2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

2021 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9077-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRA48506.2021.9561369

Deep Learning-Based Photoacoustic Visual Servoing: Using Outputs from Raw Sensor Data as Inputs to a Robot Controller

Mardava R. Gubbi[1] and Muyinatu A. Lediju Bell[2]

***Abstract*— Tool tip visualization is an essential component of multiple robotic surgical and interventional procedures. In this paper, we introduce a real-time photoacoustic visual servoing system that processes information directly from raw acoustic sensor data, without requiring image formation or segmentation in order to make robot path planning decisions to track and maintain visualization of tool tips. The performance of this novel deep learning-based visual servoing system is compared to that of a visual servoing system which relies on image formation followed by segmentation to make and execute robot path planning decisions. Experiments were conducted with a plastisol phantom, *ex vivo* tissue, and a needle as the interventional tool. Needle tip tracking performance with the deep learning-based approach outperformed that of the image-based segmentation approach by 67.7% and 55.3% in phantom and *ex vivo* tissue, respectively. In addition, the deep learning-based system operated within the frame-rate-limiting 10 Hz laser pulse repetition frequency rate, with mean execution times of 75.2 ms and 73.9 ms per acquisition frame with phantom and *ex vivo* tissue, respectively. These results highlight the benefits of our new approach to integrate deep learning with robotic systems for improved automation and visual servoing of tool tips.**

I. Introduction

The ability to visualize and track surgical tool tips is paramount to the success of multiple surgeries and procedures. Ultrasound is the one of the most commonly used imaging modalities to track tool tips due to its low cost, high frame rates, portability, and absence of harmful ionizing radiation. The combination of ultrasound imaging with either traditional techniques of visual servoing [1], [2] or recent advances in deep learning [3], [4] introduces additional layers of automation for this important task. For example, ultrasound-based visual servoing may assist with percutaneous needle insertions [5], and deep learning has the potential to improve the performance and speed of ultrasound image-based needle detection systems [6]. However, both of these automation gains rely on the ultrasound imaging process, which tends to fail in acoustically challenging environments characterized by significant acoustic clutter [7], sound scattering, and sound attenuation. Specific examples of challenging acoustic environments include transcranial imaging [8], abdominal imaging [7], spinal imaging [9], or imaging of obese patients [10].

One option to address known limitations with ultrasound imaging is to combine ultrasound imaging systems with a miniature laser system to perform intraoperative photoacoustic imaging [11]–[13], which has provided clear images of needle tips and other structures when ultrasound imaging fails [13]. Unlike ultrasound imaging, which requires the transmission and reception of sound to make images, photoacoustic imaging is implemented by transmitting light to generate an acoustic response that is received by the same ultrasound detectors used for ultrasound imaging [14], [15]. Photoacoustic imaging tends to be advantageous over ultrasound imaging in acoustically challenging environments because it only requires one-way (as opposed to round-trip) acoustic travel from the transmission source to the ultrasound receiver.

Previous work from our group demonstrated the success of using photoacoustic imaging as the computer vision component of a visual servoing system, enabling continuous monitoring of needle [13] and catheter [12] tips. The needle or catheter tip each housed an internal optical fiber as one of the key enabling modifications to the interventional setup. This optical fiber can potentially be coupled with any surgical tool tip [16]–[18] to enable photoacoustic-based visual servoing of the tool tip. Therefore, this approach was also demonstrated with a fiber that was independent of any tool, catheter, or needle tip [19].

To achieve photoacoustic-based visual servoing, raw data is typically beamformed to present a photoacoustic image that is interpretable to the human eye, followed by image segmentation to determine coordinates of interest for robot path planning. However, beamforming and other image formation approaches rely on mathematical models that do not consider all possible photoacoustic image artifact sources. Artifacts that cannot be removed with traditional amplitude-based [12], [13] or coherence-based [19] photoacoustic visual servoing approaches (e.g., reflection artifacts or coherent artifacts, respectively) are confusing for both human and robot interpretation, resulting in unreliable segmentation for photoacoustic visual servoing tasks.

In order to better discriminate sources from artifacts, we turn our attention to investigate novel input sources to the robotic system (which may not necessarily need to operate on an image that is interpretable to humans). In particular, our recent photoacoustic-based deep learning approaches for photoacoustic source detection [20]–[22] suggest that

This work is supported by NSF SCH Award IIS-2014088, NIH Trailblazer Award R21 EB025621, and NSF CAREER Award ECCS-1751522.

