Skills Classification in Cardiac Ultrasound with Temporal Convolution and Domain Knowledge Using a Low-Cost Probe Tracker

Matthew S. Holden [a], Alberto Portillo, Gerard Salame [b]

[a] School of Computer Science, Carleton University, Ottawa, Canada

[b] Denver Health Hospital, Denver, USA

Correspondence Address:

Matthew S. Holden

School of Computer Science

Carleton University

1125 Colonel By Drive

Ottawa, Ontario, Canada

K1S 5B6

Email: matthew.holden@carleton.ca

Telephone: +1-613-520-2600 x3244

ABSTRACT

As point of care ultrasound (POCUS) becomes more integrated into clinical practice, it is essential to address all aspects of ultrasound operator proficiency. Ultrasound proficiency requires the ability to acquire, interpret, and integrate bedside ultrasound images. The difference in image acquisition psychomotor skills between novice (trainee) and expert (instructor) ultrasonographer has not been described. We created an inexpensive system, called Probe Watch, to record probe motion and assess image acquisition in cardiac POCUS using an inertial measurement device and software for data recording based on open-source components. We designed a temporal convolutional network for skills classification from probe motion that integrates clinical domain knowledge. We further designed data augmentation methods to improve its generalization. Subsequently, we validated the setup and assessment method on a set of novice and expert sonographers performing cardiac ultrasound in a simulation-based training environment. The proposed methods classified participants as novice or expert with AUC 0.931 and 0.761 for snippets and trials, respectively. Integrating domain knowledge into the neural network had added value. Furthermore, we identified the most discriminative features for assessment. Probe Watch quantifies motion during cardiac ultrasound and provides insight into probe motion behavior. It may be deployed during cardiac ultrasound training to monitor learning curves objectively and automatically.

KEYWORDS

machine learning, skills assessment, cardiac ultrasound

INTRODUCTION

The advent of affordable and portable hand-held ultrasound devices has expedited the integration of point of care ultrasound (POCUS) into clinical practice. Many residency programs and medical schools have implemented this new technology into their curriculum (Schnobrich et al. 2013). Competency in POCUS requires that users are able to: obtain and optimize the appropriate ultrasound images and use them to answer a specific clinical question (Kumar et al. 2019). One important application for POCUS has been bedside cardiac ultrasonography which has been shown to enhance the detection of let ventricular systolic failure (Marbach et al. 2019). However, image acquisition in cardiac ultrasonography requires extensive hands-on training as demonstrated by Upadharsta et al. which showed that medical interns could only achieve a fair agreement for evaluating left ventricular systolic function when compared to trained attendings after several weeks of ultrasound education (Upadhrasta et al. 2019). In a recent statement in the Journal of Hospital Medicine, Soni et al. recommended creating high quality video portfolios with expert feedback (ideally given during image acquisition), with a minimum number of images based on the acquired view (Soni et al. 2019). This approach, however, does not account for variability in trainee learning curves and does not offer an objective method to assess image acquisition skill. Currently, ultrasound educators mostly rely on subjective assessments of a trainee's skill. Nielsen et al. noted that with this approach only 67% of the total observed score variance between different judges could be ascribed to differences in physician performance with an error variance exceeding 10% (Nielsen et al. 2015).

Understanding the probe movement behavior of novice and expert point of care ultrasonographers can help guide training and assess proficiency in image acquisition. To do so, we have created a probe movement tracking device aptly named Probe Watch, that maps and records various aspects of probe movement during bedside cardiac ultrasonography.

Prior Work

Over the last several decades, there has been considerable prior work on objective skills assessment from motion data (Reiley et al. 2011; Vedula et al. 2017). In particular, there has been notable prior work in point-of-care ultrasound for both diagnostic and interventional purposes (Holden 2019).

Traditionally, automated skills assessment from motion or kinematic data has been based on performance metrics or summary statistics that are derived based on clinical experts' understanding of the procedure. Summary statistics typically measure efficiency, errors, or outcomes. More recent approaches, however, also take advantage of highly informative abstract features, such as signal frequencies, symbolic representations, or entropy-based features. Time-series-based analysis approaches have also been used with success. Standard approaches include building class-wise Hidden Markov Models and using Dynamic Time Warping for alignment and comparison.

