Deep segmentation leverages geometric pose estimation in computer-aided total knee arthroplasty

Pedro Rodrigues[1] ✉, *Michel Antunes*[2], *Carolina Raposo*[2], *Pedro Marques*[3], *Fernando Fonseca*[3], *Joao P. Barreto*[1,2]

[1]*Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal*
[2]*Perceive 3D, Coimbra, Portugal*
[3]*Faculty of Medicine, Coimbra Hospital and University Centre, Coimbra, Portugal*
✉ *E-mail: prodrigues@isr.uc.pt*

Published in Healthcare Technology Letters; Received on 18th September 2019; Accepted on 2nd October 2019

Knee arthritis is a common joint disease that usually requires a total knee arthroplasty. There are multiple surgical variables that have a direct impact on the correct positioning of the implants, and an optimal combination of all these variables is the most challenging aspect of the procedure. Usually, preoperative planning using a computed tomography scan or magnetic resonance imaging helps the surgeon in deciding the most suitable resections to be made. This work is a proof of concept for a navigation system that supports the surgeon in following a preoperative plan. Existing solutions require costly sensors and special markers, fixed to the bones using additional incisions, which can interfere with the normal surgical flow. In contrast, the authors propose a computer-aided system that uses consumer RGB and depth cameras and do not require additional markers or tools to be tracked. They combine a deep learning approach for segmenting the bone surface with a recent registration algorithm for computing the pose of the navigation sensor with respect to the preoperative 3D model. Experimental validation using ex-vivo data shows that the method enables contactless pose estimation of the navigation sensor with the preoperative model, providing valuable information for guiding the surgeon during the medical procedure.

1. Introduction: Osteoarthritis is a joint disease that causes pain and stiffness due to damage to the joint cartilage and the underlying bone [1]. It is the most common joint disease in the world, with an estimated prevalence of 14% in adults with 25 years or older and 34% with 65 years or older [1]. Depending on the severity of the symptoms, the treatment options can vary from non-operative to joint arthroplasty. Total knee arthroplasty (TKA) is the principal choice for improving the quality of life of patients suffering from advanced knee arthritis. It is estimated that the demand for TKA in the United States will approach 3.5 million cases per year by 2030 [2]. Although being one of the most effective surgical options to reduce pain and restore the knee function, about 20% of the patients undergoing a TKA surgery are not satisfied [3]. As discussed in [4], there are important surgical variables (e.g. lower leg alignment and soft tissue balancing) that have a direct impact on the success of TKA, which are manually controlled by the orthopaedic surgeon. It requires experience for accurately combining all these surgical variables into an optimal implant alignment. To assist the surgeon in controlling these variables, several computer navigation systems have been developed [4].

Existing TKA navigation systems require a sensing technology for performing anatomical measurements or supporting the surgeon in following a preoperative plan of the bone resections. The most widely used technology for this purpose is optical tracking (e.g. Smith & Nephew's NAVIO Surgical System [5]). While providing accurate 3D measurements, navigation systems based on optical tracking have three main drawbacks: (i) optical tracking platforms are costly, which is one of the reasons for the high cost of existing navigation systems, (ii) they necessitate the insertion of pins in the distal femur and proximal tibia for fixing the markers to be tracked, requiring additional bone incisions and surgical time, and (iii) the trackers to be attached to the bones are bulky, interfering with the normal surgical flow.

This Letter describes and provides a proof of concept for the first contactless video-based system for computer-aided TKA that does not require any special markers to be attached to the body. A navigation sensor, integrating a consumer RGB camera and a depth camera, is used to register an anatomical model of the patient, obtained with a preoperative computed tomography (CT) scan or magnetic resonance imaging (MRI), such that the bone resections for the implant positioning can be guided according to a preoperative plan. The Letter introduces the video-based computer-aided TKA system, and describes the two main modules of the software pipeline: (i) bone surface segmentation from RGB images using a deep learning technique, and (ii) registration of a preoperative CT/MRI model with a noisy point cloud for computing the pose of the navigation sensor. Augmented reality (AR) techniques for supporting the surgeon in following a preoperative plan are then used. Experimental results in real ex-vivo data are presented.

2. Video-based computer-aided TKA: This section overviews the proposed concept for computer-aided TKA that uses a navigation sensor to perform 3D pose estimation during the open surgery. By using additional depth-sensing capabilities, we avoid the use of visual markers as in [6] (see Fig. 1).

