A. J. Boudreault, J. Spille, J. Wiltfang, A. Schlaefer, and M. Neidhardt

6-Degree Vision based Tracking of a Mandible Phantom with Deep Learning

https://doi.org/10.1515/cdbme-2024-1002

Abstract: During maxillofacial surgery, the precise placement of surgical tools is crucial for accurate implant placement. Particularly if multiple implants are needed, e.g., after cancerous bone removal, visual landmarks might not be obtainable. To this end, we propose a vision-based tracking approach with deep learning. Our markerless tracking approach is based on video streams from two cameras for tracking a mandible phantom. We study, to what extent vision-based localization using deep learning is feasible. For real-time 6D pose estimation we propose a Siamese network with a ResNet-18 subnetwork. We acquire a large training dataset with a robot and evaluate the tracking accuracy on partially occluded images. Thereby, we mimic visual information that is accessible during clinical interventions. We report a mean position error of 1.33 ± 1.14 mm and a rotation error of 0.86 ± 0.71 deg for partially occluded images. Overall we present a promising tracking approach that is marker-free and robust toward image artifacts.

Keywords: Tracking, Maxillofacial Surgery, Medical Phantom, RGB Camera, Markerless

1 Introduction

Dental implant procedures can be virtually planned accurately with CT images. During surgery, the surgeon attempts to mentally reproduce the planned procedure as closely as possible by matching visible landmarks, such as teeth and bone structure, with the CT images. While accuracy in tool placement is sufficient for single or double implants, more precise placement accuracy is desired when more consecutive implants are needed [1], e.g., after the removal of cancerous bone tissue. Despite recent advancements in bone regeneration [2], many patients still suffer from limited bone mass after undergoing a bone graft, increasing the importance of precise implant placement in these cases. To guide this intervention, vision-guided assistive systems relying on a known marker geometry visible on both the CT images and during surgery are available [3]. These systems first register the tracking marker in the CT reference frame and second, track the position of the marker during the intervention. Accordingly, a stereo camera system is mounted to the surgical tool. These systems have limitations as a line-of-sight between the camera and the marker is necessary for tracking and the marker is not rigidly fixed to the patient. Hence, the marker can move which can lead to large deviations in positioning the surgical tool. To this end, we propose a marker-free tracking approach based on RGB-images from two cameras. Markerless tracking methods have shown promising results for local high-resolution imaging [4]. In this work, we evaluate the 6D tracking accuracy provided by a deep-learning model. We generate a training data set containing a total of 110.888 images while a robot drives a mandible phantom. As a training target, we define the recorded 6D pose of the robot. We systematically study to which extent our models' performance depends on the available visual features by removing partial image information. Our results show that vision-based tracking of a mandible phantom is feasible even for partially occluded images.

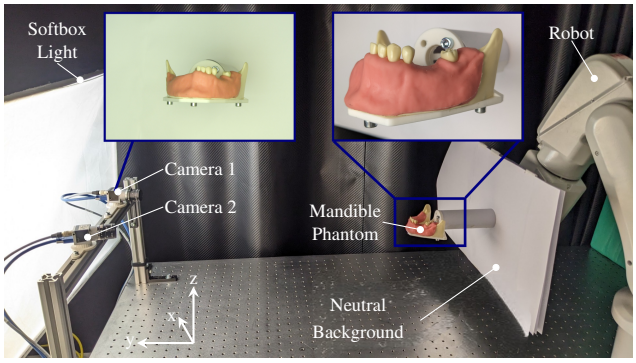


Fig. 1: Experimental Setup: Two RGB cameras record the movement of a mandible phantom. The phantom is positioned by a robot.