[1] M. R. Gubbi is with the Department of Electrical and Computer Engineering, Johns Hopkins University, 3400 N Charles St, Baltimore, MD, USA, 21218 (email: mardava.gubbi@jhu.edu)

[2] M. A. L. Bell is with the Department of Electrical and Computer Engineering, the Department of Biomedical Engineering, and the Department of Computer Science, Johns Hopkins University, 3400 N Charles St, Baltimore, MD, USA, 21218 (email: mledijubell@jhu.edu)

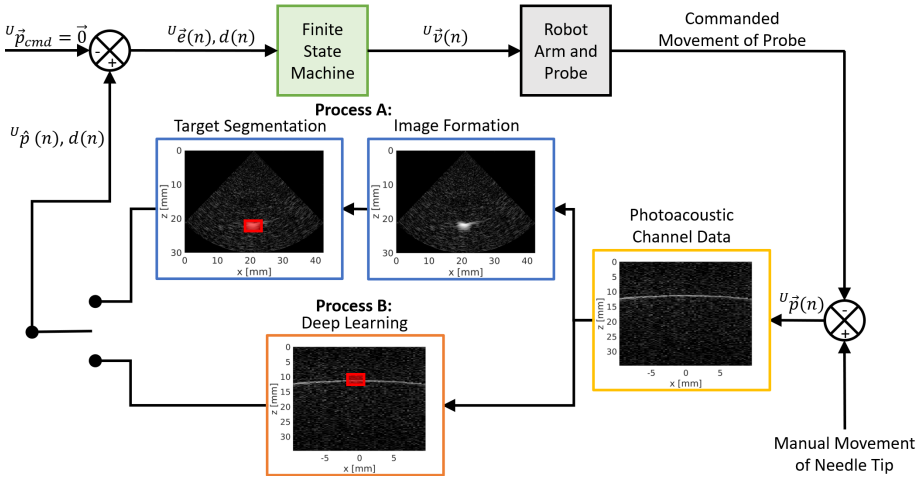

Fig. 1. Block diagram illustrating the photoacoustic visual servoing system. Process A is our previously introduced segmentation-based approach to visual servoing beamformed photoacoustic signals, after the acquisition of raw photoacoustic sensor data (also known as channel data), with representative implementations described in [12], [13], [19]. Process B is our newly introduced deep learning-based approach to visual servoing raw photoacoustic channel data. In each case, the red rectangular overlay indicates position coordinates that are input to the robot controller.

deep learning is a viable solution to address current challenges with amplitude- or coherence-based photoacoustic visual servoing. The novel concept of using deep learning to detect interventional structures of interest in raw sensor data before the application of traditional image formation techniques was previously implemented to detect needle [21], [22] and catheter [23], [24] tips. In summary, recent work from our group independently demonstrated two key advances with regard to interventional tool tip tracking: (1) photoacoustic-based visual servoing to enhance tool tip tracking and centering within the image plane [12], [13], [19] and (2) deep learning-based photoacoustic image formation from raw sensor data to improve tool tip visibility [20]–[24].

The independent demonstrations of feasibility described above suggest that the integration of deep learning with photoacoustic-based visual servoing is a superior approach to address well-known challenges with tool tip tracking. This paper presents the first known deep learning-based photoacoustic visual servoing system to address these challenges. The novelty of this contribution includes the creation and implementation of a direct pathway from the photoacoustic raw sensor data (i.e., before any image has been formed) to the robot controller, enabled by recent advances using deep learning to extract information directly from raw acoustic sensor data [20]–[25].

The remainder of this paper is organized as follows. Section II introduces our deep learning-based approach to visual servoing raw photoacoustic sensor data (also known as channel data), followed by a description of our network training process. This deep learning approach is contrasted with our previously introduced segmentation-based approach to visual servoing beamformed photoacoustic signals. Section III describes our experiments to test both approaches. Section IV presents our experimental results. Section V discusses our findings in the context of prior work. Section VI concludes the manuscript.

II. Visual Servoing System

A. System Components

Fig. 1 shows a block diagram of the photoacoustic visual servoing system used in this work. The system components include a Sawyer robot (Rethink Robotics, Boston, MA, USA), a Vantage 128 ultrasound scanner (Verasonics Inc., Kirkland, WA, USA), a Verasonics P4-2v phased array ultrasound probe, a Phocus Mobile laser (Opotek, Carlsbad, CA, USA), and a 600 μm core diameter optical fiber. One end of the optical fiber was coupled to the laser. The other end of the optical fiber was inserted into a hollow core needle, ensuring coincident fiber and needle tips to form a fiber-needle pair. The probe was attached to the end effector of the robot using a 3D-printed holder. Nanosecond laser pulses were transmitted at a rate of 10 Hz with a wavelength of 750 nm. The software components of the visual servoing system were implemented using the Robot Operating System (ROS) [26].

