Object detection to compute performance metrics for skill assessment in central venous catheterization

Olivia O'Driscoll[1], Rebecca Hisey[1], Daenis Camire[2], Jason Erb[2], Daniel Howes[2], Gabor Fichtinger[1], Tamas Ungi[1]

[1]Laboratory for Percutaneous Surgery, School of Computing, Queen's University, Kingston, Canada
[2] Department of Critical Care Medicine, Queen's University, Kingston, Canada

ABSTRACT

Purpose: As medical schools move toward competency-based medical education, they seek methods of quantifying trainee skill without human expert supervision. This study evaluates the efficacy of using object detection to track performance metrics in ultrasound-guided interventions, specifically central venous catheterization. While several studies have explored methods to automate the evaluation of these interventions, they typically rely on expensive, bulky markers. Therefore, a webcam-based approach is desirable. **Methods**: We used the Faster Region-Based Convolutional Neural Network object detection network developed by Ren et al. to track the two-dimensional path length and the usage time of seven tools used in central venous catheterization. Object detection relies solely on webcam imagery. Video data were collected from recordings of 20 central venous catheterization trials by four different medical students. Each recording was separated into individual frames, annotated, and inputted to the object detection network. Mean average precision was calculated for each fold and each tool. **Results:** The average mean average precision was 0.66. Between trials one and five, the average reduction in tool usage time was 52%, and the average reduction in 2D path length was 29%. **Conclusions**: The neural network was able to identify each tool with considerable accuracy. Furthermore, the neural network successfully computed differences in performance metrics that emerge as trainees gain experience. Faster Region-Based Convolutional Neural Network is an effective method to assess trainee skill in ultrasound-guided interventions.

Keywords: Object detection, Faster R-CNN, ultrasound-guided interventions, central venous catheterization, computer-assisted skill assessment

1. INTRODUCTION

Medical schools are currently shifting away from the traditional method of time-based instruction in favour of competency-medical education (CBME). In the time-based approach, trainees are deemed competent and can graduate after spending a specific amount of time practicing the given procedure. Whereas in CMBE, trainees attain competency once they achieve a set level of performance [1]. However, to measure trainee performance in CBME, trainees require continuous expert supervision. Current assessor-based skill-assessment methods include global rating scales (GRS), checklists, and entrustment scores. Global rating scales consist of intervention-specific markers of trainee performance. Checklists offer a series of intervention-specific steps the trainee must follow. Entrustment scores assess the degree to which the expert reviewer trusts the trainee to complete the procedure alone.

All the aforementioned evaluation methods require human experts to continuously supervise trainees. With the increasing number of medical students, CBME represents a substantial burden on experts' limited time. Because of how much expert labour is required in CBME, several studies have explored methods to automate the evaluation process. For instance, many training simulation modules have been developed for ultrasound-guided interventions within the Perk Tutor extension of 3D Slicer. 3D Slicer (www.slicer.org) is a free, open-source software platform designed for medical informatics, image analysis, and visualization. Perk Tutor (www.PerkTutor.org) is a training platform for image-guided interventions [2]. Perk Tutor modules enrich trainee education by offering a configurable design for trainees to improve their performance in numerous ultrasound-guided interventions, such as lumbar spinal procedures, prostate biopsy, and central venous catheterization (CVC). In this paper, we focus on CVC as an important candidate for computer-assisted evaluation because it is a complex procedure requiring several different tools. Additionally, it is common across many medical disciplines,

and novice complication rates are as high as 35%, making it desirable to improve trainee education through simulation-based education [3,4].

As with assessor-based approaches to quantify skill for CBME, automatic methods must be based upon standard metrics by which to evaluate trainees, with many focusing on tool motion during the procedure. Holden et al. developed a method for skill assessment in ultrasound-guided interventions based on several in-plane metrics, including the path length and the usage time of the ultrasound probe and needle [5]. However, many existing solutions to track tool motion, including that presented by Holden et al., rely on physical sensors that must be directly attached to the tool being tracked. Using EM sensors adds unnecessary bulk to the tools, which limits its representation of the actual procedure by changing the feeling it the tool in hand. Because the ultimate goal of CBME is to train future practitioners, it is desirable to create an environment as similar as possible to that in which they will practice after graduation.

