Autonomous Marker Localization for Lumbar Epidural Steroid Injection Robot

1st Zixuan Liu
Department of Computer Science
Johns Hopkins University
Baltimore, USA
zliu189@jhu.edu

2nd Shuyuan Wang
Department of Mechanical Engineering
Johns Hopkins University
Baltimore, USA
swang340@jhu.edu

3rd Depeng Liu
School of Biomedical Engineering
Shanghai Jiao Tong University
Shanghai, China
liudepeng@sjtu.edu.cn

4thGang Li, Ph.D.
Sheikh Zayed Institute for
Pediatric Surgical Innovation
Children's National Hospital
Washington DC, US
gli2@childrensnational.org

5th Iulian I. Iordachita Ph. D.
Department of Mechanical Engineering
Johns Hopkins University
Baltimore, USA
iordachita@jhu.edu

***Abstract*—In this paper, Hough Transform, Vision Transformer (ViT), and 3D U-Net are utilized to solve the difficulties of autonomous marker localization for lumbar Epidural Steroid Injection (ESI) robots in both Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) images. In CT images, the error recognized by 3D U-Net is 0.69 ± 0.28 mm, which is more accurate than the Hough Transform ($>$ 5.00 mm) and close to the manual marking error 0.63 ± 0.18 mm. In MRI images, the error in recognition achieved by the ViT plus 3D U-Net can reach 1.02 ± 0.47 mm, which is close to the result of 0.93 ± 0.44 mm from the Hough Transform, but slightly higher than the manual marking error 0.44 ± 0.43. This study has the potential to enhance clinical treatment efficiency and holds a certain value for surgical robot localization and registration in medical images.**

***Index Terms*—MRI, CT, ViT, 3D U-Net, Hough Transform, Segmentation**

I. INTRODUCTION

Low back pain poses a considerable therapeutic challenge, standing as one of the foremost five prevalent motives compelling medical consultations in the United States [1]. Traditional lumbar injections use X-ray imaging procedures for guidance, such as Fluoroscopy and CT, which however expose both patients and physicians to ionizing radiation [2]. Magnetic Resonance Imaging (MRI), on the other hand, is the ideal imaging modality for lumbar injections [3], and it provides high-resolution soft tissue contrast and anatomical details without exposing patients or clinicians to radiation, which is particularly crucial for the lumbar region and the reproductive organs of pediatric patients [4]. Nonetheless, in comparison to CT-guided lumbar injections, MRI imaging acquisition is more time-consuming and it incurs a higher cost.

Therefore, both X-ray and MRI-compatible robots have been developed with a specific focus on operating within the X-ray and MRI environments [5], [6], [7], [8]. However, most of the robot registration procedures rely on manual marking by surgeons or operators. For example, Monfaredi *et al.* [7] proposed a shoulder-mounted robot for MRI-guided needle placement, wherein the reference point of their robot had to be manually selected and marked on the MRI images. Similarly, the patient-mounted robotic platform proposed by Maurin *et al.* also required to mark the reference point on the graphic interface [8]. To simplify the clinical workflow, the previous work [9], [10] developed a body-mounted robot to perform lumbar ESI surgeries. However, they still need to mark the cylindrical fiducial markers of the robot in MRI images manually.

Some studies tried to develop autonomous marking methods. Krigger *et al.* [11] proposed both active and passive methods for the registration of cylindrical fiducial markers in MRI scans. Tokuda *et al.* [12], [13] introduced a Z-shaped frame marker technique, which eliminates the need to manually identify cylindrical markers. However, these methods require complex preparation and special materials, and cannot be easily applied to ESI robots.

In this paper, we proposed different algorithms capable of automatically recognizing the fiducial markers of lumbar ESI robots within CT and MRI images, with the aim of reducing potential errors associated with manual marking and enhancing surgical efficiency. The algorithms encompassed: 1) the use of Hough Transform for identifying and localizing metal balls and fiducial markers in CT and MRI images; 2) a 3D U-Net to segment metal balls within CT images; 3) hybrid ViT and 3D U-Net to segment fiducial markers within MRI images. To the best of our knowledge, our work represents the first attempt to achieve fully autonomous marker localization for the lumbar epidural steroid injection robot. The adaptability of different approaches in medical scenarios is discussed.