The frame U was assigned to coincide with the Verasonics P4-2v probe, with the x-, y-, and z-dimensions corresponding to the lateral, elevation, and axial dimensions of the probe, respectively. The imaging plane of the probe corresponded to the x-z plane of the frame U. The raw channel data frames acquired with the probe were processed to obtain an estimate ${}^{U}\hat{p}\,(n)$ of the position of the needle tip in the ultrasound probe frame U and a confidence measure $d\,(n) \in (0, 1)$ of the estimate. We refer to this confidence measure as the validity of the estimate.

Process A used the amplitude-based approach developed by Bell *et al.* [13] to estimate the needle tip position and assess its validity. A photoacoustic image was recreated from the acquired channel data using delay-and-sum beamforming. The beamformed image was normalized and a

binary threshold of 0.7 was applied to the normalized image. Binary erosion and dilation were performed with a 3x3 kernel to remove single pixel regions and connect segments which became disconnected during the binary threshold application. The erosion and dilation filters helped to ensure that the segmented needle tip was displayed as a single large component, rather than as multiple smaller components. Connected components were then labeled and their corresponding pixel areas were computed. If only one region was larger than 3 times the average area, then that region was assumed to be the needle tip and the centroid of that region was output as the needle tip position. Otherwise, the needle tip was assumed to be outside the field of view of the probe. For robustness, the estimated needle tip position was compared across five consecutive frames. If the needle tip was visible in each frame (i.e., $d(n) = 1$) and the estimated position of the needle tip did not change by more than 1 cm across the 5 frames, then the needle tip position was labeled as valid.

Process B used a convolutional neural network (CNN) to provide estimates of the needle tip position and corresponding confidence levels in the range 0 to 1. With a focus on proving the feasibility of integrating deep learning-based approaches with real-time photoacoustic visual servoing systems, we used the ResNet-101 architecture [27] and the Faster-RCNN detection method [28], which were previously demonstrated by Allman *et al.* [22] as an offline technique applied to photoacoustic channel data obtained with an E-CUBE 12R ultrasound scanner (Alpinion Medical Systems, Seoul, South Korea). For robustness, the estimated needle tip was compared across 5 consecutive frames as described above. If the needle tip was visible with a confidence level $d(n) > 0.7$ in each frame and the estimated position of the needle tip did not change by more than 1 cm across the 5 frames, then the needle tip position was labeled as valid.

Fig. 2 shows the finite state machine used to control the translational degrees of freedom corresponding to the lateral and elevation dimensions of the probe. Two-dimensional (2D) photoacoustic images do not contain elevation displacement information. As a result, both Process A and Process B output zeros in the y-dimension of the estimate ${}^{U}\hat{p}(n)$. In the nominal "Center" state of the FSM, the error ${}^{U}\vec{e}(n)$ in the frame U was computed using the

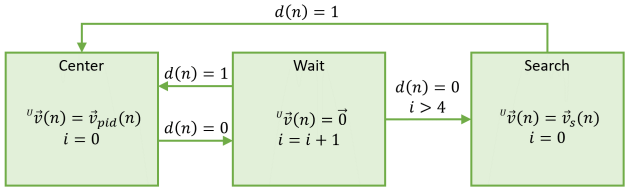

Fig. 2. Finite state machine component of visual servoing system (illustrated with validity checks, $d(n)$, corresponding to Process A).

equation:

$$ {}^{U}\vec{e}(n) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left({}^{U}\hat{p}(n) - {}^{U}\vec{p}_{cmd} \right), \quad (1) $$

where ${}^{U}\vec{p}_{cmd} = \vec{0}$ is the desired position of the needle tip in the probe frame U. This computation of ${}^{U}\vec{e}(n)$ ensures that the visual servoing system will center the probe laterally above the needle tip without changing the axial or elevation displacement between the probe and the needle tip. If the FSM was in the "Center" state and the needle tip position estimate was marked as valid (i.e., $d(n) = 1$ for Process A and $d(n) > 0.7$ for Process B), then the end effector of the robot was commanded to move along the x-axis of the probe frame U with the velocity $\vec{v}_{pid}(n)$ given by the equation

$$ \begin{aligned} \vec{v}_{pid}(n) =& K_p \, {}^{U}\vec{e}(n) + K_i \sum_{k=0}^{n} {}^{U}\vec{e}(k) \, \Delta T \\ &+ K_d \left(\frac{{}^{U}\vec{e}(n) - {}^{U}\vec{e}(n-1)}{\Delta T} \right), \end{aligned} \quad (2) $$

where K_p, K_i, and K_d are the gains of the PID controller (with values of 0.1, 0.01, and 0.001, respectively), and ΔT is the sampling time of the PID controller. The controller was executed every 0.1 s to match the pulse repetition rate of the laser.