Given a 3D model of the patient's knee, which was acquired using CT or MRI, the surgeon prepares a preoperative plan for optimising the implanting positioning, defining the resection plane parameters that will guide the proposed computer-aided TKA system. During the surgery, and at each time instant, RGB and depth data are captured from the navigation sensor and used for computing a local 3D point cloud of the knee joint. The navigation sensor is handheld and is composed of a camera and a depth sensor that are at all times, fixed together in a rigid manner. The point cloud is then used to register the 3D preoperative model, enabling the estimation of the pose of the navigation sensor with respect to the model. In this way, the location of the resection planes with respect to the navigation sensor is known, enabling to provide valuable guidance information to the surgeon.

2.1. Segmentation of bone surface: Image segmentation is a widely investigated topic in medical imaging and computer vision. As in

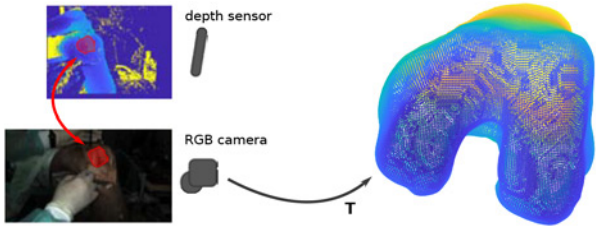

Fig. 1 *Diagram of the proposed method. Bone segmentation is performed using frames of the RGB camera. The segmentation is used to get the region of interest in the point cloud from the depth sensor. The point cloud is then registered to the anatomical model to establish the relative pose*

many medical imaging applications, in TKA, it is difficult to obtain large amounts of accurately labelled data. The main difficulty is that crowd-sourcing frameworks cannot be used, because the labelling of medical data requires expert knowledge and special confidentiality aspects.

We will explore a deep learning approach for image segmentation called U-Net [7], an encoder–decoder neural network architecture with skip-connections between the encoder and the decoder sections of the network [7]. This network was shown to achieve good results with a relatively small number of annotated images for training and for a wide variety of objects and scenarios [7, 8]. The rationale behind the skip-connections is that when doing a traditional encoder–decoder approach some fine-grained details are lost in the encoder, and as the signal is upscaled through the decoder it cannot describe the object in the input image with high resolution. These special connections aim to solve this problem.

The base neural network used was as provided by the authors in [8]. The data processing, the optimisation process, along with some other modifications, were implemented around the base U-net by Iglovikov and Svets [8]. For the encoder part of the U-Net, the weights were initialised to the weights of the VGG-11 network taking advantage of large datasets of generic training data, similar to [8]. However, since our training dataset is small, we decided to freeze the encoder weights, while the rest of the network is optimised. This led to better results and faster training. Dropout was implemented for regularisation. The hyperparameter space for the optimisation is defined by the learning rate, dropout ratio, and the number of filters in the convolutional layers of the decoder. The number of filters in the decoder decreases in a similar way to the encoder growth, but as multiple of this hyperparameter. As for the encoder, the number of filters is fixed, because the weights being used are the same as the weights of the VGG-11 network. A mini-batch of 10 was combined with the Adam optimiser for training the network.

The dataset used in this Letter contains several video sequences of two different femurs (three sequences for the first femur and one sequence for the second femur). The datasets have a wide variety of relative poses and occlusion events. In some of the sequences, a marker was attached to the bone before data acquisition. This base marker was tracked through the sequences so that it could be used to aid in the generation of ground truth data for segmentation and to compare trajectories for evaluation of pose estimation. Due to the dimensions of the dataset (~10,000 images), manual segmentation of all the images was not possible, and a semi-automated approach was used for the labelling task. This was accomplished by manually segmenting ~100 images, and propagating the segmentation to the neighbouring frames using the detected marker pose and the 3D femur model. The dataset was split into ~9000 images for training and 1000 for validation. Note that the validation dataset was taken as a different video sequence and not randomly chosen frames of the same sequences to make the validation dataset as different as possible to the training dataset. Although parts of the dataset also contain depth information, only the RGB data was used for learning to segment bone surfaces. For augmenting the variability in the training dataset, we performed particular transformations to the input images on-the-fly. Among these transformations we included image rotation, imaging flipping, and shifts to the HSV space of the images. Additionally, the base marker was masked and inpainted over to avoid implicit relationships between the poses of the marker and the femur within the learning framework and improve generalisation.