Not only are EM sensors bulky, but they are also expensive. This could pose a challenge for both medical schools requiring multiple training systems, and for low-resource hospitals in developing countries seeking to improve their medical education. To this end, optical tracking can be used as a less expensive alternative, but the markers needed for tracking are bulkier than EM sensors. Even so, not all tools retain their functionality when attached to sensors, and the number of sensors available limits the number of tools that can be tracked.

In response to the shortcomings of existing solutions outlined above, recent research has explored the avenues of computer vision and convolutional neural networks to assess trainee performance using only a webcam. Hisey et al. developed a module within Perk Tutor called Central Line Tutor, designed to provide trainees with meaningful, real-time feedback to improve their performance. Central Line Tutor used image classification to identify each step in the central venous catheterization (CVC) workflow automatically from webcam video [6]. However, workflow recognition has its limitations. While this method was able to evaluate trainees by measuring workflow compliance, image classification is unable to measure objective metrics regarding tool handling.

This is where object detection becomes useful. Not only does object detection recognize that a tool is being used, but it allows each tool's path to be tracked on-screen by placing a bounding box around each tool in the frame. While previous tool tracking methods focused on 3D tool motion, we sought to evaluate the efficacy of tracking 2D tool motion from the predicted bounding boxes of the object detection network as a way compute trainee performance metrics. Moreover, object detection is more versatile than physical tracking systems in its ability to recognize numerous tools. Therefore, object detection could be used as a method for workflow detection and tool tracking, combining the best of both current EM tracking systems and image classification solutions.

Object detection has already been explored as a method of tool tracking in laparoscopic procedures, notably as a part of the 2016 M2CAI Tool Presence Detection Challenge, using the m2cai16-tool dataset [7]. While the methods proposed in this paper focus on CVC, they could potentially apply to all ultrasound-guided interventions.

CBME requires continuous expert supervision in order to evaluate trainee progression, and automated systems have shortcomings in price, size, and versatility. We sought to evaluate the efficacy of object detection as a means to automatically track tool path lengths and usage times using only a webcam, thereby eliminating the need for expensive, bulky systems and continuous expert supervision.

2. METHODS

2.1 Dataset

To create the dataset, we recorded four different medical students performing CVC on a venous access phantom using the setup shown in Figure 1. We used the Gen I Ultrasound Central Line Training Model from CAE (Saint-Laurent, Quebec, Canada) and an ultrasound machine. The ultrasound machine was connected to a computer, displaying the Central Line Tutor interface in 3D Slicer. An RBG webcam mounted to a tripod captured video data of medical student trials. An application called Plus Server enables 3D Slicer to receive webcam data by connection through a module called Open IGT Link [8].

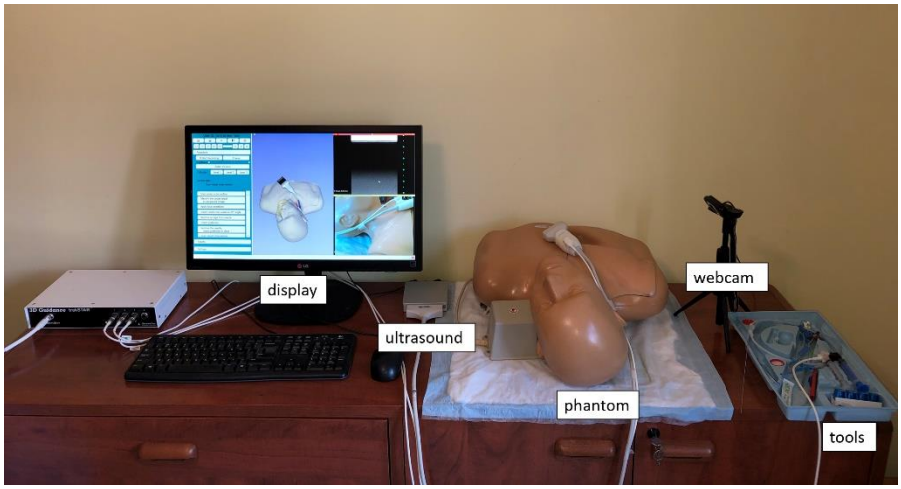

Figure 1. The setup that was used to record medical student trials.