The validity (i.e., $d(n)$) was used to indicate movement of the needle tip outside of the imaging plane of the probe. If the estimated needle tip position was marked as invalid, the FSM entered the "Wait" state. In this state, the end effector was held stationary until up to 5 frames of channel data were acquired by the photoacoustic imaging system. If a valid estimate of the needle tip position was obtained during that time, the FSM returned to the "Center" state. Otherwise, the FSM entered the "Search" state. In this state, the robot end effector was moved in a 2D spiral pattern given by

$$ \vec{v}_s(n) = \begin{bmatrix} An\cos(\omega n) \\ An\sin(\omega n) \\ 0 \end{bmatrix} \quad (3) $$

where A and ω are the parameters of the spiral search pattern.

The commanded velocity ${}^{U}\vec{v}(n)$ was then converted to the base frame B of the robot using the equation

$$ {}^{B}\vec{v}(n) = {}^{B}T_E(n) \, {}^{E}T_U \, {}^{U}\vec{v}(n) \quad (4) $$

where ${}^{E}T_U$ is the transformation from the ultrasound probe frame U to the robot end effector frame E and ${}^{B}T_E(n)$ is the instantaneous transformation from the frame E to the robot base frame B. The commanded velocity ${}^{B}\vec{v}(n)$ was then transmitted to the internal velocity controller of the robot over the ROS topic for velocity commands to which the controller subscribed.

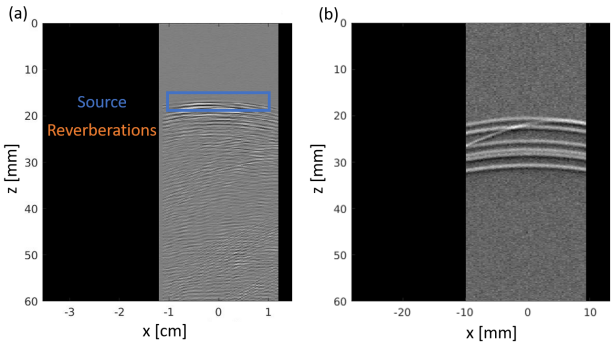

Fig. 3. Examples of (a) acquired experimental data and (b) simulated images with reverberations directly under the source. The reverberations are observed up to 5 mm deeper than the source, and laterally centered underneath the source.

TABLE I
RANGE AND INCREMENT SIZES OF SIMULATION VARIABLES

Parameter	Min	Max	Increment
Depth Position (mm)	5	55	0.3
Lateral Position (mm)	-18.9	18.9	0.189
Number of Reverberation Artifacts	3	5	1
Depth of Reverberation Artifacts under Source (mm)	1	10	0.3
Speed of Sound (m/s)	1440	1640	6
Object Intensity (Multiplier)	0.75	1.1	random
Channel SNR (dB)	-5	2	random

B. Training the Convolutional Neural Network

Simulations that mimic the physics of photoacoustic wave propagation offer the ability to generate training data without the time-intensive process of experimentally gathering and hand-labeling the large datasets [20]–[25]. This ease of data generation makes simulations a powerful tool in the context of deep learning. To train the CNN for Process B, 20,000 frames of photoacoustic channel data were generated using the k-Wave toolbox [29] in MATLAB. We simulated a single source of diameter 0.1 mm and 4-6 artifacts in each image. One of the artifacts could be anywhere in the image to simulate a reflection artifact and maintain consistency with previous implementations [22]. The remaining artifacts were constrained to the range 1 mm to 10 mm below the source to simulate the reverberation artifacts, as observed in Fig. 3(a), which shows one of the acquired channel data frames used as a reference to generate our training dataset. The ranges and increment values of our simulation variables are listed in Table I. We simulated a discrete ultrasound probe model with a sampling frequency of 11.88 MHz, an aperture of 128 elements, an element width of 0.25 mm, and an inter-element spacing of 0.05 mm. These parameters were selected to match the specifications of the Verasonics P4-2v probe to improve network performance [30]. An example of our simulated training data is shown in Fig. 3(b).

The Detectron platform [31] was utilized for training and validation. The network was initialized with pre-trained ImageNet weights [32] and trained on 80% of the simulated images. The remaining 20% of the images were used for network validation. Finally, the Detectron-ROS package [33] was utilized to incorporate the trained network into Process B of the visual servoing system.

