Investigating Keypoint Descriptors for Camera Relocalization in Endoscopy Surgery

Isabela Hernández[1*†], Roger Soberanis-Mukul[1*†], Jan Emily Mangulabnan[1], Manish Sahu[1], Jonas Winter[1], Swaroop Vedula[1], Masaru Ishii[2], Gregory Hager[1], Russell H. Taylor[1,2] and Mathias Unberath[1,2]

[1*]Johns Hopkins University, Baltimore, 21211, MD, USA.
[2]Johns Hopkins Medical Institutions, Baltimore, 21287, MD, USA.

*Corresponding author(s). E-mail(s): iherna12@jhu.edu; rsobera1@jhu.edu;
†These authors contributed equally to this work.

International Journal of Computer Assisted Radiology and Surgery

Supplementary Material

3D Descriptor Projection Analysis

We provide further insight on the selection of a single 3D point descriptor from its first projection found in the preoperative sequences. This design choice is based on the definition and training paradigms of point descriptors, which maximize their response only in the locations of true correspondences between two images. In this regard, we expect all projections of a single 3D point to refer to a single, equivalent point seen from different images. This is the case for the exemplary preoperative frames in Figure 1 (top block), showing a selection of the 2D projections corresponding to a single 3D point. In addition, the innate similarity between these descriptors should also be reflected on the response pattern on a set of unseen images. As can be seen in Figure 1 (bottom block), the individual descriptors generate similar responses on the exemplary set of intraoperative frames. The similarity is reflected on the location and numerical value of the maximum response in each query frame. In this way, the choice of a single of these projections (by convention, the first) will accomplish keypoint localization in an adequate and comparable manner to the remaining projections.

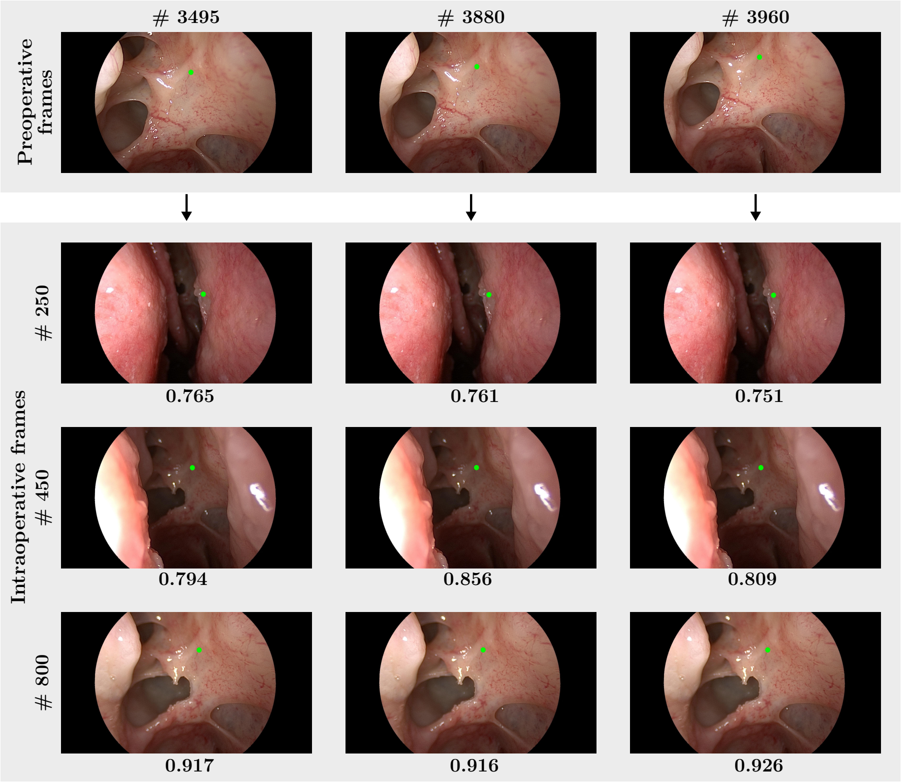

Fig. 1 Qualitative example of keypoint localization along different 3D point projections, employing [1] as main feature descriptor. Top block shows projections of an exemplary 3D point on three preoperative frames (columns), and their equivalence in representing a particular anatomical landmark. Bottom block shows the matches of each projection in three intraoperative frames and their corresponding response value (rows). The general and comparable patterns in keypoint localization support the choice of any of the projections as single representative of a given 3D point in our anatomy model. Best viewed in color.

