Feasibility of Video-Based Skills Assessment: A Study on Ultrasound-Guided Needle Insertions Using Simulated Projections

Raymond Weiming Chan*[a], Rebecca Hisey[b], Matthew S. Holden[a]
[a]School of Computer Science, Carleton University, 1125 Colonel By Drive, Ottawa, ON Canada K1S 5B6; [b]Laboratory for Percutaneous Surgery, School of Computing, Queen's University, 99 University Ave, Kingston, ON Canada K7L 3N6

ABSTRACT

PURPOSE: Automated skills assessment of ultrasound-guided needle insertions has previously been explored through 3D motion tracking data. The purpose of this study was to determine the viability of 2D motion tracking data in distinguishing between novice and expert subjects. METHODS: Perspective projection was applied to needle and ultrasound probe time series data. The resulting time series data of 2D points were used to calculate various performance metrics. Using these metrics, classifications between novice and expert were performed by random forest. This procedure was repeated with different camera positions all pointing at the reference point to examine systematically the effect of camera position on assessment. RESULTS: For in-plane needle insertions, mean AUC obtained through 3D data and mean AUC obtained through 2D data were well-matched (0.68 vs. 0.69). For out-of-plane insertions, mean AUC values from 3D and 2D data were more distant (0.86 vs. 0.77), but AUC from the optimal camera angle matched up well (0.85). CONCLUSION: 2D data is comparable to 3D data when used to perform skills assessment of ultrasound-guided needle insertions, and camera placement level with the instruments is optimal. We conclude that videos of needle insertions may be feasible for skills assessment.

Keywords: surgical skills assessment, ultrasound-guided interventions

1. INTRODUCTION

The skill level of a surgeon can be determined by cognitive capabilities, judgement and decision-making, and manual dexterity[1]. Unlike the other two categories, manual dexterity cannot be tested in traditional written and oral examinations. Skill level in technical interventions is difficult to quantify, and standardized classification has yet to be deployed for any intervention. A time-based model – one that evaluates trainees after training for a fixed amount of time – is widely used in the education of technical interventions. Time-based models are gradually being abandoned in favour of competency-based models due to its learner-centredness[2]. A competency-based approach requires trainees to practice until they are deemed competent. The drawback to this approach is that it requires continual monitoring of the trainee's skill progression. Using human experts to perform continual one-on-one monitoring of trainees is time and resource consuming. On top of that, the usage of human supervision introduces subjectivity to the skills assessment process. Research by Reznick and MacRae[3] suggests that efficient training and robust skills assessment of medical interventions is highly beneficial to healthcare. This warrants the use of computer-assisted skills assessment. A standardized and automated evaluation process should increase efficiency and quality in education of technical medical interventions.

Instrument tracking in ultrasound-guided needle insertion procedures is typically done with 3D motion trackers (e.g. electromagnetic or optical trackers). Holden *et al.* have shown that automated skills assessment of ultrasound-guided needle insertions can be done with high accuracy using data gathered from an electromagnetic pose tracker[4]. The methods used were based upon decision trees and fuzzy inference systems, so they were transparent and configurable. Additionally, they were able to provide useful feedback to the subjects. A tracker-based approach was also used in skills assessment of cardiac ultrasound[5]. Combining domain knowledge and temporal convolutional network allowed data from this tracking system to classify subjects correctly. The Imperial College Surgical Assessment Device, which uses an electromagnetic tracking system, has been validated as a reliable skills assessment tool for procedures such as labour epidural placement, spinal anesthesia, and ultrasound-guided needle insertions[6–8]. The proposed models could consistently distinguish between operators of different skill level. Tracker-based skills assessment methods have seen success in ultrasound-guided

*raymondweimingchan@cmail.carleton.ca

interventions and other procedures, but it does come with limitations.

Using motion electromagnetic trackers may not be ideal in a surgical environment. The accuracy of magnetic trackers is greatly affected by nearby metal objects and electronic equipment, so magnetic field distortion is almost completely unavoidable[9]. Reiley *et al*. suggest that data collection systems should be minimal and nonobstructive[1]. Electromagnetic pose tracking systems require an active tracker attached to the base of the ultrasound probe and the base of the needle, which may affect the ease of use of the equipment.

