Intuitive Gesture Controlled Semi-Autonomous Teleoperation System for Improved Productivity

Liyuan Qi, Olaoluwa Popoola, Jingyan Wang, Muhammad Ali Imran, Wasim Ahmad

I. INTRODUCTION

Teleoperation has recently gained prominence in robotics, finding applications in diverse industrial and commercial fields. While automated systems struggle with complex and unpredictable tasks, teleoperation capitalizes on human intelligence to enhance robotic capabilities in scenarios, such as surgery [1], [2], aerospace [3], and search & rescue [4]. Whilst it requires understanding human intention, the systems should operate in real-time to interact with the working environment effectively.

In teleoperation, human operator inputs can be categorized into four types: keyboard and joystick [5], [6], [7], [8], [9]; motion capture devices including virtual reality suites, haptic touch systems, customized motion input systems, etc [10], [11], [12], [13]; wearable sensor systems, such as EEG/EMG, IMU, and motion capture suits [14], [15], [16]; and depth or RGB camera system [17], [18], [19], [20]. Moving from the first to the last method, controlling input mechanisms becomes increasingly intuitive for the operator. Joystick and keyboard input is cost-effective and avoids complex data analysis, though it requires operators to undergo control mastery training. Motion capture devices provide a relatively natural way to record human movement within a limited workspace. Wearable sensors and camera-based systems offer a more intuitive and comfortable user interface. Despite the lack of haptic feedback, camera-based systems capture the richest data while minimizing operator burden. However, they necessitate intricate data processing to extract understandings from the captured frames. Recently, hand pointnet [21], which builds on pointnet++ [22], [23], has become one of the fundamental approaches to estimating depth from hand frames and has inspired the robotics field [20]. It converts the hand depth frames into point clouds and estimates twenty-one 3-dimensional hand joint locations. These locations could be used to analyze the $SE(3)$ hand transformation matrix and finger configuration.

In addition to human input methods, adequate computational time to compute the controlling parameters and low-latency wireless communication between the human operator and robot to transfer these parameters from the user to the robot ends are also crucial factors in teleoperation performance. Since operators would work in varied contexts, some might use mobile platforms with limited computational capacity, while others might operate the robot from long distances. In such cases, teleoperation systems may offload large computations to edge servers to expedite data processing [24], [25]. In particular, multi-access edge computing (MEC) moves the server from the centralized cloud to the edge of the network, which is closer to the robot, thereby reducing communication time. The 5G wireless communication offers ultra-reliable and low-latency transmission to link edge servers and devices, resulting in near-zero packet loss and high data transfer efficiency [26], [27].

To tackle the outlined challenges and offer a comprehensive solution, we developed and tested a gesture-based semi-autonomous teleoperation system (SATS) that incorporates computer vision, posture optimization, edge computing, and 5G wireless communication. The developed teleoperation system is presented in Fig. 1. At the operator end, a novel PointNet-based hand joint estimator uses an Intel RealSense D455 depth camera and an Nvidia RTX 3000 GPU for initial data collection and pre-processing. The extensive computation is offloaded to an MEC server equipped with Nvidia RTX 6000 GPUs. Lastly, an Nvidia Xavier NX handles the differential kinematic calculations at the robot end and runs the control software for a Franka Emika robotic arm.

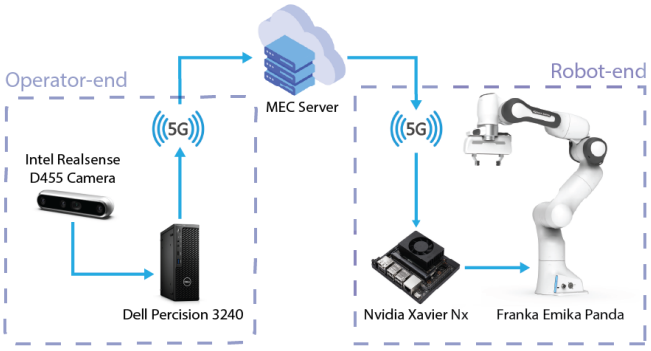

Fig. 1. An overview of the proposed teleoperation system.