2.2. Registration of a preoperative model with a noisy point cloud: 3D registration consists of aligning two models such that their overlapping areas are maximised. It is a well-studied problem in computer vision, with applications ranging from SLAM and tracking to robotics and, more recently, to medicine [9]. Some solutions for the 3D registration problem work by matching features extracted from the models and estimating the rigid transformation using RANSAC or other robust estimators [10, 11]. Such approaches perform poorly when the point clouds are too smooth and/or noisy because of the difficulty in finding repeatable features. The family of algorithms 4PCS [12–14] makes use of hypothesise-and-test schemes that randomly select sets of four coplanar points in one point cloud and find correspondences in the other for establishing alignment hypotheses. Recently, Raposo and Barreto [15] proposed an algorithm that is faster than the 4PCS family of methods and resilient to very high levels of outliers. In general terms, the method extracts pairs of points and their normals in one point cloud, finds congruent pairs of points in the other and afterwards establishes alignment hypotheses that are tested in a RANSAC scheme. The selected solution is refined using a standard ICP [16] approach.

Closely related to our work in terms of application is [9], which also employs 3D registration in the context of orthopaedic surgery for aligning a preoperative model of the targeted bone with the patient's anatomy. However, there are two important differences: (i) while Raposo and Barreto [9] include an explicit surgical step where the surgeon touches bone surface with a tracked probe for reconstructing 3D points, our method uses a depth sensor and an automatic segmentation process for reconstructing only the area corresponding to the targeted anatomy; (ii) Raposo and Barreto [9] require fiducial markers attached to the patient's body for estimating the camera pose. On the other hand, our approach accomplishes camera pose estimation by registering the segmented point cloud with the preoperative model at each frame. The registration algorithm proposed in [9] is fast, accurate and robust to outliers. However, it is not suitable for our case because it only solves the curve-versus-surface alignment problem and we require surface-versus-surface registration. As mentioned in the previous paragraph, this task is efficiently solved in [15].

Since it is reported that this method is able to handle outliers, we attempted to register the complete point cloud obtained from the depth sensor with the preoperative model. However, due to the significant levels of noise and very high percentages of outliers, the results were not satisfactory, evincing the need for proper segmentation of the bone surface. The registration parameters were tuned for accommodating the noise in the data, and qualitative and quantitative results on the registration accuracy are provided in the next section. Furthermore, some frames contain too many outliers and missing information, whether due to the sensor being too close to the knee (out of range), to specularities and/or total occlusion. Therefore a registration was deemed successful only when 80% of inliers were considered for its computation.

3. Experimental results: The camera/sensor setup used was composed of a 1080p video camera and a two-camera IR depth sensor with 480p resolution at approximately ten frames per second. The two components were fixed together and calibrated.

To evaluate the proposed method both quantitatively and qualitatively, we have three datasets: the training and validation datasets used in the machine learning framework, and an additional dataset were no markers were inserted into the bone.

The proposed architecture for femur segmentation obtained an intersection over union (IoU) metric of median 0.853 and a Dice coefficient of median 0.921. The hyperparameters used were learning rate – 0.0001, epochs – 10, number of filters for the decoder – multiples of 2. Please refer to Figs. 2 and 3 for additional results regarding femur segmentation.

To evaluate the registration, we computed the trajectory obtained in the validation dataset and compared it to the trajectory obtained by tracking the marker inserted in the femur. The two trajectories were aligned using a rigid transformation, and the results are shown in Fig. 4*a*. Fig. 4*b* shows the angular magnitude of the