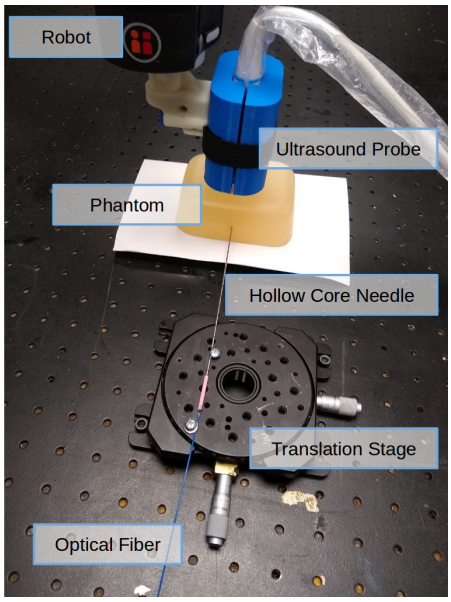

Fig. 4. Photograph of the setup for needle tracking and probe centering experiments.

III. EXPERIMENTAL METHODS

A. Probe Centering Experiment

The experimental setup for the probe centering experiments is shown in Fig. 4. These experiments were implemented to estimate the probe centering and needle tracking errors of the two processes (i.e., A and B) for needle tip detection, similar to previous experiments implemented with a segmentation-based photoacoustic visual servoing system [13]. The choices for each experimental trial included laser fluence (18.4 uJ/cm^2 or 49.5 uJ/cm^2), needle tip detection process (Process A or B), and imaging environment (plastisol phantom or *ex vivo* chicken breast). There were 9 probe centering trials per fluence, per process, per imaging environment.

At the start of each experimental trial, the translation stage was reset to 0 mm. The fiber-needle pair was inserted into the chosen imaging environment. The ultrasound probe was placed on the surface of the imaging environment, with the imaging plane of the probe placed to contain as much of the intended trajectory of the needle tip as possible. The probe was then manually displaced distances of 2-10 mm from the needle tip in 2 mm increments in the lateral probe dimension, followed by initiation of visual servoing with Process A or B.

The visual servoing system was executed to center the probe above the needle tip. If the needle tip detection process output 5 consecutive valid estimates of the needle

tip position (i.e., $d(n) = 1$ for Process A and $d(n) > 0.7$ for Process B), the trial was marked as a success. The mean of the lateral components of those 5 valid estimates ${}^{U}\hat{p}(n)$ was computed, and the magnitude of ${}^{U}\hat{p}(n)$ was output as the probe centering error. If 5 consecutive valid readings could not be obtained (i.e., $d(n) = 0$ for Process A and $d(n) < 0.7$ for Process B), the trial was marked as a failure. The mean and standard deviation of the probe centering errors were computed for each process and imaging environment.

B. Needle Tip Tracking Experiment

The same setup shown in Fig. 4 and described in Section III-A was used for the needle tip tracking centering experiments, using thee same choices for each experimental trial. There were similarly 9 needle tracking trials per fluence, per process, per imaging environment. After successfully centering the probe on the needle tip (as defined in Section III-A), the translation stage was used to move the needle tip in 2 mm increments along the lateral dimension of the probe. At each position, the output of the needle tip detection process was observed. If the process output five consecutive valid estimates of the needle tip position (i.e., $d(n) = 1$ for Process A, and $d(n) > 0.7$ for Process B), then the position was marked as a success. The needle tracking error was then computed using the equation:

$$e = \| {}^{B}\vec{p}_f - {}^{B}\vec{p}_i - {}^{B}\vec{s}_n \|, \tag{5}$$

where e, ${}^{B}\vec{p}_i$, ${}^{B}\vec{p}_f$, and ${}^{B}\vec{s}_n$ are the needle tracking error, the initial robot end effector position, the final robot end effector position, and the measured displacement of the needle tip, respectively, in the robot base frame B.

If five consecutive valid readings could not be obtained (i.e., $d(n) = 0$ for Process A and $d(n) \leq 0.7$ for Process B), the position was marked as a failure. The failure rates of Processes A and B were compared to assess the robustness of each algorithm.

IV. Results

Table II summarizes the mean and standard deviation of probe centering errors for 90 trials per Process A or B, implemented with either the plastisol phantom or the *ex vivo* tissue. For each imaging environment, the probe centering errors of Processes A and B were within 0.1 mm of each other. The probe centering errors were similarly within 0.1 mm across the two imaging environments.

Fig. 5 shows the mean and standard deviation of the needle tracking errors for 18 trials per process, imaging environment, and lateral shift value. Process A produced needle tracking errors ranging 0.59-5.36 mm with a mean of 2.63 mm across all phantom trials and ranging 1.47-2.34 mm with mean of 1.96 mm across all *ex vivo* tissue trials. Process B generally produced better needle tracking errors than that of Process A, ranging 0.65-1.03 mm with a mean of 0.85 mm across all phantom trials and ranging 0.46-1.39 mm with a mean of 0.88 mm across all *ex vivo* tissue trials.