Contemporary approaches to skills assessment from motion leverage modern deep neural networks. While several works use long short-term memory (LSTM) networks for skills assessment (Nguyen et al. 2019; Oğul et al. 2019), most works used temporal convolutional networks (TCN) to assess skill. Wang et al., Fawaz et al., and Castro et al. have proposed variants on the standard TCN for skills assessment in robot-assisted minimally invasive surgery (Castro et al. 2019; Ismail Fawaz et al. 2019; Wang and Majewicz Fey 2018). These papers report skills classification accuracies exceeding 90% on the JIGSAWS dataset (Ahmidi et al. 2017). Kim et al. showed excellent performance for TCNs in skills classification from tooltip motion in cataract surgery (Kim et al. 2019). Most relevantly, Nguyen et al. proposed an approach for skills assessment from inertial measurement unit (IMU) data, but in the context of open surgery (Nguyen et al. 2019). Their architecture combines a TCN branch with an LSTM branch, and achieves results exceeding 95% accuracy for skill classification. Furthermore, recent work from Liu and Holden has shown that TCNs

outperform LSTMs for skills classification from instrument motion data in interventional ultrasound (Liu and Holden 2020).

Several prior works have addressed skills assessment from hand or probe motion data in diagnostic ultrasound, focusing mainly on FAST ultrasound. Work from Ziesmann et al. and Zago et al. have demonstrated that efficiency-based summary statistics (i.e. time, hand or probe path lengths, number of hand or probe motions, and working volume) differ between experts and novices, and can be used for skills assessment in FAST ultrasound (Zago et al. 2019; Ziesmann et al. 2015). Bell et al. showed additionally that an outcome-based summary statistic (i.e. percentage of target points scanned) can also be incorporated into skills assessment for diagnostic ultrasound (Bell et al. 2017), and secondary analysis has shown its added value in skills assessment (Holden et al. 2017).

To the best of our knowledge, no prior work has addressed image acquisition skills assessment from motion data in cardiac ultrasound.

Objective

This work develops a low-cost and easy-to-use sensor device and data collection software to record and analyze ultrasound probe motion during bedside cardiac ultrasound. We leverage modern machine learning approaches to analyze probe motion data with the goal of assessing operators' skill in image acquisition. Our approach is validated on clinical trainees in a simulation-based training environment.

MATERIALS AND METHODS

Skills Classification Methods

Summary Statistics

In consultation with clinical experts in cardiac ultrasound, we identified a set of summary statistics that are believed to be indicative of proficiency. These summary statistics incorporate clinical domain knowledge.

We used the following summary statistics: total elapsed time (in seconds), total acceleration (in gravitational units), total rotation (in radians per second), translational movements (unitless), rotational movements (unitless), and motion smoothness (in gravitational units per second). Total acceleration and total rotation were calculated as the arc length of the acceleration and rotational velocity vectors over time. Translational and rotational movements were defined as the number of distinct periods of time where the magnitude of the acceleration and rotational velocity, respectively, exceeded a threshold value (McGraw et al. 2016). The threshold values were calibrated empirically to yield results consistent with experts' interpretation of meaningful actions in cardiac ultrasound. Motion smoothness was computed as the root-mean-square translational jerk (Stylopoulos et al. 2004). Total acceleration, total rotation, translational actions, and rotational actions were computed over each axis independently and over all axes. In total, this yields 18 summary statistics that are believed to correlate with skill in image acquisition.

We use the following definitions for probe translation and rotation described by Bahner et al. (Bahner et al. 2016): x-axis and z-axis translation is sliding, y-axis translation is pressure (perpendicular motion to the patient's chest), x-axis rotation is tilting, y-axis rotation is rotation, z-axis rotation is rocking (Figure 1).

Network Architecture

We propose a neural network architecture that incorporates both raw kinematic time series data and the summary statistics derived from consultation with experts. We conjecture that incorporating expert-defined summary statistics will improve the network's ability to generalize in low-data situations, and we conjecture that the raw time series data will allow the network to learn aspects of motion that are captured by the summary statistics. We choose a simple architecture with fewer parameters to prevent overfitting the training data.

The network's architecture comprises a temporal convolutional component to map time series data to a feature vector and a feedforward component to map the feature vector and summary statistic data to a skill

label (Figure 2). We appended the relative timestamps and the view ID as additional channels to the time series data. We use aggressive dropout ($p=0.80$) and regularization ($\lambda=10^{-5}$) throughout the network to prevent overfitting.

The TCN component of the architecture treats the time series data as a signal where each dimension is treated like a separate channel in the temporal convolution, and convolution is performed one-dimensionally across time for each channel independently. We employ two convolutional layers, where each is followed by a batch normalization layer, ReLU activation, and a max pooling layer. An average pooling layer is employed as the last layer of this component.

The multi-layer perceptron component of the architecture takes the flattened output from the TCN component and the summary statistics as input. These values are concatenated. We employ two dense linear layers, where each is followed by a batch normalization layer, and the first is followed by ReLU activation.