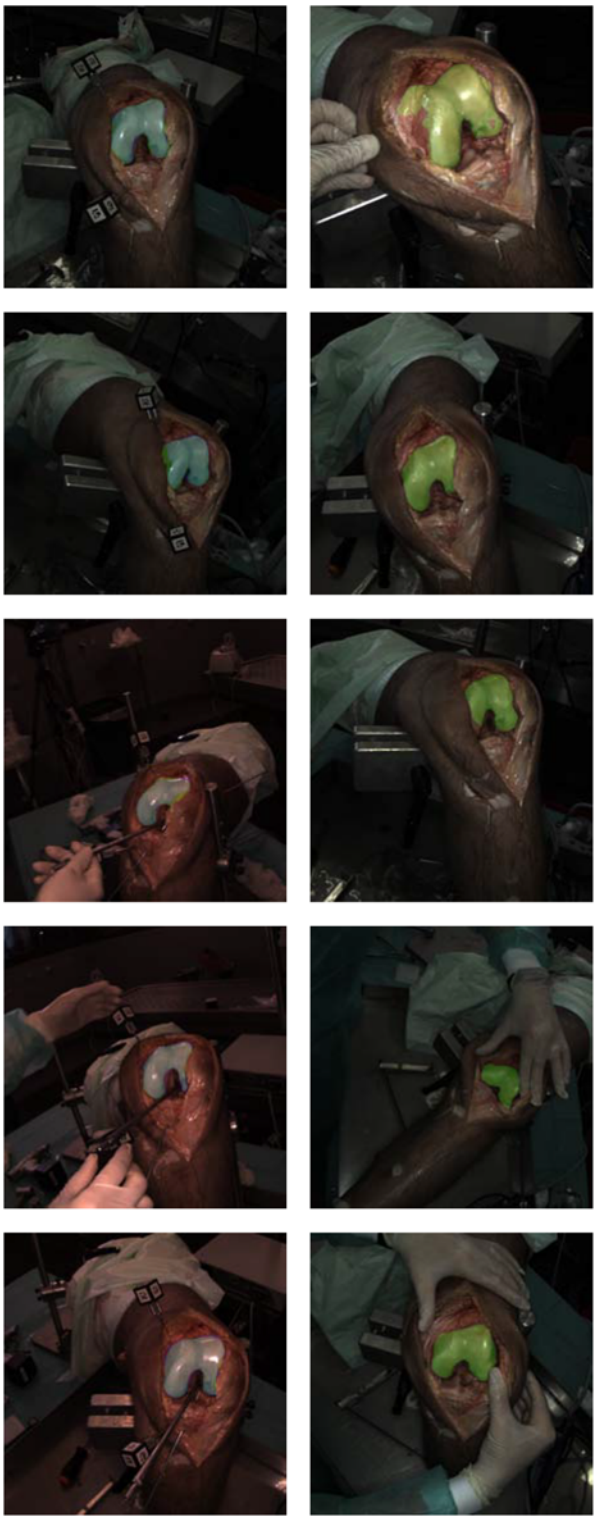

Fig. 2 *Femur segmentation results on the validation dataset (left column) and on the dataset without markers (right column), which does not have ground truth available. Green: prediction; blue: ground truth; cyan: both*

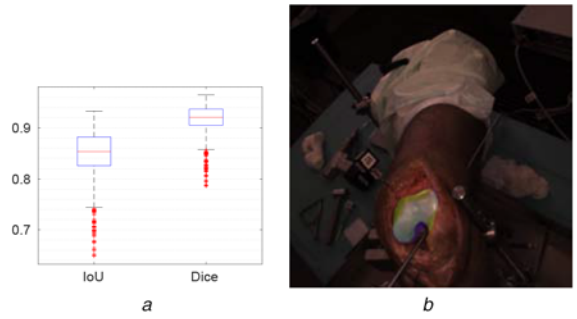

Fig. 3 *Segmentation metrics for the femur segmentation in the validation dataset*
a Metric distribution
b Frame with worst IoU metric

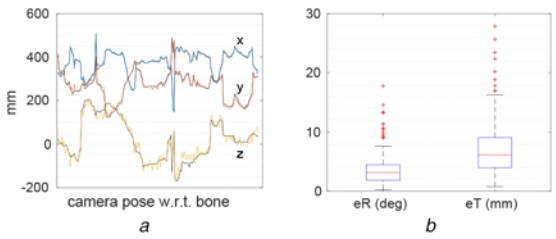

Fig. 4 *Registration results in the validation dataset*
a Comparison between the *x*, *y*, and *z* components of the trajectories of the proposed registration (colours) and marker-based tracking (black)
b Distribution of the rotation error (eR) in degrees and the translation error (eT) in millimetres

