Improving 3D Registration Results of Foot Models Dramatically with a Machine Learning Enhanced Geometric Feature Extraction

Tobias PFROMMER
ShoeFitter GmbH, Konstanz, Germany

https://doi.org/10.15221/22.43

Abstract

In this paper, a method is presented that enables a true-to-scale reconstruction of 3D foot models using the iPhone's Face ID sensor. For this purpose, multiple incoming 3D point clouds representing the foot are registered piece by piece with each other. A feature-based registration pipeline is used for pairwise registration. Geometric feature extraction in such pipelines is the first and most important step for correct registration of two 3D point clouds. For this purpose, we train and apply learned feature descriptors based on Fully Convolutional Geometric Features (FCGF) [1]. It is shown that the features computed by our trained feature extractor are more robust and faster than conventional methods. We trained FCGF using a self-generated dataset of 3D foot models augmented with synthetic data. The trained feature model was optimized with hyperparameters. For better visualization of the high-dimensional features, a t-SNE-based visualization is used to assign features that are reliably found in the same location of the foot in different models [2]. Based on the detected features, the optimal transformation of two point clouds is estimated by a feature-based RANSAC algorithm [3]. In the benchmarks, it is found that the implemented feature descriptor consistently achieves better feature matching and registration recall results than comparable feature descriptors. With the final trained model of the feature descriptor within the presented registration pipeline, a 3D reconstruction of a foot can be performed using an overlap of only 27 percent. This makes the reconstruction of the 3D model much more robust than using comparable state-of-the-art methods.

Keywords: 3d body scanning, point cloud registration, foot measurement, feature descriptor, 3d reconstruction, deep learning, fully convolutional network, 3d foot model

1. Introduction

Not least due to the trend of AR and VR, depth sensors have become increasingly popular in smartphones. This technological advancement enables the development of disruptive, mobile applications in the field of 3D reconstruction and surveying. There is a particularly great need for such applications in the body scanning industry due to trends like customization, eHealth and fashion eCommerce. Specifically, the footwear industry suffers from high size related returns because fitting is particularly problematic with shoes. Therefore, a dependable reconstruction of body parts, or in this case feet, into an accurate 3D model makes it possible to conveniently scan the body with a smartphone and extract specific measurements to provide a precise fitting recommendation.

Technologically, this is achieved by the precise alignment of individual images from the Face ID (True Depth) sensor integrated in the iPhone. With the help of this structured light sensor, three-dimensional point clouds can be reconstructed using the recorded depth data. The problem here is that the individual point clouds are positioned differently in space due to the movement of the camera. The so-called point cloud registration, which is often used in robotics, deals exactly with this problem. The goal is to resolve the perspective distortion caused by the camera movement and to align the point clouds correctly.

The precise alignment of the individual point clouds is essential in order to create a complete foot model that is as accurate as possible using depth sensors, from which user-specific foot dimensions can then be extracted. A particular challenge here is the surface of the scanned organic objects, which is very different from the typical inorganic applications of this problem. For example, the point clouds of foot models have significantly fewer geometrically distinctive surface structures such as corners or edges than classic indoor or outdoor data sets. Within this work, a method is presented with which point clouds of 3D foot models can be reliably registered. This method can be translated to other organic object and body parts.

2. Registration Pipeline

The developed registration pipeline essentially consists of two main steps. The first consists of computing feature points within a point cloud, based on which corresponding features are determined in the second step, and from which a transformation is determined to align the two point clouds. First, the data used are described, which are required for the implementation of the model for feature detection.

2.1. Dataset

For training the FCGF model, the open-source data of the MPI FAUST dataset is used [4]. This consists of 300 high-resolution scans that capture humans using various poses. Each pose contains a camera image of the pose as well as a 3D mesh of the pose relevant for the work [4]. In the first step, the 3D mesh is converted into a 3D point cloud for this purpose. Then, the lower part of the point cloud is segmented, so that the point clouds only show the feet of the individual poses. Since the dataset is topologically labeled, geometric correspondences cannot be determined. For this reason, the individual point clouds are artificially rotated and translated after the acquisition. For this purpose, the original point cloud is rotated by a random angle α in the range of 0-360 degree around the x-axis, a random angle β in the range of 0-360 degree around the y-axis and a random angle γ in the range of 0-360 degree around the z-axis. Furthermore, a random translation in x, y and z direction is applied to the point cloud.