We trained the network over 25 epochs using binary cross-entropy loss with a slow learning rate (10^{-4}) and batch size 64. We used the validation set to identify the epoch yielding the highest validation AUC (i.e. early stopping) provided the training AUC exceeded 0.85. We normalized the raw accelerometer and gyroscope data per channel for each sequence individually; we normalized the summary statistics across all samples in the training set. During training we employed significant data augmentation to improve generalization despite the small dataset.

During testing, we used the model that achieved the best validation AUC on the validation set. The raw accelerometer and gyroscope data were normalized per channel for each sequence individually; the summary statistics were normalized according to the mean and standard deviations from the training set. No data augmentation was used during testing.

Data Augmentation

Due to the small dataset size and sensor setup, we used data augmentation to facilitate generalization of our model. Data augmentation is intended to simulate transformations of the data that are expected to be present in our dataset but do not indicate a change in skill level. We employed four data augmentation methods: window slicing, window warping, white noise, and random rotation. Data augmentation was applied to the accelerometer and gyroscope data only.

We randomly resampled 60 frame snippets from the time series data. Snippets were potentially overlapping. We chose to resample the same number of snippets from each skill level to get a balanced dataset. We chose large enough number of snippets (10000 per class) to ensure full coverage. This is based on the insight that sufficiently long snippets are indicative of skill, despite not capturing the entire procedure.

We temporally warped each snippet by uniformly resampling timestamps between frames. We linearly interpolated the data between timestamps. This simulates instances where the procedure is performed at varying speeds.

We added white noise to each snippet with zero mean and fixed standard deviation. White noise was independent across time. This simulates small errors in the readings from the accelerometer and gyroscope, within their error tolerances.

We randomly rotated snippets by a fixed rotation across time. We calculated a random rotation by sampling a rotation axis from a 3D standard normal distribution and an angle of rotation from a 1D zero mean Gaussian. This simulates small rotational inconsistencies in where the Probe Watch is mounted to the ultrasound probe.

Probe Watch System Setup

We developed a low-cost tracking device, called Probe Watch, to monitor ultrasound probe motion during cardiac ultrasound. The hardware consists of a 9 degree-of-freedom inertial measurement unit (IMU)

with accelerometer, magnetometer, and gyroscope components, to record the 3 degree-of-freedom translational acceleration (in gravitational units), magnetic field (in microtesla), and rotational velocity (in radians per second), respectively. It uses TinyShield sensors and a TinyDuino Processor Board (www.tinyciruits.com). The hardware is contained within a 34 mm diameter by 23 mm thick cylindrical housing (Figure 3). The device connects via Micro USB cable and communicates by serial port. The device attaches to any ultrasound probe using an elastic strap or sticker and is placed 70 mm from the foot of the probe.

Serial data from the hardware was captured using the PLUS Toolkit (www.plustoolkit.org) (Lasso et al. 2014), and was sent to 3D Slicer (www.slicer.org) via the OpenIGTLink network protocol. We developed a data recording, storage, and analysis interface within 3D Slicer (Figure 4), based on the SlicerIGT (www.slicerigt.org) (Ungi et al. 2016) and Perk Tutor (www.perktutor.org) (Ungi et al. 2012) platforms.

The Probe Watch software facilitates user authentication, and it displays the appropriate interface depending on which group the user belongs to. The software collects and records data from the sensors, and it synchronizes with a CouchDB database (couchdb.apache.org) for storage and retrieval. The data may be visualized by graphing each component of the sensor information or through visual representation of the probe's rotation. Finally, the software is responsible for computation of expert-defined summary statistics based on the sensor data.

Study Population

Volunteers were recruited during two separate ultrasound workshops: a cardiac ultrasound workshop organized by the university of Colorado for its internal medicine residents and during the Society of Hospital Medicine's (SHM) Colorado Chapter ultrasound training workshop. Volunteers were divided into two groups based on their ultrasound expertise: trainees (novice ultrasonographers) and experts (instructors from the SHM yearly ultrasound training workshop).

The SHM Colorado Chapter ultrasound focuses on the training of medical providers with various levels of ultrasound expertise in the art of point of care ultrasonography by incorporating a combination of hands-on training and didactic lectures. Topics include basic cardiac, lung, vascular and abdominal imaging among others. Participants of this two-day pre-course are assigned into groups with a three trainee to instructor ratio at each hands-on station. Each group follows a predetermined sequence of hands-on stations spending an average of 20 minutes in each station. Each station maintained one healthy volunteer on which trainees practiced image acquisition. The University of Colorado Internal Medicine residency organized a half day workshop for its internal medicine house staff, focused on bedside cardiac ultrasonography with a similar format as the one mentioned previously. At both workshops, during the course introductory lecture, attendees were given a brief description of the study in power point format, with instructions to obtain the best possible image for each view based on the didactics given during the course. Participants amenable to join the study were assigned a subject identification number.

