Intention is what you need to estimate: Attention-driven prediction of goal pose in a human-centric telemanipulation of a robotic hand

Muneeb Ahmed [*1] **Rajesh Kumar** [*2] **Arzad Alam Kherani** [3] **Brejesh Lall** [1]

Abstract

This work entails remote telemanipulation of certain objects using Dexmo Haptic glove (DHG) and Allegro Robotic Hand (ARH). We introduce an estimation mechanism to quantify the expected goal pose of fingers of the human user, wearing the DHG, as its intent, defined in terms of the expected rotation angle of the object (about the viewing plane) that is held between the end-effectors of ARH. A significant amount of delay is observed to generate this intent due to communication and control latencies when the robot is remotely controlled. Hence, an attention based mechanism is leveraged to model the trajectory of estimated intent and predict its estimate for a lookahead of m time units from the current n^{th} estimated sample to compensate for the delays. We evaluate the performances of the estimation mechanism, and the attention mechanism on the stated robotic setup in a real-work networking scenario against some benchmark methodologies. The effect of varying lookahead is analysed against the accuracy of estimation/prediction of the intent. The testing MSE achieved in prediction of the human intent (utilizing attention model) is reported to be 0.00047 for m=1, which characterizes as $\sim 38-42$ times lesser in comparison to our previous work (utilizing LSTM).

1. Introduction

The study of teleoperated robotic systems has been witnessed in literature for numerous decades (Hokayem & Spong, 2006; D'Ettorre et al., 2021; Chao et al., 2021). These studies have resulted in significant advancements in several fields, including tele-medicine (Prokhorenko et al., 2023; Scimeca et al., 2022; Yang et al., 2020), virtual-reality /augmented-reality (Gamelin et al., 2021), precise automation (Luo et al., 2023; Zhu et al., 2022), and industrial applications, (Girbes-Juan et al., 2020; Doolani et al., 2020). Achieving accurate control of a robot situated at a remote location necessitates the conversion of motion signals emanating from the human operator onto the robot, accompanied by the provision of feedback to the human controller for its assistive or corrective followup. The repetitive cycle of human-centric control and coordination enables a robot to perform in-hand manipulation of objects with a significant accuracy. The manipulation of the object through movements across the end-effectors of the robot requires precise control of both the motion of the end-effectors and the applied force to avoid any potential undesired outcomes. However, such a move-and-wait paradigm entails communicational and control delays in conjunction to human-reaction time. Hence, it is desirable to predict the desired set of actions of the human controller ahead of time to compensate for the associated delays. Secondly, the transformation of the high-dimensional joint motion signals from the human fingers onto the robotic hand is not trivial. It requires an estimation mechanism to represent the input signals in terms of the desired action that the human intends to perform.

This work utilizes Dexmo Haptic Glove (DHG)-driven control of a remotely placed robotic hand (ARH). The control signals perceived from the fingers of the human user are captured by the DHG, and the corresponding joint motion signals from the DHG are transformed into reliable control signals for the ARH. The system is complex due to the dissimilarity in the kinematics of the robotic hand, the exoskeleton glove, and the human hand. In contrast to perceive control information from 19 odd joints in the human hand, the DHG represents such information across its 11-degrees of freedom which is to be realized by the ARH in 16-degrees of freedom. Furthermore, the limitations imposed on the torque that is perceived by the DHG, restricts the movement across all the joints. Hence, the DHG under-represents the perceived information from the human hand. In this premise, the contributions of this work are listed as follows:

- A *human-intent* template is characterized in terms of

[*]Equal contribution [1]Indian Institute of Technology Delhi, New Delhi, India [2]Addverb Robotics, Noida, India [3]Indian Institute of Technology Bhilai, Raipur, India. Correspondence to: Muneeb Ahmed <muneeb.ahmed@dbst.iitd.ac.in>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

desired goal configuration of the object in hand based on the underrepresented motion signals captured by the DHG.

- The estimated intent template is approximated using an attention-based (Vaswani et al., 2017) neural network for predicting its futuristic estimate to compensate for delays occurring in the move-and-wait model.
- A control algorithm/transformation scheme is provided for the ARH to undergo actuation towards the desired goal pose in coherence to the estimated/predicted intent.
- We analyse the performance of the attention-based (Vaswani et al., 2017) prediction with respect to varying lookahead of the prediction window.

2. Methodology

An end-to-end workflow of the proposed system towards estimation/prediction of human intent and subsequent control architecture is illustrated in Fig. 1.

