Kinematics Data Representations for Skills Assessment in Ultrasound-Guided Needle Insertion

Robert Liu and Matthew S. Holden

[1] School of Computer Science, Carleton University, Ottawa, Canada
matthew.holden@carleton.ca

Abstract. Ultrasound-guided needle insertion is a difficult skill to learn and, in the context of competency-based medical education, requires continual monitoring of trainees' performance. This work investigates two standard neural network architectures, temporal convolutional networks and long short-term memory networks, for automated classification of skill level based on kinematics data. It examines which data representations are optimal for skills assessment using the proposed architectures in low data scenarios. The data representation had significant effect on the computed results. But given the optimal data representation, the proposed architectures achieve skills classification on two simulated ultrasound-guided needle insertion tasks with better performance than summary statistics. Thus, neural networks can be an effective tool for skills assessment in ultrasound-guided interventions; however, it is recommended to search over the space of data representations when limited data is available.

Keywords: surgical skills assessment, machine learning, ultrasound-guided interventions

1 Introduction

1.1 Motivation

Ultrasound-guided interventions are difficult to learn because the operator must simultaneously manipulate the ultrasound probe and one or more instruments, interpret the noisy ultrasound image in a rotated frame of reference, and guide the instrument to the target location. Mastering this skill takes considerable practice. This requires continual skills assessment and monitoring of learning curves, to ensure that trainees achieve a minimum level of proficiency prior to graduating to the next phase of training or practice.

As an alternative to direct observation of procedural skills or video-based skills assessment, automated skills assessment has been a growing field of study. This has held true for both surgeries and ultrasound-guided interventions, in particular. Automated skills assessment reduces time commitment and costs of human preceptors, improves standardization of assessment, and increases scalability of assessment.

Recently, approaches for skills assessment have been undergoing a shift from using classical machine learning to using deep neural networks, following the ongoing trend

of machine learning. While this obviates the need to extract domain-specific features, it has effectively replaced the problem with architecture engineering. The issue of architecture engineering is especially crucial in low-data scenarios. In particular, the way in which input data is represented becomes important.

The objective of this work is to identify the optimal representations for kinematics data for skills assessment in ultrasound-guided needle insertions. This work primarily seeks to understand which representations have the greatest added value when coupled with standard sequence modeling approaches (i.e. temporal convolution networks and long short-term memory networks). To our knowledge this is the first work that has used deep neural networks to assess skills in interventional ultrasound from kinematics data.

1.2 Previous Work

Traditionally, computer-assisted training for ultrasound-guided interventions has used performance metrics or summary statistics for skills assessment [1, 2]. This involves six degree-of-freedom tracking of the hands or instrument and the anatomy. From these trackers, performance metrics or summary statistics may be computed based upon clinically relevant quantities.

Previous work in skills assessment in ultrasound-guided interventions has explored spinal anesthesia [3, 4], peripheral nerve blockade [5], lumbar puncture [6], central venous catheterization [7], and generics targeting tasks [8, 9]. This prior work has focused on skills assessment in ultrasound-guided interventions in simulation-based training environments, rather than clinical environments [10].

Current work on skills assessment from kinematics or motion data uses modern deep neural network architectures. Indeed, most approaches using deep neural networks for skills assessment use temporal convolutional networks (TCN) or long short-term memory networks (LSTM). Wang et al. [11] proposed a TCN for skills assessment in robot-assisted minimally invasive surgery (RAMIS). Fawaz et al. [12] proposed a variant on a TCN by grouping channels into clusters for RAMIS; Castro et al. [13] proposed a variant on a TCN which uses quaternion convolution for RAMIS. Recent work from Kim et al. [14] has demonstrated the utility of standard TCNs in skills assessment from tool tip motion using different representations in a cataract surgery dataset. Ogul et al. [15] have investigated an LSTM architecture for pairwise ranking of surgical skills in an activity. Nguyen et al. [16] have demonstrated that a combined CNN+LSTM architecture with squeeze and excitation blocks accurately identifies skill in opens surgery using IMU data.

Many of these previous works have focused primarily on RAMIS and used the open-source JIGSAWS dataset [17]. These works report high performance, with accuracies, F1-scores, and AUCs up to and exceeding 0.90. Although, these previous works have all used a leave-one-trial-out validation protocol, rather than a leave-one-out-user protocol to determine how well their methods works on previously unseen operators.

