Iterative Matching Point

Jiahao Li
Washington University in St. Louis
1 Brookings Dr, St. Louis, MO 63130
jiahao.li@wustl.edu

Changhao Zhang
Xi’an Jiaotong University
No.28, Xianning West Road, Xi’an, P.R. China
cvchanghao@gmail.com

Abstract

In this paper, we propose a neural network-based point cloud registration method named Iterative Matching Point (IMP). Our model iteratively matches features of points from two point clouds and solve the rigid body motion by minimizing the distance between the matching points. The idea is similar to Iterative Closest Point (ICP), but our model determines correspondences by comparing geometric features instead of just finding the closest point. Thus it does not suffer from the local minima problem and can handle point clouds with large rotation angles. Furthermore, the robustness of the feature extraction network allows IMP to register partial and noisy point clouds. Experiments on the ModelNet40 dataset show that our method outperforms existing point cloud registration method by a large margin, especially when the initial rotation angle is large. Also, its capability generalizes to real world 2.5D data without training on them.

1. Introduction

Point cloud registration is an important task in many 3D computer vision applications, such as robotics and self-driving cars. It can also be used as a post-processing step in other tasks such as 6D pose estimation. This problem is challenging in the following ways. First, unlike RGB data, point clouds are not well-organized data, which can not be easily processed using convolutions. Secondly, real world point clouds are usually very noisy. Sometimes the noise is so overwhelming that some of the local geometric structures of the point cloud is lost. Thirdly, in many cases we need to register point clouds which are only partially overlapping with each other.

The basic paradigm for solving this problem is to first find the corresponding point pairs from the two point clouds, and then move one of them such that these pairs of points are close to each other. The most well-known traditional method is Iterative Closest Point (ICP) [2]. For a given point in one point cloud, ICP assigns the closest point in the other to be its corresponding points. This method fails miserably when the rotation angle between the two point clouds is large. Recently, some works utilize deep learning techniques to extract features from the point clouds, replacing the closest point matching with feature matching.

In this paper, we propose a novel neural network architecture named Iterative Matching Point (IMP) for point cloud registration. We identify that the two point clouds usually come from different data distributions (e.g. 3D models and 2.5D data), so we propose to use two different feature extraction subnetworks for $S$ and $T$. These feature extraction subnetworks are inherently translation invariant, so the features will not be affected by large translation. The geometric features are used for finding corresponding pairs be-
between the two point clouds. However, it is not guaranteed that all the pairs are of the same quality, especially when the features are affected by noise. Therefore, we propose to solve a weighted least squares problem, where the weights are the inverse of the entropy of the probability distribution for each point.

Although the feature extraction modules are translation-invariant, they are not rotation-invariant. In fact, the larger the possible rotation angles, the more difficult it is to extract high-quality features. As a result, we propose to iteratively extract new features from the point cloud in the current pose. The model first rotates the source point cloud to be closer to the target point cloud, and then iteratively refines its pose. Experiments show that iterative refinement is crucial, especially when the point clouds are noisy.

The main contributions of this paper are summarized as follows

• We apply a loss on the network to explicitly supervise it to learn to find correct matching points.
• We assign different weights to different corresponding point pairs according to the entropies of the probability distribution obtained from matching geometric features.
• We propose to use an iterative refinement process to achieve better performance than “one-shot” registration.

3. Model

In this section, we introduce the network architecture of our proposed model (Figure 2, 3).

3.1. Notations

Suppose we have two point clouds $S$ (source) and $T$ (target), with $N_S$ and $N_T$ points respectively. $S$ represents the same shape as $T$, but is potentially noisy and incomplete. We want to find a 3D rigid body motion to transform $S$ to $S'$, such that $S'$ overlaps with $T$. The rigid body motion is represented as a rotation matrix $R \in SO(3)$ and a translation vector $t \in \mathbb{R}^3$.

3.2. Feature Extraction

We choose a variant of Dynamic Graph CNN (DGCNN) as our feature extraction module. The features are extracted independently from the source and target point cloud respectively.

Suppose we have a point cloud $P$ with $N_P$ points. For each point $p$ in $P$, the network first find the $K$ nearest neighbors of $p$ in $P$. Denote the set of $K$ nearest neighbors for $p$ as $N_p$.