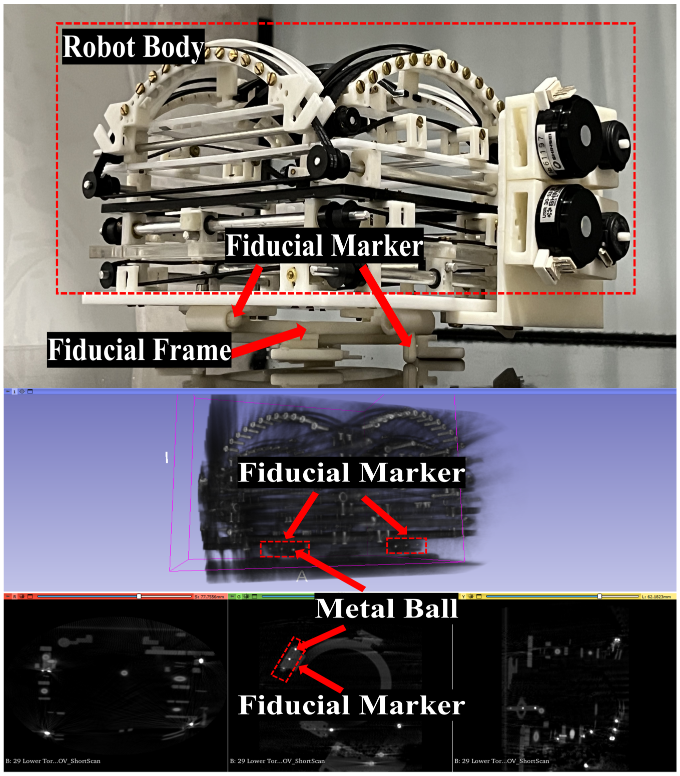

Fig. 1. Robot structure and its corresponding CT image is shown in 3D Slicer. **(a)** CT and MRI-compatible robot for low back pain surgery. The marker section, located at the bottom of the robot, consists of a circular disc surrounded by four cylindrical fiducial markers. Each cylindrical fiducial marker encompasses three metal balls internally, serving the purpose of facilitating the registration of the robot's coordinates within the image. **(b)** The 3D CT image of the robot with circular disc. The marker section is identified in red. **(c)** The CT image processed using 3D Slicer. The extracted 2D CT slices from three distinct angles: coronal, sagittal, and axial. One cylindrical fiducial marker group of metal balls is shown as an example.

II. Method

A. Image Acquisition and Pre-processing

The robot's structure is depicted in Figure 1, featuring four cylindrical fiducial markers embedded within the fiducial frame. Three of the cylindrical fiducial markers are placed horizontally and the last one is set to be vertical to help spatial registration. Each of these cylindrical fiducial markers contains three uniformly distributed metal balls. Their central positions, namely the coordinates of the middle balls from each group, served as reference points during the registration procedures. The ground truth for identifying and segmenting metal balls and cylindrical fiducial markers in both CT and MRI images is manually annotated. Based on our experience, the average time of marking the fiducial markers for one registration is longer than 20 minutes, which is similar to the laser facial registration time (16.7 ± 2.3 min) reported by Machetanz et al [14].

B. 2D Hough Transform

The Hough Transform, especially the circular variant proposed by Pedersen [15], is used to detect circles by converting the image space of circles into parameter space. In three-dimensional space, the CT image projection of a metal ball forms a circular shape, irrespective of the scanning direction. Cross-sectional views of a fiducial marker, as captured in MRI image slices, appear initially as a point, evolve into a full circle, and subsequently revert to a point. By assessing each boundary point, potential circle centers receive votes, and the center with the most votes is selected. The Hough transform localizes the center of the detected circles, denoted as (x_i,y_i) for slice i, and is calculated as follows:

$$(x,y,z) = (\frac{\sum_{i=1}^{n} x_i}{N_x}, \frac{\sum_{i=1}^{n} y_i}{N_y}, \frac{z_1+z_2}{2}) \tag{1}$$

, where x, y, and z represent the center position, N_x and N_y represent the total number of the consequent 2D slices that detected with circles.

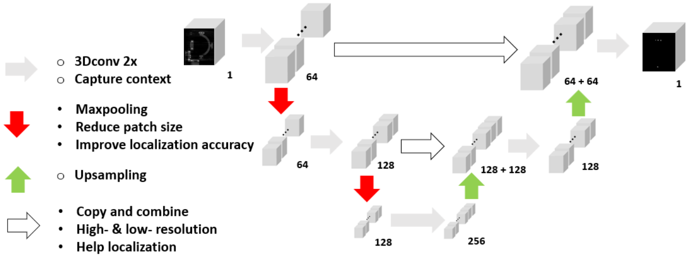

Fig. 2. We utilize the 3D U-Net architecture with a contracting and expansive path for marker localization from CT input. Skip connections between layers capture both local and global contexts, improving object segmentation accuracy.