2 Methods

2.1 Dataset

We use the ultrasound-guided needle insertion dataset originally presentation by Xia et al. [9] (**Error! Reference source not found.**). This dataset contains ultrasound-guided vascular access performed on a silicone model in a simulation-based training environment. There are 8 in-plane procedures and 7 out-of-plane procedures performed by 5 expert participants. There are 120 in-plane procedures performed by 20 novices and 114 out-of-plane procedures performed by 19 novices. Novices were medical students with no prior ultrasound-guided needle insertion experience; experts were attending emergency medicine physicians.

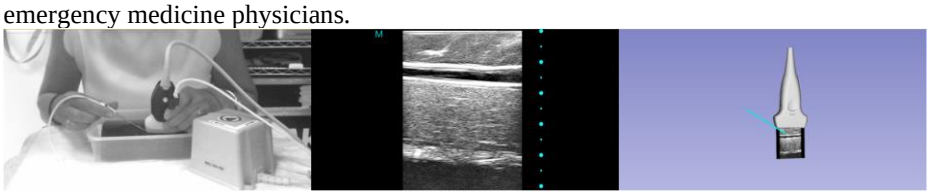

Fig. 1. Photograph, ultrasound image, and 3D visualization of participant performing ultrasound-guided needle insertion with electromagnetic trackers attached.

Kinematic data is collected through electromagnetic pose sensors (Ascension trakStar with Model 800 sensors, Northern Digital Inc., Waterloo, ON) rigidly attached to the needle, ultrasound probe, and phantom model (**Error! Reference source not found.**). Needle calibration was performed to find the needle tip's position using pivot and spin calibrations. Ultrasound calibration was performed to find the image's pose using the point-based method. Further details on the dataset may be found in the paper by Xia et al. [9].

2.2 Data Representations & Augmentation

We investigate skills assessment using four different transforms: (1) the Needle to Reference transform, (2) the Probe to Reference transform, (3) the Needle to Probe transform, and (4) the Needle Tip to Image transform. The former three transforms provide information about the needle's motion relative to the anatomy, the probe's motion relative to the anatomy, and the needle's motion relative to the probe, respectively. The latter transform, computed using the needle and probe calibration, provides information on how well the needle is visualized in ultrasound.

We investigate on four standard ways to represent the rotation of a transform. We consider (1) the Euler angle representation, (2) the axis-angle or rotation vector representation, (3) the quaternion representation, and (4) the rotation matrix representation. The Euler angle representation is a three-element vector, where we use the ZYX intrinsic rotation convention. The axis-angle or rotation vector representation is a three-element vector, where the vector's magnitude is the angle of rotation in radians. The quaternion representation is a four-element vector. The rotation matrix representation is a

six-element vector, where we use the flattened first two columns of the rotation matrix [18].

We also investigate the value of the translation and rotation information. We simulate a scenario where we only have access to the translation information (e.g. single marker infrared tracker). We simulate a scenario where we only have access to the rotation information (e.g. inertial measurement unit). This is compared to using both the translation and rotation information.

Due to the small dataset size, we employ window slicing to cut each procedure into overlapping 60 frame snippets (approximately 8 seconds). Subsequently, we randomly resample snippets to ensure an equal number of novice and expert snippets (i.e. 10000 snippets for each class). We approximate that each snippet is representative of the whole trial, and thus, focus on classifying snippets of fixed length.

2.3 Assessment Methods

As a testbed for our data representations, we use two neural network architectures that are common across prior work on skills assessment from kinematic data: TCN and LSTM. In both architectures, we employed aggressive dropout (p=0.80) and L2 regularization (λ=0.01) to prevent overfitting on our dataset; each model was trained for 100 epochs. These values were determined empirically on a small validation set.

The TCN architecture we use treats the input time series data as a signal where each dimension (i.e. translation components and rotation components) is treated as a channel. Convolution is performed one-dimensionally across time. We employ two convolutional layers, followed by an average pooling layer, and several dense layers. The network's architecture is illustrated in **Error! Reference source not found.**.