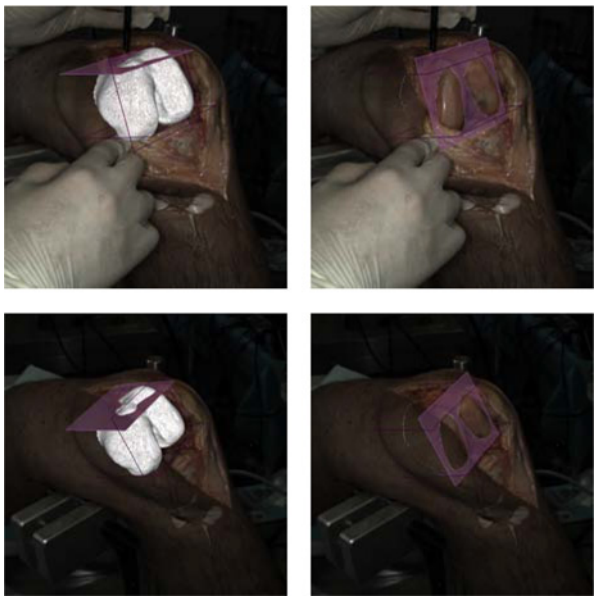

Fig. 5 *Each row corresponds to a different frame showing the markerless and contactless femur registration for AR-guided surgery using pre-operative planned cuts*

doi: 10.1049/htl.2019.0078

Fig. 6 *Femur segmentation results using video sequences of only one femur for training: good results (left column) and bad results (right column). Green: prediction; blue: ground truth; cyan: both*

residual rotation between the ground truth and the estimated rotations, in degrees, and the norm of the difference between the ground truth and the obtained translation vectors; where the ground truth was taken as the median pose of all successful registrations, thus giving a measure of robustness of the registration procedure. The obtained median rotation error (eR) is 3.17° and the median translation error (eT) is 6.18 mm. In Fig. 5, our contactless registration is used to superimpose the bone model and show the preoperative plan in an AR view.

To further test the generalisation power of our method, we performed a new hyperparameter search, this time using only one femur for training (three video sequences of the first femur) and one for testing (one video sequence of the second femur). This test aims to show that generalisation to other scenes is possible. However, using only one femur for training is not ideal, and for a fully working framework, further data is needed. Fig. 6 shows the results.

4. Discussion: The Letter proposed a new approach for navigation in TKA that avoids the need for attaching fiducials to the anatomy, which is a major problem in current navigation systems and cause surgeons to avoid these techniques. Moreover, the proposed approach uses off-the-shelf hardware and does not require any user input.

The segmentation worked under various conditions and surpassed expectations in differentiating between the bone and the adjacent tissue with similar colour and texture, even though only ∼100 images were manually segmented. The scarcity of the data required for performing machine learning tasks means that a fair evaluation of the segmentation algorithm is difficult. Further testing segmentation with additional ex-vivo knees may be necessary to confirm the generalisation power shown here.

Regarding the full registration process, the work aimed to be a proof of concept that demonstrated that it is feasible to robustly track the anatomy without the need of attaching fiducial markers. The results are encouraging, but there is still work to do to accomplish a final system that can be used in everyday clinical practice:

- Translation errors of 6.18 mm and rotation errors of 3.17°, while satisfactory for a proof of concept, are above the requirements for surgical navigation. The obtained errors can lead to critical misalignment of the planned cuts and drills. Future work will address this problem and focus on fine-tuning the registration algorithm to work under such extreme depth outlier conditions and possibly use multiple frames to perform the pose estimation. Another promising line of research is to eliminate the need to work with the depth sensor and perform pose estimation with machine learning as well. This is enticing since, in our setup, the depth map is the main source of imprecision. Another line of research with potential would be to use an end-to-end machine learning approach.
- Occlusion is currently a problem for the segmentation. However, this happens because only 100 manual segmentations were performed. Further manual segmentations can be performed for better resilience to occlusion. Another possible approach is to track the surgical tools and remove them from the segmentation maps to generate the dataset.
- Future work must comprise an extension to other anatomies. So far, we have worked only with the femur. Extension to other procedures where the anatomy is not so clearly exposed (e.g. hip arthroplasty) may not be as straightforward. Nevertheless, evaluation of accuracy in such cases may be interesting. Extension for the tibia, as required for full TKA navigation, should be feasible but must be validated as well.

Although additional testing is required for a full navigation system to be used in the OR, this work opens the possibility for a contactless registration to be used to guide the surgeon. A possible path for the application of our work is to use the contactless registration to guide the drilling of the holes for the cut guide and only then guide the distal cut. In this way, removing the necessity of the registration of the bone after the cuts, which are not contemplated by the present work. However, this approach would require the holes to be drilled before the first cut, and further testing is mandatory.