The input to the first layer of the module is the coordinates of the points. In the $n$th layer, the features are computed as follows

$$u_n(p) = \frac{1}{K} \sum_{q \in N_p} g_n(u_{n-1}(p) - u_{n-1}(q))$$

where $u_n(p)$ is the feature vector output by the $n$th layer, $g_n$ is a multi-layer perceptron.

Note that the input to $g$ is not the feature vector itself, but the difference of the feature vector of a point to those of its neighbors. For the first layer, the features are the 3D coordinates, and only the vectors from the current point to its neighbors are input to $g$. This is reasonable since the absolute values of the coordinates provide little information. What matters is the relative positions of neighboring points. As a result, the feature extraction module is translation invariant, and it will work fine even when the translation is very large.

On the other hand, our network is not “dynamic” in the sense that the neighborhood $N_p$ is not recomputed in the feature space. A point always incorporates information from its neighbors in the Euclidean space. As we stack more
layers, the receptive field for each point would be larger. As a result, not only is it able to learn local geometric features, it also has the capability to capture global information of the whole shape.

### 3.3. Matching Points

After feature extraction, we get a feature vector \( u(p) \) for each point \( p \) in the two point clouds \( S \) and \( T \). These feature vectors are fed into a registration module (Figure 3) to find correspondences between the two point clouds. A corresponding point in \( T \) to a point \( p \) in \( S \) is defined as a point \( q \) that is closest to \( p \) after \( S \) is transformed with the ground truth rigid body motion. If we can find the correct correspondence relationship, the rigid body motion can be easily obtained by Singular Value Decomposition (SVD).

Denote the \( i \)th point in \( S \) as \( p^{(i)} \) and the \( j \)th point in \( T \) as \( q^{(j)} \). We then form a similarity matrix using dot products of the features

\[
M(i, j) = u(p^{(i)}) \cdot u(q^{(j)})
\] (2)

\( M(i, j) \) is the unnormalized cosine similarity between the \( i \)th point in \( S \) and \( j \)th point in \( T \). We apply a softmax function on each row of \( M \) to obtain a probability distribution over all the points in \( T \) for each point in \( S \)

\[
P(i, j) = \text{Softmax}(M(i, :))_{[j]}
\] (3)

\( P(i, j) \) represents the probability of \( q^{(j)} \) being the corresponding point of \( p^{(i)} \). For each point \( p^{(i)} \) in \( S \), we choose the point \( p^{(j)} \) in \( T \) that has the highest probability (largest similarity) to be the corresponding point of \( p^{(i)} \) in \( T \).

Thus we have a correspondence set \( \{(p_k, q_k)\} \), and we want to solve \( R \) and \( t \) by minimizing the mean squared distance

\[
R^*, t^* = \arg \min_{R,t} \frac{1}{N_S} \sum_{k=1}^{N_S} \| R p_k + t - q_k \|^2
\] (4)

where \( N_S \) is the number of points in the source point cloud.

However, not all the correspondences are of the same quality. Some of the points, those on a flat surface for example, may have features similar to many different points. We want to give these points less weights, and pay more attention to those distinctive points that have fewer possible correspondences (such as corners). Therefore, we use the inverse of the entropy of the probability distribution as the weight

\[
w_{i} = -\frac{1}{\sum_j P(i, j) \log(P(i, j)) + \epsilon}
\] (5)

where \( \epsilon = 1 \times 10^{-8} \) is a small constant for numerical stability. Then (4) becomes

\[
R^*, t^* = \arg \min_{R,t} \frac{1}{N_S} \sum_{k=1}^{N_S} w_k \| R p_k + t - q_k \|^2
\] (6)

This optimization problem can be solved in closed form using Singular Value Decomposition (SVD). First we form the covariance matrix \( S \)

\[
S = \sum_{k=1}^{N_S} (p_k - \bar{p}) (q_k - \bar{q})^T
\] (7)

where \( \bar{p} \) and \( \bar{q} \) are the means of \( p_k \) and \( q_k \) respectively. \( W \) is a diagonal matrix with \( W(i, i) = w_i \). Apply SVD on \( S \)

\[
S = U \Sigma V^T
\] (8)
\( R^* \) and \( t^* \) can be obtained by

\[
R^* = VU^T, \quad t^* = -R^*p + q
\]

(9)

### 3.4. Iterative Refinement

Since the input to the feature extraction network is not rotation invariant, the precision of matching corresponding points is dependent on the relative poses of the input point clouds. As the initial rotation angle increases, the error of \( R^* \) also increases. This motivates the network design of iterative refinement.