2 Methods

2.1 Experimental Setup

Figure 1 presents the experimental setup used to acquire the training data, with the respective reference frame. The setup is

A. J. Boudreault, A. Schlaefer, M. Neidhardt, Hamburg University of Technology, Institute of Medical Technology and Intelligent Systems, Hamburg, Germany
email: ana.boudreault@tuhh.de
J. Spille, J. Wiltfang, Department of Gynecology and Obstetrics, University Hospitals Schleswig-Holstein Campus Kiel, Kiel, Germany

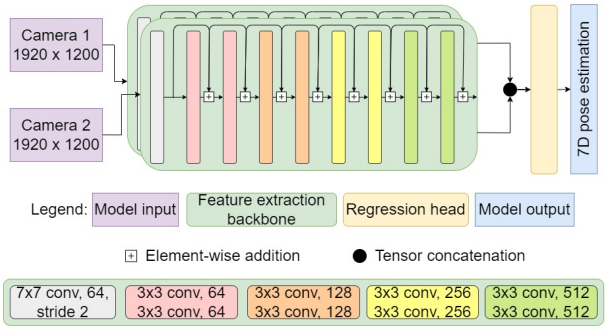
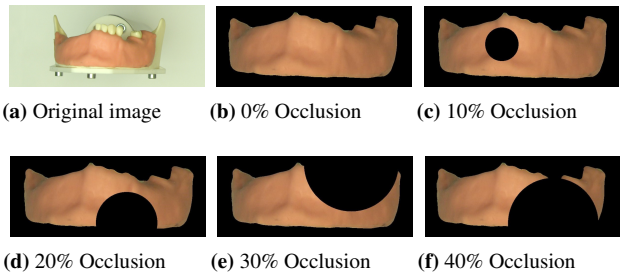


Fig. 2: Multi-path Network architecture for pose prediction from two cameras images.



(a) Original image **(b)** 0% Occlusion **(c)** 10% Occlusion

(d) 20% Occlusion **(e)** 30% Occlusion **(f)** 40% Occlusion

Fig. 3: We acquire RGB images of the phantom (a) and remove the background (b). To test the robustness of our algorithms we partially occlude the phantom between 10% and 40% (c-f) by adding black circles at random positions on the phantom. Please note, that the images have been cropped to show the phantom more clearly.

composed of a mandible phantom (A-JK OP UK, Frasaco, Germany), two RGB cameras (acA1920 - 150uc Basler, Germany), the robotic arm (IRB 120 6DOF, ABB, Switzerland) and a softbox light. The cameras are connected with a trigger cable to ensure synchronized image data acquisition. Both cameras are equipped with wide-angle lenses with a focal length of 25 mm (FL-CC2514-2M, RICOH, Japan) and the mean distance between phantom and camera is approximately 820 mm. The tracking volume, meaning the volume within which the deep-learning model is expected to be able to track the mandible pose, was determined by the camera's field of view and focus as; 120 mm in the X direction, 90 mm in Y, and 80 mm in Z. The respective axes are indicated in Figure 1. Sheets of paper on top of acrylic glass provided a neutral background so that the robot was not visible in the acquired images.

2.2 Data Acquisition

For data acquisition the robot positioned the mandible along pre-defined trajectories. Robot poses are recorded with a temporal rate of 500 Hz. Both cameras acquired images in sync with a frame rate of 40 fps and a resolution of 1920x1200 px. The frame rate was limited by the exposure time required to ensure image quality. The end-effector coordinate system was moved to approximately the center of the mandible phantom. Before data acquisition, the internal clocks of the workstations were synced with a time server to allow data matching based on the recorded time stamps. Trajectories were created to ensure an even sampling of the tracking volume in all dimensions. The poses used in the generated trajectories were constrained to the following limits; 285 < X < 405 mm, 235 < Y < 325 mm, 270 < Z < 350 mm, -5° < Rx < 5°, -10° < Ry < 10°, and -5° < Rz < 5°. The robot was tasked to drive to the limits of the desired tracking volume. Mandible rotations were added randomly to each of the intermediary poses while the robot path planning performed a linear interpolation between respective poses.

