Proceedings on Automation in Medical Engineering
Proc AUTOMED, 2026, Vol. 3, No 1, Article ID 2522, 2 Pages

Infinite Science Publishing

Automated keypoint definition for surgical instruments

L. Wiese[*,1], L. Hinz[1], M. Neuhaus[2], P. Korn[2]

[1] *Institute of Measurement and Automatic Control, Leibniz University Hannover, Garbsen, Germany*
[2] *Department of Oral and Maxillofacial Surgery, Hannover Medical School, Carl-Neuberg-Strase 1, 30625 Hannover*
[*] *Corresponding author, email: leon.wiese@imr.uni-hannover.de*

Abstract: Pose estimation in stereo vision relies on robust correspondences between image features. Instead of expensive detection and matching, regression networks can be trained to localize distinctive features directly, but manual feature annotation can bias results. This study introduces methods to increase performance by automatically selecting keypoints from 2D feature detectors. Across five surgical instruments, we compare keypoints derived from SIFT (Scale-invariant feature transform) and ORB (Oriented FAST and Rotated BRIEF). Subsequently, the number of selected keypoints with regard to the performance of subsequent keypoint regression is assessed. To avoid errors from manual annotation and to enable efficient scalability, all experiments use purely synthetic, automatically generated datasets.

I. Introduction

With the increasing use of computer-assisted methods for planning and performing surgical procedures, automated tracking of surgical instruments is gaining in importance. For high-precision applications, it is essential to ensure accurate alignment between the actual surgical procedure and the virtual representation. This is commonly achieved by attaching optical markers to the instruments, which are tracked by a stereo vision system. However, applied optical markers may restrict instrument handling and thus limits the widespread implementation of intelligent assistance systems. Moreover, it must be ensured that the markers remain visible to both cameras of the stereo vision system. In this study, we address these problems by detecting virtual markers based on the geometry on surgical instruments using a keypoint detection network. The main focus of this work is to determine the optimal number and placement of virtual markers with respect to accurate pose estimation.

Several works address keypoint regression for surgical instruments. In [1], keypoints are predicted on instruments using a visioned-language model, and the underlying networks are fine-tuned via the LoRA mechanism. A purely synthetic-data approach is presented in [3], where keypoints are defined directly on 3D models and subsequently detected using a regression network. In [2], a fully markerless method for pose estimation of medical drills and screwdrivers is proposed, but it requires the availability of a VR system. All approaches are based on manually annotated keypoint positions, which may not inevitably be optimal in terms of distinctiveness, uniqueness, visibility, and robustness. We aim to address this by deriving keypoints from different feature detectors.

II. Material and methods

In this study, five different surgical instruments are investigated, representing the classes tweezers, scissors, retractor, artery clamp, and towel clamp, for each of which a 3D model is available. These models form the basis for all subsequent processing steps. Using 3D modelling software, the instruments are placed in virtual scenes and images of the instruments are rendered from ten camera poses arranged around each instrument. On these images, features are detected using the feature detection methods SIFT [4] and ORB [4]. These image points are then projected onto the corresponding 3D geometries via camera parameters, poses and scene information. This procedure is repeated for 100 different scene configurations (illumination, etc.), rendering 10 images per configuration. In total, 1,000 synthetic images per instrument are generated and used to extract feature points.

The extracted feature points are merged and derived into so-called feature frequency maps, which match the frequency of image-based feature points to the 3D geometry. Within the feature map, keypoints are selected based on their occurrence frequency. To achieve this, the 3D geometry is partitioned into six region sets (7, 10, 15, 20, 35, and 40), and keypoints are then chosen within each

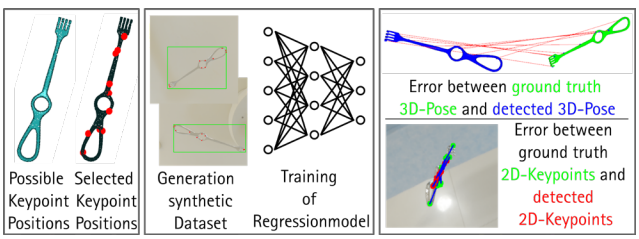

Figure 1 Data generation process

partition according to their frequency. The 3D models annotated in this way form the basis for a synthetic dataset comprising 1000 images, which are rendered and annotated in Blender using a BlenderProc pipeline [5] adapted for keypoint annotation. The parameters are chosen according to [6], with the only modification that the same scene is used for all renders. To ensure that the results are not affected by different initializations or randomly sampled parameters, the entire generation pipeline is deterministic. However, non-deterministic rendering artefacts (like noise) are currently unavoidable. A schematic overview of the entire pipeline is provided in Figure 1. Based on the generated datasets, a YOLOv11 network is trained for object detection and keypoint regression. The trained networks are evaluated using a likewise synthetically generated test dataset comprising 100 images.