In each workshop one hands-on station was dedicated for data acquisition and manned by the study investigator. Participants rotated through this station and were blinded to each other's performance. Volunteers were asked to acquire four basic cardiac images based on the recommendations from the point-of-care ultrasound certification process set forth by SHM (Soni et al. 2019): Parasternal Long Axis, Parasternal Short Axis (at the level of the mitral valve papillary muscle insertions into the left Ventricle), Apical 4 Chamber and Subxyphoid views to the best of their abilities. Participants used the Butterfly iQ ultrasound system (Butterfly Network Inc.; www.butterflynetwork.com). The probe movements needed to achieve the final image for each view were recorded (Figure 5).

This study was determined to be exempt from IRB approval by the Colorado Multiple Institute Review Board (COMIRB). All components of this study involving human participants were performed in accordance with the institution's ethical standards. All participants provided informed consent to participate in this study.

Validation Protocol

To evaluate the performance of the proposed model for classifying experts versus novices, we used trial-out k-fold cross-validation (i.e. each trial appeared in exactly one fold). We used five folds, where each fold was iteratively chosen to be in the test set, and one remaining fold was randomly chosen to be in the validation set. This was repeated 10 times with different randomly chosen folds each time. This validation scheme evaluates how well the method generalizes to a new procedure.

Due to the imbalance between the novice and expert datasets, we report the area under the curve (AUC) as the primary measure of performance. We report this measure for both each snippet (i.e. continuous segment from a trial) individually, and for each trial (i.e. single ultrasound scan of a single view) by taking a mean over all its snippets' results.

Ablation Study

We performed an ablation study to determine the added value of each component of our approach. In particular, we tested the following conditions: the full network ("Full"), the full network without data augmentation ("NoAug"), the network using just the raw time series data with no summary statistics nor data augmentation ("RawData"), and the network using just the summary statistics as input with no raw times series data nor data augmentation ("SummStat"). We used the same validation protocol, cross-validation folds, and measures of performance to evaluate each condition.

Feature Importance

To determine how important each feature (i.e. raw data component or summary statistics) is for skills assessment in image acquisition, we performed a feature importance study (Molnar 2019). For each feature during test time, we replaced the feature value for each instance with a randomly resampled feature value from another instance. This was repeated 10 times for each instance. We report the absolute change in the

snippet test AUC as a measure of how important each feature is for skills assessment in image acquisition. We used the same validation protocol and cross-validation folds to evaluate the importance of each feature.

Due to the inability of the proposed feature importance study to model feature interactions, we also compute the distance correlation between all features (Székely et al. 2007). For summary statistics, we compute the distance as the absolute value of the difference. For time series data, we compute the distance as the root-mean-square distance.

Furthermore, as evidence that the proposed summary statistics are indicative of image acquisition skill, we compute: (1) the rank correlation between each summary statistic and skill and (2) the area under the receiver operating characteristic (AUC) for skills classification using each summary statistic individually.

RESULTS

In total, after removing corrupted recordings, we captured 63 instances from four experts and seven novices. 24 recordings came from experts; 39 recordings came from novices. There were 20 parasternal long axis recordings, 13 parasternal short axis recordings, 12 apical four chamber recordings, 12 subxyphoid recordings, 5 inferior vena cava recordings, and 1 recording of unknown view.

By Shapiro-Wilk test, data was found to be non-normally distributed. Thus, we report non-parametric statistics.

Main Results

Overall, for classifying novices and experts, the proposed methods achieved a median test AUC of 0.931 for snippets, and a median test AUC of 0.761 for trials over all folds. This contrasts to the reported validation AUCs of 0.902 and 0.789 over all folds, respectively. Full results are reported in Table 1.

Ablation Study

By Kruskal-Wallis test, we found there to be a significant difference across ablation conditions for snippet test AUC. Posthoc pairwise comparisons via Dunn's test revealed that results achieved using the time series data without the summary statistics nor data augmentation (i.e. "RawData") were significantly lower than results under all other conditions. No significant differences were found for trial test AUC. Descriptive statistics indicate added value for the full approach using raw time series data, summary statistics, and data augmentation (Table 2). These differences, however, were not found to be significant.

