Point-Set Alignment Using Weak Labels

Project Report - Machine Learning for 3D Geometry

Tony Wang[1], Johannes Volk[1], Yushan Zheng[1], Yutong Hou[1]
[1]Technical University of Munich
{firstname.lastname}@tum.de

Abstract

With the advent of recent technologies, multi-view RGB-D recordings have become the prevalent way of data acquisition in the operating room (OR). The significant domain gap between standard data and OR data requires methods that are capable of effectively generalizing to this unique and challenging data domain. Therefore, previous works have established methods to leverage 3D information to detect faces in an OR multi-view RGB-D setting. These methods rely on point set registrations; however, real world 3D point clouds are often noisy and incomplete, which may yield erroneous alignments using existing point set registration methods. In this project, we aim to to address this issue by adapting a deep learning-based point-set registration method to achieve more robust rigid transformations on real-world data. We perform quantitative as well as qualitative evaluations of our proposed method and also give an outlook for future improvements. Our results show that DCP performs well on synthetic data, however, lacks the adaptability to perform well on real world data. Our code is available on GitHub[1].

1. Introduction

Recent advancements in technology have led to the widespread use of multi-view RGB-D data acquisition in operating rooms (ORs). These acquisitions could potentially enhance the workflow within healthcare institutions by facilitating informed decision making and improve the quality of care. However, the complex and dynamic environment of ORs, constituting of unusual camera angles and ubiquitous obstructions, presents significant challenges for conventional methods like face detectors [8]. Several methods have been introduced to tackle the problem of detecting faces in ORs [2, 5, 8]. The work of Bastian and Wang et al. [2] has introduced a method to leverage 3D data to detected faces in multi-view settings such as the OR. In their pipeline they perform a rigid registration between the head of their estimated human mesh and a cropped point cloud using FilterReg [6] and Iterative Closest Point (ICP) [3]. Figure 1 depicts the task of aligning the estimated head of the human mesh with the 3D data point cloud. However, existing methods tend to fail if the head of the human mesh has been estimated inadequately, or if the point cloud is too sparse. Our goal is to develop a more robust rigid point-set registration for this problem, which can leverage additional information, such as the semantic meaning of the target point cloud, by using a deep-learning based approach.

Our contributions can be summarized as follows:

- We train the deep-learning-based point set registration algorithm Deep Closest Point (DCP) [14] on a newly compiled synthetic point cloud dataset that mirrors real OR data
- We perform comparisons with different DCP architectures on our synthetic dataset
- We show qualitative as well as quantitative results of our proposed method and the baselines on real world data captured in the OR.

2. Related Work

Being a key technique for many computer vision applications, point-set registration aims at recovering the transformation between two sets of points in space such that they align with each other [10]. Among various methods proposed to solve this issue, the following discussion will focus on the ones related to our approach.

Classical Method. One of the most well-known classical registration methods is the ICP [3] algorithm. It starts with an initial transformation matrix and alternates between finding the closest matching points in two point clouds and minimizing the least-squares distance between the point corre-

[1]https://github.com/wngTn/synthetic_dcp

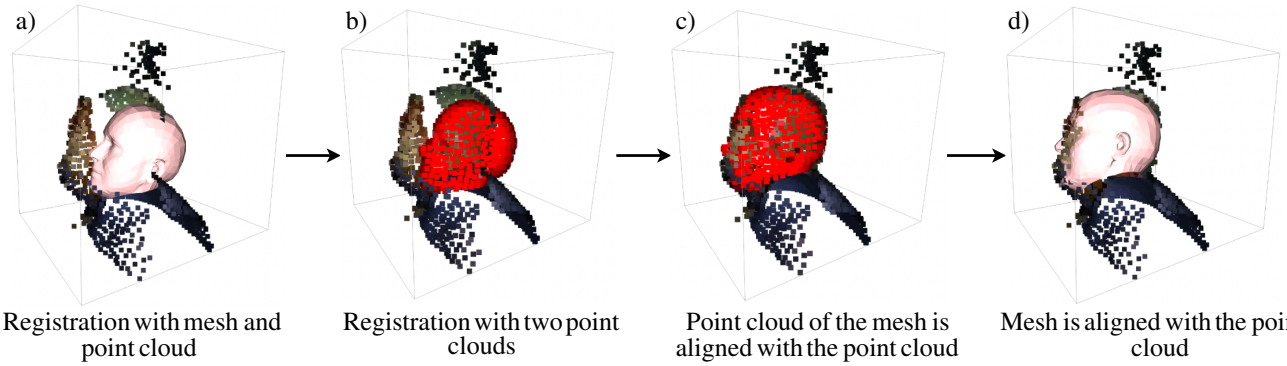

Figure 1. **Illustration of Aligning the Head of the Human Mesh With the Point Cloud.** We first extract the head of the estimated human mesh and crop the point cloud around it using a bounding box (a). We then convert the mesh into a point cloud (b) to perform a rigid point set registration between the resulting point clouds (c). The calculated transformation is then applied onto the mesh (d).

spondences to update the transformation. The alignment is iteratively refined until convergence. As ICP tends to converge to local minima and fails to perform well on datasets with a lot of outliers, overlappings, or complex structures, many variations have been proposed to address these challenges [4, 12]. However, it is challenging to achieve both optimal performance and time consumption, thus it is usually necessary to strike a balance between them [14].