As an alternative to obtrusive electromagnetic or optical trackers, video-based tracking may be used as a convenient and inexpensive technology for skills assessment. There have been many studies that successfully automate skills assessment in accordance with the Objective Structured Assessment of Technical Skills grading system using video data[10–12]. Zia *et al.* required participants to wear colored fingerless gloves[10], but using motion texture analysis[12] or computer vision[11] eliminates that requirement. O'Driscoll *et al.* had considerable success with automated tool identification and performance monitoring in central venous catheterization using webcam[13]. Similar success has also been achieved in assessment of laparoscopic skills using two webcams placed in orthogonal configuration[14]. This study used colored tapes to assist in image segmentation.

Studies like the ones mentioned above demonstrate skills assessment with video data using different techniques. They show that video recording can be reliable in skills assessment of its respective procedures, so it is reasonable to believe that similar results can be achieved for ultrasound-guided needle insertions. Furthermore, to our knowledge, most of the studies like these do not compare results from different data collection methods.

The purpose of this study is to determine the feasibility of using 2D instrument tracking data compared to using 3D data from dedicated tracker for skills assessment of ultrasound-guided needle insertions. Time series of 2D data from simulated projections imitate the data that would be gathered from video recordings of needle insertions. The accuracy of predictions with our 2D data will demonstrate the usefulness of video data with the assumption that object detection is perfect. Determining the viability of video data is important for assessing the future accessibility of computer-assisted skills assessment.

2. METHODS

Our 3D dataset consists of time series data of homogeneous transformation matrices that contain the rotation and translation of the needle and the ultrasound probe relative to a reference sensor attached to a vascular access phantom[15]. To simulate 2D data collected from video-based tracking, two easily identifiable points on the needle and four easily identifiable points on the probe were projected into 2D. This allows us to infer some information about rotation of the tools in 2D. The needle tip and the needle base were used as points. Points used on the ultrasound probe are shown in Figure 1.

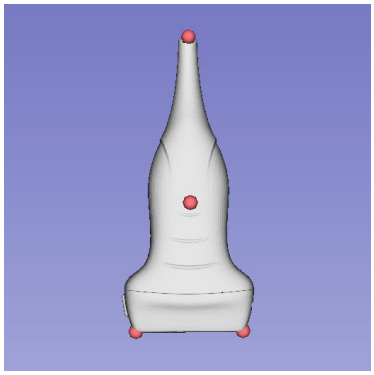


Figure 1. Ultrasound probe model (grey) and points used (red)

To project the two points on the needle and four points on the probe onto a simulated video camera image, the transform hierarchies described in equations (1) and (2) were used, respectively. The coordinates of the points, in reference to the

model of their respective instrument, were defined. $T_{ProbeModel \to ProbeSensor}$ and $T_{NeedleTip \to NeedleSensor}$ were computed by fiducial registration and pivot calibration respectively. $T_{ProbeSensor \to Reference}$ and $T_{NeedleSensor \to Reference}$ were provided by the tracking system. $T_{Reference \to Camera}$ was varied to determine the viability of different camera positions.

$$\begin{pmatrix} x_{Camera} \\ y_{Camera} \\ z_{Camera} \\ 1 \end{pmatrix} = T_{Reference \to Camera} \quad T_{ProbeSensor \to Reference} \quad T_{ProbeModel \to ProbeSensor} \quad \vec{p}_{ProbeModel} \tag{1}$$

$$\begin{pmatrix} x_{Camera} \\ y_{Camera} \\ z_{Camera} \\ 1 \end{pmatrix} = T_{Reference \to Camera} \quad T_{NeedleSensor \to Reference} \quad T_{NeedleTip \to NeedleSensor} \quad \vec{p}_{NeedleTip} \tag{2}$$