2.1. Control Methodology for transformation of pose matrix to the joint rate of robotic hand

The scenario considered in this study is to undergo rotation of the object within the grasp of the ARH as a baseline in-hand manipulation task with the premise that all other in-hand manipulation tasks are a repeated set of these rotation operations. Consider 'a_g', 'a_r', and 'a_h' to denote the number of joints in the DHG, ARH and the human hand, respectively. For each serial chain s_i , $\mathbf{v}_i$ $(1 \leq i \leq s)$ represents the vector from the centroid of the object to the point of contact on its surface with the end-effectors of the ARH, and $\tilde{\mathbf{v}}_i$ denotes its corresponding cross product matrix. Similar to the task shown in Fig. 1, the current joint configuration vector of ARH, denoted as $\mathbf{q} \in \mathbb{R}^{a_r}$, and the desired joint configuration vector of the ARH to achieve the desired goal pose, denoted as $\mathbf{q}_g \in \mathbb{R}^{a_r}$, the torque applied ($\boldsymbol{\tau}$) at the end-effectors of ARH is given as $\boldsymbol{\tau} = \mathbf{A_M}\ddot{\mathbf{q}}_\mathbf{g} + \mathbf{B}(\dot{\mathbf{q}}_\mathbf{g} - \dot{\mathbf{q}}) + \mathbf{C}(\mathbf{q_g}, \dot{\mathbf{q}}_\mathbf{g}) + \mathbf{D_g}(\mathbf{q_g}) + \mathbf{K}(\mathbf{q_g} - \mathbf{q})$, where $\mathbf{A_M}$ denotes the mass matrix of ARH, $\mathbf{K}$, $\mathbf{B}$ represent the gain matrices (Della Santina et al., 2020), and $\mathbf{D_g}(.)$ represents the gravity compensation term. It is known that $\mathbf{SE}(3)$ characterizes the current pose and the desired goal pose configuration of the held object. Hence, a pose matrix defining the state of the object at time 'n' can be represented as $\mathbf{P}_n = \begin{bmatrix} \mathcal{R}_n^T & \mathbf{0} \\ \mathbf{r}_n^T & 1 \end{bmatrix}^T$, where $\mathcal{R}_n \in \mathbb{R}^{3\times3}$, $|\mathcal{R}_n| = 1$, and ${\mathcal{R}_n}^T\mathcal{R}_n = \mathbf{I}_{3\times3}$, and $\mathbf{r}_n \in \mathbb{R}^3$. The trajectory of screw rotation, at the object level, denoted by $\mathbf{F_o}$, defines the transition of the object from its current position to the intended goal position. The twist of the end-effectors of the ARH, is represented as $\mathbf{F_{RH}} = \mathbf{J_R}\dot{\mathbf{q}}_\mathbf{g}$, where $\mathbf{J_R}$ denotes the Jacobian Matrices that establish the relationship of joint rates ($\dot{\mathbf{q}}_\mathbf{g}$) of the active joints in ARH to the observable twist in the end-effectors of ARH. Also, $\mathbf{F_{RH}} = \mathbf{J_o}\mathbf{F_o}$, where $\mathbf{F_o}$ denotes the twist in the object, and Jacobian matrices ($\mathbf{J_o}$) relate the twist in the object to the observable twist ($\mathbf{F_{RH}}$) in the ARH's end-effectors. Hence it is collated that,

$$\dot{\mathbf{q}}_\mathbf{g} = \mathbf{J_R^+}\mathbf{F_{RH}} \quad \text{and} \quad \mathbf{q_g} = \mathbf{q} + \lambda\dot{\mathbf{q}}_\mathbf{g}\Delta t \tag{1}$$

where, $\lambda \in \mathbb{R}$ controls the incremented value to the current joint configuration towards the desired goal pose in the time interval (Δt). A separate vision-based subsystem calculates the object's current pose matrix ($\mathbf{P}_c$) utilizing a position sensitive marker (such as ArUco). The pose of the object is calculated relative to the pose of the ARH. Once this subsystem is calibrated offline, the objects pose is determined by segmenting the marker on the object and calculating its relative angle with respect to the ARH, in real-time.

2.2. Estimating human intent template

We leverage superposition of the twist observed by the end-effectors of DHG onto the twist observed by the end-effectors of the ARH which is based on the premise of similar tree-type structure of their respective serial chains. The twist across the end effectors of DHG is formulated as,

