Human-in-the-loop Pose Estimation via Shared Autonomy

Zhefan Ye zhefanye@umich.edu Computer Science and Engineering, University of Michigan Ann Arbor, Michigan, USA

> Stephen Hart swhart@traclabs.com TRACLabs Inc. Webster, Texas, USA

Jean Y. Song jyskwon@kaist.ac.kr School of Computing, KAIST Daejeon, Republic of Korea

Jorge Vilchis jvilchis@umich.edu Electrical and Computer Engineering, University of Michigan Ann Arbor, Michigan, USA

Odest Chadwicke Jenkins ocj@umich.edu Computer Science and Engineering, University of Michigan Ann Arbor, Michigan, USA Zhiqiang Sui zsui@umich.edu Computer Science and Engineering, University of Michigan Ann Arbor, Michigan, USA

Walter S. Lasecki wlasecki@umich.edu Computer Science and Engineering, University of Michigan Ann Arbor, Michigan, USA



Figure 1: Overview of our approach. (a) The robot captures the scene using its depth sensor. (b) A human operator initializes the object pose using a mouse, followed by pose estimation from our machine estimator, which utilizes the object geometry model to render scenes and refine until convergence. (c) The robot performs object manipulation using the registered pose.

ABSTRACT

Reliable, efficient *shared autonomy* requires balancing human operation and robot automation on complex tasks, such as dexterous manipulation. Adding to the difficulty of shared autonomy is a robot's limited ability to perceive the 6 degree-of-freedom pose of objects, which is essential to perform manipulations those objects afforded. Inspired by Monte Carlo Localization, we propose a generative human-in-the-loop approach to estimating object pose. We characterize the performance of our mixed-initiative 3D registration approach using 2D pointing devices via a user study. Seeking

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

IUI '21, April 14–17, 2021, College Station, TX, USA

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8017-1/21/04.

https://doi.org/10.1145/3397481.3450654

an analog for Fitts's Law for 3D registration, we introduce a new evaluation framework that takes the entire registration *process* into account instead of only the outcome. When combined with estimates of registration confidence, we posit that mixed-initiative registration will reduce the human workload while maintaining or even improving final pose estimation accuracy.

CCS CONCEPTS

 Human-centered computing → Human computer interaction (HCI);
Computing methodologies → Artificial intelligence.

KEYWORDS

pose estimation, human-in-the-loop, shared autonomy, affordances, Monte Carlo localization

ACM Reference Format:

Zhefan Ye, Jean Y. Song, Zhiqiang Sui, Stephen Hart, Jorge Vilchis, Walter S. Lasecki, and Odest Chadwicke Jenkins. 2021. Human-in-the-loop Pose Estimation via Shared Autonomy. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3397481.3450654

1 INTRODUCTION & RELATED WORK

Dexterous manipulation of objects in unstructured environments requires robots to estimate object poses accurately enough to reason about the possible manipulations actions it affords. With the advent of accurate and low-cost 3D sensors, object pose estimation has become an increasingly well-studied area of research, with new methods ranging from using multiple view geometry, comparison based on object-3D reconstruction, or registration using Monte Carlo and particle filtering, having been introduced [6, 24, 31]. Despite this interest, the current limitations of pose estimation remains a critical bottleneck to creating robust and reliable robotic systems [6, 25]. Hence, human input is needed to bridge the gap.

Unfortunately, direct teleoperation by a human is not feasible because deployed systems are often intricate, making manual operation burdensome and tedious. To make matters worse, 3D registration tasks are often performed via 2D pointing devices (e.g., a joystick or mouse), which are often unintuitive and require more cognitively demanding multi-step effort to achieve the desired input – a fact that is particularly problematic during long-duration, continuous operation of the type often required of many robotic systems. This difficulty means that the skill-level necessary for human operators to reliably and efficiently estimate object poses is high, making human operators scarce and errors unavoidable, as reported by DARPA Robotics Challenge teams [2, 19].

