2021 International Conference on Culture-oriented Science & Technology (ICCST) | 978-1-6654-4254-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCST53801.2021.00105

Relative Pose Estimation for RGB-D Human Input Scans via Human Completion

Pengpeng Liu
School of Artificial Intelligence, UCAS
Institute of Automation, CAS
Beijing, China
liupengpeng2018@ia.ac.cn

Guixuan Zhang, Hu Guan, Jie Liu, Shuwu Zhang and Zhi Zengi*
Institute of Automation
Chinese Academy of Sciences
Beijing, China
{guixuan.zhang, hu.guan, jie.liu,shuwu.zhang,zhi.zeng}@ia.ac.cn

***Abstract*—Relative pose estimation for human scans enjoys a promising prospect. However, most existing methods mainly focus on indoor or outdoor scenes, requiring considerable overlap between the inputs. We present a technique for estimating the relative pose whatever the overlap between the human RGB-D input scans is. For non-overlapping scans, the insight is to take advantage of the underlying human geometry prior as much as possible. We utilize the implicit function model for human reconstruction, enriching abundant hidden cues for unseen regions, then we use the completed human geometry to get a stable pose estimation. Our evaluation shows that our approach outperforms considerably than standard pipelines in non-overlapping setting, without compromising performance over overlapping input scans.**

***Index Terms*—relative pose estimation, implicit function, human reconstruction, non-overlapping**

I. Introduction

Relative pose estimation between two RGB-D scans is a crucial problem in 3D vision and graphic. Recently, 3D reconstruction for human [1] has attracted increasing attention in both academia and industry, many application systems such as human motion capture and human performance capture [2, 3] depend heavily on the accuracy of relative pose between the input scans. Robust and efficient solution for human scans relative pose estimation will embrace a wide scope of applications beyond doubt. Examples include 3D human reconstruction from sparse views without pose parameters [4], self-calibration for systems such as human performance capture, avoiding interrupting the system to manually calibrate with chessboard, when there are disturbances to camera setups. In this paper, we are interested in relative pose estimation for human input scans, especially with non-overlapping input scans.

However, for human input scans, most existing approaches to estimate relative pose mainly have the following limitations: 1) Most researches [5, 6] focus on indoor or outdoor scenes, actually not very suitable for human specific input scene. 2) Most existing methods [7, 8] can't handle the extreme pose whose overlap between the RGB-D input scans is small or even none. These methods highly rely on accurate correspondence, which often follow a three-step paradigm [6]: feature extraction, feature matching, and rigid transform fitting with the most consistent feature correspondences. Obviously, to guarantee good performance, this paradigm requires massive overlap between the input scans. 3) Although some learning-based methods [4, 9] can regress camera parameters directly using CNNs from input images, they often fail to produce accurate poses as discussed in [10], and meanwhile, they are limited by large amounts of data and elaborate framework design to improve performance and generalization. Recent works [5, 11] focus on extreme relative pose estimation between two input RGB-D scans, Reference [5] use scan completion as an intermediate and then match the completed scans. Reference [11] propose hybrid representations which are too sophisticated, and besides, they mainly focus on indoor scene and can't be directly applied to human scans.

Taking all the above-mentioned limitations into consideration, our work presents a effective method, handling the RGB-D human input scans whatever the overlap is. Inspired from intuition that human can estimate accurately the relative pose for input pairs, even non-overlapping, leveraging the prior knowledge of the underlying geometry. We hypothesize that the key is to take advantage of the human prior knowledge for typical structure and shapes as much as possible. In this paper, we utilize the state-of-the-art RGB-D PIFu [2] to reconstruct the detail-preserving human body, enriching the underlying geometry prior knowledge. Inspired by scan completion [5], we can complete the unseen region from the visible partial, with the help of our human reconstruction, and then matching the point clouds sampled from the full mesh using the off-the-shelf optimization methods [7, 8].

To summarize, our main contributions are: 1) our relative pose estimation method for RGB-D human input scans can handle settings at arbitrary overlap, even non-overlapping. 2) our method outperforms than state-of-the-art standard optimization methods considerably, especially in small overlapping settings.