C. ViT and 3D U-Net

In this work, the metal balls in CT images are manually selected by using the 3D Slicer [16], the occupied area ratio between metal balls and the background is about 1 : 24000 for the original robot CT image and 1 : 1100 for the selected metal ball section. Hence, we introduce the Focal Tversky Loss (FTL) to address the data imbalance in image segmentation [17]. In (2), α, β, and λ are set to be 0.7, 0.3, and $\frac{4}{3}$ respectively, as we found that it can best balance the optimization direction from the True Positive (TP), False Negative (FN), and False Positive (FP) regions from the prediction and the ground truth. For segmentation, we used U-Net, a popular segmentation model in medical image segmentation [18]. As shown in Figure 2, the U-Net we use consists of an encoder-decoder structure with skip connections, allowing for effective feature extraction and preservation of spatial information.

$$FTL = (1 - \frac{TP}{TP+\alpha FN+\beta FP})^{1/\lambda} \tag{2}$$

For the MRI scans of the robot, while the cylindrical fiducial markers make up only 0.5% of the total volume, they represent less than 20% in each 2D slice. Instead of segmenting the full 3D volume with the 3D U-Net, targeting the volume segments containing the cylindrical fiducial markers is more efficient.

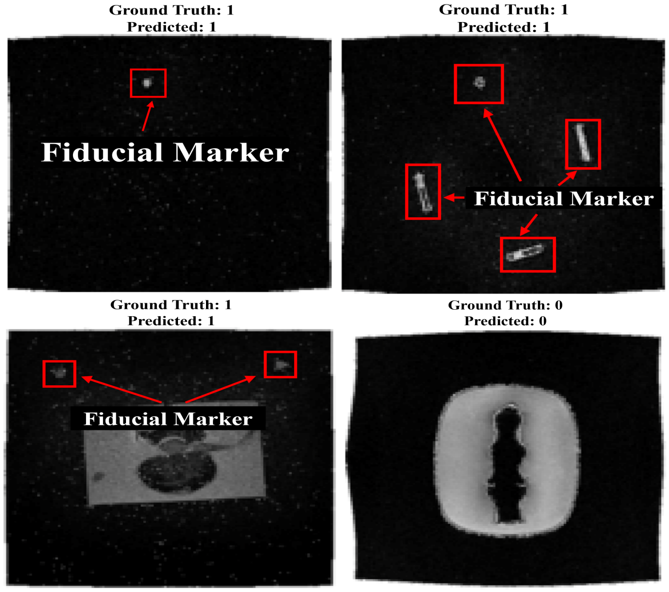

Fig. 3. ViT Sample Results with MRI imags: 1 means the current slice contains fiducial marker(s), and the locations are indicated by red rectangles, while 0 is the opposite.

Therefore, an additional step to identify the volume segments that contain the cylindrical fiducial markers is applied.

Given the fixed spatial positioning of the cylindrical fiducial markers within the robot, MRI slices used for segmentation can be selected based on this known information. Traditional CNNs cannot effectively learn the spatial information of images. Hence, we opted for the ViT for the initial slice selection [19] and classified MRI images into two categories: with and without cylindrical fiducial markers. ViT adapts the Transformer model, originally developed for natural language processing tasks, for image recognition tasks by leveraging its ability to learn intricate relationships between different regions of an image, enabling robust feature extraction and representation. The ViT achieved remarkable performance on various image classification benchmarks and demonstrated the potential of using Transformers in computer vision applications. After classifying MRI images with ViT, the 3D U-Net is used to segment areas with cylindrical fiducial markers in the identified images.

III. Experiment and Result

A. CT and MRI Dataset

To simulate surgical scenarios, CT and MRI images are collected during real surgical procedures. The CT dataset includes 28 fully scanned robot scenes captured by the Loop-X (Brainlab, Munich, Germany) in various positions and postures with the pixel size 1385×1385×1386. The data is split for training, validation, and testing at a ratio of $18 : 5 : 5$.

