IMPERIAL

SURG70006 Group Project
2024/10

Project 19
Surgical Robot Instrument Pose Estimation

Group Number: Group 8

Group Members: Jie Li, Jinling Qiu, Leen AIShekh, Yanrui Liu, Yulin Huang

Supervisor Name: Dr Stamatia (Matina) Giannarou

DEPARTMENT OF
Surgery and Cancer

Imperial College London

Contents

1 Introduction

1.1 Minimally invasive surgery and Robotic surgery(TBD)

1.2 Problem Statement

There are some vision-based methods that use external markers to track the instruments. However, these methods have major limitations; the markers must always be visible in the camera's field of view and are sensitive to background changes and occlusions[1]. In this case, a vision-based markerless instrument tracking method that does not require any modifications to the hardware setup or external markers is necessary.

1.3 Engineering Background

The da Vinci surgical robot operates with Six Degrees of Freedom (6DoF). These 6DoF refer to the robot's ability to move and rotate in three-dimensional space, including three translational movements (up/down, left/right, forward/backward) and three rotational movements (pitch, yaw, and roll)[2]. This range of motion allows the da Vinci system to replicate the complex dexterity of a surgeon's hand for precise control over surgical instruments in confined spaces. The main aim of this project is to develop a deep learning based markerless 6DoF surgical instrument pose estimation system. The system will be designed to provide highly accurate surgical instrument 6DoF estimation without relying on external markers or complex hardware.

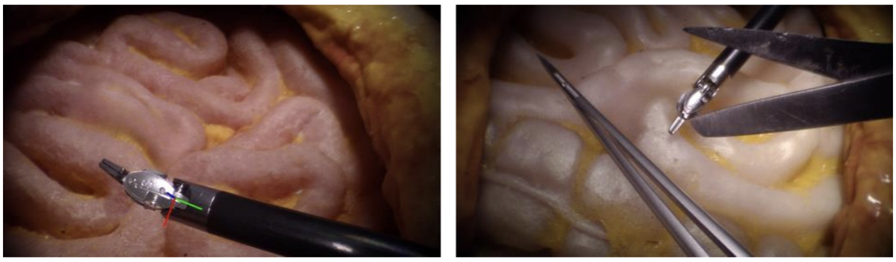


Figure 1: 6DoF surgical instrument pose estimation with (left) and without occlusion (right). [3]

Pose estimation refers to finding the transformation (translation and rotation) that relates the object (or camera) coordinates in 3D space to its projection on a 2D image. The pose estimation of surgical tools has emerged as a critical job in Robot-assisted minimally invasive surgery (RMIS). The majority of robots in RMIS are driven by cables, resulting in kinematic input that is not always precise, since the kinematic data describes the positions of the motor rather than the real position of the joints connected to the motor via a cable[1].

Optical Tracking System (OTS) and Electromagnetic Tracking System (EMTS) are well-established methods for tracking in medical applications. OTS offers high accuracy but requires a clear line-of-sight, making it prone to errors when obstructed. EMTS, while effective without line-of-sight, suffers from interference caused by metal objects and electronic devices in the operating room, leading to reduced accuracy[4].

In RMIS, a marker-based method involves placing artificial markers on surgical instruments to aid vision-based instrument tracking[5]. These markers can often be easily recognised in complex surgical environments. However, if the markers are obscured, damaged, or obscured by blood coverage, this may result in detection failure[6]. In addition, markers on the surface of surgical instruments must meet sterility requirement[7]. To address these issues, marker-less methods have been gradually proposed[8]. marker-less methods do not rely on artificial markers in endoscopic procedures, but rather on natural features of the surgical instruments for gesture estimation. This method does not require additional marking process for the instruments and is able to adapt to various environmental changes with higher flexibility. However, the current marker-less method still faces some challenges, such as being susceptible to interference from lighting conditions, blood occlusion, and instrument reflections, and may not perform as consistently as the marker-based method in complex scenes[9].

1.4 Problem Statement(TBD)

RMIS has come significantly in the last decade due to advances in surgical robotics such as artificial intelligence and the da Vinci Platform (da Vinci Platform). Pose estimation of surgical instruments has become an important task in RMIS. Nowadays there are many external devices like depth camera, electromagnetic trackers etc. available for space estimation in surgical instruments but they are not practical in in-vivo surgeries because of space and hardware constraints[10].