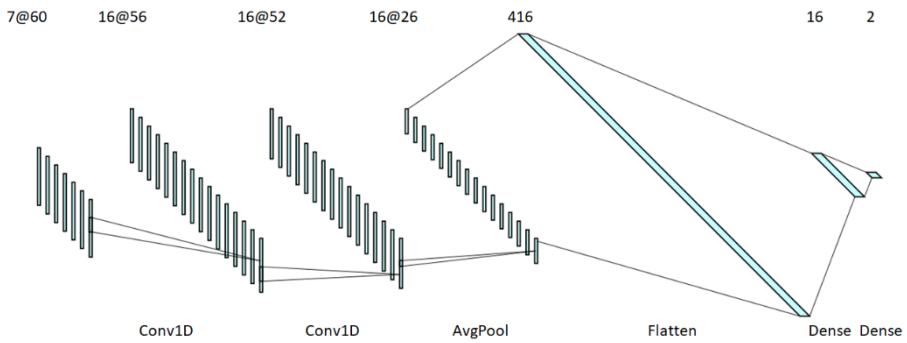

Fig. 2. Illustration of TCN architecture used for skills assessment. Top values indicate data size at each layer; bottom values indicate layer type. Input size is indicated as "number of channels @ number of frames" for each layer.

The LSTM architecture we use treats each translation component and rotation component as a feature. Recurrence is employed over time. We employ one LSTM layer, where the hidden output is used as the input to several dense layers. The network's architecture is shown in **Error! Reference source not found.**.

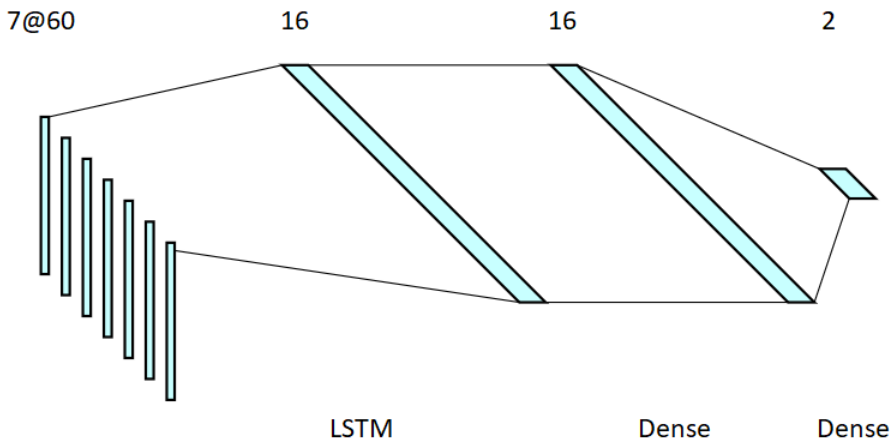

Fig. 3. Illustration of LSTM architecture used for skills assessment. Top values indicate data size at each layer; bottom values indicate layer type. Input size is indicated as "number of channels @ number of frames" for each layer.

2.4 Experimental Setup and Evaluation

To measure the performance of the assessment methods under different data representations, we performed user-out five-fold cross-validation for binary classification of novices vs. experts. We iterated over all five folds using each fold as the testing set, and randomly chose one fold from the training set as a validation set. Given there are only five expert users in the dataset, each fold contained data from exactly one expert. This cross-validation scheme best evaluates the proposed representations' generalizability to previously unseen users. Because there is class imbalance between the datasets, we report area under the curve (AUC) as the primary measure of performance. Networks for in-plane and out-of-plane data were trained separately.

3 Results

Results demonstrate that the highest performing data representation using the TCN is the NeedleTipToImage translation for in-plane insertions, with AUC of 0.83; the highest performing data representation is NeedleToReference translation for out-of-plane insertions, with AUC of 0.98. The highest performing data representation using the LSTM is the NeedleToProbe rotation matrix and translation for in-plane insertions, with AUC of 0.83; the highest performing data representation is ProbeToReference rotation matrix for out-of-plane insertions, with AUC of 0.70. Full results for all representations are reported using the TCN architecture (**Error! Reference source not found.**) and the LSTM architecture (**Error! Reference source not found.**).

By paired t-test, we found the skills assessment performance using the TCN was significantly better than using the LSTM for both in-plane ($p=0.03$; mean 0.62 vs. 0.55, respectively) and out-of-plane ($p<0.01$; mean 0.73 vs. 0.54, respectively) interventions.

By ANOVA we did not find a significant difference in skills assessment performance using any specific representation (mean AUC 0.62 for Euler & translation, 0.62 for axis-angle & translation, 0.61 for quaternion & translation, 0.64 for matrix & translation, 0.60 for Euler only, 0.58 for axis-angle only, 0.54 for quaternion only, 0.65 for matrix only, 0.64 for translation only). Nor was there any significant difference for rotation and translation, rotation only, and translation only (mean AUC 0.62, 0.59, 0.64, respectively). Likewise, we did not find a significant difference in skills assessment performance using the NeedleToReference, ProbeToReference, NeedleToProbe, or NeedleTipToImage transforms (mean AUC 0.62, 0.62, 0.61, 0.59, respectively).