II. METHODOLOGY

The first stage, shown in Fig. 2, collects RGB and Depth frame pairs from the Intel RealSense RGB-D camera. A hand detector generates a bounding box on RGB frames, and a frustum space is generated based on the bounding box in depth frame space. After filtering out the background and foreground, the hand depth frame is converted into a point cloud and uniformly downsampled to 1023 points. The Hand-VoteNet contains a standard PointNet module and a refining module. The PointNet module follows the pre-processing and

James Watt School of Engineering, University of Glasgow, Glasgow - G12 8QQ, United Kingdom
Email: l.qi.1@research.gla.ac.uk, and wasim.ahmad@glasgow.ac.uk

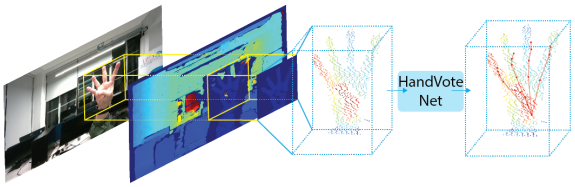

Fig. 2. Hand joint detection and estimation.

architecture of Hand PointNet [21] to get the first estimation. The refining module comprises 21 Pointnet++ modules, each dedicated to processing the subset of points in the vicinity of one of the piror estimated joints obtained. The second estimation will be the weighted voting result of all the points within each subset. The subsets could be either collected from multiple radii with a multi-scale grouping model, or from a single radius with a single-scale grouping model [23]. Finally, the refined hand joint locations are the weighted sum of these two estimations. After having the joint location, the $SE(3)$ transformation could be calculated from a singular value decomposition (SVD) based optimization [28]. The following objective function is used to minimize the least-square error between the origin hand posture $H^{'}$ and current hand posture H_i:

$$\arg\min_{R,T} \sum_{i=1}^{J} \| H_i - (RH_i^{'} + T) \|^2$$

.

We applied the offloading strategies to the hand joint estimation, reducing computational time. In order to streamline the search for the optimal strategy, we divided the code into distinct segments and conducted a sequential search to identify the configuration with the shortest execution time. Through experimentation, the optimum distribution of the tasks was found by running the RGB-D frame capturing, bounding box proposing, and point cloud converting on the operator end and the HandVoteNet posture optimisation executing on the MEC server. This strategy reduces the need to transmit large data volumes over wireless communication channels and simultaneously alleviates the burden of intensive GPU computations on the operator's device.

After receiving the transformation matrix, the desired robot joint values are calculated at the robot end by solving the differential kinematics model to find a local solution. Compared to the global inverse kinematics solution, the Jacobian-based method is more suitable for tracking a target in cartesian space. Therefore, the desired joint value is applied to the robotic arm by an impedance controller, while the human finger movements are analyzed for controlling the gripper.

III. Experiment and Results

The HandVoteNet was evaluated on the MSRA dataset. In the ablation experiment, we compared the basic pointnet model with the MSG and SSG models. The experiments were run on an edge server containing Nvidia Quadro RTX 6000 GPUs, with PyTorch framework and Open3D geometry tools, which achieved a 7.8mm mean average error, as opposed to the baseline error of 8.5mm. The individual results and their comparison are presented in Fig.3. The estimation has been improved significantly by the MSG refining model, especially for fingertips. The MEC offloading and 5G wireless communication provide high-quality data transmission from the user to the robot end with a total 79.4ms latency, measured by a round trip time experiment.

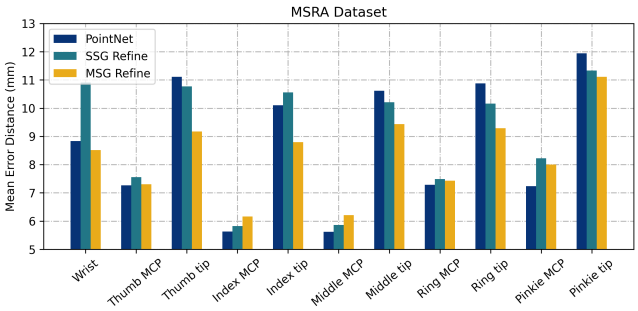

Fig. 3. Result from the MSRA experiments.