Shared autonomy grasping has been demonstrated with RViz interactive markers¹ by Gossow et al. [12]. Other work has also introduced shared autonomy approaches to assist human operators in performing various tasks [1, 4, 5, 13, 21, 26, 28, 29]. Affordancebased representations for shared autonomy have been used to operate robots when manual control is infeasible, such as over poor communication channels [7, 9, 20]. Reinforcement learning methods have been used to learn affordances [10, 15, 16]. Fitzpatrick et al. demonstrated that a robot could learn affordances for pushing and grasping objects through exploration [10]. However, learningbased approaches are often limited by specific environments and datasets, which can be costly to collect on a large scale.

LabelFusion [23], proposed by Marion et al., leverages both human and machine computation to tackle the problem of 3D object pose estimation. It allows a user to register an object model by selecting three points in the reconstructed 3D scene using a robotic visualization interface [22] and subsequently performs machine object registration based on the points selected by the user. However, one-shot point selection can pose a challenge to novice users since finding appropriate registration points may require multiple steps of inspection, especially in a cluttered 3D scene. Instead, our system considers object registration as a gradual process so that users at different skill levels can choose to operate at their own rates.

Therefore, to achieve pose estimation with less human effort, we propose a human-in-the-loop approach that balances human operation and robot autonomy by leveraging both human and machine perception. Our approach leverages the concept of mixed-initiative interaction [17] to reduce tedious and cumbersome tasks for human operators while maintaining high accuracy by allowing automation to take over when it is able. In our system, human operators begin to perform a coarse initialization until the machine is sufficiently confident that it is able to take over to refine object pose. The human operator would then be able to take back over and further refine the machine's estimate if they saw fit. The key to enabling our hybrid approach is a pose estimation method inspired by Monte Carlo localization [8] that generates multiple hypotheses of object pose initialized by a human operator and then refines them over successive iterations, akin to sequential Bayesian filtering. We evaluate our approach via a user study, which recruits both expert and novice users to validate that the benefits of our method apply to operators of different skill levels. To understand the trade-offs of approaches like ours, we propose a new way to view humanin-the-loop registration performance that takes into account the performance curve over the course of the task, instead of just the quality of the outcome.

One potential use case for our approach could be data annotation, as it can be used out-of-the-shelf and does not require the a priori training of data-driven models. When a robot can autonomously execute actions based on the object's affordances once the object's pose is determined, we believe that our results and analysis shed light on improving efficiency and ease-of-use of systems that facilitate better human-robot collaboration.

2 APPROACH: SHARED AUTONOMY POSE REGISTRATION

We first overview our shared autonomy pose estimation system, as illustrated in Fig. 1. The estimation process is composed of two steps: i) human initialization and ii) snap-to-grid (STG) registration using *Monte Carlo Localization* (MCL). Note that *localization* indicates both the translation and rotation of an object in this paper.

Human Initialization. The robot's depth sensor captures the scene (Fig. 1a), and the acquired depth image can be visualized as a point cloud. A human operator then moves an object model to align it with the corresponding object in the point cloud using a 2D pointing device (here, a mouse) and an interactive 3D viewer (Fig. 1b). The details of the initialization process are in Sec. 3.

Pose Estimation. Our pose estimator refines the object pose given the human initialization. Given a depth image, Z_d , from the robot sensor, we model the problem as a joint distribution $P(q, q_u, o, Z_d)$, where q is the six degrees of freedom (DoF) object pose in the world frame, q_u is the human-initialized pose, and o is the known object with its 3D geometry model (Fig. 1b). Therefore,

$$P(q, q_u, o, Z_d) = P(q \mid q_u, o, Z_d) P(q_u, o, Z_d) \propto P(q \mid q_u, o, Z_d)$$
(1)

where $P(q_u, o, Z_d)$ is the observation prior. Essentially, our system aims to find the optimal pose that maximizes the joint probability in Eq. 1, such that

$$q^* = \operatorname*{argmax}_{q} P(q \mid q_u, o, Z_d)$$
(2)

¹http://wiki.ros.org/rviz

Monte Carlo Localization. To estimate the optimal object pose, q^* , we use the particle-based local search, which is a simplified version of the method used in [30]. This local search method is inspired by [8], which relies on a set of M weighted samples, $\{q^{(i)}, w^{(i)}\}_{i=1}^{M}$, to represent multiple hypotheses that indicate the object poses and update the weights over iterations.