$T_{Reference \to Camera}$ was generated by placing 41 cameras 500 mm away from the reference sensor, arranged in a dome shape (Figure 2). In spherical coordinates with radial distance r, azimuthal angle θ, and polar angle φ, the cameras were placed on $\{(r, \theta, \varphi) \mid r = 500 \text{ mm}, \ \theta = \frac{a\pi}{5}, \ \varphi = \frac{b\pi}{10}\}$, where $a, b \in \mathbb{Z}$, $0 \leq a \leq 9$, and $0 \leq b \leq 4$. $T_{Reference \to Camera}$ instances have their z-axis pointing at the reference sensor, meaning the reference sensor will be in the center of the projected image.

(a) Top-down view (b) Lateral view

Figure 2. $T_{Reference \to Camera}$ instances (yellow) and example probe (grey) and needle (cyan) from an out-of-plane insertion.

Perspective transform was finally applied to the vector, with focal length f of the camera. A focal length of 50 mm was arbitrarily chosen for this study.

$$\begin{pmatrix} x_{Image} \\ y_{Image} \\ f \end{pmatrix} = \frac{f}{z_{Camera}} \begin{pmatrix} x_{Camera} \\ y_{Camera} \\ z_{Camera} \end{pmatrix} \tag{3}$$

Consistent with Xia *et al.*[15], for every trial, the following performance metrics were calculated for each of the six points: path length, motion smoothness, path inefficiency, root mean square. Other metrics that were computed for each trial were:

distance between the base of the probe to the tip of the needle, rotation between needle tip and needle base, rotation between the two points on the base of the probe, and elapsed time. We evaluated path length, motion smoothness, path inefficiency, and root mean square on the value they each provide to skills assessment by following the methods of exploratory factor analysis performed in the paper by Holden *et al.*[16] Both the 3D dataset and the 2D dataset were split by the approach style of the needle insertion (in-plane insertions and out-of-plane insertions) to be worked on separately.

Random forest was used to classify data as novice or expert using feature vectors of performance metrics, since Holden *et al.* have shown that random forests perform well on skills assessment of ultrasound-guided interventions[4]. Stratified user-out k-fold cross-validation was performed. Trials were grouped by subject, resulting in 19 novices and 5 experts for each approach. The dataset of each approach was divided into 5 folds, each containing 1 expert and either 3 or 4 novices. The folds were fed into the random forest using the method of cross-validation. Area under the ROC curve (AUC) was used as a measure of assessment performance. AUC was preferred over accuracy due to class imbalance in our dataset. Training the classifier for accuracy caused it to predict 'novice' almost exclusively. Feed-forward neural network with hyperparameter tuning has also been tested but was greatly outperformed by random forest. This may be because of the small dataset size (15 expert and 240 novice samples for both approaches combined).

Data augmentation was used to balance the classes of our training sets and validation sets and to increase our sample size for each approach within the 2D and 3D datasets from 121 to 10000. For each metric, Gaussian white noise was added with standard deviation values gathered from the corresponding metric and skill level. A factor of 0.1 was applied to the standard deviation values since the distributions of expert and novice samples had overlaps.

3D in-plane and 3D out-of-plane datasets were separately passed into a random forest through cross-validation as described above. This procedure was repeated 100 times, and mean and median AUC scores were calculated. For each camera angle, perspective projection and metrics calculations were applied to the 3D dataset to generate a 2D version. Each 2D dataset was passed into a random forest for cross-validation once.

3. RESULTS

3D in-plane and 3D out-of-plane datasets achieved 0.68 and 0.86 mean AUC respectively. 2D in-plane and out-of-plane datasets, generated from 41 different camera positions, achieved 0.69 and 0.77 mean AUC respectively. The highest AUC produced by the 2D in-plane and out-of-plane datasets were 0.83 and 0.85 respectively.

