Towards estimation of human intent in assistive robotic teleoperation using kinaesthetic and visual feedback

Muneeb Ahmed
Indian Institute of Technology Delhi
New Delhi - 110016
muneeb.ahmed@dbst.iitd.ac.in

Brejesh Lall
Indian Institute of Technology Delhi
New Delhi - 110016
brejesh@ee.iitd.ac.in

Rajesh Kumar
Addverb Technologies
Noida - 201310
rajesh.kumar01@addverb.com

Arzad A. Kherani
Indian Institute of Technology Bhilai
Raipur - 492015
arzad.alam@iitbhilai.ac.in

Abstract

The ability to predict human intent in manipulating in-hand objects is a crucial aspect of developing intelligent robotic systems that can effectively interact with and assist humans in various tasks. Due to the non-standardized nature of interfaces between different robots, it is non-trivial to establish a one-to-one mapping between the instructions provided by the human operator on to the robot, and vice-versa. Additionally, the round trip of information flow in move-and-wait teleoperation strategy for micro-instructions accumulates considerable delays in performing basic tasks, rendering the overall objective ineffective. In this context, predicting the human intent of teleoperation is a prospective strategy to mitigate the effect of such delays. This entails that a possible set of expected action(s) is to be represented ahead of time. In this study, we propose an ML-driven ensemble approach for estimating the goal pose configuration of an object of interest held within the end-effectors of a remotely connected robot using visual and kinematic measurements. We evaluate our proposed system to infer the intended action of a human operator in a real-world robotic setup involving a haptic glove and a dexterous robotic hand, on three different objects. The proposed methodology outperforms a benchmark model in literature utilizing 60 times lesser prediction time with substantially better performance. We provide a comparative analysis of intent prediction strategy using independent visual and kinaesthetic data and discuss its improvement when combining both the modalities.

1. Introduction

The study of teleoperated robotic systems has been witnessed in literature for numerous decades [6, 3, 1]. These studies have resulted in significant advancements in several fields, including tele-medicine [9, 10, 11], virtual-reality /augmented-reality [4], precise automation [7, 13], and industrial applications, [5, 2]. 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 Allegro 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 joint configuration for the ARH using a transformation/mapping mechanism (as shown in Figure 1). 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. Hence, the DHG under-represents the perceived information from the human hand. However, the magnitude of under-representation can vary across robots and utility. This work entails analysing the information retained for achieving successful teleoperation of a robot by introducing 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. Two modalities of measurements (visual and kinaesthetic) are observed in this study in achieving this objective.

Main Contributions. In this premise, the contributions of this work are listed as follows: 1) A neural network based mapping/transformation algorithm is synthesized for transforming the kinaesthetic data encoded at the joints of DHG onto the joints of the ARH towards a successful replication of actions from human to the robot. 2) We introduce a prediction mechanism to estimate the trajectory of the expected human intent in terms of the rotation of an object of interest within the grasp of the robot, using an attention-based convolutional encoder network on kinasesthetic and visual measurements discretely. 3) We introduce a stacking-based ensemble to predict the human intent using combined modalities of visual cues and kinaesthetic measurements, and provide a comparative analysis of the prediction mechanism.

1.1. Methodology

We define a scenario for performing in-hand operation of three real-world objects (shown in Figure 2) and predict the human intent template in terms of the angle of rotation of the object ahead of time to compensate for the delays that occur because of the communication/control latency in bilateral teleoperation. We leverage visual and kinaesthetic feedback to report their performance on the estimation/prediction mechanism.