Probabilistic Method. Probabilistic registration approaches like Coherent Point Drift (CPD) [10] formulate the problem of point-set alignments as probability density estimation problems and interpret one point-set as Gaussian Mixture Model (GMM) centroids and the other as observation data points, which are independently sampled from the distribution. Maximizing the likelihood of the observed data points leads to the alignment and thus the transformation between the two point sets. Probabilistic methods are more robust to noise and outliers than ICP-based algorithms, but they tend to exhibit a lack of performance efficiency. FilterReg [6] overcomes this problem by formulating the E-step in the Expectation–Maximization (EM) algorithm as a filtering problem and provides an accurate, robust, yet efficient method. However, using FilterReg in our task setting can be unreliable, due to inadequate mesh estimations or sparse and incomplete point clouds in the dataset.

Deep-learning-based Method. As a result of various successful implementations in a wide range of tasks, deep learning techniques have also been leveraged to solve the point cloud registration problem. There are two main categories of learning-based registration methods: correspondence-free methods, which extract global features for each point cloud, and correspondence-based methods, which extract features per point for the point clouds [18].

Many correspondence-free approaches [1, 13] employ PointNet [11] or its variants to extract global features and regress transformation matrices directly from the differences between them, which inevitably results in a high dependence of the performance on the feature descriptor network [18]. In other words, the generalization ability of the feature extraction step might limit the performance of these approaches and prohibit them from scaling to complex scenes.

DCP [14] is a typical correspondence-based approach, which uses DGCNN [16] to extract features for each point, followed by a transformer module to learn co-contextual information and improve the features. The mapping from one point set to the other is obtained by applying the Softmax function on the features and is then used to generate an alignment in an SVD module. It has been proven that DCP can outperform correspondence-free approaches with respect to estimation precision [18].

3. Methods

3.1. Dataset

The original implementation of DCP [14] was trained and adjusted to the ModelNet40 [17] dataset, which deviates heavily from our medical setting. Thus, we train the model on a synthetic dataset, which mirrors the data in the OR. Finally, we evaluate our approach on the synthetic as well as real-world OR data.

Test Data. We use real-world data for evaluation, which was acquired from the chair of Computer Aided Medical Procedures (CAMP) at TUM by capturing a simulated OR environment from six calibrated ceiling-mounted Microsoft Azure Kinect cameras. This dataset consists of 200 frames, each frame encapsulating six RGB images (2048x1536) and a fused point cloud of the OR. In Figure 2 we show an example of such a fused point cloud.

Synthetic Data. Our synthetic dataset is generated by utilizing a point cloud representation of a human body as the

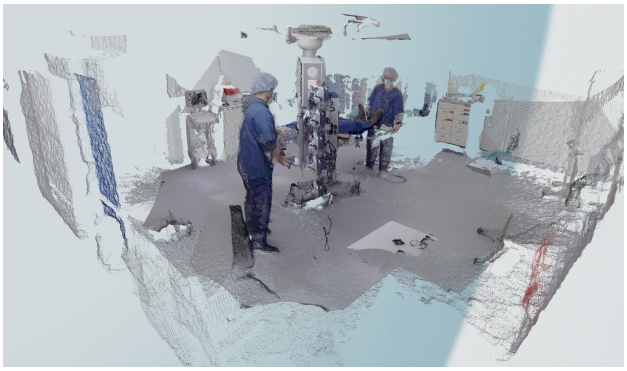

Figure 2. **Real world point cloud from a simulated operation used for evaluation.**

target and incorporating a misaligned head. The human shaped target point cloud is then cropped according to a bounding box around the misaligned head (see suppl. for an illustration on real data). As ground truth we obtain the rotation matrix and translation vector to align the misaligned head with the human shaped target point cloud. An example of the synthetic data can be seen in Figure 3.