5. Acknowledgments: The authors thank the Portuguese Science Foundation (FCT) and the COMPETE 2020 program for the generous funding through project VisArthro (ref.: PTDC/EEIAUT/3024/2014). This work was also funded by the European Union's Horizon 2020 research and innovation programmes under grant agreement no 766850. This work has also been supported by FCT under the PhD scholarship SFRH/BD/113315/2015 and by OE – national funds of FCT/MCTES (PIDDAC) under project UID/EEA/00048/2019.

6 References

[1] Scott W.N.: 'Insall & Scott surgery of the knee' (Elsevier, Oxford, UK, 2017)
[2] Kurtz S., Ong K., Lau E., *ET AL.*: 'Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030', *J. Bone Joint Surg. Am. Vol.*, 2007, **89**, pp. 780–785
[3] Bourne R.B., Chesworth B.M., Davis A., *ET AL.*: 'Patient satisfaction after total knee arthroplasty: who is satisfied and who is not?', *Clin. Orthop. Relat. Res.*, 2009, **468**, pp. 57–63
[4] Van der List J., Chawla H., Joskowicz L., *ET AL.*: 'Current state of computer navigation and robotics in unicompartmental and total

knee arthroplasty: a systematic review with meta-analysis', *Knee Surg. Sports Traumatol. Arthrosc.*, 2016, **24**, pp. 3482–3495
[5] Smith & Nephew: 'Navio surgical system'. Available at http://smithnephewlivesurgery.com/navio-surgical-system, accessed 29 March 2019
[6] Raposo C., Sousa C., Ribeiro L.L., *ET AL.*: 'Video-based computer aided arthroscopy for patient specific reconstruction of the anterior cruciate ligament'. Medical Image Computing and Computer Assisted Intervention (MICCAI), Granada, Spain, 2018
[7] Ronneberger O., Fischer P., Brox T.: 'U-net: convolutional networks for biomedical image segmentation', Lecture Notes in Computer Science, vol. 9351, in Navab N., Hornegger J., Wells W., Frangi A. (eds): 'Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015'. (Springer, Cham, 2015), pp. 234–241
[8] Iglovikov V., Shvets A.: 'Ternausnet: U-net with VGG11 encoder pre-trained on ImageNet for image segmentation', 2018. Available at http://arxiv.org/abs/1801.05746
[9] Raposo C., Barreto J.P.: '3D registration of curves and surfaces using local differential information'. The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, 2018
[10] Rusu R.B., Blodow N., Beetz M.: 'Fast point feature histograms (FPFH) for 3D registration'. IEEE Int. Conf. on Robotics and Automation, Amsterdam, Netherlands, May 2009, pp. 3212–3217
[11] Zhou Q.Y., Park J., Koltun V.: 'Fast global registration'. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), European Conf. on Computer Vision, 2016, (*LNCS*, **9906**), pp. 766–782
[12] Mohamad M., Ahmed M.T., Rappaport D., *ET AL.*: 'Super generalized 4PCS for 3D registration'. Proc. – 2015 Int. Conf. on 3D Vision, 3DV 2015, Ecole Normale Superieure, Lyon, October 2015, pp. 598–606
[13] Aiger D., Mitra N.J., Cohen-Or D.: '4-points congruent sets for robust pairwise surface registration'. SIGGRAPH'08: Int. Conf. on Computer Graphics and Interactive Techniques, ACM SIGGRAPH 2008 Papers 2008, Los Angeles, CA, USA, August 2008
[14] Mellado N., Aiger D., Mitra N.J.: 'SUPER 4PCS fast global point cloud registration via smart indexing'. Eurographics Symp. on Geometry Processing, Airela-Ville, Switzerland, 2005, vol. 33, (5), pp. 205–215
[15] Raposo C., Barreto J.P.: 'Using 2 point+normal sets for fast registration of point clouds with small overlap'. 2017 IEEE Int. Conf. on Robotics and Automation (ICRA), Marina Bay Sands, Singapore, May 2017, pp. 5652–5658
[16] Besl P.J., McKay N.D.: 'A method for registration of 3-D shapes', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1992, **14**, (2), pp. 239–256