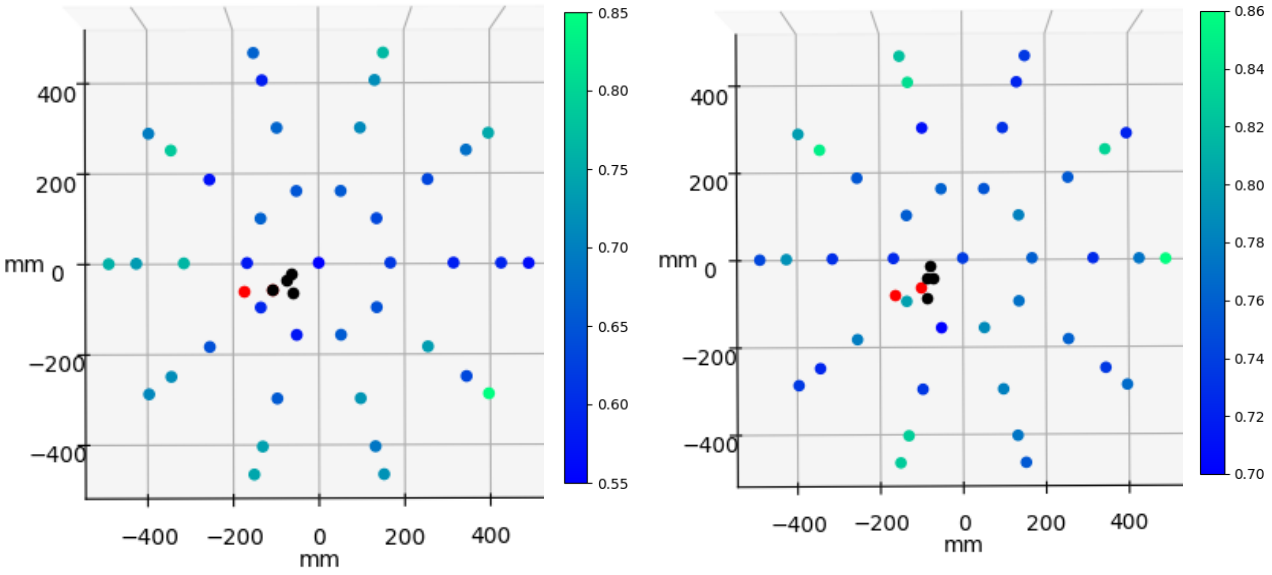
Table 1. AUC results from 3D dataset over 100 iterations of cross-validation

	Mean	Standard Deviation
In-plane	0.68	0.03
Out-of-plane	0.86	0.03

Table 2. AUC results from 2D dataset over 41 different camera positions, each with 1 iteration of cross-validation.

	Mean	Standard Deviation	Maximum	Minimum
In-plane	0.69	0.06	0.83	0.59
Out-of-plane	0.77	0.03	0.85	0.72

Figure 2. Top-down view. Cameras with low (blue) to high (green) AUC. Example probe points (black) and needle points (red).



(a) In-plane insertion (b) Out-of-plane insertion

Table 3. Mean AUC of cameras grouped by polar angle of the camera.

Polar angle (φ)	In-plane	Out-of-plane
0	0.59	0.75
$\pi/10$	0.65	0.77
$2\pi/10$	0.68	0.76
$3\pi/10$	0.70	0.79
$4\pi/10$	0.72	0.78

4. DISCUSSION

For in-plane insertions, mean AUC was similar for 2D and 3D data (Tables 1 and 2). For out-of-plane insertion, mean AUC from 3D data was higher than for 2D data. However, AUC scores were similar given optimal camera placement (Tables 1 and 2). For both insertion approaches, the best camera placements were level with the insertion target, but having a level camera placement did not guarantee high AUC, especially for in-plane insertions (Figure 2).

For in-plane insertions, as the camera moves from directly above the reference to level to the insertion target, AUC generally increases (Figure 3). This might mean that measurement of the tilt of instruments towards and away from the subject is more impactful than left and right. Also, camera angles that are on the needle side of the probe generally performed better than those on the opposite side. Our result from factor analysis showed that path length of the needle is the most important metric, so camera angles that give more amplified needle path lengths was expected to perform better. The mean AUC for in-plane insertions is well matched with mean AUC from the 3D dataset. This result, along with the maximum AUC of 0.83 (Table 2), tells us that following the general guidelines for camera placement and fine tuning the classifier allows our 2D data to perform just as well as 3D data.