II. Approach

Given a pair of human RGB-D scans S_1 and S_2 as input, the goal is output the rigid transformation T_{12} that align the two input scans. We assume that the intrinsic parameters are known, but not constraint the overlap of S_1 and S_2. To explore the hidden cues in the challenge task, we exploit RGB-D PIFu to recovery the underlying human geometry prior from a

978-1-6654-4254-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCST53801.2021.00105

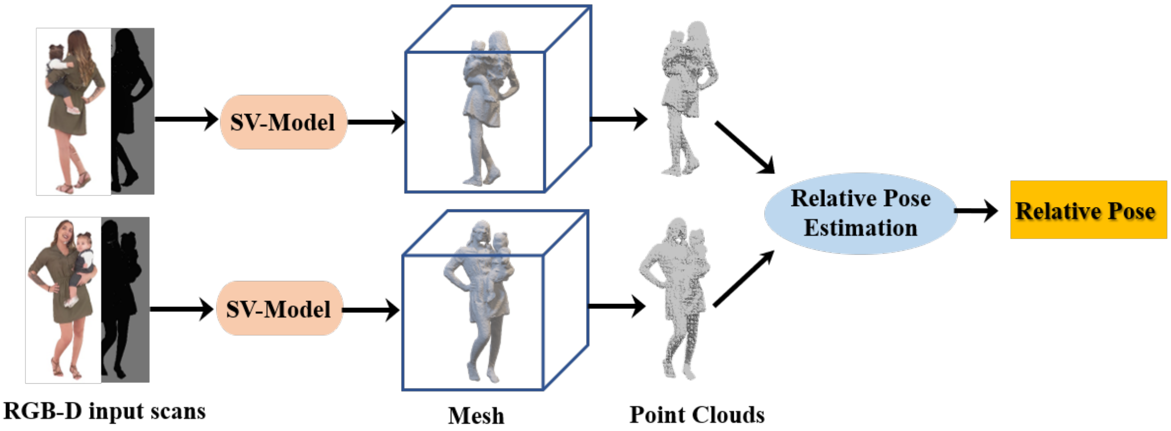

Fig. 1. **Methods Pipeline.** SV-Model means our single-view human reconstruction model. The proposed framework combines human reconstruction and relative pose estimation optimization. Given the RGB-D input scans, the SV-Model reconstruct the completed human geometry, then the optimization module estimate the relative pose.

sparse view. As illustrated in Fig. 1, our architecture consists of two modules, one is single-view human reconstruction producing a compelling and completed human geometry, the other is relative pose optimization module, to estimate relative pose from the sampling mesh.

A. Human reconstruction with implicit function

Recently, learning a continuous implicit function representing human shape [12] has been a trend, for it's not limited by output resolution and fixed topology. In these methods, RGB-D PIFu proposed by [2] has achieved a state-of-the-art result, producing a detail-preserving complete surface in real time with RGB-D scans. The human surface in [2] is defined as a level set of:

$$f(F(\Pi(q)), q_{.z}, T(q)) = s : s \in \mathbb{R} \qquad (1)$$

$$T(q) = trunc(q_{.z} - D(\Pi(q))) \qquad (2)$$

f is a continuous implicit function represented by multi-layer perceptrons(MLPs). For a query point q, $\Pi(\cdot)$ is the perspective projection function, F($\cdot$) is the feature extracted by encoder networks, $D(\cdot)$ is a bi-linear function sampling depth values on the depth image, the PSDF value of q is introduced by $q_{.z} - D(\Pi(q))$ to fully utilize the depth observation, $T(\cdot)$ is used to truncated PSDF value to $[-\sigma_i, \sigma_i]$, eliminating the ambiguities of using global depth values. The sign of s represents if the point is inside or outside the surface.

The RGB-D PIFu uses view-centric coordinate systems, extracts geometry-aware feature maps and exploits multi-view aggregation strategy for fusion, satisfying the requirements of good generalization as discussed in [13]. Actually, we experimentally testify the good generalization in real data collected by Kinect, using the trained model with 300 high-fidelity data from [2]. The reconstruction of visible parts is general detailed and lifelike, while the unseen partial completion is a little oversmooth but reasonable according to prior learning from the data, which still empowers us with abundant hidden cues. Considering the efficiency, compelling performance and good generalization in real data, we determine the RGB-D PIFu from [2] to dig out the underlying geometry prior.