For every original point cloud S from the FAUST dataset six randomly generated point clouds $S_{t1}, \dots, S_{t6}$ are generated with this method. The stored random rotation and translation can be used as ground truth transformation. The final dataset consists of 3927 point clouds of foot models. This is divided into 2639 training data, 644 test data and 644 validation data. An excerpt of the data can be seen in figure 1.

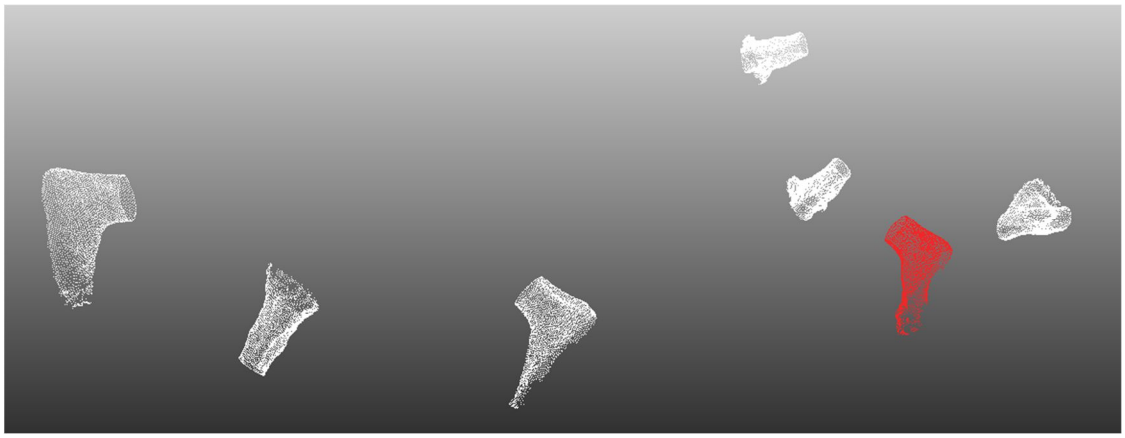


Fig. 1. Generation of a synthetic data set by copying and transformation of an original point cloud (colored in red) into several transformed point clouds.

2.2. Feature Detection with Fully Convolutional Geometric Features

For the first step in the registration pipeline, a modified model of the Fully Convolutional Geometric Feature feature descriptor is used [1]. Therefore, a custom model is trained with some changes with the help of the presented data set. The network architecture in use is based on a ResUNet architecture [1]. The input for such a network is a special sparse 4D tensor, which drastically reduces the memory and computation overhead. The output of such a network is called Fully Convolutional Geometric Features [1]. They are characterized by rotation and translation invariance and have excellent values in metrics like accuracy and number of features [1]. This makes them particularly interesting for the application described. A visualization of detected features using the t-SNE algorithm can be seen in figure 2.

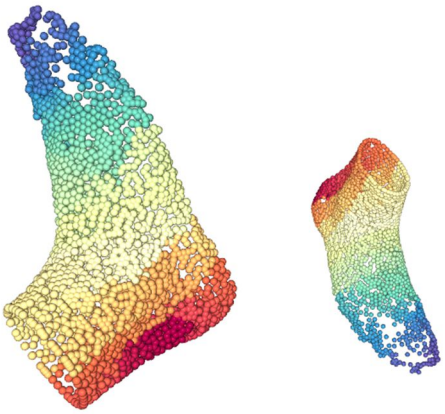


Fig. 2. Color representation of calculated feature points using two exemplary point clouds from the FAUST dataset. Similar feature points are mapped with the same color.

2.3. Point Cloud Alignment

To determine the optimal transformation of two point clouds using the computed features, the feature-based RANSAC algorithm is used [3]. The algorithm works iteratively, selecting random points from the original point cloud in each iteration. Corresponding points in the target point cloud are computed using a nearest-neighbor strategy in high-dimensional feature space [3]. Corresponding points can be seen as an example in figure 3. Based on the determined correspondences, a transformation is estimated, which is validated on the whole point cloud. The final output of the algorithm represents a final transformation that ideally superimposes two corresponding point clouds.

