EPiC Series in Health Sciences

Volume 6, 2024, Pages 90–93

Proceedings of The 22nd Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

SimPS-Net: Simultaneous Pose & Segmentation Network of Surgical Tools

Spyridon Souipas[1], Anh Nguyen[2], Stephen G. Laws[1], Brian L. Davies[1], Ferdinando Rodriguez y Baena[1]

Imperial College London, London, United Kingdom
[spyridon.souipas14, stephen.laws14, f.rodriguez, b.davies]@imperial.ac.uk
University of Liverpool, Liverpool, United Kingdom
anh.nguyen@liverpool.ac.uk

Abstract

The ability to detect and localise surgical tools using RGB cameras during robotic assisted surgery can allow for the development of various implementations, such as vision-based active constraints and refinements in robot path planning, which can ultimately lead in improved patient safety during operation. For this purpose, the proposed network, SimPS-Net capable of both detection and 3D pose estimation of standard surgical tools using a single RGB camera, is introduced. In addition to the network, a novel dataset generated for training and testing is presented. The proposed network achieved a mean DICE coefficient of 85.0%, while also exhibiting a low average error of 5.5mm and 3.3° for 3D position and orientation respectively, thus outperforming the competing networks.

1 Introduction

Image-based detection and localisation of surgical tools has received significant attention due to the development of relevant deep learning techniques, along with recent upgrades in computational capabilities [1]. Although not as accurate as optical trackers [2], image-based methods are easy to deploy, and require no surgical tool redesign to accommodate trackable markers, which could be beneficial when it comes to cheaper, "off-the-shelf" tools, such as scalpels and scissors.

In the operating room however, these techniques suffer from drawbacks due to the presence of highly reflective or featureless materials, but also occlusions, such as smoke and blood [3]. Additionally, most localisation networks focus on 2D pose estimation, which in itself cannot offer any useful feedback when it comes to robotic surgery [4]. For 3D localisation, networks often utilise tool 3D models (e.g. CAD data), not only for the purpose of point correspondence, but also for pose regression [5]. The aforementioned "off-the-shelf" tools are scarcely accompanied by such prior 3D structure data. Ultimately, in addition to the above hindrances, estimating 3D pose using a monocular camera setup, poses a challenge in itself due to the lack of depth information. Considering these limitations, SimPS-Net, a network capable of both detection and 3D pose estimation of standard surgical tools using a single RGB camera, is presented.

2 Methods and Materials

The majority of surgical tool datasets examine laparoscopic conditions, thus not involving any of the aforementioned standard tools. Furthermore, there exists a lack of 3D pose labels across these data. Hence, a novel dataset was generated, consisting of monocular RGB images, along with the 3D pose values of the tools present within each frame, in order to train the network. A total of 4 standard surgical tools were chosen, specifically a scalpel, an electric burr, a pair of forceps and a pair of scissors. 5370 images were semantically annotated, with 4027 images employed for training. The images were recorded whilst the tools were being used to operate on a cadaveric knee, thus mimicking the conditions met in the operation room.

Images were collected using a RealSense D415 (Intel, USA). The camera was rigidly mounted on the ftk500 optical tracker (Atracsys, Switzerland), which was utilised to obtain the 3D position and orientation each detected tool. Upon extrinsic calibration, the pose of each tool was converted to camera 3D coordinates, as demonstrated in Equations 1 and 2:

$$\begin{bmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x_A \\ y_A \\ z_A \\ 1 \end{bmatrix} \tag{1}$$

$$\therefore \ \mathbf{p_{cam}} = [\mathbf{R}|\mathbf{t}] \ \mathbf{p_A} \tag{2}$$

In Equation 2, the extrinsic calibration matrix is denoted by $[\mathbf{R}|\mathbf{t}]$, with $\mathbf{p_A}$ and $\mathbf{p_{cam}}$ being tooltip coordinates in optical tracker and camera frame respectively. Similarly, 3D orientation was obtained in camera coordinates. Furthermore, the employment of camera intrinsic information allowed for the projection 3D poses on 2D images, thus enabling SimPS-Net to generate inferences with pose visualisation, as shown in Figure 1.

The presented network is an expansion of Mask-RCNN [6], which incorporates two branches capable of classification and semantic segmentation respectively. The expanded architecture introduces a novel branch, capable of object 3D pose regression using the same RGB image as the other branches, as demonstrated in Figure 1.

3D pose, $\mathbf{p}$, has been characterised as the amalgamation of the position vector, $\mathbf{x}$, and the orientation vector $\boldsymbol{\theta}$. The latter is expressed in the form of quaternions. Equation 3 shows how the true pose, $\mathbf{p_{true}}$, and predicted pose, $\mathbf{p_{pred}}$, are split into position and orientation and hence used to construct the pose loss.