2 Related Work

2.1 Traditional non-learning methods

Traditional non-learning pose estimation methods are based on geometric modeling, algebraic techniques, and computer vision approaches[11]. Non-learning methods differ from deep learning methods, which require a large amount of data as input. Typically, the core of traditional non-deep learning methods is featuring extraction, which is used to obtain a unique representation of an object by identifying edges, key points, or regional features[11]. In the early days, the main methods for feature recognition include SIFT[12], which is a classical local feature descriptor that helps machines to identify and match feature points in different images, and to find key points in different scale-spaces. Another is the SURF[13], which is an improved variant of SIFT that improves the performance of feature extraction by optimising the process of feature detection and description.

After feature extraction is completed, the target model needs to be parameterised. The Perspective-n-Point (PnP) problem, introduced by Fischler in the 1980s, is a commonly used algorithm for model parameterisation,. It focuses on estimating a camera's position and orientation using 3D-to-2D point correspondences [14]. Perspective-3-Point (P3P) solves this with three such correspondences, providing up to four solutions. PnP handles larger datasets by representing n 3D points as a weighted sum of four virtual control points, reducing computational complexity [15]. Another method is the ICP algorithm[16], which computes the pose relationship between two-point clouds by minimising the distance between corresponding points. Another method is the ICP algorithm[16], which computes the pose relationship between two-point clouds by minimising the distance between corresponding points. Non-learning methods have better interpretability and can achieve more accurate results while saving computational resources. However, such methods are less robust when dealing with complex scenes and lighting changes and have certain limitations[16].

2.2 Deep Leaning Method

In recent years, with the development of deep learning, it has been gradually applied to the field of surgical instrument pose estimation. Different from the traditional way that relies on geometric models and manual feature extraction, deep learning methods can infer the complex relationship between points and points from a large amount of data [16], and according to the facts, deep learning methods are more capable of handling complex scenes with lighting changes[11]. Through the concern classification of the model, it can be divided into Holistic method and Intermediate representation method.

2.2.1 Holistic Methods

The Holistic method extracts estimated surgical instrument poses by modelling global features of the entire scene[11]. This approach does not rely on local detail features, but rather extracts pose information from global features making the Holistic method highly robust to complex scene variations. In 2015, Alex Kendall and his team proposed PoseNet, a deep learning method that directly regresses camera pose from monocular RGB images, enabling end-to-end position and orientation end-to-end estimation[17]. In 2020, Yannick Bukschat et al. proposed EfficientPose, an end-to-end 6D multi-target pose estimation method. The model is capable of simultaneously detecting the 2D bounding boxes of multiple targets in a monocular image and regressing their complete 6D poses in 3D space[18]. In 2022, Bo Chen and colleagues developed the ROPE framework, which introduces a new occlusion enhancement technique and a multi-precision supervised mechanism, aiming to learn deep features that are robust to occluded environments, thus improving the accuracy of pose estimation in object-occluded scenes[19]. The Holistic method is able to capture the overall features of an object directly from the whole image, with a low dependence on feature points, without the need for precise positioning of feature points or additional feature extraction steps, which makes the model structure more concise[19]. However, the Holistic method is less accurate in dealing with local details, and when surgical instruments are occluded, it is difficult to recover the occluded instrument information from the overall features[20].