Given an object class o and its geometry model g, we can render a depth image with the intrinsic matrix of the robot's sensor. The depth image is further converted to a point cloud, Z_r . Initial hypotheses for this search are sampled from a Gaussian distribution. The mean is set to the human initialized pose, q_u . To evaluate each sample hypothesis, we first crop the point cloud Z_r based on the corresponding sample pose. The hypotheses are then measured by the agreement between the rendered point cloud and the observed point cloud. We define the similarity between two point clouds as the following *Inliers* function,

$$\text{Inliers}(p, p') = \begin{cases} 1, & \text{if } \|p - p'\|_2 < \epsilon \\ 0, & \text{otherwise} \end{cases}$$
(3)

where p is a point in the observed point cloud Z_d , and p' is a point in the rendered point cloud, Z_r , and ϵ is a pre-determined threshold. This *Inliers* function is applied to the entire point cloud such that:

$$I^{(i)} = \sum_{a, b \in z^{(i)}} \text{Inliers}(Z_r(a, b), Z_d^{(i)}(a, b))$$
(4)

where *i* indicates a sample, *a* and *b* are 2D indices in the observed point cloud Z_d . Here, the weight *w* of each pose hypothesis, $q^{(i)}$, is defined as,

$$W(q^{(i)}) = \alpha * \frac{I^{(i)}}{N_b} + \beta * \frac{I^{(i)}}{N_r}$$
(5)

where N_b is the number of matched points in the observed point cloud Z_d , N_r is the number of matched points in the rendered point cloud Z_r , and α , β are pre-determined coefficients. The first term in Eq. 5 estimates the similarity between the rendered point cloud and the observed point cloud. The second term mitigates the noise caused by human initialization and environment by comparing the number of matched points with itself. By employing *importance sampling* [32], we reassign a new weight to each sample at each iteration.

3 EXPERIMENTAL SETUP

Our implementation is built upon the Robot Operating System (ROS) [27]. We use affordance template framework (AT) [14, 18], which is based on Rviz [14], to perform object registration tasks².

To compute the trajectory between two end-effector waypoints, we utilized TracIK planner [3]. Our manipulation experiments are performed by a Fetch mobile manipulator³. For particle-based local search, we implement our approach using CUDA⁴ and use a total number of 625 particles to best utilize CUDA memory. We run 640 iterations for each registration task.

We conducted a user study with 16 undergraduate and graduate students at the University of Michigan. Eight of them are considered to be *experts* due to their familiarity with RViz or other relevant 3D

³http://fetchrobotics.com/research/

⁴https://developer.nvidia.com/about-cuda

registration tools. The other eight participants are considered to be *novices* since they have no prior experience with 3D registration or manipulation tools. We instructed the participants to complete the registration tasks in four scenes, and for each scene, they were asked to perform the task in two modes — *accurate* mode and *quick* mode. In accurate mode, the participants need to register the 3D object model as accurately as possible, while in quick mode, they were asked to register as quickly as possible. We randomized the order of the tasks to avoid learning effects.

Because 3D controls, such as the ones used by RViz, only allow human operators to move along one DoF at a time, extracting the raw human registered pose at each timestep will not properly reflect the overall progression of the human registration process. Therefore, we broke down the human registration process according to click actions (since the participants need to use click-and-drag to perform registration) and reconstruct them based on their DoF. To obtain the ground truth pose of each scene, we calculated the average of the final human registered poses in accurate mode.