Table 1. Area under the curve for skills classification using TCN architecture. For each cell, the top and bottom values indication performance for in-plane and out-of-plane insertions. Bolded results indicate best performance for each approach; underlined results indicate best performance for each transform per approach.

Rep.	Rotation Rep.	Transform			
		NeedleTo Reference	ProbeTo Reference	NeedleTo Probe	NeedleTip ToImage
Rotation + Translation	Euler	0.53 ± 0.13 0.77 ± 0.12	0.64 ± 0.15 0.74 ± 0.10	0.64 ± 0.13 0.83 ± 0.08	0.68 ± 0.09 0.81 ± 0.13
	Axis-Angle	0.71 ± 0.07 0.94 ± 0.02	0.76 ± 0.07 0.60 ± 0.11	0.43 ± 0.08 0.56 ± 0.12	0.70 ± 0.07 0.86 ± 0.11
	Quaternion	0.40 ± 0.14 0.49 ± 0.12	0.65 ± 0.14 0.78 ± 0.12	0.73 ± 0.12 0.76 ± 0.16	0.64 ± 0.13 0.73 ± 0.09
	Matrix	0.70 ± 0.13 0.76 ± 0.11	0.56 ± 0.14 0.73 ± 0.15	0.78 ± 0.10 0.50 ± 0.14	0.60 ± 0.08 0.69 ± 0.18
Rotation	Euler	0.66 ± 0.17 0.77 ± 0.07	0.51 ± 0.12 0.74 ± 0.10	0.43 ± 0.12 0.70 ± 0.10	0.66 ± 0.10 0.69 ± 0.09
	Axis-Angle	0.72 ± 0.07 0.59 ± 0.09	0.82 ± 0.05 0.43 ± 0.14	0.35 ± 0.10 0.73 ± 0.10	0.30 ± 0.09 0.82 ± 0.09
	Quaternion	0.51 ± 0.17 0.53 ± 0.13	0.55 ± 0.17 0.60 ± 0.11	0.69 ± 0.13 0.81 ± 0.10	0.55 ± 0.05 0.59 ± 0.17
	Matrix	0.79 ± 0.08 0.64 ± 0.11	0.74 ± 0.09 0.89 ± 0.06	0.60 ± 0.11 0.91 ± 0.05	0.44 ± 0.07 0.82 ± 0.12
Translation	-	0.56 ± 0.11 **0.98 ± 0.01**	0.64 ± 0.18 0.82 ± 0.09	0.81 ± 0.10 0.79 ± 0.13	**0.83 ± 0.07** 0.75 ± 0.12

Table 2. Area under the curve for skills classification using LSTM architecture. For each cell, the top and bottom values indication performance for in-plane and out-of-plane insertions. Bolded results indicate best performance for each approach; underlined results indicate best performance for each transform per approach.

Rep.	Rotation Rep.	Transform			
		NeedleTo	ProbeTo	NeedleTo	NeedleTip

		Reference	Reference	Probe	ToImage
Rotation + Translation	Euler	0.62 ± 0.14 0.42 ± 0.10	0.64 ± 0.09 0.40 ± 0.13	0.59 ± 0.15 0.54 ± 0.15	0.55 ± 0.10 0.48 ± 0.13
	Axis-Angle	0.61 ± 0.06 0.60 ± 0.04	0.51 ± 0.13 0.48 ± 0.06	0.69 ± 0.07 0.60 ± 0.05	0.48 ± 0.12 0.35 ± 0.16
	Quaternion	0.60 ± 0.04 0.61 ± 0.14	0.59 ± 0.11 0.67 ± 0.08	0.52 ± 0.09 0.56 ± 0.12	0.38 ± 0.11 0.58 ± 0.12
	Matrix	0.65 ± 0.07 0.62 ± 0.13	0.65 ± 0.09 0.68 ± 0.12	**0.83 ± 0.03** 0.51 ± 0.10	0.41 ± 0.05 0.63 ± 0.13
Rotation	Euler	0.58 ± 0.12 0.66 ± 0.06	0.58 ± 0.08 0.65 ± 0.17	0.41 ± 0.15 0.64 ± 0.07	0.45 ± 0.07 0.42 ± 0.14
	Axis-Angle	0.53 ± 0.09 0.63 ± 0.06	0.56 ± 0.15 0.56 ± 0.09	0.58 ± 0.10 0.55 ± 0.14	0.58 ± 0.07 0.52 ± 0.15
	Quaternion	0.53 ± 0.10 0.41 ± 0.12	0.59 ± 0.09 0.48 ± 0.09	0.48 ± 0.12 0.34 ± 0.12	0.35 ± 0.09 0.66 ± 0.09
	Matrix	0.60 ± 0.07 0.56 ± 0.12	0.60 ± 0.13 **0.70 ± 0.05**	0.50 ± 0.08 0.49 ± 0.15	0.64 ± 0.11 0.50 ± 0.08
Translation	-	0.60 ± 0.12 0.52 ± 0.03	0.49 ± 0.07 0.45 ± 0.13	0.47 ± 0.06 0.56 ± 0.15	0.53 ± 0.08 0.43 ± 0.12