Similar to in-plane insertions, AUC from out-of-plane insertions generally increase as the camera angle moves towards being level to the insertion target. For out-of-plane insertions, well-performing camera placements are clustered together easily identifiable from others in the graph (Figure 2.b). We can clearly see that camera angles that are laterally between the general angle of the needle and the general angle of the probe perform much better than others. This allows us to

confidently provide the guideline of placing the camera in the said general area. Due to this, the performance of out-of-plane insertions 2D data is better represented by the maximum AUC of 0.85 rather than the mean AUC of 0.77 (Table 2).

According to the exploratory factor analysis, path length of the probe, path length of the needle, and path efficiency of the probe had the most added value in skills assessment. Root mean square position of the probe had added value only for in-plane insertions. We observe that measuring multiple points on the same instrument did not have any value for skills assessment according to the exploratory factor analysis. On average, the non-parametric estimate of effect size of skill level on performance metrics is higher for out-of-plane insertions, which is consistent with the result of out-of-plane insertions having a higher mean AUC. Intuitively, classification of out-of-plane insertions should be easier since the distance between the needle tip and the ultrasound image plane is a good indicator of skill level.

The results suggest that using videos of needle insertions for skills assessment may be possible without reducing performance. This study, however, has several limitations: (1) it does not model occlusions of the needle or probe, (2) it does not account for image segmentation errors, and (3) it uses appointment status as ground-truth skill level.

In a real surgical assessment environment, there will always be some obstructions of the probe and the needle from the view of the camera. This is a major challenge of using a camera to gather positional data of the instruments. For in-plane insertions, there does not seem to be a specific section of camera placements that significantly outperforms others. This means that, in a real-world setup, it should be beneficial to prioritize the visibility of the instruments when choosing a camera placement.

Another difficulty that may arise when processing video data from a camera is the possibility of image segmentation errors. Depending on the transparency of the phantom, it may be especially difficult to track the needle accurately. For in-plane insertions, one can use Hough transform on the ultrasound image for needle recognition[17]. This can provide additional information on the needle that otherwise may not be captured on video.

Future work involves assessing the accuracy of these methods using point segmentation from video. This work has shown than video from camera has great potential in skills assessment of ultrasound-guided needle insertions, but the aforementioned limitations leave room for potential pitfalls. The next step is to attempt to use cameras to gather data and perform skills assessment. The establishment of ground-truth skill level by appointment status also leaves more to be desired. Determining ground-truth skill level by expert review would imitate the process of skills assessment that could happen in the real world. Another avenue for future work involves skills assessment directly from time series data using a temporal convolutional network. This might improve the results since calculating performance metrics takes away some information that might be useful to the classifiers. Aggregating samples from all camera angles could be another way to see classifiers interact with this dataset. Feeding aggregated data from all camera angles to classifiers will allow us to evaluate robustness to different camera angles and to obtain a better gauge of general accuracy.

5. CONCLUSION

Simulated 2D instrument tracking data is comparable to using 3D instrument tracking data when used to perform metrics-based skills assessment of ultrasound-guided needle insertions. This suggests that video-based tool tracking could be comparable to traditional instrument tracking for skills assessment. Placement of camera influences accuracy of assessment; we found camera placement level with the instruments was optimal.

ACKNOWLEDGEMENTS

Raymond Weiming Chan was supported by the Natural Sciences and Engineering Research Council of Canada Undergraduate Student Research Award. This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada grant RGPIN-2020-05582.

REFERENCES

[1] Reiley, C. E., Lin, H. C., Yuh, D. D. and Hager, G. D., "Review of methods for objective surgical skill evaluation," Surgical Endoscopy **25**(2) (2011).