To evaluate our system, we use the human registration result as our baseline condition, which consists of two modes (accurate and quick) and no autonomy. Given the human initialized poses from human registration, our shared autonomy system approximates the target pose using particle-based local search, and we call this condition snap-to-grid, or STG. Instead of measuring only the final result, we take the *entire* registration process into account. This analysis is inspired by Fitts's law [11], which relates human performance on a pointing task (measured as time-taken) with the difficulty of the pointing task itself (measured as a function of target size and distance from cursor) via a logarithmic function. Our framework defines an equivalent task by fixing the object's distance and defining a target as the registration error margin, which has 6 DoF. In other words, Fitts's law does not directly describe the mouse movement in our registration task because the human operators must interleave the controls in all DoF during the registration tasks due to the lower dimensionality of the input device (2 DoF) compared to the input task (6 DoF). Further, target size - which does not formally exist in the problem of pose estimation - is really a measure of "input equivalence" (e.g., anywhere on a button is equally valid to click). We thus use an acceptable error margin in place of target size in our analysis. Fig. 2 shows the performance curves with our adaptation of Fitts's law. The error margin represents the tolerance from the ground truth pose, and the timestep indicates the amount of time required to reach the error margin. Given the reconstructed registration process, we can capture the pose at every 0.5 seconds and measure the difference between the current pose and the ground truth pose as the error margin. By computing Euclidean distance of translation and geodesic distance of rotation, we observe how the registration progression is gradually approaching the ground truth pose. Note that all curves in Fig. 2 are averaged across all scenes and all participants within each subject group.

4 RESULTS

We first compare the STG condition (blue curves in Fig. 2) with the baseline condition (red curves in Fig. 2) and find that STG can consistently improve translation registration performance in the early stage of registration. Fig. 2a and Fig. 2b illustrate that

²http://traclabs.com/projects/affordance-templates/



Figure 2: We compare the result of human registration (red curves) with STG (blue curves) for two subject groups (novice and expert) in both *quick* mode and *accurate* mode. The x-axis represents the registration error margin from the ground truth pose, and the left y-axis represents the amount of time needed to reach an error margin. The shaded area, together with the right y-axis, indicates the remaining number of human operators at a given error margin.

the performance of STG performs at least equal to or better than human registration. The vertical curves in Fig. 2 suggest that the registered pose never reaches the specified accuracy level in the given condition, as most of the novice operators in Fig. 2a cannot reach 0.05 error margin (red shaded area in Fig. 2a). Our shared autonomy system can take over the *quick* registration task and help the novice operators improve registration accuracy, as indicated by the blue curve and blue shaded area in Fig. 2a.

For rotation error, our approach did not help improve the final accuracy, as shown in Fig. 2c and Fig. 2d. However, STG reduces rotation error after the object reaches 90 degree error margin. This is due to the fact that the 3D sensor only captured a portion of the object in the point cloud, whereas STG uses the complete 3D object model to perform particle-based local search. Therefore, if the rotation error margin is too large or too small, our inliers function (Eq. 4) cannot properly measure the difference between the object model and the object point cloud, potentially leading STG to get stuck in a local optimum.

This result suggests that the best practice for the human operators is to focus on rotation registration, while our particle-based local search method helps reduce translation error in order to balance between human effort and registration accuracy. We believe the analysis, which is inspired by Fitts's law and made possible by Fig. 2, can be applied to other tasks that involve 3D registration with 2D pointing devices. Human-in-the-loop Pose Estimation via Shared Autonomy

5 CONCLUSION

We present a human-in-the-loop approach to improving object pose estimation. Our work addresses the critical bottleneck of estimating object pose in unstructured environments for manipulation. Our experiments show that Monte Carlo Localization (MCL) method can utilize human initialization to bootstrap the registration process and estimate object pose by proposing multiple hypotheses and resampling based on observation. Through shared autonomy, we can balance between human effort and pose estimation accuracy. We demonstrate this via a performance analysis approach inspired by Fitts' Law, which indicates where and when our new approach can be expected to positively augment human performance. Here, analysis helps to inform the design of systems using our approach, but it can also be used to assess future shared autonomy systems with similar goals.