1.2. Grasping/Mapping Algorithm

The input ($Q_{DHG} \in \mathcal{R}^{11}$) to the mapping mechanism (as shown in Figure 1) is scaled vector of 11-DOF (degrees of freedom) kinaesthetic measurements from the encoding joints of DHG represented as Q_{DHG}. The output is a scaled vector of 16-DOF joint configuration to be actuated by the ARH, represented as Q_{ARH}. The mapping mechanism demonstrated is a fully-connected neural network to perform the transformation, $\tau : \mathcal{R}^{11} \rightarrow \mathcal{R}^{16}$, where the intrinsic architectural structure of this model is shown in Figure 1. The motivation to leverage hidden layers each having k nodes, $k \in \mathcal{R}^{6}$, to represent the embeddings of kinaesthetic input is taken from a result from the following observation. It was observed that the PCA-decomposition of the encoded input, Q_{DHG} across the objects of interest in the study yields a knee-point of the magnitude of variance in principal components around 6^{th} principal component (as shown in Figure 3). This is analogous to the fact that the state configuration of unrestrained rigid bodies could be represented exactly in 6-DOF [12]. Taking this observation, it was seen that 6 dimensions can represent almost 98% of the variance in the principal components. Hence, leading to the choice of hidden layer dimensions. The data to train this network was curated in the form of a pairwise $Q_{DHG} \rightarrow Q_{ARH}$ joint state configurations upon manipulating the objects of interest (shown in Figure 2) and perform random joint angle configurations on both the devices simultaneously, yielding reliable pairs of ground truth for performing this mapping. Since the robots vary in parameters, various mapping schemes can emerge as a result. Here, neural network assumes a black-box approach of this mapping scheme, the significant results of which are shown in the Results section.

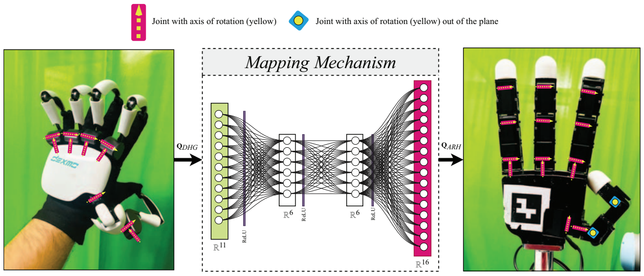

Figure 1. Showing the neural network based mapping mechanism for transforming joint configuration of DHG to ARH with respective axes of rotation.

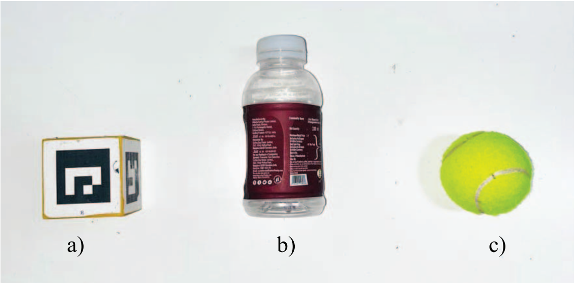

Figure 2. a) Cuboidal b) Cylindrical c) Spherical object used in the study.

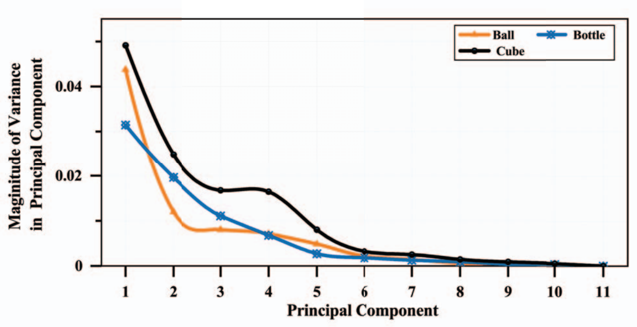

Figure 3. Magnitude of variance in principal components across dimensions of HG.

1.3. Mechanism for estimating and predicting human intent from visual cues

Based on the premise that the joint angle configurations can represent the overall angle of rotation of the in-held object at any arbitrary time n, this mechanism uses the relative position of the ArUco marker with respect to the ARH, as a determinant of the current state of the object being manipulated. There exist two such markers, one placed on the ARH and the other placed on the object of interest. The system is calibrated offline initially. Once calibration is complete, the expected angle of the object (about the viewing plane) is a representative of the intent of the human (defined by goal pose of the fingers). This is achieved by segmenting the ArUco marker on the object and calculating its angle with respect to the marker on the ARH, yielding the current pose of the object at time n, as ψ_n. Consider a sequence of previous r pose values, as $\mathbf{A_{vis}} = \{\psi_{n-r+1}, \psi_{n-r+2}, \dots, \psi_n\}$. This is passed to the transformer encoder block that yields a predicted estimate of the pose ($\hat{y}_{vis} = \hat{\psi}_{n+m}$) based on the distribution of trajectory $\mathbf{A_{vis}}$ with a lookahead of m units. A schematic diagram of this mechanism is shown in Figure 4. The architecture of the attention-based convolutional encoder (labelled as Transformer encoder) is illustrated in Figure 5.

