Keypoint Matching for Instrument-Free 3D Registration in Video-based Surgical Navigation

Tânia Baptista[1,2](✉), Carolina Raposo[1,2], Miguel Marques[2], Michel Antunes[1,2], and Joao P. Barreto[1,2]

[1] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal
tania.baptista@isr.uc.pt
[2] Perceive3D, Coimbra, Portugal

Abstract. Video-based Surgical Navigation (VBSN) inside the articular joint using an arthroscopic camera has proven to have important clinical benefits in arthroscopy. It works by referencing the anatomy and instruments with respect to the system of coordinates of a fiducial marker that is rigidly attached to the bone. In order to overlay surgical plans on the anatomy, VBSN performs registration of a pre-operative model with intra-operative data, which is acquired by means of an instrumented touch probe for surface reconstruction. The downside is that this procedure is typically time-consuming and may cause iatrogenic damage to the anatomy. Performing anatomy reconstruction by using solely the arthroscopic video overcomes these problems but raises new ones, namely the difficulty in accomplishing keypoint detection and matching in bone and cartilage regions that are often very low textured. This paper presents a thorough analysis of the performance of classical and learning-based approaches for keypoint matching in arthroscopic images acquired in the knee joint. It is demonstrated that by employing learning-based methods in such imagery, it becomes possible, for the first time, to perform registration in the context of VBSN without the aid of any instruments, i.e., in an instrument-free manner.

Keywords: Instrument-free registration · Surgical navigation · Feature matching · Arthroscopy

1 Introduction

Video-based Surgical Navigation (VBSN) [17] inside the articular joint has proven to be effective in overlaying surgical plans with the patient's anatomy for guiding the surgeon during arthroscopic procedures. It leverages a fiducial marker, referred to as the world marker (WM), that is rigidly attached to the anatomy for determining the pose of the arthroscope at every frame-time instant. Additionally, a probe instrumented with a different fiducial is used to digitize the bone surface and reconstruct a 3D point for each frame in which its pose is determined. If both WM and probe are in the field-of-view of the arthroscope, their relative pose can be determined and reconstructed 3D points can be represented in WM coordinates. VBSN further includes a registration step that aligns a pre-operative model of the bone, which can be obtained from CT or MRI, with the data acquired intra-operatively. This step is crucial for the surgical navigation stage as it will enable planning information extracted from the pre-operative model to be represented in the patient's anatomy. The quality of the reconstructed 3D points will impact registration performance. Allied to the fact that arthroscopic scenarios are challenging, namely because there is limited maneuverability and visibility inside the joint, there exist floating particles and tissue, and the arthroscopic lens induces high image distortion, the digitization process can become time-consuming and error prone. Additionally, having to physically touch anatomical parts comes with the risk of causing iatrogenic damage.

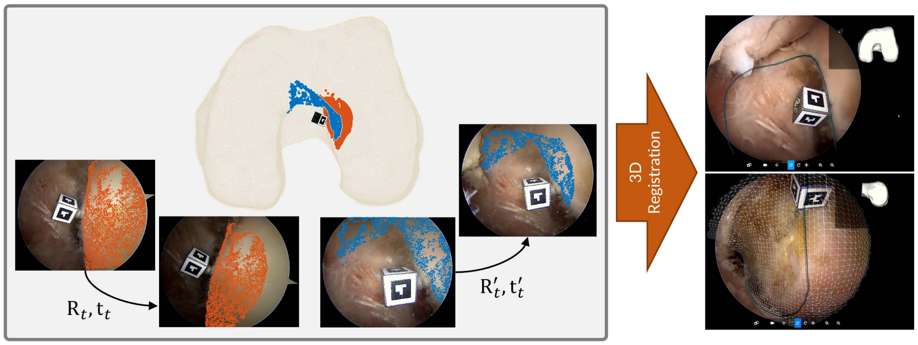


Fig. 1: 3D reconstruction and registration pipeline for assessing the different feature matching approaches: in an arthroscopic video sequence, pairs of frames are selected and their relative pose is determined by tracking the WM. Triangulation is then performed and the reconstructed points are represented in WM coordinates. A final step of 3D registration allows to overlay the pre-operative mesh model with the patient's anatomy using augmented reality.