The MRI dataset consists of images covering coronal, sagittal, and axial planes obtained from five different scans, each conducted at a distinct robotic angle. These images were acquired using a 1.5T MRI scanner (Aera, Siemens, Germany) with an original pixel resolution of 288×384, containing nine slices per MRI image. The dataset has been divided into training, validation, and testing subsets in a $5 : 3 : 2$ ratio.

TABLE I
Evaluation of multiple methods

Method	CT Error / mm	Method	MRI Error / mm
Hough Transform	> 5.00	Hough Transform	0.93 ± 0.44
3D Unet	0.69 ± 0.28	ViT + 3D Unet	1.02 ± 0.47
Manual Marking	0.63 ± 0.18	Manual Marking	0.44 ± 0.43

B. Learning-based Workflow and Experiment

In the case of 3D CT images, each of them is divided into multiple 3D sub-patches with size 70, and these patches are inputted into the 3D U-Net architecture to extract metal ball information from a global perspective. A median filter is applied to the output model to eliminate potential noise signals misclassified as the metal ball. For the 3D U-Net, it is trained by using Google Colab's A-100 GPU. For MRI images, the approach involves sending 2D CT slices to a ViT to identify the fiducial marker section within the global picture. The ViT is trained by a computer with a GPU RTX 3090Ti and CPU Intel i9-13900K. Subsequently, the slices that are classified with fiducial markers are reconstructed into a 3D model and processed by the same 3D U-Net for fiducial marker localization. This approach improves computational efficiency by reducing the computational load associated with 3D convolutions.

The 3D-Unet is trained for 100 epochs. The Adam optimizer and the FTL discussed before are utilized with a fixed learning rate of 0.001. On the other hand, the ViT training is conducted with images resized to 128×128 pixels, and 16×16 pixel patches. A total of 100 epochs are employed for training, using the cross entropy loss function. The training utilizes a StepLR scheduler with a step size of 1 and a gamma value of 0.7, alongside a learning rate of 0.0002.

C. Results

The error in manual marking is calculated by repeating the same marking action 30 times, while the error in other algorithms is calculated directly against the average value of manual marking, the ground truth. For CT images by using the 3D U-Net, the metal ball localization task takes an average of approximately 11.2 seconds to complete. The average registration error is 0.69 mm, as shown in Table I.

After 45 epochs of ViT training, the highest accuracy in classifying whether MRI slices contain cylindrical fiducial markers reaches 85.2%, with no false negative cases in the test dataset. On average, this procedure takes 6.1 seconds to complete and yields an error of approximately 1.02 mm. Sample results are illustrated in Figure 3.

IV. Discussion and Future Work

We proposed and compared different algorithms that enable robot marker localization in medical images. They have different advantages in different scenarios, which can reduce the human error rate for manual marking and make the surgical process more efficient. The advantage of the Hough Transform is its speed or rapid processing. However, even if the Hough Transform can identify the center coordinates of the marker with easy implementation, its results tend to be more error-prone. Additionally, it is difficult to manage scenarios where the fiducial marker is oriented differently, as the Hough Transform cannot detect cylindrical fiducial markers when they are not oriented perpendicular to the MRI images. Moreover, when the Hough Transform is applied to the CT image, other components of the robot in the CT image give tremendous interference for the metal ball detection, which causes a large error in localization. The Learning-based method offers increased reliability compared to traditional computer vision methods. However, a significant concern lies in its generalizability to other surgery robots and its application in multiple image types (e.g., ultrasound). While Hough methods are more accommodating to marker detection in various tasks, the versatility of ViT heavily relies on the training set. To enhance our model's robustness, we plan to use transfer learning and expand our medical image dataset. Balancing localization accuracy and context in training the 3D U-Net is challenging: larger patches provide better context but compromise accuracy. The network sometimes misidentifies robot parts, requiring potential median filter application or ViT preprocessing. Given the GPU RAM computational demands, we'll refine hyperparameters and enhance hardware to optimize model performance.

V. Conclusion

This study introduces both learning-based techniques and Hough Transform for marker localization in CT and MRI contexts and specifies the best practices for their use in different scenarios. The Hough Transform provides quick and accurate results with less interference in medical images, while the learning-based methods may perform better in such conditions. For MRIs focusing on speed and simplicity, the Hough Transform is suitable. However, for most other situations, ViT and 3D Unet are recommended; they quickly and accurately auto-annotate markers than manual marking, which are 11.2 seconds for a CT image and 6.1 seconds for a set of MRI images that contain the whole robot, and much shorter than manual marking time that is more than 20 minutes.