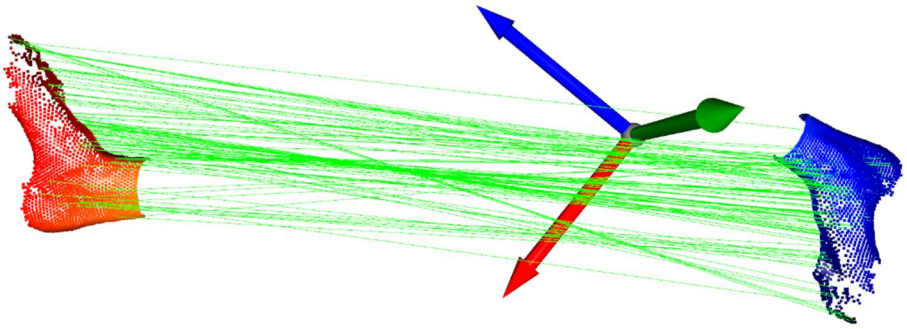


Fig. 3. Corresponding points of two point clouds based on which the final transformation is estimated with the feature-based RANSAC algorithm.

3. Evaluation

In order to evaluate the developed method quantitatively and visually, first a method is presented to map high dimensional features on a low dimensional color scale. Then, two evaluation metrics are described that can be used to qualitatively evaluate features. The presented approach is compared with the Fast Point Feature Histograms (FPFH) feature descriptor, a geometric traditional approach which often is used in such application scenarios [5].

3.1. Color coding for visualization of features

Since a visual representation of the features in a 32- or 64-dimensional space is very difficult, the t-Distributed Stochastic Neighbour Embedding (t-SNE) is used to make a visual evaluation of the feature detection possible [2]. The t-SNE consists of two main steps. In the first step a probability distribution is laid over pairs of high dimensional objects. This results in similar high-dimensional objects being assigned a high probability, while dissimilar objects are assigned a low probability. In the second step a similar probability distribution is laid over the points in the low-dimensional map and tries to minimize the Kullback-Leibler-Divergence which is an indicator for the similarity of two probability distributions [2]. Similarities are finally mapped to a color space which means that similar feature points in two

corresponding point clouds have the same color (see figure 4). As shown, features in the two corresponding point clouds are reliably detected. Only in the area of the heel there are minimal deviations.

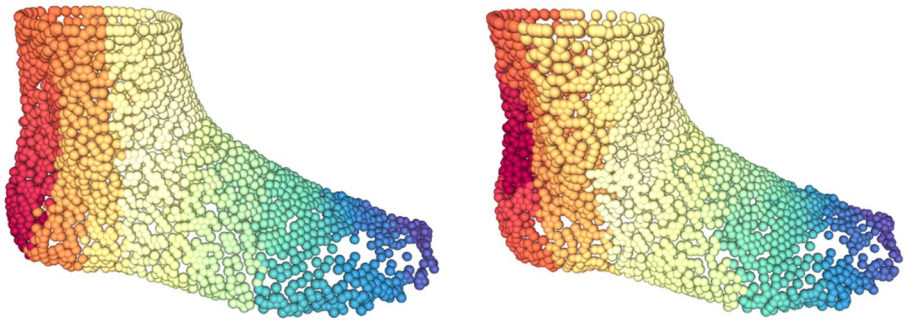


Fig. 4. Color-coded features of two different foot point clouds using the t-SNE algorithm.

3.2. Feature Match Recall

To measure the quality of the implemented features, two different evaluation metrices are used. The first one is the feature match recall. Therefore, it is assumed that a pair of corresponding point cloud fragments S and St are aligned by a rigid transformation T. To map input points to the feature space, a nonlinear feature function is used. The feature match recall is computed by equation 1 [1].

$$R = \frac{1}{M}\sum_{s=1}^{M} \mathbb{1}\left(\left[\frac{1}{|\Omega_s|}\sum_{(i,j)\in\Omega_s} \mathbb{1}(\| T^* x_i - y_i \| < \tau_1)\right] > \tau_2\right) \qquad (1)$$

M indicates the number of point cloud fragments. $\Omega_s$defines a set of correspondences within a pair of point cloud fragments s. The variables x and y represent 3D coordinates of the two point cloud fragments. T^* describes the ground truth transformation. τ_1 is a metric limit, the so-called inlier distance threshold, which specifies how large the Euclidean distance of pairs to be matched may be after the ground truth transformation. τ_2 defines the inlier threshold, which specifies the percentage of matches that are detected as true matches [1].

The Feature Match Recall measures the quality of features within a system, where a set of points are registered pairwise [1]. The results of the implemented system can be seen using figure 5 and figure 6. The developed feature descriptor model is compared with the FPFH feature descriptor. We ran different test scenarios on our test dataset with different inlier ratios in the range of [0 – 0.2] (figure 5). It can be seen that the implemented feature descriptor outperforms the FPFH feature descriptor. Even with a high inlier ratio of 0.15, 81% of the corresponding pairs still could be associated.