Feature Importance

The feature with the largest computed feature importance was rotational actions in the y-axis. The most important time series component was the relative timestamp; the most important summary statistic was rotational actions in the y-axis. By ranksum test, we found that summary statistics had significantly greater feature importance than raw time series data ($p < 0.001$). Full feature importance results are reported in Table 3. Distance correlations between features are illustrated in Figure 6. Rank correlation and AUC for skill class versus summary statistics are given in Table 4; visualizations of the five summary statistics with the greatest AUC are provided in Figure 7.

DISCUSSION

Using open-source software and affordable components, we were able to create a device that objectively classifies ultrasound operators into novice and expert groups. The proposed assessment method performs skills classification for image acquisition on motion snippets with median AUC of 0.931.

Traditionally, automated assessment methods have been based upon expert-defined summary statistics; however, most modern approaches use deep neural networks for assessment. Our approach combines these. The raw time series data includes information not present in the summary statistics, while the summary

statistics allow for better generalization when the training set is small. The data augmentation further facilitates generalization. The ablation study demonstrates that this approach has added value. Furthermore, the feature importance study demonstrates which summary statistics are most important, providing insight for educational coaching settings. Although the approach does not capture complex feature interactions.

Although prior literature has put forth reproducible and validated scores to assess the quality of the final image acquired (Skinner et al. 2016), there is a paucity of data to describe the image acquisition process itself. Ultrasound educators rely mostly on subjective methods to assess a trainee's proficiency in image acquisition which, as stated earlier, is fraught with errors. Other researchers have relied on one aspect of the acquisition process such as time to image acquisition as a marker of proficiency (Smith et al. 2018). In contrast, Probe Watch offers insight into the various translational and rotational movements that make up an operator's probe behavior and using that data to categorize their level of expertise. For example, based on the feature importance results, rotation along the y axis which represents the rotation motion in POCUS, was the most predictive in distinguishing novice versus expert sonographers suggesting that novices tend to rotate their probe much more than experts. In addition, the importance of total acceleration suggests that novices had more erratic movements than their expert counterparts. These insights can help ultrasound educators address and anticipate common errors among their trainees and identify patterns of probe behavior among students to create more individualized feedback.

Although other studies have assessed hand movement of ultrasound operators to assess image acquisition skill (Bell et al. 2017; Zago et al. 2019; Ziesmann et al. 2015), these methods use expensive and cumbersome technologies that are not readily available to ultrasound educators. Our hardware costs approximately $225 USD to build, and it uses commercially available accelerometers and gyroscopes in a compact format that is placed directly onto the ultrasound probe. Furthermore, our software is built upon free, open-source components. Having access to this technology and using it in conjunction with image

quality assessments and objective structured clinical skills exams would allow evaluation of a trainee's ability to: (1) obtain and optimize the image and (2) identify and integrate their findings into their clinical assessment. Thereby addressing all the facets of POCUS proficiency discussed earlier.

While this work has demonstrated the effectiveness of the proposed methods in cardiac ultrasound, they could be readily extended to image acquisition assessment in other diagnostic ultrasound applications. This would only require modification of the summary statistics according to application-specific domain knowledge.

Limitations

Our study was done in a structured educational environment using healthy volunteer patients. This is a common approach to ultrasound training in these educational settings and is done to: (1) familiarize trainees to normal cardiac anatomy and (2) help trainees feel successful and promote further POCUS practice in their daily clinical work. It is unclear how the assessment performance of Probe Watch would change in a clinical setting given patient specific variables such as body habitus, underlying clinical illness, and comorbidities that can influence the process of image acquisition. In addition, our study focused on a dichotomous categorization of ultrasound operators into either novices or experts based solely on appointment status, without further subcategorization of skill level. Furthermore, our sample size of 11 operators and 63 total instances of image acquisition is small, and it is unclear how well the results will apply to a larger population.

We trained the assessment method to maximize loss on training snippets. We compute the trial-wise measures of performance as a mean over all snippets within the trial. As a result, we find greater snippet-wise performance metrics than trial-wise performance metrics. Using an appropriate consensus layer to compute trial-wise assessment and backpropagating over it would improve trial-wise performance (Funke et al. 2019).

Future Work

More research is needed to understand the ability of the proposed approach to monitor changes in skill level during ultrasound training and if it can be used to help educators guide their trainees. This would facilitate its use in competency-based medical education programs.

The focus of this work was on skills assessment in the image acquisition component of cardiac ultrasound. We also propose integrating complementary approaches to automated skills assessment. While our approach focuses on assessment from motion data, other works have shown automated assessment is feasible from ultrasound image data (Abdi et al. 2017; Mazomenos et al. 2018) or video data (Tyrrell and Holden 2020). These combined approaches may provide more robust and more wholistic skills assessment, without requiring additional hardware or expense.