$$\mathbf{F_{DHG}} = \mathbf{J_E}\dot{\mathbf{q}}' \tag{2}$$

$\mathbf{J_E}$ represents the Jacobian Matrices that establish the relationship of joint rates ($\dot{\mathbf{q}}'$) of the active joints of DHG to the twist ($\mathbf{F_{DHG}}$) of its end-effectors. Also, the vector $\mathbf{F_o} = [\dot{\mathbf{x}}, \dot{\boldsymbol{\psi}}]$ comprising the object's linear velocity ($\dot{\mathbf{x}} \in \mathbb{R}^3$) and angular velocity ($\dot{\boldsymbol{\psi}} \in \mathbb{R}^3$) of the object. Similarly, $\mathbf{F_{DHG}} = \mathbf{J_o}\mathbf{F_o}$, where $\mathbf{F_o}$ denotes the twist in the object, and Jacobian matrices ($\mathbf{J_o}$) relate the velocities in the object to the twist in the end-effectors of DHG. It is collated that $\mathbf{F_o} = {\mathbf{J_o}}^+\mathbf{F_{DHG}}$. Hence, the displacement ($\mathbf{x} \in \mathbb{R}^3$) and the angular displacement ($\boldsymbol{\psi} \in \mathbb{R}^3$) in the object upto time 't' can be determined as $\mathbf{x} = \int_0^t \dot{\mathbf{x}}dt$ and $\boldsymbol{\psi} = \int_0^t \dot{\boldsymbol{\psi}}dt$. Since, the contemplated intent is a rotation, only the angular velocity terms ($\boldsymbol{\psi} = \begin{bmatrix}\psi_x & \psi_y & \psi_z\end{bmatrix}^T$) contribute to the contemplated intent of the human. The estimated intent captured in the form of angular displacement, is further processed by a recurrent neural network (discussed in the next section) that yields a predicted estimate of the intent ($\hat{\boldsymbol{\psi}} = \begin{bmatrix}\hat{\psi}_x & \hat{\psi}_y & \hat{\psi}_z\end{bmatrix}^T$). Finally, the pose matrix for the predicted intent is formulated as, $\mathbf{P}_o = \begin{bmatrix} \mathcal{R}^T & \mathbf{0} \\ \mathbf{x}^T & 1 \end{bmatrix}^T$, where $\mathcal{R} = \mathbf{E}^2(1 - \cos(\|\hat{\boldsymbol{\psi}}\|_2) + \mathbf{E}(\sin(\|\hat{\boldsymbol{\psi}}\|_2) + \mathbf{I}_{3\times3}$, and $\mathbf{E} = \begin{bmatrix} 0 & \hat{\psi}_z & -\hat{\psi}_y \\ -\hat{\psi}_z & 0 & \hat{\psi}_x \\ \hat{\psi}_y & -\hat{\psi}_x & 0 \end{bmatrix}^T$.

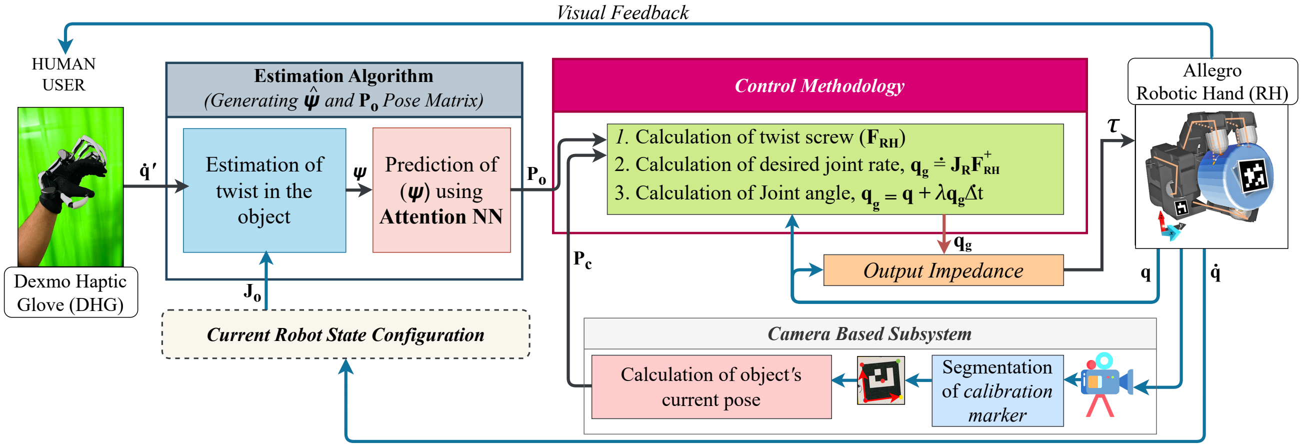
Visual Feedback
HUMAN USER
Dexmo Haptic Glove (DHG)
$\dot{\mathbf{q}}'$
Estimation Algorithm
(Generating $\hat{\boldsymbol{\psi}}$ and $\mathbf{P_o}$ Pose Matrix)
Estimation of twist in the object
$\boldsymbol{\psi}$
Prediction of ($\boldsymbol{\psi}$) using Attention NN
$\mathbf{J_o}$
Current Robot State Configuration
$\mathbf{P_o}$
Control Methodology
1. Calculation of twist screw ($\mathbf{F_{RH}}$)
2. Calculation of desired joint rate, $\mathbf{q_g} \doteq \mathbf{J_R F^{+}_{RH}}$
3. Calculation of Joint angle, $\mathbf{q_g} = \mathbf{q} + \lambda \mathbf{q_g} \Delta t$
$\mathbf{q_g}$
Output Impedance
$\mathbf{P_c}$
τ
Allegro Robotic Hand (RH)
$\mathbf{q}$
$\dot{\mathbf{q}}$
Camera Based Subsystem
Calculation of object's current pose
Segmentation of calibration marker