Denote the rotation and translation output by iteration \( m \) as \( R^*_m \) and \( t^*_m \). In iteration \((m+1)\), we transform the source point cloud using \( R^*_m \) and \( t^*_m \) to get a new point cloud \( S_m \). Typically, \( S_m \) has a smaller angle gap to the target \( T \). \( S_m \) is then fed to a new feature extraction module with the same structure but different weights to extract new features in this new pose. Again, the similarity matrix \( M \) is formed, correspondences are found and new \( R^*_m+1 \) and \( t^*_m+1 \) are solved by SVD. Since \( S_m \) is already close to \( T \), the correspondences found by matching features extracted under this new pose would be more accurate. Figure 1 shows an examples of iterative refinement.

The final prediction is formed by compositing all the intermediate \( R^*_m \) and \( t^*_m \).

### 3.5. Loss Function

Although the final output of the network is \( R^* \) and \( t^* \), we do not directly impose a loss on them. The most critical part of the network is actually the probability matrix \( P \), which determines the quality of correspondences. As a result, our loss is defined on the probability matrix to explicitly force the network to learn that features of corresponding points should have larger inner products.

Denote the probability matrix formed in the \( m \)th iteration as \( P_m \). We use the the negative log likelihood of the values of the correct correspondences as the loss

\[
L(S, T) = -\frac{1}{N_S} \sum_m \sum_{i=1}^{N_S} \log(P_m(i, j^{(i)})) \tag{10}
\]

where \( j^{(i)} \) is the ground truth index of the point in \( T \) that corresponds to the \( i \)th point in \( S \). During training, this correspondence can be easily obtained by setting \( j^{(i)} \) to the index of the nearest point in \( T \) to the \( i \)th point in \( S \) when \( S \) is transformed using the ground truth rotation and translation.

### 4. Experiments

In this section, we demonstrate the effectiveness of our proposed method by showing results of extensive experiments under different problem settings.
4.3. Metrics

We follow [14] and use four different metrics to compare the above methods. \( \text{RMSE}(R) \) is the square root of the mean squared error of the initial angles of \( R \). \( \text{MAE}(R) \) is the mean absolute error of the initial angles of \( R \). \( \text{RMSE}(t) \) is the square root of the mean squared error of \( t \). \( \text{MAE}(t) \) is the mean absolute error of \( t \).

4.4. Results

The results of registering two identical point clouds are shown in Table 1. We can see that our model outperforms all the other methods under all the metrics by a large margin. ICP and PointNetLK intrinsically make the assumption that the two point clouds are already close enough to each other, so they do not perform well when the initial angles are large (60°). FGR and DCP are global registration methods and they have fairly good results on this task. Our method performs better than all the other methods on this task. Most of the metrics are nearly zeros, meaning the two registered point clouds overlap exactly with each other.

To demonstrate the robustness of our method, we add gaussian noise with standard deviation 0.02 to the source point cloud during training and testing. This way, the two point clouds cannot overlap exactly any more. The results are shown in Table 2. It can be seen that ICP and DCP are quite robust to noise, but the performance of FGR and PointNetLK worsens significantly. Our model performs the best, and the average precision is still under 1°.

The upper two graphs of Figure 7 show the \( \text{MAE}(R) \) of different models with different maximal initial angles. We can see that when there is no noise, our method has almost no degradation in performance when the initial angle increases. When noise is added, our model is still the most robust. The performance starts to get worse only when the angle is very large (90°). Note that our model is trained on data with a maximal initial angle 60°, but it turns out to have the ability to generalize to the case of even larger angles.

4.5. Partial Registration

We also conducted experiments in the settings where the source point cloud is part of the target. The data is generated
Figure 6: Examples of registering real world 2.5D point clouds, where the model is trained on 20 categories of ModelNet40. The top row is the input and the bottom row is the output. Note that the point clouds are resampled so that they do not overlap exactly.