4 Discussion

Performance varies considerably due to representation, and the optimal representation varies depending on network architecture and task. Thus, we do not recommend an optimal data representation in general, but rather, searching over the space of data representations to find the optimal representation for a given task. This is crucial in tasks with small datasets, where deep neural networks do not generalize well.

The results show the proposed TCN outperforms the proposed LSTM across most data representations, although their results are not directly comparable due to different numbers of parameters used. Furthermore, assessment performance for the out-of-plane dataset exceeds assessment performance for the in-plane dataset.

This study has several limitations. Primarily, the sample size is small; thus, the uncertainty associated with the calculated measures of performance are high. This study also is limited by using operators' self-proclaimed appointment status (i.e. emergency medicine physicians or medical student). Ideally the ground-truth skill level should be determined by an expert rater using a scale with evidence of validity. This would account for variation in performance due to extraneous factors. Finally, the dataset comes from a simulation-based training situation, and therefore may not be directly applicable to a skills assessment in a clinical scenario.

The results shown here are consistent with the work from Kim et al. [14] which demonstrate that data representation of kinematics data for skills assessment important. Their work identifies that using tooltip velocities has benefit over using tooltip positions. The performance achieved for binary skills classification exceeds the results from Holden et al. [19] (0.83 vs. 0.82 AUC for in-plane insertions; 0.98 vs. 0.94 AUC for out-of-plane insertions), when the appropriate data representation is used. We highlight,

however, that the numeric results reported in this comparison are not indicative of performance on a new dataset. These numeric results should be seen more like performance on a validation set, as the representation has been tuned on the testing set.

In many applications, it is desirable to use less expensive and less obtrusive hardware for data collection. For example, using an inertial measurement unit to determine rotation information or using a single marker optical tracker to determine translation information reduce the hardware needs. Results indicate that these setups may be possible in some cases without affecting skills assessment performance.

Future work involves using further data augmentation strategies (e.g. white noise, window warping, etc.) and representation strategies (e.g. multiple transforms with multiple representations) to determine their added value in skills assessment in scenarios with limited data. We also intend to collect a larger dataset so that we may get a more precise measure of performance for different representations. Finally, we wish to try more modern neural network architectures.

5 Conclusion

In this work, we have demonstrated that standard neural network architectures can be used for skills assessment in ultrasound-guided needle insertion. The proposed TCN outperforms previous work using machine learning on summary statistics for skills assessment [19], when the appropriate data representation is used. On the other hand, this work has found that data representation can significantly affect skills assessment performance. We found that the optimal representation can vary depending on the task and network architecture. Thus, it is recommended to search over the space of data representations when choosing a neural network architecture for skills assessment in interventional ultrasound when limited data is available.

6 Acknowledgement

This research was enabled in part by support provided by Compute Ontario (www.computeontario.ca) and Compute Canada (www.computecanada.ca).

References

1. Reiley, C.E., Lin, H.C., Yuh, D.D., Hager, G.D.: Review of methods for objective surgical skill evaluation. Surg. Endosc. 25, 356–366 (2011). https://doi.org/10.1007/s00464-010-1190-z.
2. Holden, M.S.: Computer-Assisted Assessment and Feedback for Image-Guided Interventions Training, (2019).
3. Hayter, M.A., Friedman, Z., Bould, M.D., Hanlon, J.G., Katznelson, R., Borges, B., Naik, V.N.: Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement. Can. J. Anesth. 56, 419–426 (2009). https://doi.org/10.1007/s12630-009-9090-1.