2.2.2 Intermediate Representations

The Intermediate Representation method decomposes the complex pose estimation task into multiple more manageable subtasks by introducing a finer-grained intermediate description of the target. By extracting local features, the method effectively solves the problem that it is difficult to accurately estimate the pose of surgical instruments when they are occluded[21]. In 2017, Yu Xiang and his team proposed PoseCNN for pose estimation in complex scenes. The method decomposes the pose estimation task into multiple components that deal with 3D translations and rotations of images separately. In addition, PoseCNN introduces a novel loss function that allows the network to better handle objects with symmetry[22]. Subsequently, in 2019, Sida Peng and his team proposed PVNet[23]. this approach uses a pixel-level voting network that significantly improves pose estimation accuracy in occluded and truncated scenes by predicting vectors from each pixel to a key point, combined with a RANSAC-based voting mechanism. In 2020, Masakazu Yoshimura and his team developed a deep learning model based on an improved SSD-6D architecture[24]. The model utilises a manually generated dataset of single-frame endoscopic images combined with data enhancement techniques to effectively address occlusion and perspective distortion problems common in surgical environments. In 2022, Mitchell Doughty and his team proposed HMD-EgoPose[24]. The method uses the EfficientDet-D0 network for multi-scale feature extraction and combines rotational, translational, and hand sub-networks to achieve 6-degree-of-freedom markerless pose estimation in monocular RGB images. In 2024, Jihun Park and his team introduced a new occlusion-aware loss function based on the YOLOv8 model, which dramatically improved the accuracy of precise detection and pose estimation of key points of surgical instruments in complex occlusion environments[25]. The research team trained the model on a real surgical dataset, which significantly improved its robustness in real surgical scenarios. The Intermediate representation method makes the task much less difficult by decomposing the complex pose estimation task into multiple, more manageable subtasks. At the same time, Intermediate Representation models local features so that the model can still maintain high stability in complex scenes. However, this method requires high accuracy in data labelling, and the accumulation of errors may affect the accuracy of the final results due to the inclusion of multiple intermediate steps[7][26].

3 Goals and Objectives

3.1 Objectives of the Project

This project aims to leverage state-of-the-art deep learning models for pose estimation to accurately determine the rotation and position of surgical tools present in the surgical environment during RMIS. The detailed objectives of the project are as follows:

1. **Dataset Analysis**
   The first objective is to analyze the datasets which include high-quality images captured by the Da Vinci Si endoscopic stereo camera and accurate and consistent ground truth data obtained from the Hamlyn Centre.

2. **Model Conversion**
   This project will convert the PyTorch model to an ONNX model to enable its deployment on ARM architectures, such as the NVIDIA Jetson AGX platform.

3. **Robust Pose Estimation**
   This project also need to devise novel approaches to ensure accurate and robust pose estimation in the presence of challenges such as partial tool visibility, occlusions, and other variations encountered during surgery.

4. **Performance Evaluation**

   Finally, the project will evaluate and validate the performance of the applied models under various degrees of occlusion, ensuring their reliability in practical surgical scenarios.

4 Proposed Methodology

4.1 Model for Pose Estimation

4.2 Model Conversion

In the process of deploying deep learning models on the NVIDIA AGX platform, PyTorch models first need to be converted to the Open Neural Network Exchange (ONNX) format to ensure cross-platform compatibility. The ONNX format is a common intermediate representation that enables model migration and transforming between different frameworks and hardware, so that PyTorch models can be used directly in the embedded environment. glsonnx models can be optimized on the NVIDIA Jetson AGX through TensorRT. TensorRT could improve the inference speed and computational efficiency, especially for embedded platforms such as NVIDIA's Jetson AGX. This optimization process includes multi-threading, pipelining, buffer assignment, and network duplication[27][28]. In addition, the use of docker containers for model deployment is required, allowing model training environment transformation in deploying complex models[29].

4.3 Model Integration with da Vinci

4.4 Hardware and Software Requirements TBD

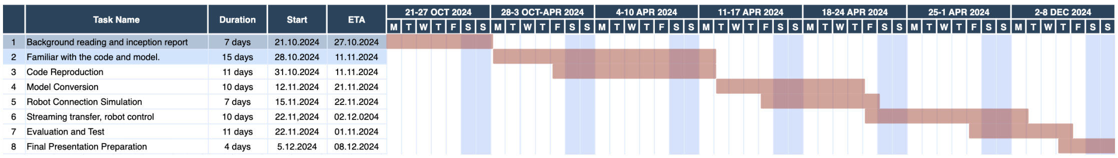
The project requires the NVIDIA AGX Platform for high-performance computing in applications and autonomous machines. It uses Python, C, and C++ languages, and tool packages like PyTorch for deep learning framework and OpenCV for image processing. Additionally, Robot Operating System (ROS) provides a flexible framework for developing and running robot software across multiple platforms. Docker helps the deployment and management of applications across different platforms or systems.