There exist alternative ways of performing surface reconstruction in arthroscopic/endoscopic/laparoscopic scenarios without having to touch the patient's anatomy. These typically involve structured light, in which light with a known pattern is projected onto the anatomy and detected by the camera for performing some type of triangulation [2,12]. Besides not having the risk of damaging the patient's anatomy, these solutions typically allow a broader access than the touch probe, which facilitates the registration process. However, existing designs present several disadvantages: i) they are too large to be inserted into the arthroscopic portals that typically measure about 6 mm [6], and ii) they must be constructed and calibrated, which is not a straightforward process [12,18,21].

This paper concerns the problem of performing 3D surface reconstruction and registration without using any instrumentation for digitization. This brings important advantages with respect to the existing approaches that include i) no risk of causing damage to the anatomy, ii) no need to open larger than usual portals, iii) accessibility to anatomical regions that cannot be reached with touch probes or structured light systems and iv) a fast digitization step that only involves anatomy inspection with the arthroscopic camera. Unfortunately, the aforementioned difficulties associated with working in arthroscopic environments become more evident when trying to perform 3D reconstruction solely from the arthroscopic video, to which the very low texture of bony surfaces is an added challenge. Previous efforts in performing visual SLAM using arthroscopic sequences did not provide acceptable results [17].

Recent advances in keypoint matching have demonstrated superior performance compared to traditional methods, particularly in textureless environments [7,22]. Traditional feature matching approaches involve sparse keypoint and descriptor extraction followed by matching [13]. However, reliable keypoint extraction remains a challenge when working in low textured scenes, often resulting in very sparse and inaccurate reconstructions. Recently, two approaches have emerged as robust solutions for textureless scenes: (i) semi-sparse methods, which do not rely on direct keypoint detection [22], and (ii) dense feature matching, which captures all matches between views.

This paper assesses, for the first time, the possibility of performing 3D surface reconstruction and registration in arthroscopic scenes in which the only required instrumentation is a WM rigidly attached to the anatomy. To properly assess the performance of state-of-the-art approaches for two-view matching, this assessment is conducted using a straightforward reconstruction pipeline that employs tracking of the WM for determining the camera motion and simple triangulation for 3D point reconstruction (Fig. 1). It is demonstrated that despite being a difficult problem with no solution in the literature, if the camera motion is known, the recent advances in feature matching enable 3D surface reconstruction in arthroscopic scenarios with sufficient quality for accomplishing 3D registration that meets the medical requirements.

2 Instrument-free 3D Registration in Arthroscopy

In the context of VBSN [17], a fiducial marker (WM) is rigidly attached to the anatomy and works as the world reference frame. By tracking this marker, the camera pose is known at all times with respect to a coordinate system that is static with respect to the anatomy. This allows not only to determine the camera motion between any two frames but also to represent 3D measurements in world coordinates. Using this information, 3D reconstruction without the aid of any additional instrumentation can be performed in a straightforward manner by establishing keypoint correspondences, filtering out incorrect matches using the known epipolar geometry and triangulating the remaining ones. A final step of 3D registration aligns the reconstructed points with the pre-operative model of the patient's anatomy, enabling surgical plans to be overlaid in the targeted anatomical part. Fig. 1 illustrates this pipeline using the example of a distal femur. The WM shown in the figure is a metal 3 mm-cube with an attached thread that is screwed into bone and provides submillimetric tracking accuracy. Its implantation is invasive but by being placed in bone (and not in cartilage or soft tissue), surgeons are not concerned about permanent damage to the anatomy due to the ability of bone to self-regenerate. This procedure does not disrupt the normal course of the medical procedure as it typically lasts less than 30 seconds. Fig. 1 further evinces the fact that 3D reconstructions obtained from pairs of frames observing different regions of the anatomy can be represented in the same coordinate system (the WM), allowing a large spread of reconstructed points across the whole anatomy. This benefits 3D registration algorithms as a larger overlap between the pre-op model and intra-op data leads to higher accuracy [1]. In this work we use the method proposed in [14] for initial global registration and a standard ICP [3] for refinement.

