Acknowledgments: Burçin Buket Oğul was financially supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under 2214-A program.

References

1. Burges, C. J., Shaked, T., Renshaw, E. et al.: Learning to rank using gradient descent. Int. Conf. in Machine Learning, 89–96 (2005).
2. Doughty, H., Damen, D. and Mayol-Cuevas, W.: Who's better? who's best? pairwise deep ranking for skill determination. IEEE Conf. on Comp. Vis. Pattern Recog. (2018).
3. Fard, M.J., Ameri, S., Darin, E.R. et al.: Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int. J. Med. Robot. and Comput. Assist. Surg. 14(1): e1850 (2018).
4. Fawaz, I. H., Forestier, G., Weber, J. et al.: Evaluating surgical skills from kinematic data using convolutional neural networks. MICCAI Workshop, 214–221 (2018).
5. Funke, I., Mees, S. T., Weitz, J. and Speidel, S.: Video-based surgical skill assessment using 3D convolutional neural networks. arXiv preprint arXiv:1903.02306 (2019).
6. Gao, Y., Vedula, S. S., Reiley, C. E. et al.: JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modelling. MICCAI Workshop (2014).
7. Grantcharov, T. P., Bardram, L., Funch-Jensen et al.: Assessment of technical surgical skills. Euro. J. Surg. 168, 139–144 (2002).
8. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. Int. Conf. in Artif. Neural Netw. (ICANN) 3697, 799–804 (2005).
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9, 1735–1780 (1997).
10. Li, Z., Huang, Y., Cai, M., and Sato, Y.: Manipulation-skill assessment from videos with spatial attention network. arXiv preprint arXiv:1901.02579 (2019).
11. Martin, J., Regehr, G., Reznick, R.et al.: Objective structured assessment of technical skill (osats) for surgical residents. Br. J. Surg. 84, 273–278 (1997).
12. Peters, B.S., Armijo, P.R., Krause C. et al.: Review of emerging surgical robotic technology. Surg. Endosc. 32(4):1636–1655 (2018).
13. Wang, Z., Fey A. I.: SATR-DL: Improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. IEEE Conf. Eng. Med. Biol. Soc. 1793-1796 (2018).
14. Wang, Z., Fey, A.M.: Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int. J. Comput. Assist. Radiol. Surg. 13, 1959–1970 (2018).
15. Zia, A., Essa, I.: Automated surgical skill assessment in RMIS training. Int. J. Comput. Assist. Radiol. Surg. 13, 731–739 (2018).