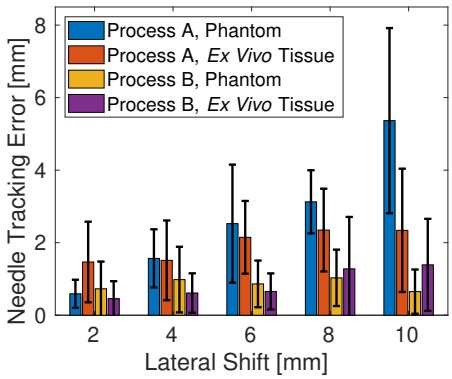

Fig. 5. Mean needle tip tracking errors as functions of the lateral shift for Processes A and B in the phantom and *ex vivo* tissue. The black error bars represent the standard deviation of each set of measured errors.

Table III lists the failure rates during needle tip tracking. For multiple trials with Process A, the needle tip was incorrectly labeled a reflection artifact that formed a larger bright region than the needle tip. This mislabeling caused a majority of the observed failures of Process A in both the phantom and the *ex vivo* tissue. Process B generally produced lower failure rates than Process A, with a mean improvement of 60.6% across all lateral shifts and both imaging environments. The highest failure rates of 1.85% and 3.70% for Process B were observed at a lateral offset of 10 mm in the phantom and *ex vivo* tissue environments, respectively. These failure locations are consistent with previous reports demonstrating increased CNN failure rates as lateral offset from the center of an image increases, due to a reduced number of source examples on the image periphery [24].

We additionally note the <100 ms execution time requirement for Process B in order to achieve the same 10 Hz frame rate previously demonstrated with visual servoing systems using iterations of Process A [12], [13]. This

TABLE II
Probe Centering Errors

Process	Test Case	Mean Error [mm]	Std Dev [mm]
A	Phantom	0.11	0.12
A	*Ex Vivo* Tissue	0.16	0.14
B	Phantom	0.19	0.16
B	*Ex Vivo* Tissue	0.18	0.17

TABLE III
Needle Tracking Failure Rates

Lateral Shift [mm]	Process A		Process B	
	Phantom	*Ex Vivo*	Phantom	*Ex Vivo*
2	3.51%	0.00%	0.00%	0.00%
4	7.02%	0.00%	0.00%	1.85%
6	7.02%	0.00%	0.00%	0.00%
8	5.26%	1.85%	0.00%	1.85%
10	3.51%	1.85%	1.85%	3.70%

requirement is dictated by the 10 Hz laser pulse repetition frequency of the photoacoustic imaging system. The mean $\pm$ one standard deviation of execution times from 36 trials of both experiments described above with Process B were 75.2 ± 12.8 ms and 73.9 ± 13.2 ms per channel data frame with the phantom and *ex vivo* tissue, respectively.

V. Discussion

The results presented in this manuscript highlight the potential of a CNN to provide an alternative input to command robotic visual servoing systems. This potential was demonstrated with a system composed of a Verasonics ultrasound engine, which was not used in any previous work testing similar CNN architectures [22]–[24], [30]. It is promising that the presented needle tracking errors are comparable to the 0.40 $\pm$ 0.22 mm point source location errors obtained by Allman *et al.* in [22] with an Alpinion E-CUBE 12R ultrasound scanner and an L3-8 probe. This similar success indicates that the previously proposed deep learning methods for photoacoustic point source detection are generalizable across multiple imaging system platforms.

We identified three advantages of integrating this novel deep learning approach with a robotic photoacoustic visual servoing system, when compared to the amplitude-based image segmentation approach: (1) lower tool tip tracking failure rates in the presence of reflection artifacts, (2) reduced needle tracking errors across different imaging environments, and (3) maintenance of 10 Hz frame rates despite increased algorithmic complexity. Regarding the first advantage, the sensitivity of the beamforming techniques to reflection artifacts (e.g., caused by bone) can lead to uncertain and potentially hazardous robot arm movements during surgical procedures, which is a major concern for the steps required to complete the segmentation-based visual servoing approach (i.e., Process A). Instead of adding successive layers of complexity to the beamformer or segmentation algorithm to account for these artifacts and features, the deep learning approach (i.e., Process B) trains a CNN to distinguish between true sources and reflection artifacts in the raw channel data, thus mitigating the introduction of misclassification errors.

To appreciate the second advantage, the reduced mean needle tracking errors with the deep learning approach (i.e., 0.85 mm and 0.88 mm in phantom and *ex vivo* tissue, respectively) can be compared to the needle tracking errors obtained with the segmentation approach (i.e., 2.63 mm and 1.96 mm in the phantom and *ex vivo* tissue, respectively). The overall mean improvement with Process B translates to 67.7% and 55.3% reductions in needle tracking errors in the phantom and *ex vivo* tissue, respectively.