2.1 Feature Matching Algorithms

For the past 20 years, since the appearance of SIFT [13], the literature in feature extraction and matching has evolved significantly, with recent methods typically making use of deep learning models [5,7,8,20,22]. These methods report dramatically better performances when compared to the classical approaches [13], in particular in the presence of low-textured scenes and large viewpoint and illumination changes [7,22]. In light of these results, we selected the state-of-the-art two-view feature matching algorithms for evaluation under real arthroscopic scenarios. For the sake of completeness, and given the success of SIFT, we also included it in the assessment. The list of methods used as keypoint matchers in the described 3D reconstruction and registration pipeline is the following: (i) SIFT+NN [13], where SIFT combines a feature detector and descriptor, and mutual nearest neighbor matching (NN) is used to obtain candidate correspondences; (ii) DISK [23]+NN, where DISK is a CNN-based approach for detecting and describing keypoints, and its output is used as input for the NN matching algorithm; (iii) SuperPoint+LightGlue [5,11] (SP+LG), where SP is a self-supervised approach to extract feature points and descriptors, while LG performs matching; (iv) LoFTR [22], a semi-sparse image matching method; and two recently dense feature matches, (v) RoMa [8] and (vi) DKM [7]. For each method, we iteratively fine-tuned its parameters using a chosen representative arthroscopic sequence and subsequently applied the optimized parameters to the remaining arthroscopic sequences. The chosen parameters for each method can be found in the Supplementary Material (Table 3).

2.2 Semantic Segmentation in Arthroscopy

One of the main challenges associated with arthroscopic imagery is the existence of tissue connected to the rigid anatomical parts, such as bone and cartilage, that is non-rigid and usually not visible in a CT or MRI. Since the pre-operative model contains only bone and cartilage structures, 3D points should be reconstructed mostly on these structures to minimize the existence of outliers that hamper the 3D registration process. While correspondences established on highly non-rigid regions can be eliminated through verification of the epipolar geometry, correspondences belonging to anatomical parts that are less non-rigid, such as ligaments, cannot be discarded in a straightforward manner.

To tackle this problem, we developed a deep-learning model for automatic semantic image segmentation that identifies which pixels belong to bone or cartilage structures. The chosen architecture to perform bone segmentation is a standard U-Net [19] with loss function $loss = 1 - L_{\text{DICE}}$, where dice loss L_{DICE} measures the overlap between the inferred segmentation and the ground-truth segmentation.

3 Experiments and Results

This section assesses the performance of the 6 feature matching algorithms described in Section 2.1 in terms of quantity and quality of the 3D reconstructed points and the ability to perform 3D registration with the obtained reconstructions. Results with and without considering the automatic segmentation model are reported.

3.1 Dataset

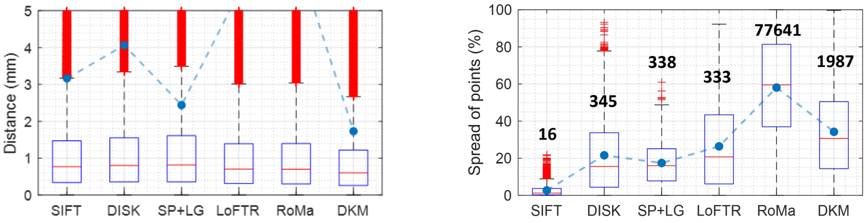
Uncompressed arthroscopic video with 1080p resolution was acquired at a frame rate of 18 fps in 7 different cadaver lab experiments of ACL reconstruction procedures, in which the distal femur is the targeted anatomy. A total of about 20k images was obtained and manually annotating all of them for generating the bone+cartilage segmentation masks would be impractical. Thus, images for training the segmentation model described in Section 2.2 were generated as described in [9] by first manually annotating a subset of frames, ranging from 3% to 30% of the total number of images, for each cadaver specimen sequence and then training a model with the architecture as described in Section 2.2 for each specimen. The objective is that the trained model overfits the input data so that it can accurately predict the labels for the remaining images of the same sequence. This approach was employed for 5 out of the 7 specimens. Three out these 5 specimen sequences, comprising a total of 10387 images were then used for training the general segmentation model and segmentation masks for the remaining 2 out of 7 specimens were generated by model inference. Table 1 in the Supplementary Material details the dataset in terms of how segmentation masks were obtained. Figure 1 in the Supplementary Material shows examples of segmentation masks predicted by the general model for Specimens 5 and 6, which were exclusively allocated for testing. The model is able to generalize to unseen data and accurately identifies bone+cartilage and background regions.