5 Risk Assessment

Risk	**Contingencies**	**Likelihood**	**Impact**
Pose estimation unstable in complex scenes (e.g., occlusion, dynamic background)	Optimizing the dataset with more occlusion scene data	High	Very High
Low Frame Rates	Optimizing model efficiency, invest in high-quality hardware	High	Medium
Model computation time is too long for real-time simulation	Optimizing model efficiency and using hardware acceleration	High	Medium
Reliance on specific deep learning frameworks, leading to migration difficulties	Reduced binding to specific frameworks	Medium	High
Data damage or lost	Using GitHub or external storage devices to make backup	Medium	Very High
Data Privacy	Implementing robust data encryption and comply with data protection laws such as GDPR	Medium	Very High
Ethical and Regulatory	Consulting with regulatory experts and following related regulations	Very Low	High
Timeline delay	Making a detailed timeline and a thorough monitoring plan	High	High

Table 1: Risk Assessment Table

6 Project Timeline



	Task Name	Duration	Start	ETA
1	Background reading and inception report	7 days	21.10.2024	27.10.2024
2	Familiar with the code and model.	15 days	28.10.2024	11.11.2024
3	Code Reproduction	11 days	31.10.2024	11.11.2024
4	Model Conversion	10 days	12.11.2024	21.11.2024
5	Robot Connection Simulation	7 days	15.11.2024	22.11.2024
6	Streaming transfer, robot control	10 days	22.11,2024	02.12.0204
7	Evaluation and Test	11 days	22.11.2024	01.11.2024
8	Final Presentation Preparation	4 days	5.12.2024	08.12.2024

Figure 2: Gantt Chart

7 Project Management

7.1 Progress Monitoring

For code part, we use GitHub for monitoring and progress management. We use GitHub's version control to branch the code (each team member manages a branch independently) to ensure smooth team collaboration. We also manually record project logs(such as meeting minutes and group activities) and combine them with timelines to ensure the project run smoothly.

7.2 Rules of Group Members

References

[1] H. Xu, M. Runciman, J. Cartucho, C. Xu, and S. Giannarou, "Graph-based pose estimation of texture-less surgical tools for autonomous robot control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2731–2737. DOI: 10.1109/ICRA48891.2023.10160287.

[2] L. Bai, J. Yang, X. Chen, *et al.*, "Solving the time-varying inverse kinematics problem for the da vinci surgical robot," *Applied Sciences*, vol. 9, no. 3, 2019, ISSN: 2076-3417. DOI: 10.3390/app9030546. [Online]. Available: https://www.mdpi.com/2076-3417/9/3/546.

[3] S. R. I. P. E. (SurgRIPE), *Surgical robot instrument pose estimation (surgripe)*, Synapse Project, SynID: syn51471789. Available: https://www.synapse.org/Synapse:syn51471789/wiki/622255, Accessed: 2024-10-22, 2024.

[4] A. Sorriento, M. B. Porfido, S. Mazzoleni, *et al.*, "Optical and electromagnetic tracking systems for biomedical applications: A critical review on potentialities and limitations," *IEEE Reviews in Biomedical Engineering*, vol. 13, pp. 212–232, 2020. DOI: 10.1109/RBME.2019.2939091.

[5] F. P. Villani, M. Di Cosmo, Á. B. Simonetti, E. Frontoni, and S. Moccia, "Development of an augmented reality system based on marker tracking for robotic assisted minimally invasive spine surgery," in *International Conference on Pattern Recognition*, Springer, 2021, pp. 461–475.

[6] L. Ma and B. Fei, "Comprehensive review of surgical microscopes: Technology development and medical applications," *Journal of biomedical optics*, vol. 26, no. 1, pp. 010 901–010 901, 2021.

[7] H. Xu, M. Runciman, J. Cartucho, C. Xu, and S. Giannarou, "Graph-based pose estimation of texture-less surgical tools for autonomous robot control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 2731–2737.

[8] R. Reilink, S. Stramigioli, and S. Misra, "3d position estimation of flexible instruments: Marker-less and marker-based methods," *International journal of computer assisted radiology and surgery*, vol. 8, pp. 407–417, 2013.

[9] J. Hein, M. Seibold, F. Bogo, *et al.*, "Towards markerless surgical tool and hand pose estimation," *International journal of computer assisted radiology and surgery*, vol. 16, pp. 799–808, 2021.