To evaluate the working of the proposed teleoperation system, shown in Fig. 4, three operators tested the developed system for four tasks completion: 1) Pick and Place, 2) Human handover, 3) Candy Pouring, and 4) Cube stacking, achieving 77.27%, 86.67%, 67.86% and 89.47% success rate respectively and a total success rate of 80.87%.

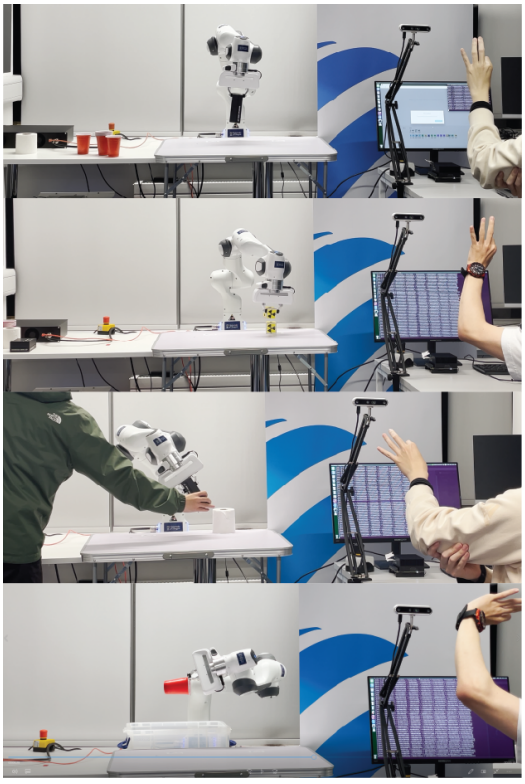

Fig. 4. From top to bottom are four tasks to evaluate the teleoperation system: The boxes picking and placing task, the cube stacking task, the various angles human handover task, and the candy pouring task.

REFERENCES

[1] G. T. Sung and I. S. Gill, "Robotic laparoscopic surgery: a comparison of the da vinci and zeus systems," *Urology*, vol. 58, no. 6, pp. 893–898, 2001.

[2] D. Zhang, Z. Wu, J. Chen, R. Zhu, A. Munawar, B. Xiao, Y. Guan, H. Su, W. Hong, Y. Guo, G. S. Fischer, B. Lo, and G.-Z. Yang, "Human-robot shared control for surgical robot based on context-aware sim-to-real adaptation," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 7694–7700.

[3] G. Hirzinger, B. Brunner, J. Dietrich, and J. Heindl, "Rotex-the first remotely controlled robot in space," in *Proceedings of the 1994 IEEE international conference on robotics and automation(ICRA)*. IEEE, 1994, pp. 2604–2611.

[4] C. G. Atkeson, B. P. W. Babu, N. Banerjee, D. Berenson, C. P. Bove, X. Cui, M. DeDonato, R. Du, S. Feng, P. Franklin, *et al.*, "No falls, no resets: Reliable humanoid behavior in the darpa robotics challenge," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 623–630.

[5] D. P. Losey, K. Srinivasan, A. Mandlekar, A. Garg, and D. Sadigh, "Controlling assistive robots with learned latent actions," 2020.

[6] W. Pryor, B. P. Vagvolgyi, A. Deguet, S. Leonard, L. L. Whitcomb, and P. Kazanzides, "Interactive planning and supervised execution for high-risk, high-latency teleoperation." Institute of Electrical and Electronics Engineers Inc., 10 2020, pp. 1857–1864.

[7] S. S. White, K. W. Bisland, M. C. Collins, and Z. Li, "Design of a high-level teleoperation interface resilient to the effects of unreliable robot autonomy." Institute of Electrical and Electronics Engineers Inc., 10 2020, pp. 11 519–11 524.