Image Pairs Selection The algorithms considered in Section 2.1 receive as input a pair of frames and output a set of keypoint correspondences. It is well known that small baselines between the frames may lead to large reconstruction errors [10], and very large baseline may cause significant viewpoint changes that cause matching algorithms to fail. Since we have continuous and known camera motion acquired at high frame rate, we can select appropriate baselines that favour 3D triangulation. For each arthroscopic sequence, 150 pairs of images with baselines between 1 and 5 mm were selected. In order to guarantee that the selected pairs present variability in terms of camera motion, clusters of frames are identified and the pairs are selected such that they belong to different clusters. Cluster selection is performed using a graph theory-based scheme. First, relative poses between all frames are computed, followed by the generation of a graph. In this graph, nodes represent camera poses, and an edge exists between nodes if the relative pose to another node falls within a threshold of 1.5 mm and $20°$. Then, each connected component of the graph is assigned to a cluster. However, in cases where the camera movement is gradual, ensuring adequate variability between clusters becomes challenging as connected components may encompass camera poses capturing different image content. To address this issue, an iterative approach is employed as follows: (i) computing the centroid of each connected component, (ii) calculating the distance of individual nodes within the component connected to the centroid, (iii) forming a new cluster if this distance exceeds 5 mm, and (iv) repeating the procedure until no further divisions occur.

Registration Data Generation Since in arthroscopic scenarios the working volume is small, the camera is typically close to the anatomy and the visible region is restricted. For this reason, for a particular pair of frames, only a small portion of the targeted anatomy becomes reconstructed. Successfully registering intra-op data with a pre-op model, requires a good spreading of reconstructed points. In order to accomplish this, data for performing registration is selected by considering N pairs of frames, with $N = 5, 10, 30$ that belong to different clusters. This selection is done in a random manner for creating 5 registration sets for each value of N.

3.2 3D Reconstruction Evaluation

For each pair of frames obtained as described in Section 3.1, keypoint correspondences are generated using each of the 6 considered algorithms and filtered by considering a threshold of 5 pixels for the epipolar error. Standard triangulation is then employed for reconstructing 3D points. In order to assess the performance of the algorithms in reconstructing points on the regions of interest (bone and cartilage), the segmentation masks are also applied for considering only those points. Two evaluation metrics are considered: spreading of points on the region of interest and distance of the reconstructed points to the pre-operative model. Spreading of points is measured by identifying all pixels in the regions of interest within a distance of 10 pixels from each matching point. This metric is quantified by determining the ratio of unique identified pixels to the total number of pixels in the regions of interest. Using the registration obtained with the original VBSN method [17] as ground truth, all reconstructed points were aligned with the pre-operative model and the distance of each point to the model was measured. Results for all methods and specimens are shown in Fig. 2. On top of each boxplot in Fig. 2b the median number of total matches is shown. Results show that the accuracy of reconstruction is similar among the methods, with DKM [7] being overall the best performer. In general, reconstruction errors are satisfactory, presenting median values below 1 mm for all methods. The most different aspect between the methods is level of spreading and the number of points they are able to reconstruct. As expected, SIFT [13] provides very local and sparse reconstructions. On the other hand, learning-based methods are able to reconstruct orders of magnitude more points, in particular RoMa [8] that is fully dense, and, more importantly, provide 3D points with good spreading.



(a) Distribution of distances of reconstructed points to the pre-op model. Blue dots represent the average. The average distances for LoFTR and RoMa are about 5 mm and 6 mm, respectively.

(b) Distribution of spreading of reconstructed points. Blue dots represent the average, and the median number of total matches is shown in bold.

Fig. 2: Assessment of reconstruction performance for different knee specimens obtained for the 6 considered keypoint matchers in terms of point a) distance to the pre-operative model and b) spreading.