With regard to the third advantage of achieving 10 Hz frame rates, the deep learning approach has a higher algorithmic complexity compared to the image segmentation approach (i.e., $\mathcal{O}(MNS)$ for Process A vs. $\mathcal{O}\left(N_{\text{conv}} D^2 N_{\text{ch}}^2 MN\right)$ for Process B, where M, N, S, and N_{conv} are the numbers of receiving elements, samples acquired per frame, scan lines in the beamformed image, and convolutional layers in the CNN, respectively, D is the size of the largest convolutional kernel, and N_{ch} is the maximum filter dimension). This constraint would ordinarily compromise the maximum achievable frame rate. However, the use of GPUs combined with recent deep learning advances allow us to benefit from the deep learning approach without compromising the desired frame rate dictated by the 10 Hz laser pulse repetition frequency.

Future possible improvements to the proposed deep learning visual servoing system include mitigating tracking errors obtained with larger lateral displacements from the center of the probe and increasing the number of degrees of freedom for the motion of the robot end effector. Regarding tracking error mitigation, an increase in the lateral displacement of the source from the center of the probe during the *ex vivo* experiments resulted in needle tracking errors increasing from 0.46 mm to 1.39 mm (Fig. 5) and needle tracking failure rates increasing to a maximum of 3.70% (Table III). This increase in error may potentially be resolved by improving the training process and by increasing the number of training images containing sources with large lateral offsets [24]. Regarding the tracking degree of freedom, the nominal motion of the robot end effector is limited to 1 dimension in our visual servoing system, and a second dimension is used to search for and find the tool tip when it is not in the imaging plane of the probe. While these two degrees of freedom sufficiently achieve the desired end result, future work will determine the extent to which additional degrees of freedom are necessary to achieve more complicated path planning outcomes with the proposed deep learning-based photoacoustic visual servoing system.

VI. Conclusion

This work is the first to demonstrate the integration of deep learning-based techniques with photoacoustic-based robotic visual servoing of needle tips. The deep learning-based needle tip detection process is more accurate (e.g., 0.46-1.39 mm needle tracking errors) and produces lower failure rates (e.g., 0-3.70%) when compared to the alternative photoacoustic image segmentation-based visual servoing system (which produced tracking errors and failure rates of 0.59-5.36 mm and 0-7.02%, respectively). The deep learning-based system additionally maintains the frame rates achieved with the segmentation-based approach. Overall, these results demonstrate the promise of a robotic photoacoustic visual servoing system that bypasses traditional image formation and segmentation steps, instead supplying robot controller input based on details contained within raw photoacoustic sensor data. While this work focuses on tracking needle tips, the system described herein can be extended to track the tips of catheters and a multitude of other surgical tools that are critical to automated surgeries and interventional procedures.

REFERENCES

[1] F. Chaumette and S. Hutchinson, "Visual servo control. I. Basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[2] ——, "Visual servo control. II. Advanced approaches [Tutorial]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 109–118, 2007.

[3] A. Pourtaherian, F. G. Zanjani, S. Zinger, N. Mihajlovic, G. C. Ng, H. H. M. Korsten, and P. H. N. de With, "Robust and semantic needle detection in 3D ultrasound using orthogonal-plane convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 9, pp. 1321–1333, 2018.

[4] L. A. Groves, B. VanBerlo, T. M. Peters, and E. C. S. Chen, "Deep learning approach for automatic out-of-plane needle localisation for semi-automatic ultrasound probe calibration," *Healthcare Technology Letters*, vol. 6, no. 6, pp. 204–209, 2019.

[5] K. Mathiassen, K. Glette, and O. J. Elle, "Visual servoing of a medical ultrasound probe for needle insertion," in *Proceedings of the IEEE International Conference on Robotics and Automation*. IEEE, 2016, pp. 3426–3433.

[6] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Convolution neural networks for real-time needle detection and localization in 2D ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 5, pp. 647–657, 2018.

[7] M. A. Lediju, M. J. Pihl, J. J. Dahl, and G. E. Trahey, "Quantitative assessment of the magnitude, impact and spatial extent of ultrasonic clutter," *Ultrasonic Imaging*, vol. 30, no. 3, pp. 151–168, 2008.

[8] G. T. Clement and K. Hynynen, "A non-invasive method for focusing ultrasound through the human skull," *Physics in Medicine & Biology*, vol. 47, no. 8, p. 1219, 2002.

[9] E. A. Gonzalez, A. Jain, and M. A. L. Bell, "Combined ultrasound and photoacoustic image guidance of spinal pedicle cannulation demonstrated with intact ex vivo specimens," *IEEE Transactions on Biomedical Engineering*, 2020.

