Calibration Tolerance in Ultrasound-Guided Simulation Training

Jason Au, Matthew S. Holden
Carleton University, Ottawa, Canada

Introduction: Objective assessment of surgical skill using sensor-derived motion data is increasingly common in simulation and training [1]. In image-guided tasks like ultrasound-guided needle insertion, tracking systems require calibrations for needle and probe to synchronize instruments and imaging. We evaluated whether simulated calibration perturbations meaningfully degrade machine learning (ML) skill score prediction.

Methods: We analyzed an ultrasound-guided needle insertion dataset inspired by Xia et al. [2], consisting of 38 novices and 5 experts performing in-plane (IP) and out-of-plane (OOP) insertions. Performance was rated using an OSATS-lik rubric (score range 5 to 25). Trials were segmented into overlapping sequences of 60 consecutive frames (sliding windows) to capture short-term motion dynamics. We simulated seven calibration-error scenarios by perturbing the static transformations (ReferenceToVessel, NeedleTipToNeedle, or ImageToProbe) with rotational and translational noise. These include domain-shift conditions from phantom-scale to human-scale setups (1500–2720 mm translation, 0–360° rotation), full-phantom perturbations (50–150 mm, 0–360° rotation), and small random miscalibration drifts (10 mm; 0–15°). We compared a temporal deep model (TCN) to ROCKET (RandOm Convolutional KErnel Transform), a time-series feature extraction method using randomized convolutional kernels [3], using user-out 5-fold cross-validation. Performance was measured using mean absolute error (MAE) in rating unit and root mean squared error (RMSE). Lower MAE indicates closer agreement between predicted and expert-assigned skill scores.

Results: On unperturbed data (control), ROCKET outperformed TCN across all metrics. In OOP trials, ROCKET achieved trial MAE ≈ 1.9–2.1, comparable to inter-expert variability (MAE = 1.95). Across 20 runs per experiment, perturbations did not degrade performance and occasionally improved MAE relative to control (negative effect size).

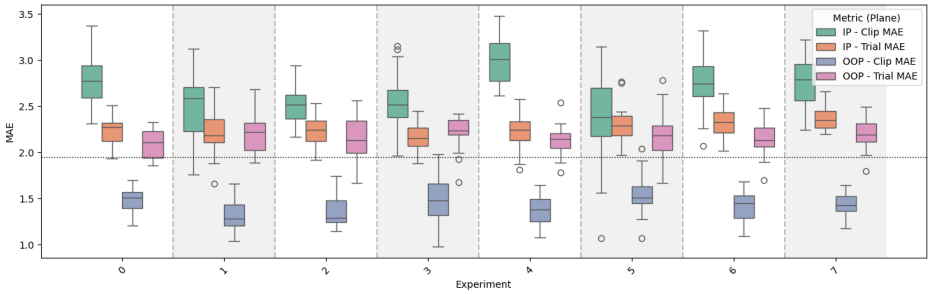

Figure 1: Box-and-whisker plot of MAE values across perturbation experiments (n = 20 runs per condition).

Conclusions: Contrary to the expectation that frequent recalibration is essential, ML-based skill score prediction remained stable under substantial simulated calibration errors. The results suggest models may rely more on relative kinematic patterns than on absolute pose in a global coordinate frame, and that perturbations may act as implicit regularization. These findings support reduced deployment time for sensor-based assessment systems. Future work will evaluate real miscalibrations.

Acknowledgements: Supported in part by NSERC (RGPIN-2020-05582). Computational resources were provided by Compute Ontario and the Digital Research Alliance of Canada.

References

[1] D. Kulik, C. R. Bell, and M. S. Holden, "Fast skill assessment from kinematics data using convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–7, 2023.

[2] S. Xia, Z. Keri, M. S. Holden, R. Hisey, H. Lia, T. Ungi, C. H. Mitchell, and G. Fichtinger, "A learning curve analysis of ultrasound-guided in-plane and out-of-plane vascular access training with perk tutor," in *Proc. SPIE Medical Imaging*, vol. 10576, p. 1057625, 2018.

[3] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1454–1495, 2020.