The medical students had no previous experience in performing the procedure. Each student performed five procedural trials, totaling 20 recordings. The webcam was set up so that only the tools, the phantom and the students' hands were in the frame, preserving student anonymity. The recordings were separated into individual frames, using a 3D Slicer module designed to training image collection (https://github.com/SlicerIGT/aigt/tree/master/DeepLearnLive) [5]. The individual frames were annotated manually with bounding boxes as per Figure 2. Bounding boxes encompassed the entire visible portion of each tool that was in frame. This meant that multiple different tools could be present in the same frame.

The seven tools labelled were anesthetic, catheter, dilator, guidewire, guidewire casing, scalpel, and syringe, as shown in Figure 2. A total of 53748 images were collected, from which 73895 annotations were created. The representation of each tool in the dataset is proportional to the amount of time it is used in each procedure, so there was considerable variability in the number of images per tool. We balanced the dataset by randomly selecting an equal number of frames from each class. After balancing, around 300 images were used for training and 80 for validation in each fold.

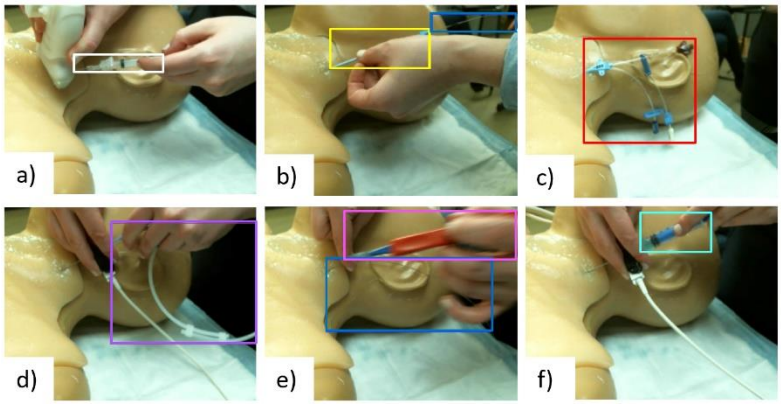

Figure 2. Ground truth labels for each class, a) anesthetic; b) dilator (yellow) and guidewire (blue); c) catheter; d) guidewire casing; e) scalpel (pink) and guidewire (blue); f) syringe

2.2 Object Detection

After creating a dataset, we had to decide which of the two most common object detection network families would best suit the application needs: You Only Look Once (YOLO) or Region-based Convolutional Neural Network (R-CNN). R-

CNNs combine a convolutional neural network (CNN) with a region proposal network (RPN). R-CNN uses the RPN to group similar pixels and then uses the CNN to classify these groupings. Unlike R-CNN, YOLO uses only one neural network. The YOLO technique is faster and favored for real-time approaches. [9] However, R-CNN's are more accurate and are desirable for model performance. This study favors accuracy over speed and uses the R-CNN technique. Specifically, we use Faster R-CNN, developed by Ren et al. [10] Faster R-CNN first uses a feature network to identify features of the image. It then replaces the selective search used in R-CNN and Fast R-CNN with an RPN. The RPN determines regions of interest (ROI) from the identified features. Finally, a fully connected network classifies the ROIs as per the classes on which it was trained.

The feature network used was ResNet-50. To improve model performance, we leveraged transfer learning, which is an optimization where knowledge learned from one problem is applied to a different problem. We used image classification weights that were pre-trained on the ImageNet dataset as the starting point for training, which allowed for increased performance and faster progress when it came to training our own model.

Additionally, tool kits for CVC never contain more than one copy of any individual tool, so we reduced false positive by permitting the model to only apply one bounding box per tool per frame. However, there were often multiple tools visible by the webcam at a time, so the only limit to the number of different classes that could be detected in a single frame was the number of classes themselves.