Figure 7: Comparison of the performance of different models with different maximal initial angles. **Left**: Registration of two identical point clouds. **Middle Left**: Registration of two point clouds with one of them noisy. **Middle Right**: Registration of noiseless partial point clouds. **Right**: Registration of noisy partial point clouds.

|                | RMSE(R) | MAE(R) | RMSE(t) | MAE(t) |
|----------------|---------|--------|---------|--------|
| ICP[2]         | 14.89   | 12.43  | 0.033   | 0.029  |
| FGR[19]        | 2.77    | 2.24   | 0.0048  | 0.0042 |
| PNLK[1]        | 18.19   | 15.62  | 0.026   | 0.023  |
| DCP[14]        | 6.75    | 5.83   | 0.014   | 0.012  |
| Ours           | **0.00049** | **0.00043** | **0.00** | **0.00** |

Table 1: Results of registering two identical point clouds within 60° on the ModelNet40 dataset.

|                | RMSE(R) | MAE(R) | RMSE(t) | MAE(t) |
|----------------|---------|--------|---------|--------|
| ICP[2]         | 14.95   | 12.51  | 0.033   | 0.028  |
| FGR[19]        | 15.75   | 13.74  | 0.033   | 0.029  |
| PNLK[1]        | 26.91   | 22.08  | 0.058   | 0.050  |
| DCP[14]        | 9.26    | 8.08   | 0.037   | 0.032  |
| Ours           | **0.48** | **0.42** | **0.0053** | **0.0046** |

Table 2: Results of registering noiseless point clouds with noisy point clouds within 60° on the ModelNet40 dataset.

by truncating the source point cloud in half with a random plane. This partial data is fed to the network during both training and testing. The maximal initial angle is 60°. Table 3 shows the results for noiseless partial point cloud registration. Table 4 shows the results for adding gaussian noise to the partial source point cloud. The lower two graphs of Figure 7 show the results with different maximal initial angles. ICP and PointNetLK perform significantly worse, maybe because when only part of the source point cloud is visible, it is easier to get stuck in a local minimum. Both FGR and DCP are good at registering noiseless point clouds, but DCP is robust to noise and FGR is not. The results of our method are still the best among all these models, and is only slightly worse than registering two whole point clouds.

4.6. Real World Data

Although the results shown above are obtained by evaluating on unseen categories of ModelNet40, it is still not
Table 3: Results of registering complete point clouds with partial point clouds within 60° on the Modelnet40 dataset.

|       | RMSE(R) | MAE(R) | RMSE(t) | MAE(t) |
|-------|---------|--------|---------|--------|
| ICP[2] | 25.23   | 21.21  | 0.11    | 0.099  |
| FGR[19]| 5.73    | 5.06   | 0.028   | 0.023  |
| PNLK[1]| 27.06   | 23.38  | 0.11    | 0.093  |
| DCP[14]| 10.02   | 8.68   | 0.062   | 0.054  |
| Ours   | 0.11    | 0.092  | 0.00072 | 0.00062|

Table 4: Results of registering noiseless complete point clouds with noisy partial point clouds within 60° on the Modelnet40 dataset.

|       | RMSE(R) | MAE(R) | RMSE(t) | MAE(t) |
|-------|---------|--------|---------|--------|
| ICP[2] | 25.21   | 21.28  | 0.11    | 0.098  |
| FGR[19]| 32.04   | 28.59  | 0.15    | 0.12   |
| PNLK[1]| 27.24   | 23.60  | 0.099   | 0.086  |
| DCP[14]| 10.51   | 9.10   | 0.064   | 0.056  |
| Ours   | 1.12    | 0.96   | 0.012   | 0.011  |

Figure 8: Comparison of the speed of different models.

Figure 9: Comparison of different ways to determine correspondences.

4.7. Speed

Figure 8 compares the speed of the above models. Among all the models, ICP is the fastest and PointNetLK is the slowest. Our method is several times slower than ICP and almost twice slower than DCP. This is mainly due to the iterative refinement process where we have to re-extract the features and form the probability matrices.

5. Ablation Study

In this section, we show the results of ablation experiments on our model to show the effectiveness of each component.

5.1. Correspondences

As described in Section 3.3, we assign the point in $T$ with the largest probability to be the corresponding point for each point in $S$. Alternatively, we can form a “virtual” point on $T$, whose coordinates are the weighted average of the coordinates of all the points in $T$, where the weights are the probabilities.

On the other hand, when solving (6), we assign the weights to be the entropy of the probability distribution. Intuitively, it puts more weight on those points that are “certain” about their correspondences, which is better than (4). We do experiments to confirm this hypothesis.