Our approach does not require 3D scene reconstruction, nor does it require any training data. Thus, it can be used out-of-the-box regardless of whether the scene is known in advance. We believe that our findings will facilitate better shared autonomy design between humans and robots.

ACKNOWLEDGMENTS

This research was supported in part by NSF award IIS-1638047. This research was also supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2020R1I1A1A0107238511), and the Toyota Research Institute (TRI).

REFERENCES

- Nicholas Alunni, Calder Phillips-Grafflin, Halit Bener Suay, Daniel Lofaro, Dmitry Berenson, Sonia Chernova, Robert W Lindeman, and Paul Oh. 2013. Toward a userguided manipulation framework for high-dof robots with limited communication. In *IEEE TePRA*. IEEE, 1–6.
- [2] Christopher G Atkeson, BPW Babu, N Banerjee, D Berenson, CP Bove, X Cui, M DeDonato, R Du, S Feng, P Franklin, et al. [n.d.]. What happened at the darpa robotics challenge, and why. ([n.d.]).
- [3] Patrick Beeson and Barrett Ames. 2015. TRAC-IK: An Open-Source Library for Improved Solving of Generic Inverse Kinematics. In *IEEE-RAS Humanoids*. Seoul, Korea.
- [4] Jeffrey P Bigham, Walter S Lasecki, and Jacob O Wobbrock. 2015. Target Acquisition and the Crowd Actor. *Human Computation* 2, 2 (2015), 135–154.
- [5] D. J. Butler, S. Elliot, and M. Cakmak. 2017. Interactive scene segmentation for efficient human-in-the-loop robot manipulation. In *IEEE IROS*. 2572–2579. https://doi.org/10.1109/IROS.2017.8206079
- [6] Alvaro Collet, Manuel Martinez, and Siddhartha S Srinivasa. 2011. The MOPED framework: Object recognition and pose estimation for manipulation. *IJRR* (2011), 0278364911401765.
- [7] Mathew DeDonato, Velin Dimitrov, Ruixiang Du, Ryan Giovacchini, Kevin Knoedler, Xianchao Long, Felipe Polido, Michael A. Gennert, Taşkın Padır, Siyuan Feng, Hirotaka Moriguchi, Eric Whitman, X. Xinjilefu, and Christopher G. Atkeson. 2015. Human-in-the-loop Control of a Humanoid Robot for Disaster Response: A Report from the DARPA Robotics Challenge Trials. *Journal of Field Robotics* 32, 2 (2015). https://doi.org/10.1002/rob.21567
- [8] Frank Dellaert, Dieter Fox, Wolfram Burgard, and Sebastian Thrun. 1999. Monte Carlo Localization for Mobile Robots. In IEEE ICRA.
- [9] Maurice; Fallon, Scott; Kuindersma, Sisir; Karumanchi, Matthew; Antone, Toby; Schneider, Hongkai; Dai, Claudia; Perez D'Arpino, Robin; Deits, Matt; DiCicco, Dehann; Fourie, Twan; Koolen, Pat; Marion, Michael; Posa, Andres; Valenzuela, Kuan-Ting; Yu, Julie; Shah, Karl; Iagnemma, Russ; Tedrake, and Seth Teller. 2014. *Affordance-based Perception and Whole-body Planning in the DARPA Robotics Challenge*. Technical Report. MIT.
- [10] Paul Fitzpatrick, Giorgio Metta, Lorenzo Natale, Sajit Rao, and Giulio Sandini. 2003. Learning about objects through action-initial steps towards artificial cognition. In *IEEE ICRA*, Vol. 3. IEEE, 3140–3145.