We used a leave-one-user-out cross-validation scheme. To create our test set for each fold, we reserved all five videos from a single medical student. For our validation set, we randomly selected one video from each of the remaining three participants. The remaining 12 videos were used for training.

We used mean average precision (mAP) to measure the network's ability to identify each tool. To calculate mAP, we must first define what is considered a correct sample. Because there are more factors at play than just the classification, we must incorporate the accuracy of the bounding boxes with that of the classification in order to create binary categories of true positive and false positive. To this end, we used intersection over union (IOU) to evaluate the accuracy of the bounding boxes. IOU is calculated as the area of overlap between the ground truth and predicted bounding box, divided by the area of intersection. Because the probability of the predicted bounding box sharing the exact coordinates of the ground truth bounding box is extremely low, IOU is beneficial as it rewards bounding box predictions based on their overlap. Each sample with a correct classification and an IOU of greater than 0.5 was considered a true positive. If the IOU was less than or equal to 0.5, the sample was considered a false positive. Ground truth bounding boxes that the network did not detect were considered false negatives. All other samples were considered true negatives.

Once we divided samples into true positives, true negatives, false positives, and false negatives, we calculated the mAP. The mAP can be defined as the area under the precision/recall curve for all classes. Precision is the number of true positives divided by the sum of true and false positives. In contrast, recall is the number of true positives divided by the sum of true positives and false negatives. More generally, is the mean of the average precisions calculated for each class. Therefore, mAP and IOU serve as effective metrics by which to evaluate the predicted classifications and bounding boxes.

Using mAP as a metric for network performance is standard in object detection studies and challenges. In the 2016 M2CAI Tool Presence Detection Challenge, the top three performing networks had mAPs ranging from 0.638-0.525. More recently, Zhang et al used a Modulating Anchoring Network based on Faster R-CNN to achieve an mAP of 0.696 on the m2cai16-tool dataset [11].

2.3 Trainee Performance Assessment

To evaluate the efficacy of using an object detection network to assess trainee skill, we computed each tool's two-dimensional path length and usage time. These performance metrics have both been shown to decrease with trainee competency [5]. The number of frames in which each tool is present defines the usage time and is calculated as per Equation 1, where t is the usage time, and n is the number of frames in which the tool appears.

$$t = \frac{n}{frame\ rate} \tag{1}$$

The two-dimensional path length is the movement of the center of the bounding box across sequential frames and is defined as the sum of Euclidean distances between the center of bounding boxes. It is calculated as per Equation 2, where p is the path length, d is the distance between sequential bounding boxes, i is the frame number, and N is the total number of frames.

$$p = \sum_{i=0}^{N} d(i, i+1) \tag{2}$$

On the test set of each fold, we calculated the two-dimensional path lengths and usage times for each tool, which are easily computed from the predicted bounding boxes. We compared these values from the first and fifth trials to determine if a discernible difference emerged with trainee experience, and therefore to evaluate the efficacy of using Faster R-CNN to calculate them.

To calculate the reduction in these metrics between the first and fifth trials, we used Equation 3, where d is percentage decrease, t_1 is the calculated value of the given metric in trial one, and t_5 is that for trial 5.

$$d = \frac{(100\%)(t_5 - t_1)}{t_1} \tag{3}$$

3. RESULTS

The neural network's average mAP was 0.66. The most accurate fold was fold 3, with an mAP of 0.71, while the least accurate fold was fold 0 with an mAP of 0.59, as shown in Table 2.

Table 2. Mean average precision of each fold

Fold	**Mean Average Precision**
0	0.59
1	0.68
2	0.67
3	0.71
average	**0.66**

The syringe had the highest mAP, at 0.92, while the catheter had the lowest mAP at 0.49, as shown in Table 3.

Table 3. Mean average precision of each tool

Tool	**Mean Average Precision**
Anesthetic	0.64
Catheter	0.49
Dilator	0.71
Guidewire	0.51
Guidewire casing	0.86
Scalpel	0.50
Syringe	0.92

After training, we calculated the usage time and path lengths for each of the seven tools used in the procedure. Doing so allowed us to assess the efficacy of using Faster R-CNN to compute performance metrics for skill assessment in CVC. On average, the tool usage time in the fifth trial was 52% shorter than that of trial one, as shown in Figure 3. The average 2D tool path length in the fifth trial was 29% shorter than that of the first trial, as represented in Figure 4.