There are four possible combinations: **mean-no-weight**, **max-no-weight**, **mean-weight** and **max-weight**, whose meanings are clear by their names. Figure 9 shows the performance of these different choices. These results are for the task of registration when the source point cloud is partial and noisy. We can see that using “max” is consistently better than “mean”. Also, in either way, using a weighted objective as in (6) is better than using uniform weights (no weights).
5.2. Iterative Refinement

In Section 3.4, we explain how the model iteratively refines the predicted $R$ and $t$. Figure 1 shows a qualitative example of this refinement process. In this section, we further demonstrate its effectiveness by showing some quantitative results. Figure 10 shows the results of our model registering two whole point clouds with and without noise. We can see that when the point cloud is noiseless, the model performs perfectly even without iterative refinement, so the two curves overlap with each other. The advantage of iterative refinement reveals itself, however, when the source point cloud is noisy. The performance margin between the two models get even more prominent when the maximal initial angle is large. This is reasonable since matching features extracted from noisy point clouds is more difficult. The iterative refinement process allows the different feature extraction networks to learn to handle different situations (different angles), thus improves the overall performance.

5.3. Loss

As mentioned in Section 5.1, for a given point in $S$, we can use the weighted mean of the points in $T$ as the corresponding point. In this way, the coordinates of these “virtual” points are differentiable functions of the probability matrix. Also, the SVD layer in Pytorch is differentiable. As a result, we can directly impose a loss on $R$ and $t$, instead of on the probability matrix. We expect the network to learn the correspondence relationship by itself. This approach is adopted by [14]. In the following, we show that imposing a loss on the probability matrix is better than that on $R$ and $t$.

We do two experiments with two ways of defining loss respectively, and denote them as $MAT$ and $RT$. The loss on $R$ and $t$ is

$$L(S, T) = \| R^T R_{gt} - I \|^2 + \| t - t_{gt} \|^2$$  \hspace{1cm} (11)$$

where $I$ is the identity matrix, $R_{gt}$ and $t_{gt}$ are the ground truth rotation and translation. For a fair comparison, we use mean-weight (Section 5.1) option for both of the experiments. Iterative refinement is disabled for simplicity. The task we evaluate on is the same as Table 1.

The top graph of Figure 11 shows the variation of the mean entropy of the rows of the probability matrix. Bottom: Performance comparison of imposing losses on the probability matrix or $R$, $t$.

6. Conclusion

In this paper, we propose Iterative Matching Point (IMP), a novel neural network based method for point cloud registration. IMP directly learns to find correspondences between the two point clouds, and iteratively refines the prediction. Experiments show that our model is able to register point clouds with large rotation angles. Its capability can also generalize to real world 2.5D data, using only half of the data in ModelNet40 for training.
References

[1] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2, 4, 6, 7

[2] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell., 14(2):239–256, 1992. 1, 2, 4, 6, 7

[3] S. Bouaziz, A. Tagliasacchi, and M. Pauly. Sparse iterative closest point. In Proceedings of the Eleventh Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, SGP ’13, pages 113–123, Aire-la-Ville, Switzerland, Switzerland, 2013. Eurographics Association. 2

[4] C. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks, 2019. 2

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 4

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015. 4

[7] H. M. Le, T.-T. Do, T. Hoang, and N.-M. Cheung. Sdrsac: Semidefinite-based randomized approach for robust point cloud registration without correspondences. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2

[8] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI’81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. 2

[9] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 4

[10] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2017. 2

[11] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413, 2017. 2

[12] A. Segal, D. Hhnel, and S. Thrun. Generalized-icp. In J. Trinkle, Y. Matsuoka, and J. A. Castellanos, editors, Robotics: Science and Systems. The MIT Press, 2009. 2

[13] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz. Splatnet: Sparse lattice networks for point cloud processing, 2018. 2

[14] Y. Wang and J. M. Solomon. Deep closest point: Learning representations for point cloud registration, 2019. 2, 4, 5, 6, 7, 8

[15] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds, 2018. 2

[16] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015. 4

[17] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (RSS), 2018. 7

[18] J. Yang, H. Li, D. Campbell, and Y. Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. 2016. 2

[19] Q. Zhou, J. Park, and V. Koltun. Fast global registration. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II, pages 766–782, 2016. 2, 4, 6, 7