Figure 1. The proposed system workflow for estimation & the prediction of human intent with corresponding control setup.

2.3. Prediction of Intended Goal pose using Attention-based neural network

To pursue mitigation of the effect of these delays, an attention-based prediction strategy is introduced in the estimation algorithm, that is effective to preserve the temporal features in the human behavior. Our proposal is based in the observation that humans frequently employ prehensile manipulation in tasks that involve repetitive actions, necessitating the utilization of sequence learning techniques. The architectural specifications of the attention model are illustrated in Fig. 2. The encoder examines a sequence of preceding $r = 20$ samples of $\boldsymbol{\psi}$ in order to generate an approximation $\hat{\boldsymbol{\psi}}$ of the desired outcome. Consider $\boldsymbol{\mathcal{A}}_n \in \mathbb{R}^r$ a sequence of previous integral solutions of angular displacement (about certain axis) at any arbitrary time $n \geq r$, as $\boldsymbol{\mathcal{A}}_n = \{\boldsymbol{\psi}[n-r+1], \boldsymbol{\psi}[n-r], \ldots, \boldsymbol{\psi}[n]\}$. This can be considered for all axes of rotation (depending upon the need). A sequence of $\boldsymbol{\mathcal{A}}_n$ is given as input to the encoder block. Here, we do not explicitly add any positional embedding due to the afixed serialized nature of the input. However, convolutional layers are added after the attention block to capture spatial variance. The kernels tend to extract the variance in across the subsamples in its input (Ahmed et al., 2021). Since the input is a 1-dimensional sequence, the spatial variance captured correlates to temporal variance. Hence, some temporal characteristics could be preserved. The encoder block is cascaded 4 times. The output is processed by two cascaded fully connected layers. The output is a vector $\hat{\boldsymbol{\psi}} \in \mathbb{R}^m$, $m \in \mathbb{Z}^+$. It is seen in the results later that an ablation study is sought to analyse the performance of the discussed mechanism by predicting $n + m$ futuristic values of the input.

3. Experimental Results

(Dataset). The motion signals generated from DHG are 11-dimensional in contrast to 16-dimensional joint angle configuration of the ARH. These data from these signals along with the estimated representation of the intent (in terms of the angle observed by the object being manipulated) are curated to form a dataset. Being high-dimensional representations, there is no direct mapping of the signals. Fig. 3 illustrates a snapshot of the joint motion signals observed in the dataset during a random manipulation of an object.

(Prediction of Intent). The hyperparameters set during training the attention-based prediction network are illustrated in Table 1. The performance of the proposed system is illustrated in Fig.4.

Table 1. Hyperparameters space of training the prediction network.

Hyperparameter	Space
Learning Rate	**0.0001**,
Optimizer	Adam
Loss	MSE (Mean Squared Error)
Epochs	200, **200**, 500
Batch-size	8, **16**

4. Discussion

(Ablation Study). An ablation study was considered to vary the length of the output vector ($\hat{\boldsymbol{\psi}} \in \mathbb{R}^m$, $1 \leq m \leq 20$). The number of nodes in the second fully connected layer are equal to m. The pose matrix $\mathbf{P_o}$ is framed using the value corresponding to the m^{th} element of the predicted output vector. The effect of changing the dimensions of the output is observed empirically. Variation of the predicted