References

[1] Ivan Urits, Aaron Burshtein, Medha Sharma, Lauren Testa, Peter A Gold, Vwaire Orhurhu, Omar Viswanath, Mark R Jones, Moises A Sidransky, Boris Spektor, et al. Low back pain, a comprehensive review: pathophysiology, diagnosis, and treatment. *Current pain and headache reports*, 23:1–10, 2019.

[2] Umile Giuseppe Longo, Mattia Loppini, Luca Denaro, Nicola Maffulli, and Vincenzo Denaro. Rating scales for low back pain. *British medical bulletin*, 94(1):81–144, 2010.

[3] John Bui and Nikolai Bogduk. A systematic review of the effectiveness of ct-guided, lumbar transforaminal injection of steroids. *Pain Medicine*, 14(12):1860–1865, 2013.

[4] Kevin A Smith and John Carrino. Mri-guided interventions of the musculoskeletal system. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(2):339–346, 2008.

[5] Hao Su, Ka-Wai Kwok, Kevin Cleary, Iulian Iordachita, M. Cenk Cavusoglu, Jaydev P. Desai, and Gregory S. Fischer. State of the art and future opportunities in mri-guided robot-assisted surgery and interventions. *Proceedings of the IEEE*, 110(7):968–992, 2022.

[6] Septimiu E Salcudean, Hamid Moradi, David G Black, and Nassir Navab. Robot-assisted medical imaging: A review. *Proceedings of the IEEE*, 110(7):951–967, 2022.

[7] Reza Monfaredi, Iulian Iordachita, Emmanuel Wilson, Raymond Sze, Karun Sharma, Axel Krieger, Stanley Fricke, and Kevin Cleary. Development of a shoulder-mounted robot for mri-guided needle placement: phantom study. *International journal of computer assisted radiology and surgery*, 13:1829–1841, 2018.

[8] Benjamin Maurin, Bernard Bayle, Olivier Piccin, Jacques Gangloff, Michel de Mathelin, Christophe Doignon, Philippe Zanne, and Afshin Gangi. A patient-mounted robotic platform for ct-scan guided procedures. *IEEE Transactions on Biomedical Engineering*, 55(10):2417–2425, 2008.

[9] Gang Li, Niravkumar A Patel, Weiqiang Liu, Di Wu, Karun Sharma, Kevin Cleary, Jan Fritz, and Iulian Iordachita. A fully actuated body-mounted robotic assistant for mri-guided low back pain injection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5495–5501. IEEE, 2020.

[10] Gang Li, Niravkumar A Patel, Jan Hagemeister, Jiawen Yan, Di Wu, Karun Sharma, Kevin Cleary, and Iulian Iordachita. Body-mounted robotic assistant for mri-guided low back pain injection. *International journal of computer assisted radiology and surgery*, 15(2):321–331, 2020.

[11] Axel Krieger, Robert C Susil, Cynthia Ménard, Jonathan A Coleman, Gabor Fichtinger, Ergin Atalar, and Louis L Whitcomb. Design of a novel mri compatible manipulator for image guided prostate interventions. *IEEE Transactions on Biomedical Engineering*, 52(2):306–313, 2005.

[12] Junichi Tokuda, Gregory S Fischer, Simon P DiMaio, David G Gobbi, Csaba Csoma, Philip W Mewes, Gabor Fichtinger, Clare M Tempany, and Nobuhiko Hata. Integrated navigation and control software system for mri-guided robotic prostate interventions. *Computerized Medical Imaging and Graphics*, 34(1):3–8, 2010.

[13] Junichi Tokuda, Sang-Eun Song, Kemal Tuncali, Clare Tempany, and Nobuhiko Hata. Configurable automatic detection and registration of fiducial frames for device-to-image registration in mri-guided prostate interventions. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part III 16*, pages 355–362. Springer, 2013.

[14] Kathrin Machetanz, Florian Grimm, Martin Schuhmann, Marcos Tatagiba, Alireza Gharabaghi, and Georgios Naros. Time efficiency in stereotactic robot-assisted surgery: an appraisal of the surgical procedure and surgeon's learning curve. *Stereotactic and Functional Neurosurgery*, 99(1):25–33, 2021.

[15] Simon Just Kjeldgaard Pedersen. Circular hough transform. *Aalborg University, Vision, Graphics, and Interactive Systems*, 123(6):2–3, 2007.

[16] 3d slicer. In *https://www.slicer.org/*, 2019.

[17] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.