3.3 3D Registration Evaluation

Registration was performed using points reconstructed from sets of pairs of frames as described in Section 3.1. For Specimens 2, 5 and 7, a full inspection of the femur was not performed, causing the acquired arthroscopic sequences to be restricted to a single condyle. This yielded reconstructions with small overlap with the corresponding pre-operative model, precluding registration from working properly. For the remaining specimens, registration was performed and results considering semantic segmentation masks are shown in Fig. 3. Results without semantic segmentation can be found in the Supplementary Material (Fig. 3). Since the arthroscopic data was acquired in the context of ACL surgery, for which there are post-op scans showing the location of the drilled tunnel, results are given in terms of tunnel placement accuracy as a distance between entry points and an angle between the directions of the intra-op and post-op tunnels.

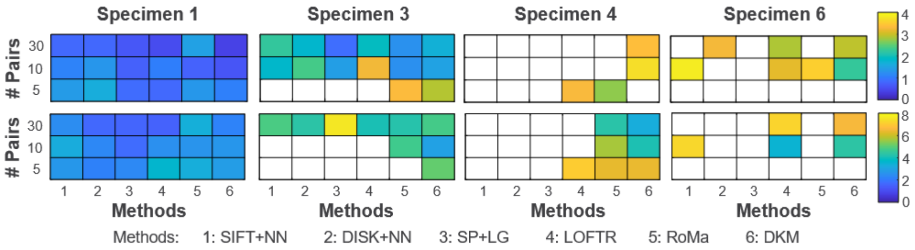


Fig. 3: Assessment of the registration accuracy in terms of tunnel placement for reconstructions obtained with the 6 different methods combined with semantic segmentation for 5, 10 and 30 pairs of images. The top row represents the median error in the entry point, in mm, and the bottom row represents the median error in tunnel direction, in degrees. White cells correspond to failure cases or errors larger than 4 mm or 8°. The color version of the image is available online.

Different specimens present significantly different performances mainly because of the different levels of texture associated with the anatomies. Examples of arthroscopic frames acquired for each specimen evincing this aspect are shown in Fig. 2 of the Supplementary Material. White cells represent failure cases[3] or errors larger than 4 mm or 8°. The most important observation of this study is that when using at least 10 pairs of images and semantic segmentation, DKM [7] yields errors below 4 mm and 8° for all specimens, achieving average errors in the entry point of 2.0 mm and in tunnel direction of 3.5°. These errors are comparable with the ones reported in the literature, with state-of-the-art works on navigated surgery reporting errors in the entry point of 2.17 mm using a robotic system [15] and average errors of 8.5° [4] and 6.74° [16] in tunnel direction. This finding demonstrates that if the camera motion is known, registration without the aid of any instrumentation is possible. Figure 4 in the Supplementary Material exemplifies the alignments of the 3D model and reconstructed points obtained for the 6 different methods using a randomly selected registration solution.

All other methods present significantly poorer performances overall and failed at least once in our experiments (refer to Table 2 in the Supplementary Material). Also, not using segmentation masks precludes 3D registration from working satisfactorily as the percentage of outliers becomes prohibitive.

4 Conclusions

This paper studies, for the first time, the possibility of performing instrument-free 3D registration in the context of arthroscopy in which a fiducial marker has been rigidly attached to the anatomy. A pipeline including image pair selection,

[3] Registration is considered to fail if it is unable to find a solution with at least 30% of inliers at 1 mm.

keypoint matching, semantic segmentation, triangulation and 3D registration is considered for evaluating the most relevant classical and deep learning-based feature matching methods, and no optimization or refinement of keypoints is performed. It has been demonstrated that by using DKM [7] as the feature matcher, instrument-free registration that places ACL tunnels with accuracy that meets the medical requirements is achieved.

We believe that by training the deep learning models with specific arthroscopic imagery can significantly improve the matching performance, which we intend to test in the future.

Acknowledgments. Tânia Baptista acknowledges the Portuguese Science Foundation (FCT) for supporting her PhD through grant 2022.12136.BDANA. The authors are also grateful to FCT for generously funding the Institute of Systems and Robotics through project UIDB/00048/2020 with DOI 10.54499/UIDB/00048/2020 (https://doi.org/10.54499/UIDB/00048/2020).

Disclosure of Interests. The authors declare no conflicts of interest relevant to the content of this article.