2.3 Deep Learning Model

Our deep learning model architecture is depicted in Figure 2. We train our model to predict the mandible pose reported by the robot from a set of RGB images. As a baseline model, we use a multi-path network which has successfully been employed for the 6D object pose estimation from depth-resolved images [5]. The model can be split into two sections: the feature extraction backbone, and the regression head. The feature extraction backbone consists of two copies of the same subnetwork based on the ResNet-18 classifier, both sharing the same weights. Each subnetwork takes the image from one of the two cameras as input and outputs feature maps. Before transitioning to the regression head, we apply a global average pooling (GAP) layer. These values are then concatenated and become the inputs to the first layer of the regression head. A hidden layer of size 256, and a 7-unit (3D position vector, and 4D orientation quaternion) output layer round out this section of the model and produce a pose estimation. The loss was defined as the mean-squared error, and the model was trained on an NVIDIA GeForce RTX 4090 GPU, using CUDA 11.8.

2.4 Datasets and Training

For training our deep learning network, two cameras recorded 110.888 images while a robot systematically sampled the desired tracking volume. We segment the gum (red part) of the phantom by thresholding color values in the RGB image (Figure 3a), thereby removing unnatural visual features, e.g., metal screw heads and the phantom mounting as depicted in Figure 3b. Next, we create partial occlusions on the image by adding black circles at random locations on the segmentation. In total 6 networks were trained: (1) on the original RGB image, and adding randomly placed occlusions covering (2) 0%, (3) 10% (4) 20%,

(5) 30%, and (6) 40% of the segmentation mask. Respective example images are depicted in Figure 3. We randomly split our dataset into a training set (80% of the images) and a test set (20% of the images). The resulting training set was further split into a validation set (25% of the images) and the remaining 75% created the final training set. The image split was fixed for all experiments performed. We evaluated the performance of our networks on an additional trajectory, created by randomly driving the mandible to various poses within the tracking volume, generating in total 3534 additional images. Before being used as input to the model, pixel values were normalized and the mandible position coordinates were scaled to values between -1 and 1.

2.5 Evaluation Metrics

We define the mean absolute error (MAE) to estimate the positioning accuracy as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - \hat{p}_i|$$

where p_i is the position reported by the robot, $\hat{p}_i$ are the predicted values, and N is the number of samples. We also define the Hamilton product to estimate the rotation error on the orientation quaternions. The resulting quaternion is then converted to Euler angles to get a more interpretable error metric. While the quaternion to Euler angle conversion isn't deterministic, it is assumed that the angle prediction error will be small enough that singularities will be omitted.

3 Results and Discussion

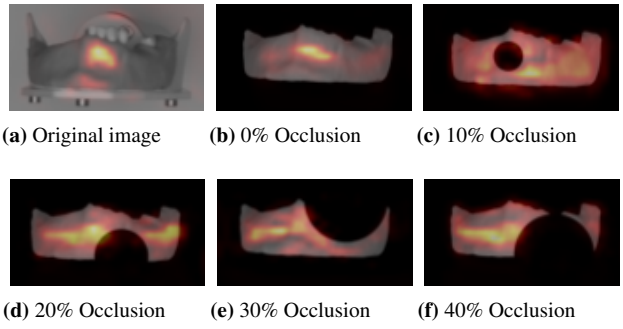
Table 1 presents the average position and rotation prediction error on the test set for each of the models trained. While both the RGB and the 0% occlusion model are trained on occlusion-free images, the latter performance is higher due to isolating the relevant information in the original image. Overall, our results indicate the best performance for the 0% occlusion model with a position error of 0.98 ± 0.94 mm and a rotation error of 0.29 ± 0.24 deg. As expected, a general drop in the tracking performance can be observed as the relative occlusion size increases.