B. Relative Pose Estimation Module

Our approach essentially falls into the optimization category although we utilize the deep learning to enrich the underlying human geometry prior. This module serves as a relative pose estimation, ensuring a fair good and stable pose estimation included non-overlapping input setting. We expect this module as simple and effective as possible. For the overlap between the input RGB-D scans is larger than 0.80, we can get a good initial via traditional optimization methods[7, 14, 15]. However, the standard pipeline of first extracting features from RGB-D scans and then matching the correspondences features is not suitable for small-overlapping or non-overlapping scans. For this challenge setting, we can assume the T_i is the identity matrix, using the single-view RGB-D PIFu as abovementioned to complete the unseen part, fully leveraging the human geometry prior,. It has a large overlap between sampling point clouds from the completed meshes. Thus, many simple and effective optimization methods such as global registration [8] or geometric registration [16, 17] can be choosed, to estimate the relative pose from these sampling point clouds. Actually, our relative pose estimation approach is efficient, because the major time-consuming 3D human reconstruction via [2] can be finished in time.

III. Experiments

In this section, we present an experimental evaluation of the proposed approach. We firstly describe our evaluation dataset and evaluation metrics. Then, we compare our method with several baseline techniques, assessing performance quantitively on different overlapping rate RGB-D input pairs.

TABLE I
THE RELATIVE POSE ESTIMATION EVALUATION ON OUR DATASET.

	overlap(0-5%)		**overlap(10%-30%)**		**overlap(40%-70%)**		**overlap($\geq$ 80%)**	
	Ratation	**Trans.**	**Ratation**	**Trans.**	**Ratation**	**Trans.**	**Ratation**	**Trans.**
Super4PCs[14]	126.08	1.715	103.95	1.468	38.99	0.574	4.46	0.084
Greg[15]	124.41	1.679	92.50	1.538	64.47	1.109	18.44	0.325
ICP[7]	165.97	2.252	122.02	1.943	67.86	1.208	27.75	0.540
Greg+ICP	126.72	1.664	103.88	1.461	35.10	0.508	**1.27**	0.0262
Ours	**7.15**	**0.138**	**6.54**	**0.130**	**5.90**	**0.106**	1.29	**0.0259**

A. *Dataset*

We perform experimental evaluation on two types of data. One is synthesized data, rendering 200 high-quality scans from 60 views with rotation and random shifts. Note that, to keep consistent with the real data, for the color image, we use the PRT-based render as in [1], for the depth image, we first render the ground truth depth maps, and then adding the TOF sensors noise on top of them following [18]. Finally, we synthesize RGB-D data with resolution 512×512. The other is real data collected by multi-view Kinects, included large poses, various clothes, different people. We firstly segment the human with mask provided by Kinect. Then, we align the color image and depth image with pose parameters of Kinect depth camera, getting the final real RGB-D data with resolution 640×576.

For a more comprehensive and detailed analysis, we classify all the data into four categories according to the overlap rate: overlap rate 0-5%, overlap rate 10%-30%, overlap rate 40%-70%, and overlap rate over 80%. Then, we select about 400 characteristic data pairs for each category, besides half each for the real data and the synthetic data, as our final evaluation data.

Note that, for the RGB-D PIFu, we train it use 500 high-quality scans following [2].

B. *Evaluation Metrics*

We evaluate the rotation matrix R and translation part t of the relative pose $T = (R, t)$ respectively. We follow the standard protocol of reporting the rotation angle error $arccos(\frac{tr(R^* R^T) - 1}{2})$ and the translation error $\|t - t^*\|_2$, let (R^*, t^*) be the ground truth relative pose and (R, t) be the predicted pose.

C. *Quantitative Evaluation*

We consider the four baseline approaches: Super4PCS [14], Greg [15], ICP [7], and combine Greg with ICP. Super4PCS is a widely used global scan matching method between two 3D point clouds. Greg is another state-of-the-art global registration, which combining cutting-edge feature and reweighted least squares for rigid pose registration. ICP(Iterative Closet Point) , a local optimization algorithm, has been a mainstay of geometric registration in both research and industry. In this paper, we use point-to-plane ICP which has a faster convergence. We also combine global registration with local optimization as a baseline, the former provides an initial pose, the later refines the pose.