[2] Frank, J. R., Snell, L. S., Cate, O. ten, Holmboe, E. S., Carraccio, C., Swing, S. R., Harris, P., Glasgow, N. J., Campbell, C., Dath, D., Harden, R. M., Iobst, W., Long, D. M., Mungroo, R., Richardson, D. L., Sherbino, J., Silver, I., Taber, S., Talbot, M., et al., "Competency-based medical education: Theory to practice," Medical Teacher **32**(8) (2010).

[3] Reznick, R. K. and MacRae, H., "Teaching Surgical Skills — Changes in the Wind," New England Journal of Medicine **355**(25) (2006).

[4] Holden, M. S., Xia, S., Lia, H., Keri, Z., Bell, C., Patterson, L., Ungi, T. and Fichtinger, G., "Machine learning methods for automated technical skills assessment with instructional feedback in ultrasound-guided interventions," International Journal of Computer Assisted Radiology and Surgery **14**(11) (2019).

[5] Holden, M. S., Portillo, A. and Salame, G., "Skills Classification in Cardiac Ultrasound with Temporal Convolution and Domain Knowledge Using a Low-Cost Probe Tracker," Ultrasound in Medicine and Biology **47**(10) (2021).

[6] Chin, K. J., Tse, C., Chan, V., Tan, J. S., Lupu, C. M. and 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," Regional Anesthesia and Pain Medicine **36**(3) (2011).

[7] Corvetto, M. A., Fuentes, C., Araneda, A., Achurra, P., Miranda, P., Viviani, P. and Altermatt, F. R., "Validation of the imperial college surgical assessment device for spinal anesthesia," BMC Anesthesiology **17**(1) (2017).

[8] Hayter, M. A., Friedman, Z., Bould, M. D., Hanlon, J. G., Katznelson, R., Borges, B. and Naik, V. N., "Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement," Canadian Journal of Anesthesia **56**(6) (2009).

[9] Nakamoto, M., Nakada, K., Sato, Y., Konishi, K., Hashizume, M. and Tamura, S., "Intraoperative magnetic tracker calibration using a magneto-optic hybrid tracker for 3-D ultrasound-based navigation in laparoscopic surgery," IEEE Transactions on Medical Imaging **27**(2) (2008).

[10] Zia, A., Sharma, Y., Bettadapura, V., Sarin, E. L., Clements, M. A. and Essa, I., "Automated assessment of surgical skills using frequency analysis," [Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)] (2015).

[11] Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A. and Radwin, R. G., "Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating," Annals of Surgery **269**(3) (2019).

[12] Sharma, Y., Plötz, T., Hammerld, N., Mellor, S., McNaney, R., Olivier, P., Deshmukh, S., McCaskie, A. and Essa, I., "Automated surgical OSATS prediction from videos," 2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI 2014 (2014).

[13] O'Driscoll, O., Hisey, R., Camire, D., Erb, J., Howes, D., Fichtinger, G. and Ungi, T., "Object detection to compute performance metrics for skill assessment in central venous catheterization," 2021.

[14] Pérez-Escamirosa, F., Chousleb-Kalach, A., Hernández-Baro, M. del C., Sánchez-Margallo, J. A., Lorias-Espinoza, D. and Minor-Martínez, A., "Construct validity of a video-tracking system based on orthogonal cameras approach for objective assessment of laparoscopic skills," International Journal of Computer Assisted Radiology and Surgery **11**(12) (2016).

[15] Xia, S., Keri, Z., Holden, M. S., Hisey, R., Lia, H., Ungi, T., Mitchell, C. H. and Fichtinger, G., "A learning curve analysis of ultrasound-guided in-plane and out-of-plane vascular access training with Perk Tutor," 2018.

[16] Holden, M. S., Keri, Z., Ungi, T. and Fichtinger, G., "Overall Proficiency Assessment in Point-of-Care Ultrasound Interventions: The Stopwatch is not Enough," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **10549 LNCS** (2017).

[17] Hong, J. S., Dohi, T., Hasizume, M., Konishi, K. and Hata, N., "A motion adaptable needle placement instrument based on tumor specific ultrasonic image segmentation," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **2488** (2002).