Figure 4 presents the saliency maps for all 6 model variations. For each input image, all feature maps output by the ResNet-18 backbone are averaged, and all negative values are set to 0. Figure 4a shows that the RGB model uses non-mandibular elements to produce a prediction, e.g., the phantom mounting in the background. These features can be hidden by the mandible itself, which might also be related to the higher tracking error indicated in table 1. Overall the model focuses on features located at the center of the mandible.

Figure 5 compares the accuracy in tracking the position and orientation for each model on an additional acquired random trajectory. Again, both the error and the error range increase with the relative occlusion size. The largest inaccuracies occur

	Position Error			
Data Set	X **[mm]**	Y **[mm]**	Z **[mm]**	**Mean**
Original	3.48 ± 2.18	2.00 ± 1.94	1.49 ± 1.15	2.32 ± 1.76
0%	1.63 ± 1.64	0.55 ± 0.48	0.77 ± 0.69	0.98 ± 0.94
10%	2.13 ± 1.79	0.84 ± 0.78	1.01 ± 0.85	1.33 ± 1.14
20%	2.32 ± 1.96	1.17 ± 1.00	1.00 ± 0.86	1.50 ± 1.27
30%	2.92 ± 2.67	1.28 ± 1.12	1.06 ± 0.85	1.75 ± 1.55
40%	2.89 ± 2.46	1.05 ± 0.90	1.40 ± 1.16	1.78 ± 1.51
	Rotation Error			
Data Set	Rx **[deg]**	Ry **[deg]**	Rz **[deg]**	**Mean**
Original	2.04 ± 0.77	2.08 ± 1.02	2.33 ± 0.98	2.15 ± 0.92
0%	0.30 ± 0.25	0.32 ± 0.26	0.25 ± 0.22	0.29 ± 0.24
10%	0.91 ± 0.76	1.30 ± 1.05	0.37 ± 0.32	0.86 ± 0.71
20%	0.43 ± 0.34	1.70 ± 1.28	0.41 ± 0.33	0.85 ± 0.65
30%	0.75 ± 0.64	0.85 ± 0.75	0.53 ± 0.43	0.71 ± 0.61
40%	1.06 ± 0.90	1.69 ± 1.22	0.57 ± 0.48	1.11 ± 0.86

Tab. 1: Tracking performance given as mean absolute error (MAE, $\mu \pm \sigma$) for different scaled occlusions on the phantom.



(a) Original image **(b)** 0% Occlusion **(c)** 10% Occlusion

(d) 20% Occlusion **(e)** 30% Occlusion **(f)** 40% Occlusion

Fig. 4: Saliency maps overlaid on top of the camera 1 image. Outputs have been cropped, to show the mandible more clearly.

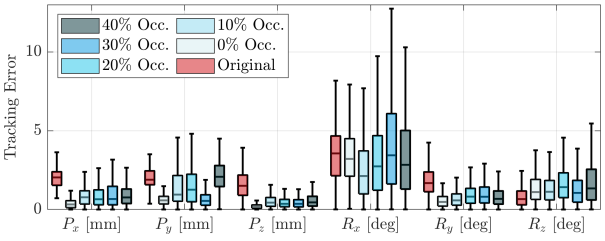
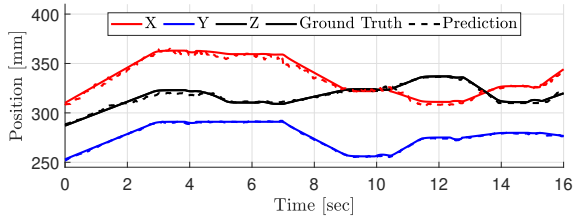
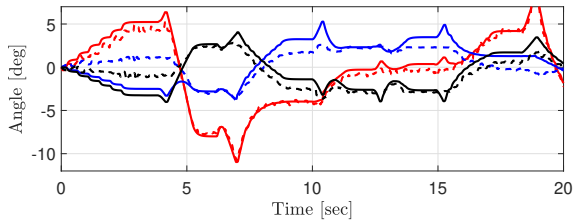


Fig. 5: Accuracy in tracking position (P) and rotation (R) with varying occlusion ratios on a random trajectory.