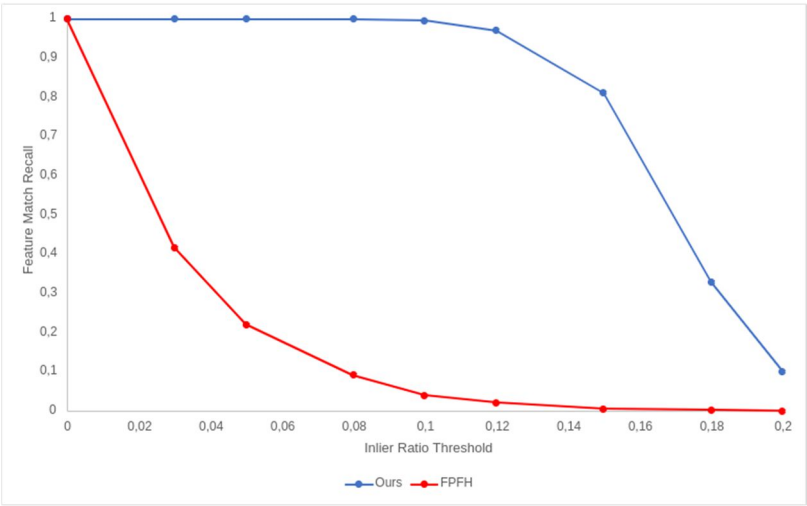


Fig. 5. Feature match recall for different limits of inlier ratio. The developed FCGF descriptor model compared to FPFH.

Different scenarios with varying inlier distance threshold were also tested. The results can be seen in figure 6. It is shown that the implemented descriptor is extremely accurate and calculates features within less than a millimeter, while the FPFH can have more deviations in the centimeter range.

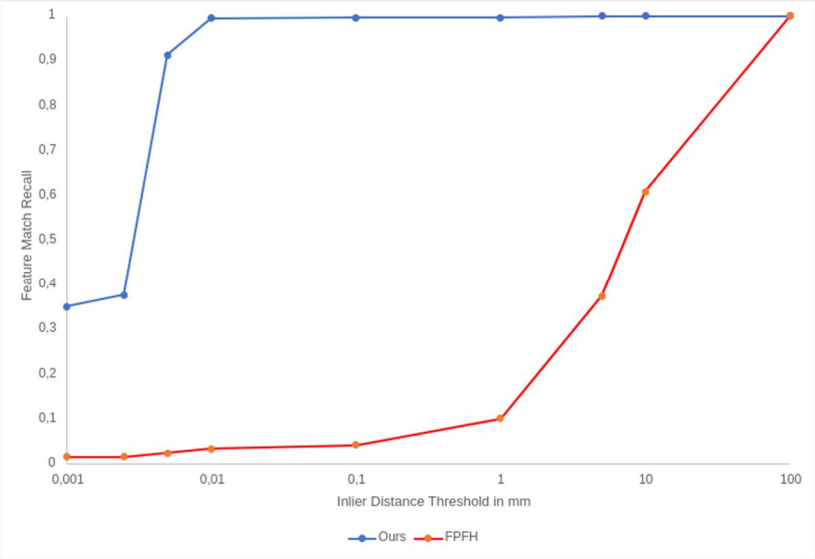


Fig. 6. Feature match recall for different limits of inlier distance. The developed FCGF descriptor model compared to FPFH.

3.3. Registration Recall

The registration Recall calculates on a set of overlapping point cloud fragments on which a ground truth is denoted, how many overlapping fragments can be retrieved by a matching algorithm. The Registration Recall uses the error metric in equation 2 to calculate true positives [1].

$$E_{RMSE} = \sqrt{\frac{1}{\Omega*} \sum_{(x^*, y^*) \in \Omega*} \| T_{i,j} x^* - y^* \|^2} \quad (2)$$

Thereby $\{i, j\}$ is a pair of predicted fragments with the predicted transformation T. $\Omega*$ denotes a set of corresponding ground truth pairs within the fragments $\{i, j\}$. x^* and y^* are the 3D coordinates of the ground truth pair [1]. For fragment pairs, that have an overlap of at least 99% the registration is denoted as a correct pair. Due to the nature of our data set, the theoretical overlap is 100%, so the value is chosen accordingly high. Table 1 shows the results of the registration recall in the test scenario. The implemented feature descriptor model is compared with the FPFH feature descriptor again. For all experiments the feature-based RANSAC algorithm is used to estimate the transformation. For both datasets, the implemented FCGF feature descriptor model achieved slightly better results.