Ethical approval All studies involving post-mortem subjects followed the procedures for informed consent that are described in the Declaration of Helsinki.

References

1. Aiger, D., Mitra, N.J., Cohen-Or, D.: 4-points congruent sets for robust pairwise surface registration. In: ACM SIGGRAPH 2008 Papers. Association for Computing Machinery (2008)
2. Baptista, T., Marques, M., Raposo, C., Ribeiro, L., Antunes, M., Barreto, J.P.: Structured light for touchless 3D registration in video-based surgical navigation. Int J CARS (2024)
3. Besl, P.J., McKay, N.D.: Method for registration of 3-D shapes. In: Schenker, P.S. (ed.) Sensor Fusion IV: Control Paradigms and Data Structures. vol. 1611, pp. 586–606. International Society for Optics and Photonics, SPIE (1992)
4. Cho, W.J., Kim, J.M., Kim, D.E., Lee, J.G., Park, J.W., Han, Y.H., Seo, H.G.: Accuracy of the femoral tunnel position in robot-assisted anterior cruciate ligament reconstruction using a magnetic resonance imaging-based navigation system: A preliminary report. The International Journal of Medical Robotics and Computer Assisted Surgery **14**(5), e1933 (2018)
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)
6. Edgcumbe, P., Pratt, P., Yang, G.Z., Nguan, C., Rohling, R.: Pico Lantern: Surface reconstruction and augmented reality in laparoscopic surgery using a pick-up laser projector. Medical Image Analysis **25**(1), 95–102 (2015)
7. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: DKM: Dense kernelized feature matching for geometry estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (2023)
8. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust Dense Feature Matching. arXiv preprint arXiv:2305.15404 (2023)

9. Félix, I., Raposo, C., Antunes, M., Rodrigues, P., Barreto, J.P.: Towards markerless computer-aided surgery combining deep segmentation and geometric pose estimation: application in total knee arthroplasty. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **9**(3), 271–278 (2021)
10. Gallup, D., Frahm, J.M., Mordohai, P., Pollefeys, M.: Variable baseline/resolution stereo. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008)
11. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: ICCV (2023)
12. Long, Z., Nagamune, K.: Underwater 3D Imaging Using a Fiber-Based Endoscopic System for Arthroscopic Surgery. JACIII **20**(3), 448–454 (2016)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)
14. Mellado, N., Aiger, D., Mitra, N.J.: Super 4PCS Fast Global Pointcloud Registration via Smart Indexing. Computer Graphics Forum **33**(5), 205–215 (2014)
15. Na, G., Wang, T., Wei, M., Hu, L., Liu, H., Wang, Y., Yang, B., Yu, G.: An ACL reconstruction robotic positioning system based on anatomical characteristics. International Journal of Advanced Robotic Systems **17**(1), 172988141988616 (2020)
16. Park, S.H., Moon, S.W., Lee, B.H., Park, S., Kim, Y., Lee, D., Lim, S., Wang, J.H.: Arthroscopically blind anatomical anterior cruciate ligament reconstruction using only navigation guidance: a cadaveric study. The Knee **23**(5), 813–819 (2016)
17. Raposo, C., Sousa, C., Ribeiro, L., Melo, R., Barreto, J.P., Oliveira, J., Marques, P., Fonseca, F.: Video-Based Computer Aided Arthroscopy for Patient Specific Reconstruction of the Anterior Cruciate Ligament. In: MICCAI. pp. 125–133. Springer-Verlag, Berlin, Heidelberg (2018)
18. Reiter, A., Sigaras, A., Fowler, D., Allen, P.K.: Surgical structured light for 3D minimally invasive surgical imaging. In: IROS. pp. 1282–1287. IEEE (2014)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
20. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning Feature Matching With Graph Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
21. Schmalz, C., Forster, F., Schick, A., Angelopoulou, E.: An endoscopic 3D scanner based on structured light. Medical Image Analysis **16**(5), 1063–1072 (2012)
22. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-Free Local Feature Matching With Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8922–8931 (June 2021)
23. Tyszkiewicz, M., Fua, P., Trulls, E.: DISK: Learning local features with policy gradient. In: Advances in Neural Information Processing Systems. vol. 33, pp. 14254–14265. Curran Associates, Inc. (2020)