(a) Tracking of the Position



(b) Tracking of the Rotation

Fig. 6: Tracking performance with 10% occlusion of the phantom given as (a) position and be (b) rotation angle.

when predicting the rotation along the X axis, as those produce the fewest changes in the mandible's appearance from the camera's point of view. An example prediction obtained from the 10% occlusion model on the random trajectory is depicted in Figure 6.

Our results show that vision-based tracking with deep learning can achieve relatively high accuracy, with a position error below 2.5 mm and a rotation error below 1.15 deg despite occluding 40% of the tracking information. Still, tracking on unoccluded images with marker-based tracking methods for maxillofacial surgery [6] have yet to be matched ($\pm$0.05 mm; $\pm$0.25 deg). Our approach is however in line with other markerless methods, such as [7] ($\pm$2.5 mm; $\pm$1.25 deg), which relies on the bounded ICP algorithm. Furthermore, our model has an average inference time of 1.75 ms (NVIDIA GeForce RTX 4090 GPU), making real-time tracking feasible with the limiting factor coming from the camera acquisition frame rate. Our study has some limitations, such as parts of the phantom do not move independently which simplifies our acquired data. Further deviations from a clinical setting include the black background, the circular black occlusions, and the visibility of some mandible sections that would normally be hidden, e. g. the bottom edge. The robustness of the tracking performance in a more realistic scenario still needs to be evaluated. However, our training data only contains a small crop of the available visual information. Hence, if more information, e.g., the head of the patient is available the tracking accuracy might increase further. Moreover, the ability to generalize to different patient heads and mouths is also not represented in our study. Increasing the variability in our training data set by including multiple patients can also further increase robustness during tracking.

4 Conclusion

We present a markerless 6D pose estimation approach based on the multi-path network architecture using two RGB images as input. Our results indicate, that our method remains able to provide a fairly accurate pose prediction despite having partially occluded images as input. Further studies with multiple patients need to be performed in the future.

Author Statement

Research funding: This work was partially funded by the Interdisciplinary Competence Center for Interface Research (ICCIR) and the Forschungszentrum Medizintechnik Hamburg (FMTHH; grant 01fmthh2019). Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Matsumura, A.e.a.. Multivariate analysis of causal factors influencing accuracy of guided implant surgery for partial edentulism: a retrospective clinical study. Int J Implant Dent 2021;7(1). URL: http://dx.doi.org/10.1186/s40729-021-00313-2. doi:10.1186/s40729-021-00313-2.

[2] Wagner, J.e.a.. Bone regeneration in critical-size defects of the mandible using biomechanically adapted cad/cam hybrid scaffolds: An in vivo study in miniature pigs. Journal of Cranio-Maxillofacial Surgery 2024;52(1):127–135.

[3] mininavident AG, . Denacam - 3d planung digital umgesetzt. https://www.mininavident.com/; 2024. Accessed: 2024-04-16.

[4] Schlüter, M.e.a.. Concept for markerless 6d tracking employing volumetric optical coherence tomography. Sensors 2020;20(9):2678.

[5] He, Y.e.a.. Fs6d: Few-shot 6d pose estimation of novel objects. In: CVPR. 2022, p. 6814–6824.

[6] Hemm, S.e.a.. Accuracy investigation of dual mode markers for navigated dental implant surgery with a new 3d realtime navigation system denacam. In: EPiC Series in Health Sciences. EasyChair; 2017,URL: http://dx.doi.org/10.29007/hw4v. doi:10.29007/hw4v.

[7] Hu, X.e.a.. Occlusion-robust visual markerless bone tracking for computer-assisted orthopedic surgery. IEEE Trans Instrum Meas 2022;71:1–11. doi:10.1109/TIM.2021.3134764.