Table 1. Registration Recall of the implemented FCGF compared to FPFH for the FAUST validation and test data set.

Feature Descriptor	Validation dataset	Test dataset
FPFH	0.947	0.986
Ours	**0.992**	**0.993**

4. Registration results

To test the presented registration pipeline as practically as possible, it is used in a real registration scenario. For this purpose, some point clouds of the test data set are segmented piece by piece (see figure 7). The segmentation is intended to represent the mapping of a point cloud fragment as well as possible through the recording by a camera. Different segmentations are tested. For each pair of registrations, it is calculated how high the overlapping part of the point clouds is. Basically, the smaller

the overlap between the point clouds to be registered, the higher the difficulty to align the point clouds. The individual point clouds are registered in pairs. The individual point clouds can be seen in Figure 7. The pairs to be registered are: P1(yellow) - P2 (blue), P2 - P3(green) and P3 - P4(red). (P1, P2) have an overlap of 73%, (P2, P3) an overlapping area of 27% and (P3, P4) an overlap of 49%. The results of the registration with varying overlap can be seen in the figure 8a. It is presented that the point cloud pairs could be registered very reliably and extremely precisely with each other, although the overlapping area between P2 and P3 is only 27%. For comparison, features in the same point cloud fragments are calculated using the FPFH feature detector. The result of this registration can be seen in figure 8b.

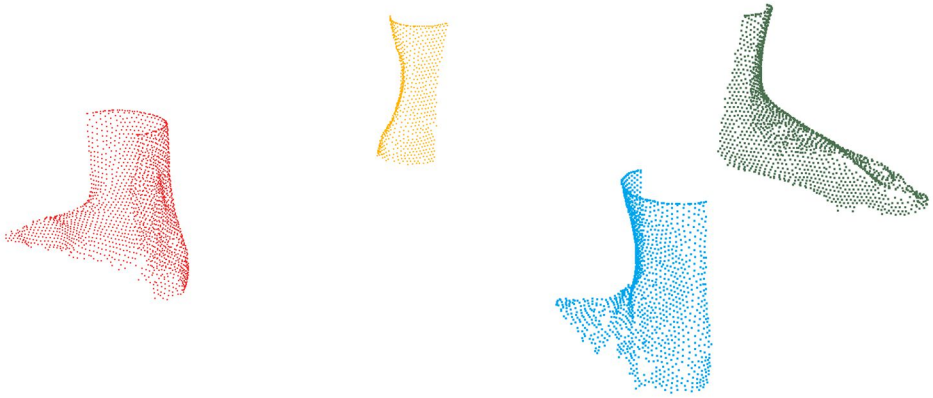
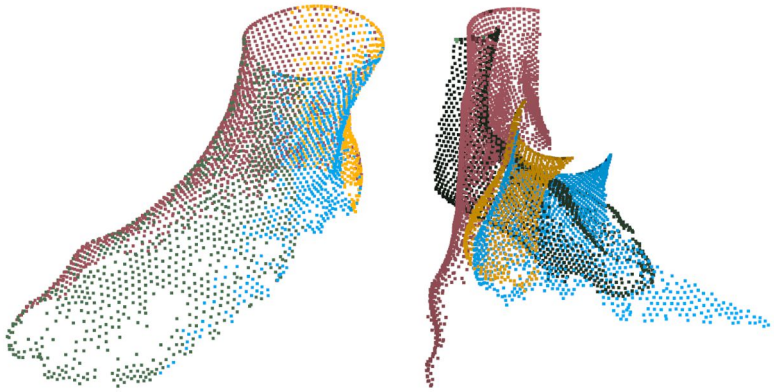


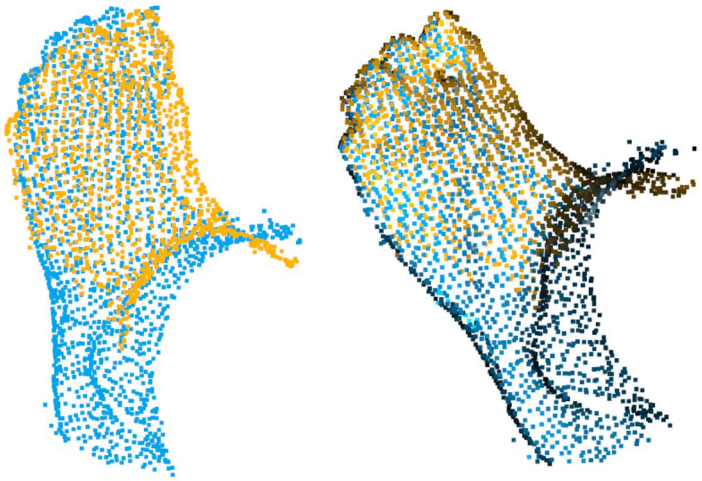
Fig. 7. Partial fragments of a 3D foot model from the FAUST test dataset with their spatial orientation.