[10] L. Gesualdo, L. Cormio, G. Stallone, B. Infante, A. M. D. Palma, P. D. Carri, M. Cignarelli, O. Lamacchia, S. Iannaccone, S. D. Paolo, L. Morrone, F. Aucella, and G. Carrieri, "Percutaneous ultrasound-guided renal biopsy in supine antero-lateral position: a new approach for obese and non-obese patients," *Nephrology Dialysis Transplantation*, vol. 23, no. 3, pp. 971–976, 2008.

[11] M. T. Graham, J. Huang, F. Creighton, and M. A. L. Bell, "Simulations and human cadaver head studies to identify optimal acoustic receiver locations for minimally invasive photoacoustic-guided neurosurgery," *Photoacoustics*, p. 100183, 2020.

[12] M. Graham, F. Assis, D. Allman, A. Wiacek, E. Gonzalez, M. Gubbi, J. Dong, H. Hou, S. Beck, J. Chrispin, and M. A. L. Bell, "In vivo demonstration of photoacoustic image guidance and robotic visual servoing for cardiac catheter-based interventions," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, p. 1015–1029, 2020. [Online]. Available: http://dx.doi.org/10.1109/TMI.2019.2939568

[13] M. A. L. Bell and J. Shubert, "Photoacoustic-based visual servoing of a needle tip," *Scientific Reports*, vol. 8, no. 1, p. 15519, 2018.

[14] M. Xu and L. V. Wang, "Photoacoustic imaging in biomedicine," *Review of Scientific Instruments*, vol. 77, no. 4, p. 041101, 2006.

[15] M. A. L. Bell, "Photoacoustic imaging for surgical guidance: Principles, applications, and outlook," *Journal of Applied Physics*, vol. 128, no. 6, p. 060904, 2020.

[16] B. Eddins and M. A. L. Bell, "Design of a multifiber light delivery system for photoacoustic-guided surgery," *Journal of Biomedical optics*, vol. 22, no. 4, p. 041011, 2017.

[17] J. Shubert and M. A. L. Bell, "A novel drill design for photoacoustic guided surgeries," in *Photons Plus Ultrasound: Imaging and Sensing 2018*, vol. 10494. International Society for Optics and Photonics, 2018, p. 104940J.

[18] M. Allard, J. Shubert, and M. A. L. Bell, "Feasibility of photoacoustic-guided teleoperated hysterectomies," *Journal of Medical Imaging*, vol. 5, no. 2, p. 021213, 2018.

[19] E. A. Gonzalez and M. A. L. Bell, "GPU implementation of photoacoustic short-lag spatial coherence imaging for improved image-guided interventions," *Journal of Biomedical Optics*, vol. 25, no. 7, p. 077002, 2020.

[20] A. Reiter and M. A. L. Bell, "A machine learning approach to identifying point source locations in photoacoustic data," in *Photons Plus Ultrasound: Imaging and Sensing 2017*, vol. 10064. International Society for Optics and Photonics, 2017, p. 100643J.

[21] D. Allman, A. Reiter, and M. A. L. Bell, "A machine learning method to identify and remove reflection artifacts in photoacoustic channel data," in *Proceedings of the IEEE International Ultrasonics Symposium*. IEEE, 2017, pp. 1–4.

[22] ——, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1464–1477, 2018.

[23] D. Allman, F. Assis, J. Chrispin, and M. A. L. Bell, "Deep neural networks to remove photoacoustic reflection artifacts in ex vivo and in vivo tissue," in *Proceedings of the IEEE International Ultrasonics Symposium*. IEEE, 2018, pp. 1–4.

[24] ——, "A deep learning-based approach to identify in vivo catheter tips during photoacoustic-guided cardiac interventions," in *Photons Plus Ultrasound: Imaging and Sensing 2019*, vol. 10878. International Society for Optics and Photonics, 2019, p. 108785E.

[25] M. A. L. Bell, "Deep learning the sound of light to guide surgeries," in *Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XVII*, vol. 10868. International Society for Optics and Photonics, 2019, p. 108680G.

[26] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an open-source Robot Operating System," in *ICRA Workshop on Open Source Software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[29] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of Biomedical Optics*, vol. 15, no. 2, p. 021314, 2010.

[30] D. Allman, A. Reiter, and M. A. L. Bell, "Exploring the effects of transducer models when training convolutional neural networks to eliminate reflection artifacts in experimental photoacoustic images," in *Photons Plus Ultrasound: Imaging and Sensing 2018*, vol. 10494. International Society for Optics and Photonics, 2018, p. 104945H.

[31] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron," https://github.com/facebookresearch/detectron, 2018.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[33] J. Suri, "Detectron-ROS," https://github.com/justice-suri/detectron_ros, 2019.