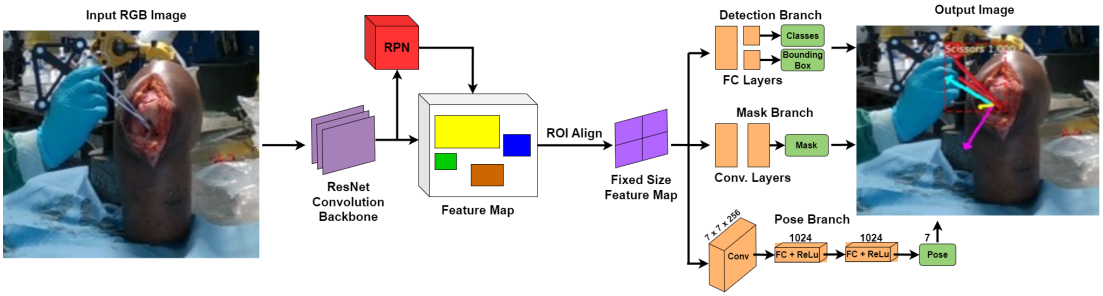

Figure 1: SimPS-Net Architecture

$$\mathcal{L} = \alpha \ \|\mathbf{x_{true}}, \mathbf{x_{pred}}\|_2 + \beta \ \|\boldsymbol{\theta_{true}}, \boldsymbol{\theta_{pred}}\|_2 \tag{3}$$

Relevant work has demonstrated that an orientation constant, β, can improve inference results. [7]. Additionally, α has been incorporated to address the scale discrepancy between orientation and position values.

3 Results

For testing purposes, 806 previously unseen images were utilised for detection and pose estimation. Upon testing various permutations, the optimal constant values were determined to be $\alpha = 700$ and $\beta = 300$.

Detection success was quantified using the mean average precision (mPA) and mean DICE coefficient (mDICE). Pose errors were quantified as the average error along each axis in mm for position and degrees for orientation. Inference results are presented in Table 1 for SimPS-Net. For comparison, three networks, namely PoseNet [7], ROPE [8], and GDR-Net [9], were trained using the generated dataset, with testing results being also listed.

As noted in Table 1, the examined architecture achieves auspicious results, with each position and orientation metric being lower or at least comparable to the state of the art. Conclusively, the average positional and orientation errors were calculated as 5.5mm and 3.3° respectively.

Table 1: SimPS-Net Results Comparison against Literature

Source	**PoseNet** [7]	**ROPE** [8]	**GDR-Net** [9]	**SimPSNet**
mAP (%)	NA	56.8	58.5	62.9
mDICE (%)	NA	80.2	83.7	85
X (mm)	18.4 (11.6)	8.4 (3.5)	6.1 (5.2)	5.2 (4.5)
Y (mm)	18.6 (13.1)	11.4 (6.2)	5.0 (2.4)	4.0 (4.3)
Z (mm)	13.4 (9.2)	9.2 (5.2)	7.3 (3.4)	6.3 (6.0)
Pitch (deg)	2.3 (1.8)	3.2 (2.5)	2.6 (3.1)	2.4 (2.8)
Yaw (deg)	1.3 (1.0)	1.8 (2.1)	2.3 (2.6)	1.5 (1.5)
Roll (deg)	28.2 (28.2)	8.3 (16.2)	6.7 (10.7)	6.1 (37.3)

4 Discussion

Monocular methods for 3D pose estimation are scarce in the context of surgical tools. In the proposed solution, 3D pose estimation is achieved without incorporating prior knowledge, such as 3D structure or shape assumptions. Despite pose metrics not outperforming the capabilities of optical trackers, the presented camera-based solution exhibits minimal footprint and high ease of deployment, thus making it a suitable option for applications such as robot path planning outside the body, where sub-milimiter accuracy is not a prerequisite. Additionally, the robust

results of the network along the depth axis (Z) and the roll orientation suggest that SimPS-Net is indeed suitable for 3D pose application.

Nevertheless, some areas require further investigation. With image occlusions being a significant hindrance in the operating room, an occlusion handling method should be implemented across the detection branch in order to improve relevant metrics. In addition, the novel dataset should be expanded to include images with multiple tools within the same frame. Finally, by employing improved hardware, the network could be deployed in real-time, thus allowing for the development of a vision-based active constraint, which would allow for improved path planning and obstacle avoidance during manipulation of a robotic platform.

References

[1] Congmin Yang, Zijian Zhao, and Sanyuan Hu. Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature, 1 2020.

[2] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical Image Analysis*, 35:633–654, 2017.

[3] Zijian Zhao, Zhaorui Chen, Sandrine Voros, and Xiaolin Cheng. Real-time tracking of surgical instruments based on spatio-temporal context and deep learning. *Computer Assisted Surgery*, 24(sup1):20–29, 2019.

[4] Iro Laina, Nicola Rieke, Christian Rupprecht, Josué Page Vizcaíno, Abouzar Eslami, Federico Tombari, and Nassir Navab. Concurrent Segmentation And Localization For Tracking Of Surgical Instruments. *Lecture Notes in Computer Science*, 10434 LNCS:664–672, 2017.

[5] Md Kamrul Hasan, Lilian Calvet, Navid Rabbani, and Adrien Bartoli. Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry. *Medical Image Analysis*, 70:101994, 2021.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, 2017.

[7] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2938–2946, 2015.

[8] Bo Chen, Tat Jun Chin, and Marius Klimavicius. Occlusion-Robust Object Pose Estimation with Holistic Representation. In *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, pages 2223–2233. Institute of Electrical and Electronics Engineers Inc., 2022.

[9] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, 2021.