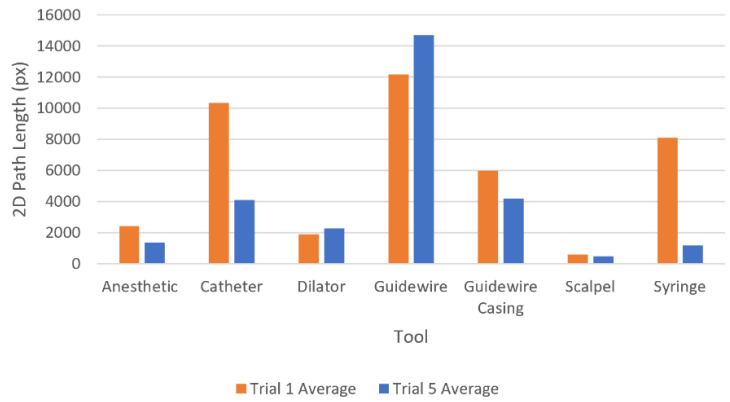

Figure 3. A comparison of tool usage time between the first and fifth trials.

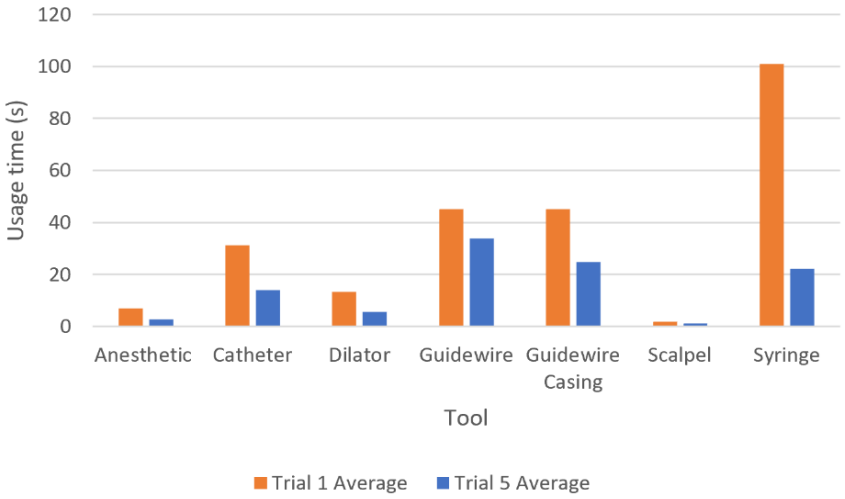

Figure 4. A comparison of 2D tool path length between the first and fifth trials

DISCUSSION

While the network's mAP of 0.66 is on par with mean detection studies, there was justifiable variability of mAP across tools [7,11]. The syringe is easy to locate and classify because of its bright color and consistent orientation during the procedure. The catheter changes shape, size, and color during the procedure, contributing to its lower mAP.

The only performance metrics that, on average, did not improve between trial one and trial five were the two-dimensional path lengths of the guidewire and dilator. To preserve the phantom, the dilator is not inserted into the vein as it would be during a procedure on a live patient. Therefore, it is understandable that measuring the dilator's 2D path length is not an effective way to assess trainee skill in this simulation setup. Because of the guidewire's variable shape, the center of the bounding box varies substantially. It is therefore reasonable that the guidewire's 2D path length is also an ineffective

method to assess trainee skill. Otherwise, the usage time and 2D tool path length, as measured by object detection, remain useful performance metrics for measuring trainee skill.

We sought to develop a low-cost, lightweight solution for computer-assisted skill assessment that would eliminate the need for continuous expert supervision in CBME. We did this by evaluating the efficacy of using Faster R-CNN to assess trainee skill in ultrasound-guided interventions based only on video data from a webcam. The initial results presented in this paper reveal that Faster R-CNN was able to compute the improvements in performance metrics that emerge with trainee skill. Therefore, Faster R-CNN shows promise as method of computer-assisted skill assessment.