CONCLUSIONS

We have developed a low-cost tool, Probe Watch, that is able to quantify the motion associated with image acquisition during cardiac ultrasound. This allows educators insight into a trainee's probe motion behavior. The proposed approach combines summary statistics, raw time series data, and data augmentation to classify skill level in image acquisition. Despite the inexpensive setup and small dataset size, the approach accurately classified image acquisition skill level in a simulation-based training scenario. Probe Watch may be easily deployed in medical education settings and will allow changes in trainees' image acquisition skill level to be tracked over time and create more personalized instructions for students.

ACKNOWLEDGMENTS

This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada grant RGPIN-2020-05582.

CONFLICT OF INTEREST

All authors declare that they have no conflict of interest.

REFERENCES

Abdi AH, Luong C, Tsang T, Allan G, Nouranian S, Jue J, Hawley D, Fleming S, Gin K, Swift J, Rohling R, Abolmaesumi P. Automatic Quality Assessment of Echocardiograms Using Convolutional Neural Networks: Feasibility on the Apical Four-Chamber View. IEEE Trans Med Imaging 2017;

Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, Hager GD. A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. IEEE Trans Biomed Eng 2017;64:2025–2041.

Bahner DP, Blickendorf JM, Bockbrader M, Adkins E, Vira A, Boulger C, Panchal AR. Language of Transducer Manipulation: Codifying Terms for Effective Teaching. J Ultrasound Med 2016;

Bell CR, McKaigney CJ, Holden M, Fichtinger G, Rang L. Sonographic Accuracy as a Novel Tool for Point-of-care Ultrasound Competency Assessment. Uijtdehaage S, ed. AEM Educ Train 2017;1:316–324.

Castro D, Pereira D, Zanchettin C, MacEdo D, Bezerra BLD. Towards Optimizing Convolutional Neural Networks for Robotic Surgery Skill Evaluation. Proc Int Jt Conf Neural Networks 2019.

Funke I, Mees ST, Weitz J, Speidel S. Video-based surgical skill assessment using 3D convolutional neural networks. Int J Comput Assist Radiol Surg Springer, 2019;14:1217–1225.

Holden MS. Computer-Assisted Assessment and Feedback for Image-Guided Interventions Training. ProQuest Diss. Theses. Queen's University, Kingston, Canada, 2019.

Holden MS, Keri Z, Ungi T, Fichtinger G. Overall Proficiency Assessment in Point-of-Care Ultrasound Interventions: The Stopwatch is not Enough. Imaging Patient-Customized Simulations Syst Point-of-Care Ultrasound Int Work BIVPCS 2017 POCUS 2017 Cham: Springer International Publishing, 2017. pp. 146–153.

Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. Int J Comput Assist Radiol Surg 2019;

Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int J Comput Assist Radiol Surg 2019;

Kumar A, Kugler J, Jensen T. Evaluation of Trainee Competency with Point-of-Care Ultrasonography (POCUS): a Conceptual Framework and Review of Existing Assessments. J. Gen. Intern. Med. 2019.

Lasso A, Heffter T, Rankin A, Pinter C, Ungi T, Fichtinger G. PLUS: Open-source toolkit for ultrasound-guided intervention systems. IEEE Trans Biomed Eng 2014;61:2527–2537.

Liu R, Holden MS. Kinematics Data Representations for Skills Assessment in Ultrasound-Guided Needle Insertion. Med Ultrasound, Preterm, Perinat Paediatr Image Anal Springer, 2020. pp. 189–198.

Marbach JA, Almufleh A, Di Santo P, Jung R, Simard T, McInnes M, Salameh JP, McGrath TA, Millington SJ, Diemer G, West FM, Domecq MC, Hibbert B. Comparative accuracy of focused cardiac ultrasonography and clinical examination for left ventricular dysfunction and valvular heart disease. Ann. Intern. Med. 2019.

Mazomenos EB, Bansal K, Martin B, Smith A, Wright S, Stoyanov D. Automated Performance Assessment

in Transoesophageal Echocardiography with Convolutional Neural Networks. Int Conf Med Image Comput Comput Interv 2018. pp. 256–264.

McGraw R, Chaplin T, McKaigney C, Rang L, Jaeger M, Redfearn D, Davison C, Ungi T, Holden M, Yeo C, Keri Z, Fichtinger G. Development and Evaluation of a Simulation-based Curriculum for Ultrasound-guided Central Venous Catheterization. CJEM 2016;1–9.

Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Book 2019;

Nguyen XA, Ljuhar D, Pacilli M, Nataraja RM, Chauhan S. Surgical skill levels: Classification and analysis using deep neural network model and motion signals. Comput Methods Programs Biomed 2019;

Nielsen DG, Jensen SL, O'Neill L. Clinical assessment of transthoracic echocardiography skills: A generalizability study. BMC Med Educ 2015;

Oğul BB, Gilgien MF, Şahin PD. Ranking robot-assisted surgery skills using kinematic sensors. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2019.

Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. Surg Endosc Springer-Verlag, 2011;25:356–366.

Schnobrich DJ, Gladding S, Olson APJ, Duran-Nelson A. Point-of-Care Ultrasound in Internal Medicine: A National Survey of Educational Leadership. J Grad Med Educ 2013;

Skinner AA, Freeman R V, Sheehan FH. Quantitative feedback facilitates acquisition of skills in focused cardiac ultrasound. Simul Healthc LWW, 2016;11:134–138.

Smith CJ, Morad A, Balwanz C, Lyden E, Matthias T. Prospective evaluation of cardiac ultrasound

performance by general internal medicine physicians during a 6-month faculty development curriculum. Crit Ultrasound J 2018;

Soni NJ, Schnobrich D, Matthews BK, Tierny DM, Jensen TP, Dancel R, Cho J, Dversdal RK, Mints G, Bhagra A, Reierson K, Kurian LM, Liu GY, Candotti C, Boesch B, LoPresti CM, Lenchus J, Wong T, Johnson G, Maw AM, Franco-Sadud R, Lucas BP. Point-of-Care Ultrasound for Hospitalists: A Position Statement of the Society of Hospital Medicine. J Hosp Med 2019;

Stylopoulos N, Cotin S, Maithel SKK, Ottensmeye M, Jackson PGG, Bardsley RSS, Neumann PFF, Rattner DWW, Dawson SLL, Ottensmeyer M, Jackson PGG, Bardsley RSS, Neumann PFF, Rattner DWW, Dawson SLL. Computer-enhanced laparoscopic training system (CELTS): bridging the gap. Surg Endosc 2004;18:782–789.

Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat 2007;

Tyrrell RE, Holden MS. Ultrasound video analysis for skill level assessment in FAST ultrasound. Comput Methods Biomech Biomed Eng Imaging Vis Taylor & Francis, 2020;1–5.

Ungi T, Lasso A, Fichtinger G. Open-source platforms for navigated image-guided interventions. Med. Image Anal. 2016.

Ungi T, Sargent D, Moult E, Lasso A, Pinter C, McGraw RC, Fichtinger G. Perk Tutor: An Open-Source Training Platform for Ultrasound-Guided Needle Insertions. IEEE Trans Biomed Eng 2012;59:3475–3481.

Upadhrasta S, Raafat MH, Conti RAS. Reliability of focused cardiac ultrasound performed by first-year

internal medicine residents at a community hospital after a short training. J community Hosp Intern Med Perspect Taylor & Francis, 2019;9:373–376.

Vedula SS, Ishii M, Hager GD. Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. Annu Rev Biomed Eng Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, California 94303-0139, USA, 2017;19:301–325.

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;

Zago M, Sforza C, Mariani D, Marconi M, Biloslavo A, Greca A La, Kurihara H, Casamassima A, Bozzo S, Caputo F, Galli M, Zago M. Educational impact of hand motion analysis in the evaluation of fast examination skills. Eur J Trauma Emerg Surg 2019;

Ziesmann MT, Park J, Unger B, Kirkpatrick AW, Vergis A, Pham C, Kirschner D, Logestty S, Gillman LM. Validation of hand motion analysis as an objective assessment tool for the Focused Assessment with Sonography for Trauma examination. J Trauma Acute Care Surg 2015;79:631–637.

FIGURE CAPTIONS LIST

Figure 1. Definition of translational (top) and rotational (bottom) probe motion along each axis.

Figure 2. Neural network architecture for binary classification of skill from raw time series data and summary statistics. Input raw time series data is indicated in blue; input summary statistics are indicated in orange. Temporal convolutional component is indicated in grey; fully connected component is indicated in green. Input size for each layer is indicated as number of channels @ number of frames.

Figure 3. Photograph (left) and diagram (right) of Probe Watch hardware mounted to an ultrasound probe.

Figure 4. Probe Watch software interface depicting recording (top) and assessment (bottom) functionality. The left frame provides the recording and analysis controls; middle frame provides a graphical display of acceleration and rotational velocity; the right frame provides a 3D visualization of rotation.