Pose sampling. To create human shaped target point clouds we adopt the parametric mesh model SMPL [9], which provides a realistic representations of human bodies with a wide variety of shapes poses. In real world settings, the target human point cloud has different poses, as individuals look in different directions and carry out varying tasks. We, therefore, generate realistic pose parameters for SMPL by fitting the pose parameters to existing 3D human key-points from the HumanAct12 dataset [7]. As result we acquire 9000 different SMPL pose parameters that resemble realistic poses.

Data Augmentation. The synthetic dataset should reflect the real world OR data as closely as possible, however, the SMPL [9] model depicts the human head in a bare state without any accessories. In OR scenes people usually wear masks and surgical caps, which may lead to a mismatch and degrade the performance. Thus, in order to imitate the human bodies in the OR, we augment the original SMPL model by optionally adding a mask, a surgical cap and glasses to the head part. Adding these accessories is done by extracting basis vectors with respect to the head and wearable objects. After changing the basis and applying translations, the meshes align as desired and are robust to head rotation and position in the scene. Since the meshes, especially those of the surgical cap, cover parts of the underlying SMPL mesh, the hidden inner points are removed by indices. Additionally, we replicate the inherent imperfections of real life data, by randomly sampling points on the augmented mesh and adding gaussian noise to them. The resulting output are two point clouds that can be used for training: one misaligned head and one augmented point cloud that resembles an individual of our test set.

3.2. Training and Implementation Details

For each item in our synthetic dataset we sample a random pose and apply also a random rotation and transformation on the head segment to obtain our misaligned head. This method prevents overfitting to a specific transformation or pose. We use the same model architecture, loss and hyperparameters as proposed by Wang et al. [14]. We set the length of our synthetic train dataset to contain 1000 randomly sampled poses per epoch.

4. Evaluation and Discussion

4.1. Evaluation

Quantitative evaluation of our approach on real-world data is challenging, due to missing ground truth and costly 3D annotations. To address this, we therefore adopt the method introduced by Bastian and Wang et al. [2]. We annotated face bounding boxes on 2D images and render the face segment of the head mesh back into the 2D images to generate face bounding boxes (see suppl. for more details). We report the recall and precision scores at an intersection over union (IoU) of 0.4 to quantify registration quality and compare the performance of our fine-tuned model with FilterReg and the estimated head meshes without point registration.

Furthermore, we also compare our proposed method and our baselines on our synthetic data. As we have the ground truth transformations available, we report the Root-Mean-Squared Error (RMSE) and the Mean Absolute Error (MAE) for both rotation and translation (Table 2).

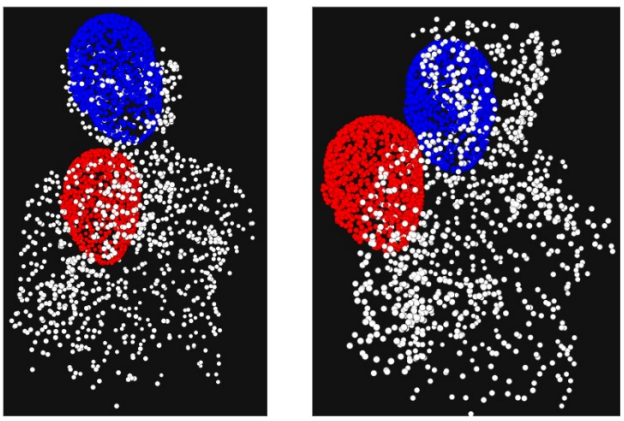

Figure 3. Result of overfitting synth. DCP-v2 on a single pose using a bounding box of $[0.25, 0.25, 0.25]$. The red point cloud denotes the misaligned head, blue the predicted transformation.

Camera Id	No Registration		FilterReg [6]		Synth. DCP-v2		Num. GT Detections
	Recall	Precision	Recall	Precision	Recall	Precision	
cn01	21.81	35.33	31.28	51.01	29.63	30.64	243
cn02	13.30	17.01	22.34	28.00	18.62	15.09	188
cn03	17.44	24.59	25.58	38.60	17.44	12.61	86
cn04	27.71	36.51	43.37	56.25	13.25	10.58	166
cn05	22.78	27.33	22.78	27.33	24.44	18.80	180
cn06	10.28	26.19	33.64	85.71	28.04	32.97	107

Table 1. Quantitative comparison between vanilla baseline model without point registration, FilterReg and our on the synthetic dataset fine-tuned DCP model. We report the recall and precision rates for all six camera angles at an IoU of 0.4. Synth. DCP-v2 denotes the DCP architecture with transformer trained on synthetic dataset.

registration method	Rotation		Translation	
	RMSE	MAE	RMSE	MAE
no registration	105.83	74.70	0.0651	0.0614
FilterReg [6]	23.59	16.79	**0.0305**	**0.0306**
DCP-v1 [14]	29.95	39.81	0.1336	0.1692
DCP-v2 [14]	23.02	28.47	0.0332	0.0427
synth. DCP-v1	25.20	20.70	0.0716	0.0570
synth. DCP-v2	**20.36**	**16.32**	0.0428	0.0317

Table 2. Quantitative error comparison on synthetic data between no registration, FilterReg and DCP (both originally proposed versions with/-out transformer; out-of-the-box pretrained on ModelNet40 [17] vs. trained on our synthetic dataset). The synthetic data is generated with a bounding box of $[0.1, 0.1, 0.1]$.