- [11] Ken Goldberg, Siamak Faridani, and Ron Alterovitz. 2015. Two large open-access datasets for Fitts' law of human motion and a succinct derivation of the squareroot variant. *IEEE THMS* 45, 1 (2015), 62–73.
- [12] David Gossow, Adam Leeper, Dave Hershberger, and Matei Ciocarlie. 2011. Interactive markers: 3-d user interfaces for ros applications [ros topics]. *IEEE RAM* 18, 4 (2011), 14–15.
- [13] Sai Gouravajhala, Jean Y Song, Jinyeong Yim, Raymond Fok, Yanda Huang, Fan Yang, Kyle Wang, Yilei An, and Walter S Lasecki. 2017. Towards Hybrid Intelligence for Robotics. CI (2017).
- [14] Stephen Hart, Paul Dinh, and Kimberly Hambuchen. 2015. The affordance template ROS package for robot task programming. In *IEEE ICRA*.
- [15] Stephen Hart and Rod Grupen. 2009. Intrinsically motivated affordance learning. (2009).
- [16] Stephen Hart and Roderic Grupen. 2013. Intrinsically motivated affordance discovery and modeling. In *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer, 279–300.
- [17] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In ACM CHI. ACM, 159–166.
- [18] Joshua James, Yifan Weng, Stephen Hart, Patrick Beeson, and Robert Burridge. 2015. Prophetic goal-space planning for human-in-the-loop mobile manipulation. In *IEEE-RAS Humanoids*. IEEE, 1185–1192.
- [19] Matthew Johnson, Brandon Shrewsbury, Sylvain Bertrand, Tingfan Wu, Daniel Duran, Marshall Floyd, Peter Abeles, Douglas Stephen, Nathan Mertins, Alex Lesman, et al. 2015. Team IHMC's lessons learned from the DARPA robotics challenge trials. J FIELD ROBOT 32, 2 (2015), 192–208.
- [20] Twan Koolen and Jesper Smith. 2013. Summary of Team IHMC's Virtual Robotics Challenge Entry. IEEE-RAS, Atlanta, Georgia.
- [21] Walter S Lasecki, Raja Kushalnagar, and Jeffrey P Bigham. 2014. Legion scribe: real-time captioning by non-experts. In Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility. ACM, 303–304.
- [22] Pat Marion. 2015. Director: A robotics interface and visualization framework. http://github.com/RobotLocomotion/director
- [23] Pat Marion, Peter R. Florence, Lucas Manuelli, and Russ Tedrake. 2017. Label-Fusion: A Pipeline for Generating Ground Truth Labels for Real RGBD Data of Cluttered Scenes. arXiv:1707.04796 [cs] (Sep 2017). http://arxiv.org/abs/1707.04796 arXiv: 1707.04796.
- [24] Venkatraman Narayanan and Maxim Likhachev. 2016. PERCH: Perception via Search for Multi-Object Recognition and Localization. In Robotics and Automation (ICRA), 2016 IEEE International Conference on. IEEE.
- [25] Chavdar Papazov, Sami Haddadin, Sven Parusel, Kai Krieger, and Darius Burschka. 2012. Rigid 3D geometry matching for grasping of known objects in cluttered scenes. *IJRR* (2012), 0278364911436019.
- [26] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.
- [27] Morgan Quigley, Josh Faust, Tully Foote, and Jeremy Leibs. [n.d.]. ROS: an open-source Robot Operating System.
- [28] Jean Y Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, and Walter S Lasecki. 2018. Two Tools are Better Than One: Tool Diversity as a Means of Improving Aggregate Crowd Performance. In ACM IUI.
- [29] Jean Y Song, Stephan J Lemmer, Michael Xieyang Liu, Shiyan Yan, Juho Kim, Jason J Corso, and Walter S Lasecki. [n.d.]. Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In ACM IUI.
- [30] Zhiqiang Sui, Odest Chadwicke Jenkins, and Karthik Desingh. 2015. Axiomatic particle filtering for goal-directed robotic manipulation. In *IEEE IROS*. IEEE, 4429–4436.
- [31] Zhiqiang Sui, Lingzhu Xiang, Odest C Jenkins, and Karthik Desingh. 2017. Goaldirected robot manipulation through axiomatic scene estimation. *The International Journal of Robotics Research* 36, 1 (2017), 86–104.
- [32] Surya T Tokdar and Robert E Kass. 2010. Importance sampling: a review. WIREs Comp Stats 2, 1 (2010), 54–60.