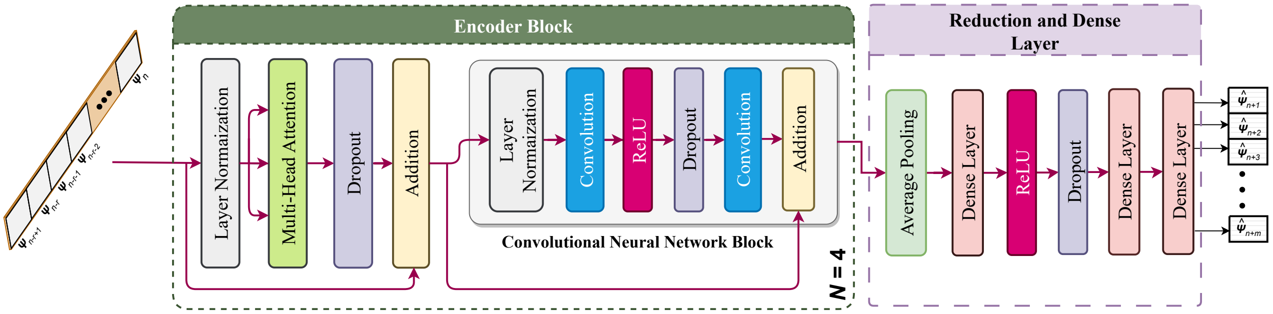

Figure 2. Proposed attention-based neural network based architecture for intent prediction.

value ($\hat{\psi}(n)$) from the ground truth value ($\psi(n)$ about a certain axis (y) at time n is quantified as MSE, given as $\frac{1}{d}\sum_{k=1}^{d}(\hat{\psi}_y(k) - \psi_y(k))^2$, where d is the number of samples in the dataset. It is observed that the mean squared error, across training, validation and test samples, increases with the increasing value of lookahead (as illustrated in Fig. 6). The test MSE is reported to be 0.00047, 0.00047, 0.0013, 0.0007, 0.0016, 0.00415, 0.0025, 0.00346, 0.0026, 0.0040, 0.0138, 0.0184, for lookahead = 1,2,3,4,5,6,7,8,9,10,15,20, respectively. In comparison to our previous work (Kumar et al., 2021) utilizing LSTM, the test MSE was reported to be 0.018, 0.06, 0.12, 0.11, 0.26, 0.42, 0.30, 0.54, 0.71, 0.40, 0.68, 0.78, for lookahead =1,2,3,4,5,6,7,8,9,10,15,20, respectively. As seen from these results, the proposed attention model surpasses the performance of the LSTM model in the same scenario by having about about $\sim 38 - 42$ times lesser error in prediction.

(Profiling of delays). It is experimentally observed motion in the object at the ARH lags in time with respect to the motion at the DHG owing to the delay in processing, control, and communication. However, by the introduction of the prediction network, the effect of delays is mitigated. There exists a trade-off between the value of m and the error observed. The integral value of m is reported to compensate for equal number of round-trip delay components, albeit with a subsequent increase in error as the value of m increases. The system is tested on a real-world 4G-network with an average round-trip latency of ~ 70 ms (over Robotic Operating System - ROS setup). The channel latency delay is dependent on the network dynamics (which is beyond the scope of this study). However, it is desirable to know the mitigation influence of the prediction mechanism in compensating for the delays observed. Hence, the proposed system consumes a total time of $\sim 11.1 - 32.25$ ms to and compensate the effect of delay of approximately $76.25 - 100$ ms (without the human reaction time), which would otherwise occur in singular round trip of control-feedback signals.

(Observation of Human Behavior). We present the observed understanding that when manipulating objects by fingertips, humans intent tends to exhibit a sigmoidal behavior (as illustrated in Fig.6). This behavior entails an initial phase of acceleration accompanied by marginal displacement of the object, followed by a prolonged period of substantial displacement, and subsequently a phase of small motion and significant deceleration. The cumulative effect of the human intention, in relation to the goal pose, becomes apparent only once the deceleration phase commence, which cause an apparent lag in the motion of the ARH with respect to the DHG in addition to the communication latency.

(Generalizability, Extension and Comparison). In this work, the intent is quantified in terms of the desired goal pose configuration undergoing rotation within the grip of ARH that is controlled by DHG. All other motions (such as, in-hand translation) could be modelled as a composition of these rotations. Since, the estimation mechanism is dependent on the forwards kinematics of the robot, the proposed methodology is robust to variation in object's shape. The generalizability across human subjects is inherent to the design of DHG. We compare our proposed methodology with benchmark work (McGhan et al., 2015) in the literature, where a vision dataset of 11200 samples is modelled using Markov Decision Processes, to classify the data across 8 different motion types with a true positive rate of 93.5% in ~ 625 ms. While as, the proposed approach in our work is generalized as the predicted output is a continuous value instead of categories. It predicts/estimates the intent of motion with an MSE of 0.0018 within $11.1 - 32.5$ ms time-frame.