One shortcoming of the methods presented in this paper is that they only provided information about the 2D motion of the tools. In future studies, we hope to circumvent this by using a three-dimensional camera to provide more information about trainee tool handling, and enrich trainee education by providing more detailed, meaningful feedback. To improve the accuracy of the calculated metrics, we will not only expand our dataset, but leverage more data from each class by using a new method for balancing our dataset, rather than simply random undersampling.

Our next step is to integrate Faster R-CNN into Central Line Tutor to provide real-time feedback based on both workflow detection and tool handling. In addition, we will need to optimize the network for speed to maximize performance in real-time applications. Finally, we hope to explore the possibility of using Faster R-CNN on new datasets as a method for skill assessment in other ultrasound-guided procedures. This project's code can be found at (https://github.com/SlicerIGT/aigt).

ACKNOWLEDGEMENTS

This work was funded in part by NIH/NIBIB and NIH/NIGMS (via grant 1R01EB021396-01A1), by CANARIE's Research Software Program, and is supported as a Collaborative Health Research Project (CHRP #127797) by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Institutes of Health Research (CIHR). R. Hisey is supported by the Queen Elizabeth II Graduate Scholarships in Science and Technology (QEII-GSST). G. Fichtinger is supported as a Canada Research Chair.

REFERENCES

[1] Halman, S., Fu, A. Y. N., Pugh, D., "Entrustment within an objective structured clinical examination (OSCE) progress test: Bridging the gap towards competency-based medical education," Medical Teacher 42(11), 1283–1288 (2020).

[2] Ungi, T., Sargent, D., Moult, E., Lasso, A., Pinter, C., Mcgraw, R. C., Fichtinger, G., "Perk Tutor: An Open-Source Training Platform for Ultrasound-Guided Needle Insertions," IEEE Transactions on Biomedical Engineering 59(12), 3475–3481 (2012).

[3] Kumar, A., and Alwin, C. "Ultrasound guided vascular access: efficacy and safety," Best Practice & Research Clinical Anaesthesiology, 299-311 (2009)

[4] Mcgaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H., Wayne, D. B., "Does Simulation-Based Medical Education With Deliberate Practice Yield Better Results Than Traditional Clinical Education? A Meta-Analytic Comparative Review of the Evidence," Academic Medicine 86(6), 706–711 (2011).

[5] Holden, M. S., Xia, S., Lia, H., Keri, Z., Bell, C., Patterson, L., Ungi, T., 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), 1993–2003 (2019).

[6] Hisey, R., Ungi, T., Holden, M., Baum, Z., Keri, Z., McCallum, C., Howes, D. and Fichtinger, G. "Real-time workflow detection using webcam video for providing real-time feedback in central venous catheterization training," Proc. SPIE 10576, Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, 1057620 (2018)

[7] "Tool Presence Detection Challenge Results.," Workshop and Challenges on Modeling and Monitoring of Computer Assisted Interventions, <http://camma.u-strasbg.fr/m2cai2016/index.php/tool-presence-detection-challenge-results./> (13 December 2020).

[8] Tokuda, J., Fischer, G. S., Papademetris, X., Yaniv, Z., Ibanez, L., Cheng, P., Liu, H., Blevins, J., Arata, J., et al., “OpenIGTLink: an open network protocol for image-guided therapy environment,” The International Journal of Medical Robotics and Computer Assisted Surgery 5(4), 423–434 (2009).

[9] Redmon, J., Santosh, D., Girshick, R., Ali, F., “You Only Look Once: Unified, Real-Time Object Detection,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788 (2016).

[10] Ren, S., He, K., Girshick, R. and Sun, J. "Faster r-cnn: Towards real-time object detection with region proposalnetworks," Advances in neural information processing systems, 91-99 (2015)

[11] Zhang, B., Wang, S., Dong, L., Chen, P., “Surgical Tools Detection Based on Modulated Anchoring Network in Laparoscopic Videos,” IEEE Access 8, 23748–23758 (2020).