III. Results and discussion

Based on five different number of keypoints and the two feature detectors SIFT and ORB, a total of 12 networks are trained for each of the five instruments on a synthetic

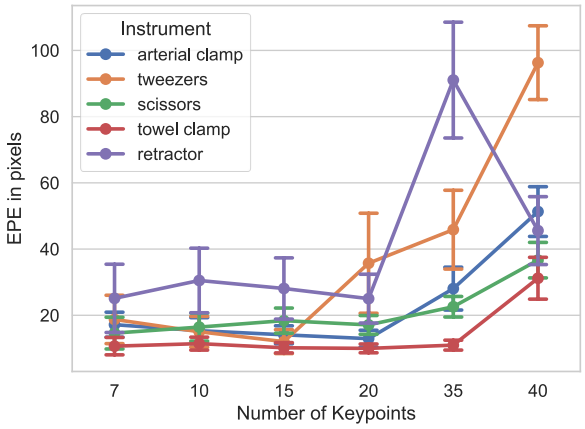

Figure 2: Euclidean Point Error for SIFT Detector

training dataset and subsequently evaluated on an equally synthetic test dataset. Performance is assessed using the euclidean distance between the predicted and the ground-truth keypoints, referred to as the Euclidean Point Error (EPE) in pixels.

Figure 2 depicts the EPE versus the number of selected keypoints for the SIFT feature detector across all five instruments. It can be observed that, for most instruments, the distance between predicted and ground-truth keypoints increases with the number of keypoints. One possible

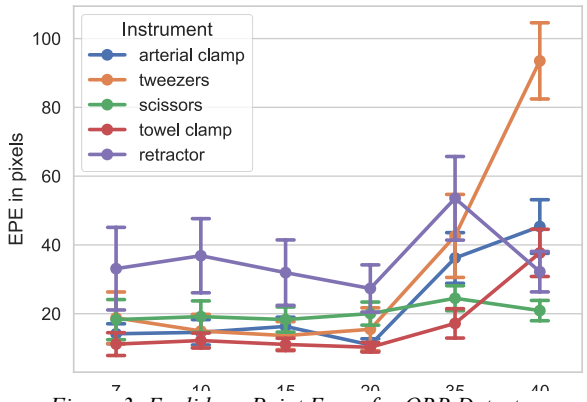

Figure 3: Euclidean Point Error for ORB Detector

explanation is that with many keypoints the regression networks increasingly exhibit ambiguous assignments. This is related to the training objective: Keypoints are considered correct if they fall within a tolerance region around the ground-truth keypoint, whose size is determined by the parameter sigma. At high keypoint densities, these tolerance regions can overlap, causing multiple keypoints to fall into the same region and leading to erroneous assignments. When comparing the individual instruments, it becomes apparent that highly symmetric geometries tend to exhibit larger errors. This is particularly pronounced for the retractor and the forceps. Symmetry induced ambiguities in keypoint assignment may also explain the comparatively wide error intervals observed for these two instruments. A similar trend can also be observed for the ORB detector, as shown in Figure 3. In a direct comparison, the following mean values averaged across all instruments can be determined for the different numbers of keypoints:

Table 1: Mean EPE per KP

Det./ nKP	7	10	15	20	35	40
SIFT	17.26	17.75	16.51	20.156	39.70	52.21
ORB	19.10	19.55	18.25	16.804	34.81	45.93

IV. Conclusions

This study demonstrates the successful conditioning of a YOLOv11 architecture for object detection and keypoint regression based on 2D features obtained with the SIFT and ORB algorithms. It was further observed that the error tends to increase with a growing number of keypoints. However, no universal conclusion can be drawn regarding the most suitable feature detector, as the outcome strongly depends on the instrument under consideration. Based on this study, further investigations are to be conducted. Beyond the YOLO architecture, alternative models, should be evaluated. Moreover, SIFT and ORB represent only a small subset of available methods for feature detection. Therefore, it remains to be explored whether other detectors and descriptors may provide a more suitable basis for robust and accurate keypoint derivation.

Author's Statement

This research was funded by Leibniz Young Investigator Grants program by the Leibniz University Hannover grant number 11-76251-114/2022.

References

[1] DUANGPROM, Krit; LAMBROU, Tryphon; BHATTARAI, Binod. Estimating 2D Keypoints of Surgical Tools Using Vision-Language Models with Low-Rank Adaptation. In: MICCAI Workshop on Data Engineering in Medical Imaging. Cham: Springer Nature Switzerland, 2025. S. 201-211.

[2] HEIN, Jonas, et al. Next-generation surgical navigation: Marker-less multi-view 6DoF pose estimation of surgical instruments. Medical Image Analysis, 2025, S. 103613.

[3] ABOUKHADRA, Ahmed Tawfik, et al. SurgeoNet: Realtime 3D Pose Estimation of Articulated Surgical Instruments from Stereo Images Using a Synthetically-Trained Network. In: DAGM German Conference on Pattern Recognition. Cham: Springer Nature Switzerland, 2024. S. 199-211.

[4] HARTLEY, Richard; ZISSERMAN, Andrew. Multiple view geometry in computer vision. Cambridge university press, 2003.

[5] DENNINGER, Maximilian, et al. Blenderproc2: A procedural pipeline for photorealistic rendering. Journal of Open Source Software, 2023, 8. Jg., Nr. 82, S. 4901.

[6] WIESE, Leon, et al. Detection of Surgical Instruments Based on Synthetic Training Data. *Computers*, 2025, 14. Jg., Nr. 2, S. 69.