Keypoint Relocalization Response Threshold: Sensitivity Analysis

To support the dynamic threshold approach described in Section 2.3 of the main manuscript, we conduct an additional experiment with the contrasting strategy of fixing the response threshold for 2D-3D point correspondence selection for all query images. We vary the value of the threshold between 0.99 and 0.9, in order to evaluate its effect on the localization performance of this version of our method. Figure 2 reports the results of this experiment.

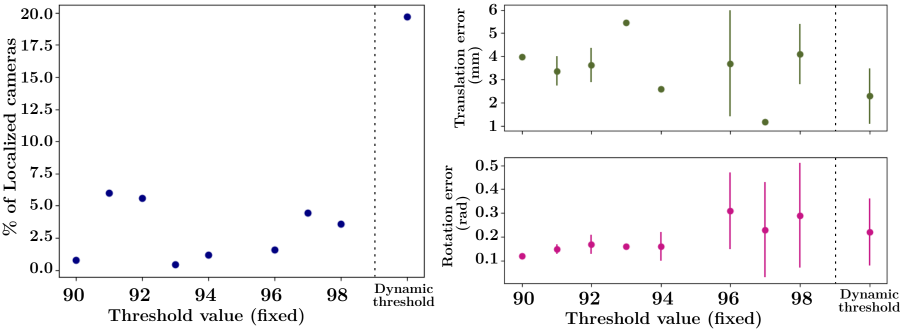

Fig. 2 Performance comparison using different response threshold strategies on one estimated trajectory (Subject # 1 - Progression Step # 1). Translation and rotation error metrics variate, but are comparable to the results with a dynamic threshold. However, there is a considerable difference in the localization rate between strategies, with more adequate results achieved with the dynamic threshold approach. Best viewed in color.

We observe considerable variation in the performance of our method along consequent threshold values. In terms of translation and rotation error, the dynamic threshold performs similarly to the fixed threshold case. However, there is a significant difference in the number of localized cameras between these approaches. While the dynamic threshold case achieves 19.68% of localized cameras, the fixed threshold case achieve an average of 2.96%, and reaches a minimum of one (1) camera localized for $\tau_q = 0.93$. Note that for $\tau_q = 0.99$ and $\tau_q = 0.95$, no predictions were obtained. A fundamental reason for this behavior is the generation of different ranges of response in each individual frame, which cannot be described with a single fixed threshold. With high response threshold values, the latter approach leads to the selection of too little point samples, and do not satisfy the minimum requirement for the PnP stage. Moreover, for lower response threshold values, this approach leads to the selection of higher number of point samples, introducing potential false positives and leading to incorrect camera poses (further eliminated during post-processing stages). These results motivate the use of our dynamic threshold for tailored frame-level responses.

Spatial Distribution of Localized Cameras

Besides reporting an average of 21.86% of localized cameras along the main query sequences, we also include observations on the distribution of these correct localizations in spatial domain (see subsection **Spatial Distribution of Localized Cameras** in Section 4 of the main manuscript). To complement the main remarks in this section, we depict the distribution of localized cameras in the time dimension of the two exemplary query sequences (*Subject # 1 - Undisturbed Anatomy* and *Progression Step # 1*), and include two additional estimated trajectories (*Progression Step # 2* and *Progression Step # 1* using SIFT [2] as descriptor for keypoint recognition).