[10] J. Cartucho, C. Wang, B. Huang, D. S. Elson, A. Darzi, and S. Giannarou, "An enhanced marker pattern that achieves improved accuracy in surgical tool tracking," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 10, no. 4, pp. 400–408, 2022. DOI: 10.1080/21681163.2021.1997647. eprint: https://doi.org/10.1080/21681163.2021.1997647. [Online]. Available: https://doi.org/10.1080/21681163.2021.1997647.

[11] K. Fan, Z. Chen, Q. Liu, G. Ferrigno, and E. De Momi, "A reinforcement learning approach for real-time articulated surgical instrument 3d pose reconstruction," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[12] K. D. Lakshmi and V. Vaithiyanathan, "Image registration techniques based on the scale invariant feature transform," *IETE Technical Review*, vol. 34, no. 1, pp. 22–29, 2017.

[13] W. Wijesinghe, "Speed up robust features in computer vision systems," 2010.

[14] X. X. Lu, "A review of solutions for perspective-n-point problem in camera pose estimation," *Journal of Physics: Conference Series*, vol. 1087, no. 5, p. 052 009, Sep. 2018. DOI: 10.1088/1742-6596/1087/5/052009. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1087/5/052009.

[15] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *Int. J. Comput. Vision*, vol. 81, no. 2, pp. 155–166, Feb. 2009, ISSN: 0920-5691. DOI: 10.1007/s11263-008-0152-6. [Online]. Available: https://doi.org/10.1007/s11263-008-0152-6.

[16] B. Bellekens, V. Spruyt, R. Berkvens, and M. Weyn, "A survey of rigid 3d pointcloud registration algorithms," in *AMBIENT 2014: the Fourth International Conference on Ambient Computing, Applications, Services and Technologies, August 24-28, 2014, Rome, Italy*, 2014, pp. 8–13.

[17] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[18] Y. Bukschat and M. Vetter, "Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach," *arXiv preprint arXiv:2011.04307*, 2020.

[19] B. Chen, T.-J. Chin, and M. Klimavicius, "Occlusion-robust object pose estimation with holistic representation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2929–2939.

[20] T. L. Watson and R. A. Robbins, *The nature of holistic processing in face and object recognition: Current opinions*, 2014.

[21] C. Song, J. Song, and Q. Huang, "Hybridpose: 6d object pose estimation under hybrid representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 431–440.

[22] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[23] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4561–4570.

[24] M. Yoshimura, M. M. Marinho, K. Harada, and M. Mitsuishi, "Single-shot pose estimation of surgical robot instruments' shafts from monocular endoscopic images," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2020, pp. 9960–9966.

[25] J. Park, J. Hong, J. Yoon, B. Park, M.-K. Choi, and H. Jung, "Towards precise pose estimation in robotic surgery: Introducing occlusion-aware loss," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 639–648.

[26] M. Allan, S. Ourselin, D. J. Hawkes, J. D. Kelly, and D. Stoyanov, "3-d pose estimation of articulated instruments in robotic minimally invasive surgery," *IEEE transactions on medical imaging*, vol. 37, no. 5, pp. 1204–1213, 2018.

[27] E. Jeong, J. Kim, and S. Ha, "Tensorrt-based framework and optimization methodology for deep learning inference on jetson boards," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 21, pp. 1–26, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246288038.

[28] L. S. Karumbunathan, *Nvidia jetson agx orin series*, 2022.

[29] A. Khoshsirat, G. Perin, and M. Rossi, "Divide and save: Splitting workload among containers in an edge device to save energy and time," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2023, pp. 134–138. DOI: 10.1109/ICCWorkshops57953.2023.10283807.

Glossary

6DoF Six Degrees of Freedom 2

da Vinci Platform da Vinci Platform 3

EMTS Electromagnetic Tracking System 3

ICP Iterative Nearest Point 4

ONNX Open Neural Network Exchange 7

OTS Optical Tracking System 3

P3P Perspective-3-Point 4

PnP Perspective-n-Point 4

RMIS Robot-assisted minimally invasive surgery 2, 3, 6

ROS Robot Operating System 7

SIFT Scale Invariant Feature Transform 4

SURF Speeded Up Robust Feature 4