Figure 5. Photograph of participant scanning a volunteer patient with Probe Watch attached to the ultrasound probe.

Figure 6. Distance correlation between all input raw time series data and summary statistics.

Figure 7. Rank correlation between individual summary statistics and ground-truth skill level for the summary statistics with the five largest areas under the receiver operating characteristic. Novices are indicated in blue; experts are indicated in red.

TABLES

Table 1. Performance measures for classification of experts and novices. Reported as median (inter-quartile range).

Snippet Validation AUC	0.902 (0.722 – 0.983)
Trial Validation AUC	0.789 (0.675 – 0.889)
Snippet Test AUC	0.931 (0.776 – 0.971)
Trial Test AUC	0.761 (0.667 – 0.855)

Table 2. Performance measures under different conditions of ablation for classification of experts and novices. Reported as median (inter-quartile range). Bold indicates condition with best performance.

Condition	Snippet Test AUC	Trial Test AUC
Full	**0.931** **(0.776 – 0.971)**	**0.761** **(0.667 – 0.855)**
NoAug	0.899 (0.793 – 0.965)	0.722 (0.667 – 0.854)
RawData	0.784 (0.715 – 0.848)	0.725 (0.667 – 0.850)
SummStat	0.907 (0.774 – 0.957)	0.745 (0.606 – 0.839)

Table 3. Feature importance for each of the features used in skills classification. Blue indicates summary statistics; yellow indicates raw time series data. Reported as median (inter-quartile range).

Feature	Snippet Test AUC
Rotational Actions Y	7.70E-02 (2.67E-02 - 1.32E-01)
Total Acceleration Z	5.97E-02 (2.37E-02 - 2.40E-01)
Total Acceleration	3.30E-02 (5.84E-03 - 1.27E-01)
Elapsed Time	3.19E-02 (8.78E-03 - 8.45E-02)
Translational Actions X	2.93E-02 (1.37E-02 - 6.64E-02)
Total Acceleration X	2.74E-02 (1.11E-02 - 7.67E-02)
Motion Smoothness	2.65E-02 (6.73E-03 - 5.48E-02)
Translational Actions Z	2.62E-02 (6.15E-03 - 5.42E-02)
Translational Actions Y	2.61E-02 (9.76E-03 - 6.29E-02)
Total Rotation Y	2.51E-02 (1.05E-02 - 7.31E-02)
Rotational Actions X	2.40E-02 (9.05E-03 - 6.50E-02)
Total Rotation Z	1.47E-02 (5.99E-03 - 4.61E-02)
Total Acceleration Y	1.43E-02 (3.41E-03 - 2.39E-02)
Total Rotation X	1.38E-02 (5.25E-03 - 3.38E-02)
Rotational Actions	1.35E-02 (3.67E-03 - 4.11E-02)
Rotational Actions Z	1.30E-02 (2.94E-03 - 3.68E-02)
Total Rotation	1.08E-02 (2.87E-03 - 3.13E-02)
Relative Timestamp	2.77E-03 (7.12E-05 - 7.70E-03)
Region ID	2.26E-04 (0.00E+00 - 4.67E-03)
Accelerometer X	2.15E-06 (8.33E-17 - 7.19E-04)

Accelerometer Z	1.40E-06	(0.00E+00 - 1.02E-03)
Gyroscope Y	9.34E-08	(0.00E+00 - 9.99E-05)
Accelerometer Y	1.11E-16	(0.00E+00 - 2.38E-04)
Gyroscope X	1.11E-16	(0.00E+00 - 1.43E-04)
Gyroscope Z	1.11E-16	(0.00E+00 - 1.84E-04)
Translational Actions	0.00E+00	(0.00E+00 - 0.00E+00)

Table 4. Rank correlation and area under the curve (AUC) for individual summary statistics versus ground-truth skill level.

Summary Statistic	Rank Correlation	AUC
Elapsed Time	-0.26	0.66
Total Acceleration	-0.44	0.77
Total Rotation	-0.40	0.75
Total Acceleration X	-0.47	0.79
Total Acceleration Y	-0.31	0.69
Total Acceleration Z	-0.39	0.74
Total Rotation X	-0.30	0.68
Total Rotation Y	-0.46	0.79
Total Rotation Z	-0.41	0.76
Translational Actions	0.00	0.50
Translational Actions X	-0.32	0.69
Translational Actions Y	0.08	0.45
Translational Actions Z	-0.29	0.65

Motion Smoothness	-0.35	0.72
Rotational Actions	-0.34	0.71
Rotational Actions X	-0.16	0.60
Rotational Actions Y	-0.35	0.72
Rotational Actions Z	-0.26	0.66

1