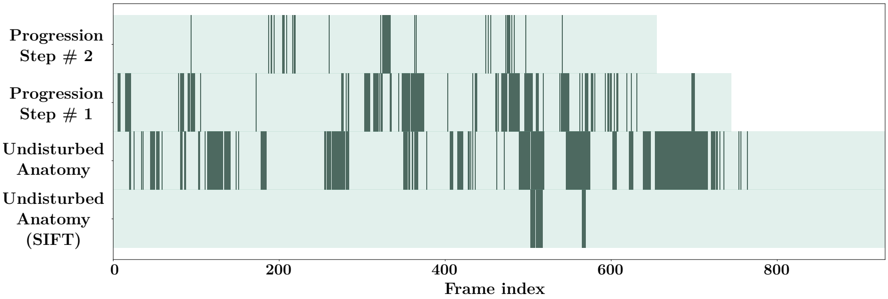

Fig. 3 Spatial distribution of the localized cameras (dark bars) across the test sequence progression (light green) employing [1] as main feature descriptor. The results show that the learning-based, dense descriptor localizes cameras that distribute over the entire sequence. In contrast, SIFT-based localizations collapse into a single region of the sequence.

Effect of Noisy 2D-3D Correspondences in PnP Localization

In the main manuscript, we introduce the main results of the effect that noisy correspondences have in the pose estimation process with PnP. In this supplementary section, we expand this analysis by presenting the complete series of error measures obtained in response to the different levels of noise in the correspondences.

As mentioned before, the pose estimation process is conducted through a set of 2D-3D correspondences estimated with a learning-based model. Given that the estimation of correspondences could be subject to errors in the matching process, we explore the robustness of PnP against noisy inputs. In this respect, we propose an experiment using the 2D-3D correspondence set estimated by COLMAP-SfM as a reference for ideal correspondences. We apply additive Gaussian noise to the 2D-3D correspondences to simulate a mismatch between the 3D points and the predicted 2D location. We repeat this experiment for different values of standard deviation of the Gaussian noise to account for varying levels of localization error, over a subset of 106 images uniformly sampled from *Progression Step # 1.*

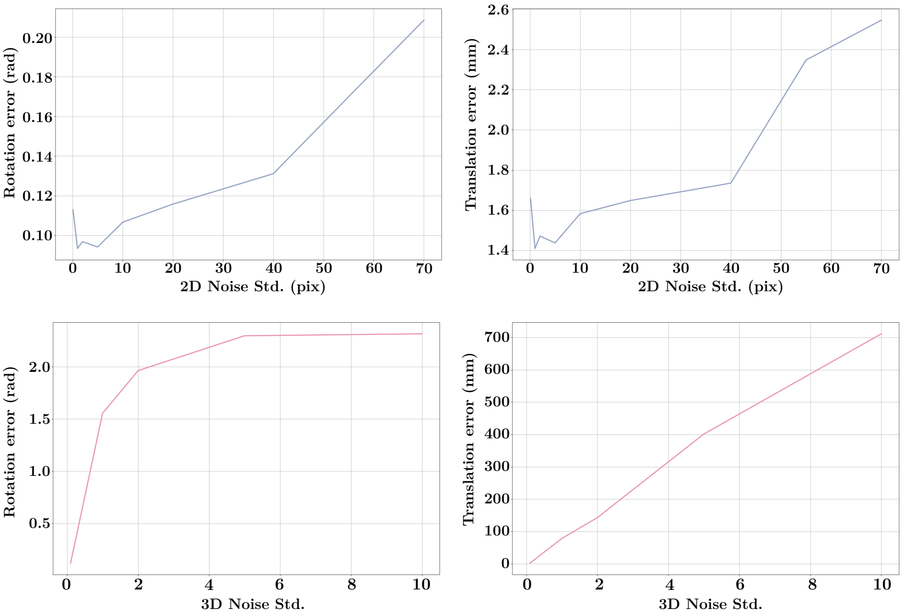

Fig. 4 PnP sensitivity to noise in the 2D-3D matches. Rotation and translation errors obtained when a correct 3D point projected into a noisy 2D location in the image (top row), and when a noisy 3D position is projected into the image (bottom row). Noise is generated with different standard deviations.