5. Conclusion

This paper presents an initial advancement in the towards control of a robotic hand by a differently structured exoskeleton glove. Firstly, a methodology is proposed to model the underrepresented actuation of the human fingers via DHG in terms of the intended motion of the object. This is augmented by an attention-driven neural network that

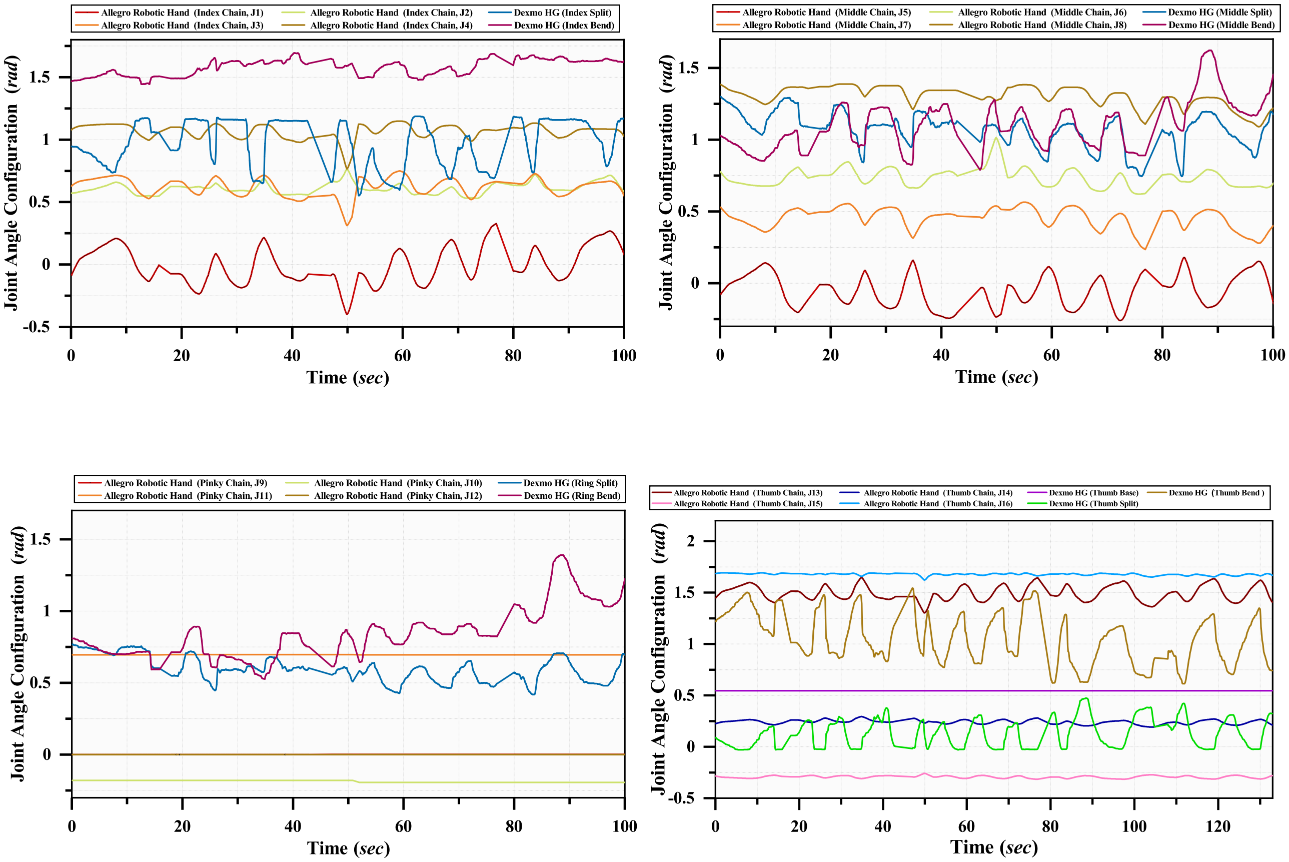

Figure 3. A snapshot of the Joint Motion signal in the dataset during arbitrary motion of an object signifying a complex relationship of the mapping algorithm.