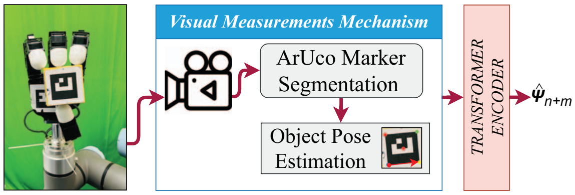

Figure 4. Mechanism of estimating human-intent in terms of angle of rotation of the object taken from visual cues.

1.4. Mechanism for estimating and predicting human intent from kinaesthetic measurements

The current pose of the object can also be estimated by the transformation of kinaesthetic measurements emerging from the joints of ARH. Due to the non-linear mapping of the pair $\mathbf{Q}_{ARH} \rightarrow \psi_n$, we leverage a neural network approach. The input to the fully connected neural network is 16-DOF vector, $\mathbf{Q}_{ARH}$, parsed by ReLU activation and mapped to the pose value ψ_n of the object at an arbitrary time n. The ground truth of the current pose of the object (while training this network) is fed using the visual measurements (same as defined in previous section). Upon generating a vector of r pose values, $\mathbf{A_{kin}} = \{\psi_{n-r+1}, \psi_{n-r+2}, \dots, \psi_n\}$, the transformer encoder predicts the m^{th} pose value of the object, $\hat{y}_{kin} = \hat{\psi}_{n+m}$, similar to the scheme defined in previous section.

1.5. Hybrid Model combining the information from visual and kinaesthetic measurements

A stacking based ensemble approach defined in Algorithm 1 to utilize the predictions generated from both modalities. The individual pre-trained modes are used to generate a vector of two units (signifying the individual prediction), as $[\hat{y}_{vis}, \hat{y}_{kin}$. Then, a single layer perceptron network is trained on this input as $Y = \alpha\hat{y}_{vis} + \beta\hat{y}_{kin}$, where α and β represent the weights of the ensemble model that are trained using a suitable learning algorithm. The Y value thus obtained is the predicted value of the current pose of the object when combining both the modalities.

2. Results and Discussion

2.1. Results from mapping mechanism

The results from the mapping mechanism has yielded significant accuracy in replicating the pose generated from DHG to ARH with a root mean squared error of 0.0872665-0.0933 radians across all joints. The results are shown in Figure 7.

2.2. Results from modelling visual cues independently

Here we discuss the results obtained from marker-driven visual cues, the angle of the ArUco marker is taken relative to the marker on the ARH. By noting the any two corner coordinates of the marker, the angle between the vectors obtained using the two points on each marker is synthesized into a sequence for learning the prediction. The initial lag of angle due to calibration is removed already. The transformer encoder that trains on such sequences, is set with the hyperparameters mentioned in Table 1. The training data is sequenced with $r = 10$ units (depicting the window size for input), and $m = 1$ (depicting the lookahead) with a stride

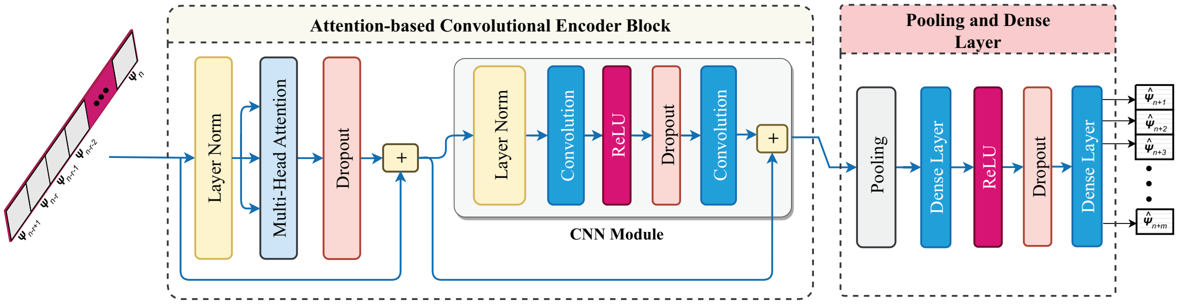

Figure 5. Transformer encoder for prediction of human intent from previous r pose values.