4. Corvetto, M.A., Fuentes, C., Araneda, A., Achurra, P., Miranda, P., Viviani, P., Altermatt, F.R.: Validation of the imperial college surgical assessment device for spinal anesthesia. BMC Anesthesiol. 17, (2017). https://doi.org/10.1186/s12871-017-0422-3.
5. Chin, K.J., Tse, C., Chan, V., Tan, J.S., Lupu, C.M., Hayter, M.: Hand motion analysis using the imperial college surgical assessment device: validation of a novel and objective performance measure in ultrasound-guided peripheral nerve blockade. Reg. Anesth. Pain Med. 36, 213–219 (2011). https://doi.org/10.1097/AAP.0b013e31820d4305.
6. Clinkard, D., Moult, E., Holden, M., Davison, C., Ungi, T., Fichtinger, G., McGraw, R.: Assessment of Lumbar Puncture Skill in Experts and Nonexperts Using Checklists and Quantitative Tracking of Needle Trajectories: Implications for Competency-Based Medical Education. Teach. Learn. Med. 27, 51–56 (2015). https://doi.org/10.1080/10401334.2014.979184.
7. Clinkard, D., Holden, M., Ungi, T., Messenger, D., Davison, C., Fichtinger, G., McGraw, R.: The development and validation of hand motion analysis to evaluate competency in central line catheterization. Acad. Emerg. Med. 22, 212–218 (2015). https://doi.org/10.1111/acem.12590.
8. Tabriz, D.M., Street, M., Pilgram, T.K., Duncan, J.R.: Objective assessment of operator performance during ultrasound-guided procedures. Int. J. Comput. Assist. Radiol. Surg. 6, 641–652 (2011). https://doi.org/10.1007/s11548-010-0541-5.
9. Xia, S., Keri, Z., Holden, M.S., Hisey, R., Lia, H., Ungi, T., Mitchell, C.H., Fichtinger, G.: A learning curve analysis of ultrasound-guided in-plane and out-of-plane vascular access training with Perk Tutor. In: Webster, R.J. and Fei, B. (eds.) Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling. p. 66. SPIE (2018). https://doi.org/10.1117/12.2293789.
10. Vedula, S.S., Ishii, M., Hager, G.D.: Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. Annu. Rev. Biomed. Eng. 19, 301–325 (2017). https://doi.org/10.1146/annurev-bioeng-071516-044435.
11. Wang, Z., Majewicz Fey, A.: Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int. J. Comput. Assist. Radiol. Surg. (2018). https://doi.org/10.1007/s11548-018-1860-1.
12. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. (2019). https://doi.org/10.1007/s11548-019-02039-4.
13. Castro, D., Pereira, D., Zanchettin, C., MacEdo, D., Bezerra, B.L.D.: Towards Optimizing Convolutional Neural Networks for Robotic Surgery Skill Evaluation. In: Proceedings of the International Joint Conference on Neural Networks (2019). https://doi.org/10.1109/IJCNN.2019.8852341.
14. Kim, T.S., O'Brien, M., Zafar, S., Hager, G.D., Sikder, S., Vedula, S.S.: Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int. J. Comput. Assist. Radiol. Surg. (2019). https://doi.org/10.1007/s11548-019-01956-8.
15. Oğul, B.B., Gilgien, M.F., Şahin, P.D.: Ranking robot-assisted surgery skills using kinematic sensors. In: Lecture Notes in Computer Science (including subseries Lecture

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2019). https://doi.org/10.1007/978-3-030-34255-5_24.

16. Nguyen, X.A., Ljuhar, D., Pacilli, M., Nataraja, R.M., Chauhan, S.: Surgical skill levels: Classification and analysis using deep neural network model and motion signals. Comput. Methods Programs Biomed. (2019). https://doi.org/10.1016/j.cmpb.2019.05.008.
17. Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D.: A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. IEEE Trans. Biomed. Eng. 64, 2025–2041 (2017). https://doi.org/10.1109/TBME.2016.2647680.
18. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2019). https://doi.org/10.1109/CVPR.2019.00589.
19. Holden, M.S., Lia, H., Xia, S., Keri, Z., Ungi, T., Fichtinger, G.: Configurable Overall Skill Assessment in Ultrasound-Guided Needle Insertion. In: 16th Annual Imaging Network Ontario Symposium (ImNO) (2018).