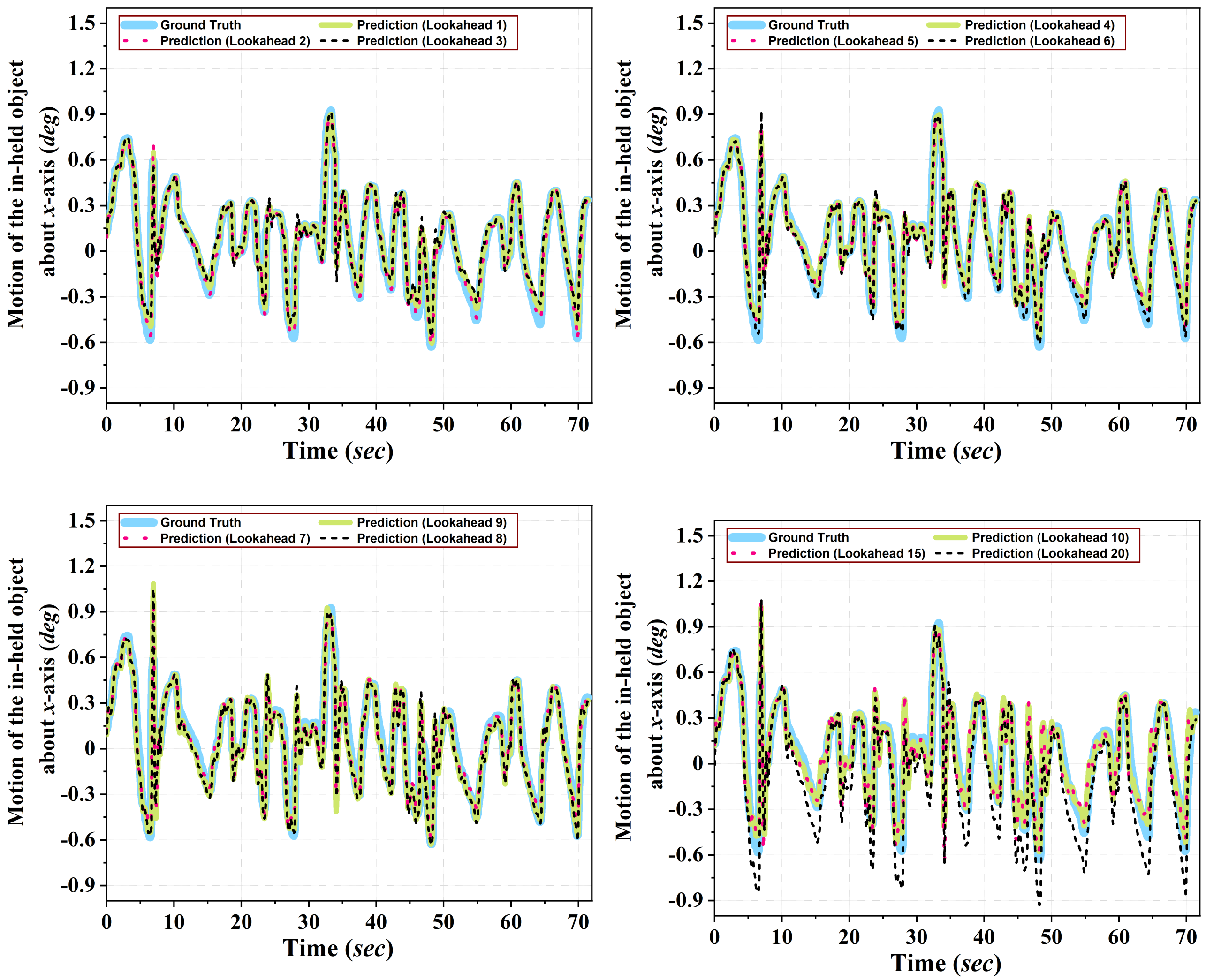

Figure 4. Error with respect to lookahead (m).

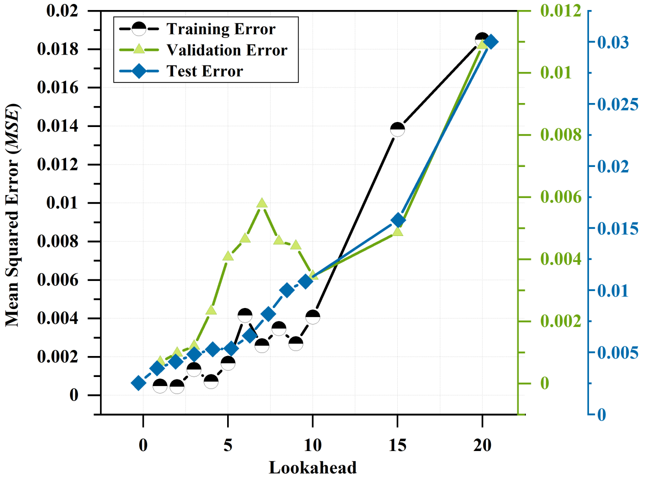

Figure 5. Performance graph showing Training, Test and Validation MSE with varying lookahead (m).