TABLE I provides quantitative results of our approach and baseline methods. We show the mean error for rotation and translation components for overlapping rate $(0-5\%, 10\%-30\%, 40\%-70\%, \geq 80\%)$ scan pairs, respectively. Overall, we found that the less overlap rate, the greater advantage of our method. Our approach outperforms baseline approaches considerably in small-overlapping or almost overlapping settings, and performs slightly better in large-overlapping setting. In the four baselines, combining Greg with ICP is much better than others, especially in significant overlap scene, while in small-overlapping setting, all the baselines perform badly, making no difference.

Small overlap(overlapping rate 10%-30%) or almost no overlap(overlapping rate 0-5%). All the baselines perform very badly, with over 120 rotation errors and over $1.5m$ translation errors. These methods rely on accurate correspondence in overlap region, thus they can't handle the small-overlapping or non-overlapping settings. Even so, our approach performs much better, with mean errors in rotation/translation $7.15/0.138m$ for almost no overlap and $6.54/0.130m$ for small overlap. It fully demonstrates the effectiveness of our proposed approach for non-overlapping setting, thanks to the coarse-to-fine optimization strategy.

Middle overlap(overlapping rate 40%-70%). Although the Greg with ICP is much better than other baselines, with rotation/translation errors $35.10/0.508m$, our approach achieves much better results, with corresponding errors $5.90/0.106m$.

Significant overlap(overlapping rate over 80%). There is no significant difference between all the methods. All the baselines perform fair good, especially, the Greg with ICP with small rotation/translation errors $1.27/0.0262m$. It further shows that the standard baselines require the input scans possessing considerable overlapping regions for good performance. However, Our approach is still competitive, with mean rotation/translation errors $1.29/0.0259m$.

From this experiment, all the four baselines rely highly on large overlap of input scans, not able to handle the extreme pose setting, while our approach performs stable and good in all settings.

IV. Conclusion

In this paper, we have proposed an approach for estimating the relative pose between two RGB-D scans. For the input, we focus on human scene where there are plenty of applications in reality and don't limit the input scans to have large overlap. We leverage implicit function model for human reconstruction with the RGB-D input, fully exploring the underlying human geometry prior for unseen parts. Through evaluation on different overlap data, our method considerably outperforms the state-of-the-art baselines, especially for non-overlapping scans.

Acknowledgment

Thanks very much to all the teachers and students who helped me finish this paper. This work was supported by the National Key RD Program of China (2018YFB1403900). It was also the research achievement of the Key Laboratory of Digital Rights Services.

References

[1] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2304–2314.

[2] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu, "Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5746–5756.

[3] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7287–7296.

[4] X. Li, S. Liu, K. Kim, S. De Mello, V. Jampani, M.-H. Yang, and J. Kautz, "Self-supervised single-view 3d reconstruction via semantic consistency," in *European Conference on Computer Vision*. Springer, 2020, pp. 677–693.

[5] Z. Yang, J. Z. Pan, L. Luo, X. Zhou, K. Grauman, and Q. Huang, "Extreme relative pose estimation for rgb-d scans via scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4531–4540.

[6] G. K. Tam, Z.-Q. Cheng, Y.-K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X.-F. Sun, and P. L. Rosin, "Registration of 3d point clouds and meshes: a survey from rigid to nonrigid," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 7, pp. 1199–1217, 2012.

[7] S. Bouaziz, A. Tagliasacchi, and M. Pauly, "Sparse iterative closest point," in *Computer graphics forum*, vol. 32, no. 5. Wiley Online Library, 2013, pp. 113–123.

[8] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," in *European Conference on Computer Vision*. Springer, 2016, pp. 766–782.

[9] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Relative camera pose estimation using convolutional neural networks," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2017, pp. 675–687.

[10] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.

[11] Z. Yang, S. Yan, and Q. Huang, "Extreme relative pose network under hybrid representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2455–2464.

[12] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3d shape reconstruction and completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6970–6981.

[13] M. A. Bautista, W. Talbott, S. Zhai, N. Srivastava, and J. M. Susskind, "On the generalization of learning-based 3d reconstruction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2180–2189.

[14] N. Mellado, D. Aiger, and N. J. Mitra, "Super 4pcs fast global pointcloud registration via smart indexing," in *Computer Graphics Forum*, vol. 33, no. 5. Wiley Online Library, 2014, pp. 205–215.

[15] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[16] B. PaulJ and M. NeilD, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[17] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and vision computing*, vol. 10, no. 3, pp. 145–155, 1992.

[18] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect v2 for mobile robot navigation: Evaluation and modeling," in *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 388–394.