4.2. Discussion

In Table 1 we show the recall and precision of each method for each camera. Our proposed method under-performs in certain cameras than the baseline without registration. Furthermore, it does not attain the performance of the probabilistic method FilterReg [6].

This can be attributed to several factors. Firstly, the partial point cloud registration problem in real-world 3D data, characterized by noisy, sparse, and incomplete point clouds, might present a more challenging scenario than the synthetically generated scenario. Moreover, the limited correspondences and partial overlap makes it difficult for deep learning-based methods to learn the alignment between point sets. As the one-shot construction of DCP [14] limits it from refining the estimated alignment, the authors of DCP propose PRNet [15] to address the partial-to-partial registration problem. PRNet could be applied iteratively to enable coarse-to-fine refinement of the registration process and thus might better fit our task setting.

In Table 2 we denote the RMSE and MAE of both rotation and translation of our proposed method and our baselines. Here, we further evaluate with different architectures: without transformer (DCP-v1) and with transformer (DCP-v2). These results show that our fine-tuned model yields comparable with FilterReg [6] on synthetic data. The subpar performance on the real data can be thus partly attributed to the still large domain gap between our synthetic dataset and the real data.

In Figure 3 we show qualitative results of an experiment to verify whether DCP is able to understand the semantic meaning, i.e., the anatomy of a human shaped point cloud. We test this by trying to overfit DCP onto a specific pose with only random small transformations. To incorporate more of the human shaped point cloud we increase the bounding box size to $[0.25, 0.25, 0.25]$. However, since the head is not always successfully aligned at the head position, we conclude that DCP does not learn the underlying anatomy of the human shaped point cloud. Further qualitative examples are visualized in the suppl. materials.

5. Conclusion

This project aims to align point clouds of heads to a human shaped point cloud of real life data. This process can be used to detect human faces in multi-view RGB-D settings like OR datasets as established by previous methods. To accomplish this, we train DCP [14] on synthetic data that resembles real-world data by simulating human poses and incorporating medical accessories. We further evaluate its performance with other baselines on both the synthetic and real data. The quality of the alignment is quantified by rendering the aligned 3D head meshes back into 2D images. Our results indicate that the one-shot DCP method may not be ideal for our task, given the partial-to-partial registration nature of the problem and the dissimilarity between the real and synthetic data. Hence, incorporating other methods, such as PRNet [15], or using alternative architectures that better recognize the semantic meaning of point clouds would be needed to improve the alignment results in future work.

References

[1] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7163–7172, 2019.

[2] Lennart Bastian, Tony Danjun Wang, Tobias Czempiel, Benjamin Busam, and Nassir Navab. Disguisor: Holistic face anonymization for the operating room. Manuscript submitted for publication, 2022.

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[4] Ben Eckart, Kihwan Kim, and Jan Kautz. Fast and accurate point cloud registration using trees of gaussian mixtures. *arXiv preprint arXiv:1807.02587*, 2018.

[5] Evangello Flouty, Odysseas Zisimopoulos, and Danail Stoyanov. Faceoff: Anonymizing videos in the operating rooms. *CoRR*, abs/1808.04440, 2018.

[6] Wei Gao and Russ Tedrake. Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11095–11104, 2019.

[7] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[8] Thibaut Issenhuth, Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy. Face detection in the operating room: Comparison of state-of-the-art methods and a self-supervised approach. *CoRR*, abs/1811.12296, 2018.

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[10] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.

[11] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[12] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001.

[13] Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. Pcrnet: Point cloud registration network using pointnet encoding. *arXiv preprint arXiv:1908.07906*, 2019.

[14] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3523–3532, 2019.

[15] Yue Wang and Justin M Solomon. Prnet: Self-supervised learning for partial-to-partial registration. *Advances in neural information processing systems*, 32, 2019.

[16] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[17] Zhirong Wu, Shuran Song, Aditya Khosla, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets for 2.5d object recognition and next-best-view prediction. *CoRR*, abs/1406.5670, 2014.

[18] Zhiyuan Zhang, Yuchao Dai, and Jiadai Sun. Deep learning based point cloud registration: an overview. *Virtual Reality & Intelligent Hardware*, 2(3):222–246, 2020.