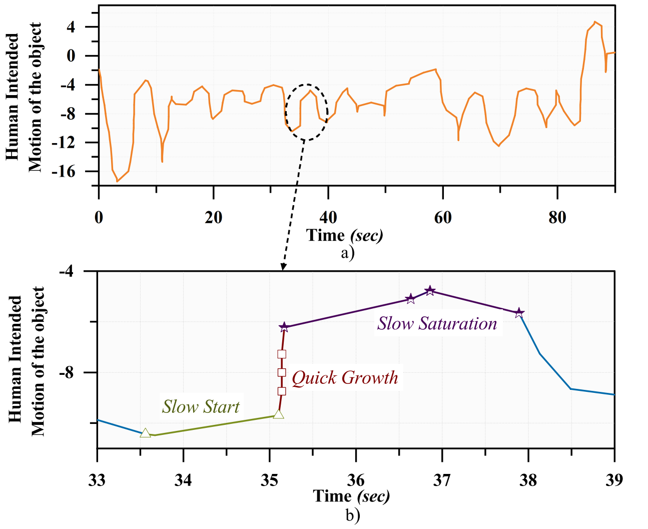

Figure 6. (a) Exemplar human intended motion of the object (b) Detailed analyses of a single wavelet.

predicts the motion of the object to compensate for processing/communication delays. Subsequently, a reliable control algorithm is introduced for the ARH enabling it to manipulate an object and rotate it to a desired pose. Experiments are carried out to analyse the trade-off between the accuracy of the proposed methodology in predicting the human intent by varying the span of predicted values. The challenges related to mitigation of delays, patterns in human behaviour, and generalizability are discussed. The code and additional results are available **here**.

References

Ahmed, M., Masood, S., Ahmad, M., and Abd El-Latif, A. A. Intelligent driver drowsiness detection for traffic safety based on multi cnn deep model and facial subsampling. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19743–19752, 2021.

Chao, Y.-W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y. S., Van Wyk, K., Iqbal, U., Birchfield, S., et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9044–9053, 2021.

Della Santina, C., Catalano, M. G., Bicchi, A., Ang, M., Khatib, O., and Siciliano, B. Soft robots. *Encyclopedia of Robotics*, 489, 2020.

Doolani, S., Wessels, C., Kanal, V., Sevastopoulos, C., Jaiswal, A., Nambiappan, H., and Makedon, F. A review of extended reality (xr) technologies for manufacturing training. *Technologies*, 8(4):77, 2020.

D'Ettorre, C., Mariani, A., Stilli, A., y Baena, F. R., Valdastri, P., Deguet, A., Kazanzides, P., Taylor, R. H., Fischer, G. S., DiMaio, S. P., et al. Accelerating surgical robotics research: A review of 10 years with the da vinci research kit. *IEEE Robotics & Automation Magazine*, 28(4):56–78, 2021.

Gamelin, G., Chellali, A., Cheikh, S., Ricca, A., Dumas, C., and Otmane, S. Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments. *Personal and Ubiquitous Computing*, 25: 467–484, 2021.

Girbes-Juan, V., Schettino, V., Demiris, Y., and Tornero, J. Haptic and visual feedback assistance for dual-arm robot teleoperation in surface conditioning tasks. *IEEE Transactions on Haptics*, 14(1):44–56, 2020.

Hokayem, P. F. and Spong, M. W. Bilateral teleoperation: An historical survey. *Automatica*, 42(12):2035–2057, 2006.

Kumar, R., Gandotra, P., Lall, B., Kherani, A. A., and Mukherjee, S. Estimation and prediction of deterministic human intent signal to augment haptic glove aided control of robotic hand. *CoRR*, abs/2110.07953, 2021. URL `https://arxiv.org/abs/2110.07953`.

Luo, J., Liu, W., Qi, W., Hu, J., Chen, J., and Yang, C. A vision-based virtual fixture with robot learning for teleoperation. *Robotics and Autonomous Systems*, 164:104414, 2023.

McGhan, C. L., Nasir, A., and Atkins, E. M. Human intent prediction using markov decision processes. *Journal of Aerospace Information Systems*, 12(5):393–397, 2015.

Prokhorenko, L., Klimov, D., Mishchenkov, D., and Poduraev, Y. Modular robot interface for a smart operating theater. *Journal of Robotic Surgery*, pp. 1–13, 2023.

Scimeca, L., Hughes, J., Maiolino, P., He, L., Nanayakkara, T., and Iida, F. Action augmentation of tactile perception for soft-body palpation. *Soft robotics*, 9(2):280–292, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yang, G., Lv, H., Zhang, Z., Yang, L., Deng, J., You, S., Du, J., and Yang, H. Keep healthcare workers safe: application of teleoperated robot in isolation ward for covid-19 prevention and control. *Chinese Journal of Mechanical Engineering*, 33(1):1–4, 2020.

Zhu, C., Yang, C., Jiang, Y., and Zhang, H. Fixed-time fuzzy control of uncertain robots with guaranteed transient performance. *IEEE Transactions on Fuzzy Systems*, 2022.