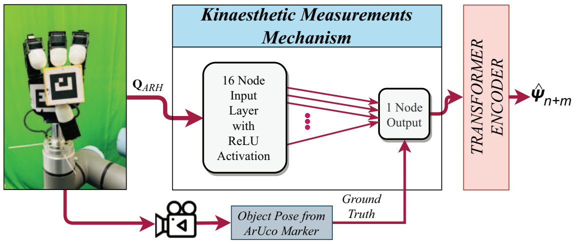
Kinaesthetic Measurements Mechanism
Q_{ARH}
16 Node Input Layer with ReLU Activation
1 Node Output
TRANSFORMER ENCODER
$\hat{\psi}_{n+m}$
Object Pose from ArUco Marker
Ground Truth

Figure 6. Kinaesthetic measurement mechanism.

Input: Joint state configuration of ARH (i.e.,$\mathbf{Q}_{ARH}$); Joint state configuration of DHG (i.e.,$\mathbf{Q}_{DHG}$)

Output: Predicted value of the human-intent Y representing the predicted angle of rotation of the object.

```
$n \leftarrow$ currentFrameNumber          ▷ a constant
$m \leftarrow 1$                            ▷ Lookahead

while True do
    $\hat{y}_{vis} =$ Vision_Measurements_Mechanism($\mathbf{Q}_{ARH}, n, m$)
    $\hat{y}_{kin} =$ Kinaesthetic_Measurements_Mechanism($\mathbf{Q}_{DHG}, n, m$)
    $Y = \alpha\hat{y}_{vis} + \beta\hat{y}_{kin}$
    $n \leftarrow n + 1$
    return $Y$
end
```

Algorithm 1: Stacking ensemble for generating prediction(s) of human intent using vision and kinaesthetic measurements.

of 1 using sliding window. The data is split into train, test and validation in the ratio of 0.8:0.1:0.1. Hence, using this visual input modality, we report the MSE training loss, validation loss, and test loss of 0.0026, 0.0021, respectively.

Hyperparameter	Space
Learning Rate	0.0001
Loss	MSE
Batch-size	16
Scaling	Min-Max Normalization
Epochs	200
Optimizer	Adam

Table 1. Hyperparameters set during training the transformer encoder model.

2.3. Results from modelling kinaesthetic cues independently

Here we discuss the effect on performance of training when using kinaesthetic data as input to the transformer encoder. The output from the kinaesthetic measurement mechanism is used to generate the current pose of the object, a sequence of which is parsed by the transformer encoder to generate a prediction with $m = 1$ lookahead. Similar to the process in previous experiment, the hyperparameters of the encoder network are set according to Table 1. The value of $r = 20$ was empirically chosen to generate the sequence at any given time n, using sliding window, with stride=1. Such a stride was chosen to reduce the information while training the network. The data is split into train, test and validation in the ratio of 0.8:0.1:0.1. The training loss, and validation loss, was observed as 0.0027, 0.001, respectively.

2.4. Results from hybrid mechanism

From the previous experiments we realize that the validation error in training the encoder on kinaesthetic inputs is lesser than using visual modality independently. This result is observed due to the variability of detecting the corners of ArUco which may be susceptible to noise/artifacts that arise in sensing the environment. It is also necessary to mention that kinaesthetic measurements are sensitive to physical conditions of the environment. Hence, in order to leverage the utility of both modalities, an ensemble was uti-