[8] S. Gholami, F. Tassi, E. D. Momi, and A. Ajoudani, "A reconfigurable interface for ergonomic and dynamic tele-locomanipulation." Institute of Electrical and Electronics Engineers Inc., 2021, pp. 4260–4267.

[9] M. K. Zein, A. Sidaoui, D. Asmar, and I. H. Elhajj, "Enhanced teleoperation using autocomplete," 2020.

[10] C. Pohl, K. Hitzler, R. Grimm, A. Zea, U. D. Hanebeck, and T. Asfour, "Affordance-based grasping and manipulation in real world applications." Institute of Electrical and Electronics Engineers Inc., 10 2020, pp. 9569–9576.

[11] M. Risiglione, J. P. Sleiman, M. V. Minniti, B. Cizmeci, D. Dresscher, and M. Hutter, "Passivity-based control for haptic teleoperation of a legged manipulator in presence of time-delays." Institute of Electrical and Electronics Engineers Inc., 2021, pp. 5276–5281.

[12] M. Wonsick, T. Kelestemur, S. Alt, and T. Padir, "Telemanipulation via virtual reality interfaces with enhanced environment models." Institute of Electrical and Electronics Engineers Inc., 2021, pp. 2999–3004.

[13] L. S. Yim, Q. T. Vo, C.-I. Huang, C.-R. Wang, W. McQueary, H.-C. Wang, H. Huang, and L.-F. Yu, "Wfh-vr: Teleoperating a robot arm to set a dining table across the globe via virtual reality." IEEE, 10 2022, pp. 4927–4934. [Online]. Available: https://ieeexplore.ieee.org/document/9981729/

[14] L. Chen, A. Swikir, and S. Haddadin, "Drawing elon musk: A robot avatar for remote manipulation." Institute of Electrical and Electronics Engineers Inc., 2021, pp. 4244–4251.

[15] L. S. Uiterkamp, F. Porcini, G. Englebienne, A. Frisoli, and D. Dresscher, "Emg-based feedback modulation for increased transparency in teleoperation." IEEE, 10 2022, pp. 599–604. [Online]. Available: https://ieeexplore.ieee.org/document/9981162/

[16] A. Padmanabha, Q. Wang, D. Han, J. Diyora, K. Kacker, H. Khalid, L.-J. Chen, C. Majidi, and Z. Erickson, "Hat: Head-worn assistive teleoperation of mobile manipulators." IEEE, 5 2023, pp. 12 542–12 548. [Online]. Available: https://ieeexplore.ieee.org/document/10160431/

[17] S. Li, J. Jiang, P. Ruppel, H. Liang, X. Ma, N. Hendrich, F. Sun, and J. Zhang, "A mobile robot hand-arm teleoperation system by vision and imu." Institute of Electrical and Electronics Engineers Inc., 10 2020, pp. 10 900–10 906.

[18] B. Xie, M. Han, J. Jin, M. Barczyk, and M. Jägersand, "A generative model-based predictive display for robotic teleoperation," vol. 2021-May. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 8551–8557.

[19] R. Chandra, V. H. Giraud, M. Alkhatib, and Y. Mezouar, "Dual quaternion based dynamic movement primitives to learn industrial tasks using teleoperation." IEEE, 5 2023, pp. 3757–3763. [Online]. Available: https://ieeexplore.ieee.org/document/10160970/

[20] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9164–9170.

[21] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8417–8426.

[22] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017, pp. 652–660.

[23] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[24] B. Kehoe, A. Matsukawa, S. Candido, J. Kuffner, and K. Goldberg, "Cloud-based robot grasping with the google object recognition engine," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 4263–4270.

[25] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1957–1964.

[26] F. Hu, Y. Deng, H. Zhou, T. H. Jung, C.-B. Chae, and A. H. Aghvami, "A vision of an xr-aided teleoperation system toward 5g/b5g," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 34–40, 2021.

[27] H. Zhu, M. Sharma, K. Pfeiffer, M. Mezzavilla, J. Shen, S. Rangan, and L. Righetti, "Enabling remote whole-body control with 5g edge computing," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 3553–3560.

[28] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, vol. PAMI-9, pp. 698–700, 1987.