(a) *(b)*

Fig. 8. Result of the reconstruction of the partial fragments of a 3D foot model shown in figure 8 using the implemented FCGF Feature descriptor model and the feature-based RANSAC (a). Result of the reconstruction with the FPFH Feature descriptor (b).

Finally, a test is performed in which the feature descriptor is applied to real data from the True Depth sensor that is very different from the synthetically generated training data in terms of density, noise, and point distribution. Here, a comparison of the registration with the FPFH feature descriptor is also performed. As can be seen from the figure 9b, registration based on FPFH features works better in this test scenario than the registration result with the implemented feature model which is shown in figure 9a.



(a) *(b)*

Fig. 9. Result of the reconstruction of real data of the True Depth sensor. On the left the transformation using FCGF, on the right using FPFH.

5. Conclusions and future work

Within this work a method is presented which allows the 3D reconstruction of 3D foot models with a few single point cloud fragments. For this purpose, a learning-based 3D Feature descriptor was developed. The presented results show that 3D point cloud registration of 3D foot models can be successfully performed using the Fully Convolutional Geometric Features. The trained feature descriptor model achieved consistently satisfactory results within the test scenario in feature match recall and registration recall. Thus, the implemented model still achieved satisfactory results even at an extremely low Inlier Distance threshold. Using t-SNE visualization, it can be seen that features are reliably detected in different point clouds at the same location of the foot, allowing reliable transformation estimation to be performed.

The registration results confirm the impression given by the evaluation metrics. The results show that individual point cloud fragments of a foot can be successfully registered using the trained FCGF model despite their less distinctive geometric structures, even if the overlap between adjacent point clouds is small.

The tests on the real data with the True Depth sensor system show the problems of the presented model. Thus, the results for the registration of the point clouds taken with the True Depth sensor system. One probable reason for this is the very different structure of the synthetic data used for training the learning-based features compared to the point clouds acquired with the True Depth sensor. Especially with regard to the density of the point clouds, which have a much higher resolution for the True Depth Sensor (50 000 – 100 000 points) than the point clouds of the FAUST data set (about 2000 points). Another problem is the exceptionally smooth surface of the synthetic data compared to the data of the True Depth sensor, which contain noise.

To improve the performance for real data, it will be necessary to create a new data set in the future. This data set should ideally be acquired with the True Depth Sensor or a similar Structured Light Sensor. The learning-based feature descriptor should be trained again with the new dataset to better understand geometric structures that are more similar to the True Depth Sensor data and to calculate feature points from them more precisely. This could make it possible to detect features in noisy and denser point clouds more reliably than is possible in the current model, thus ensuring more reliable registration.

6. Acknowledgements

I would like to thank the HTWG Konstanz (University of applied science Konstanz) and especially the Institute of Optical Systems (IOS) around the team of Georg Umlauf, as well as Jakob Raible for their support and supervision of this work. Also, I want to thank the whole ShoeFitter team for their support that I received during the work.

References

[1] C. Choy, J. Park and V. Koltun, "Fully Convolutional Geometric Features," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8957-8965, doi: 10.1109/ICCV.2019.00905.

[2] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.

[3] “Global registration — Open3D 0.15.1 documentation.” http://www.open3d.org/docs/release/tutorial/pipelines/global_registration.html#RANSAC (accessed Aug. 31, 2022).

[4] F. Bogo, J. Romero, M. Loper, and M. J. Black, “FAUST: Dataset and evaluation for 3D mesh registration.” Accessed: May 10, 2021. [Online]. Available: http://faust.is.tue.mpg.de.

[5] R. B. Rusu, N. Blodow, and M. Beetz, “Fast Point Feature Histograms (FPFH) for 3D registration,” in 2009 IEEE International Conference on Robotics and Automation, Aug. 2009, pp. 3212–3217, doi: 10.1109/robot.2009.5152473.