Results suggest a certain level of robustness to noise in the estimated 2D keypoint location. In contrast, errors in the 3D counterpart of the correspondences lead to higher errors in the estimated pose with PnP. These findings suggest the occurrence of false positive 3D matches in the query images as one of the error sources for the camera pose estimation, as these lead to noisy 3D projections in the correspondence set employed to solve the PnP problem. Supplementary Figure 5 shows qualitative examples of true and false positive matches as seen in the image domain. Furthermore, it is possible that the training strategies employed with the descriptors generate a high sensitivity matching process. This allows strong correspondences to be found in a local vicinity (favorable for SfM), but may present a low specificity in long-range matches. Low specificity may contribute to errors when employing the matching process to relocalize keypoints in a different sequence of the same patients, suggesting that additional constraints must be taken into account during the definition of the learning base descriptor.

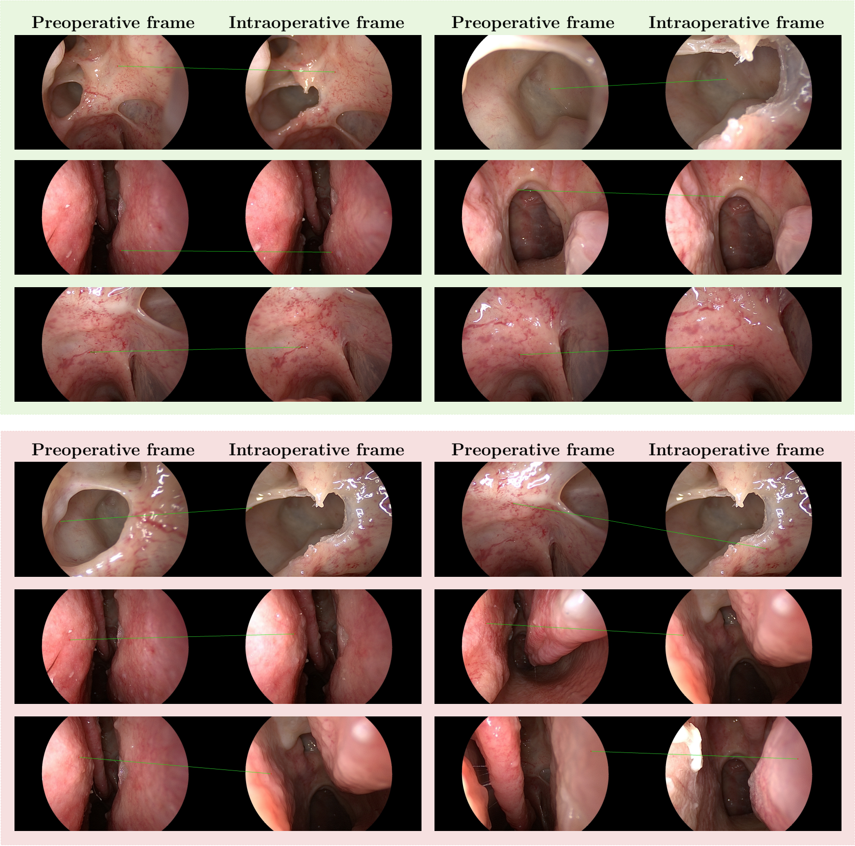

Fig. 5 Qualitative examples of keypoint localization employing [1] as main feature descriptor. Top block (green) shows true positive matches between preoperative and intraoperative frames of the same anatomy. Bottom block (red) shows false positive matches between anatomical landmarks, identified as a source of incorrect 2D-3D correspondence sets and camera pose estimates. Best viewed in color.

References

[1] Liu, X., Zheng, Y., Killeen, B., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Extremely Dense Point Correspondences using a Learned Feature Descriptor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4847–4856 (2020)

[2] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94