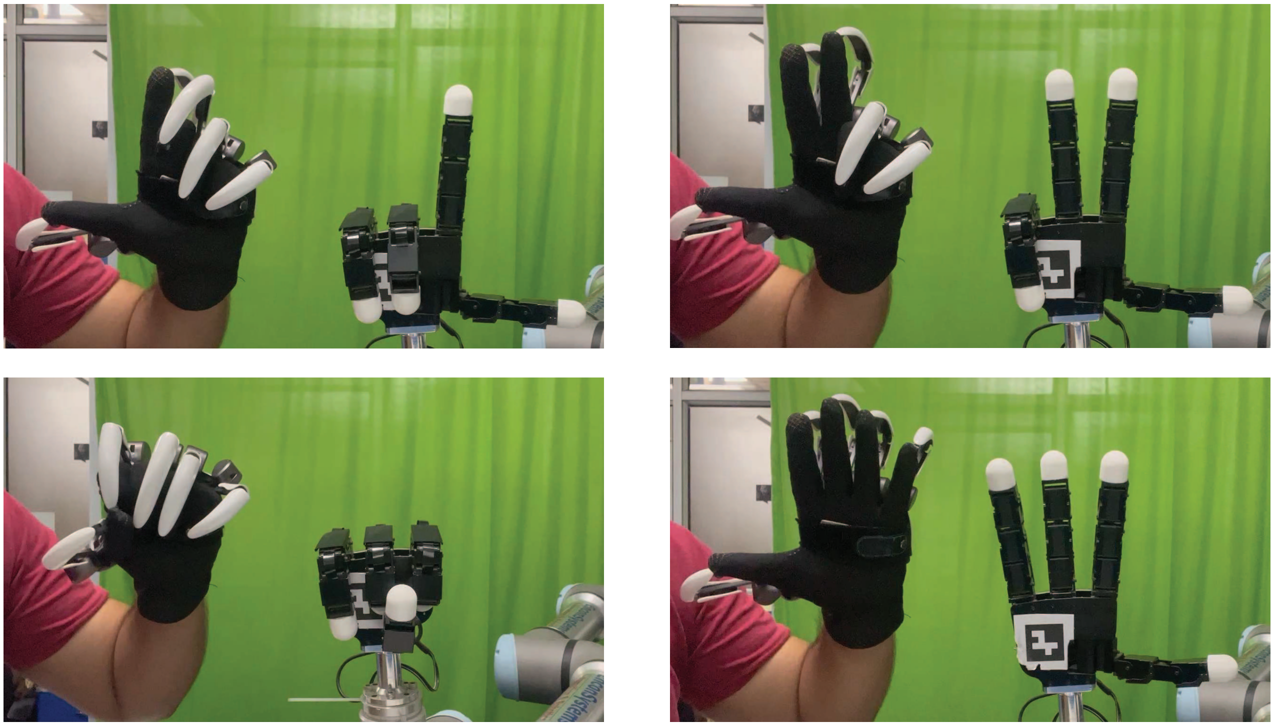

Figure 7. Visualization of mapping mechanism while performing different poses from DHG to ARH.

lized in this experiment to observe any prospective increase in performance. We report an improvement when combining the modalities of input. The trajectory of true and predicted output value of human intent is shown in Figure 8. Table 2 illustrates the test mean squared error (MSE) and test mean average error (MAE) on individual models and the improvement on the stacking model.

Model	Test MSE	Test MAE
Kinaesthetic	0.001619	0.0144
Visual	0.006	0.02
Ensemble Model	0.000937	0.008641

Table 2. Showing comparative analysis of the stacking ensemble, visual measurements model, and kinaesthetic measurements model trained from visual and kinaesthetic cues.

Analysis of delays. The system was experimentally tested on a real-world 4G network. It was 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, the introduction of a prediction network helps reduce the impact of these delays. There is a trade-off between the value of parameter m and the observed error. The value of m is used to compensate for an equal number of round-trip delay components, but this leads to an increase in error as the value of the lookahead window (m) increases. The latency of the channel depends on network dynamics, which is not the focus of this study. However, it is important to understand how the prediction mechanism helps mitigate the observed delays. Therefore, the proposed system takes around $11.1 - 32.25$ ms to compensate for an approximately $76.25 - 100$ ms delay (excluding human reaction time) that would otherwise occur in a single round trip of control-feedback signals.

Further comparison. We compare our proposed methodology with benchmark work [8] 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 $\sim 625\ ms$. While as, the proposed approach in our work is generalized as the predicted output is a continuous value instead of categories. It predicts the intent of motion with an MSE of 0.0009 in $10.1 - 35.5\ ms$ timeframe.

Ablation Study. We proceed to vary the length of the lookahead window m to observe the performance of the prediction mechanism on the actual data. On a novel sample taken from a similar distribution, the results are shown in Figures 8. It is observed that the error in prediction increases as the lookahead increases, which signifies an intuitively expected result.

Visualized Results. We analyze the visualization of the prediction mechanism on the aforementioned objects of interest, the same is illustrated in Figure 9.

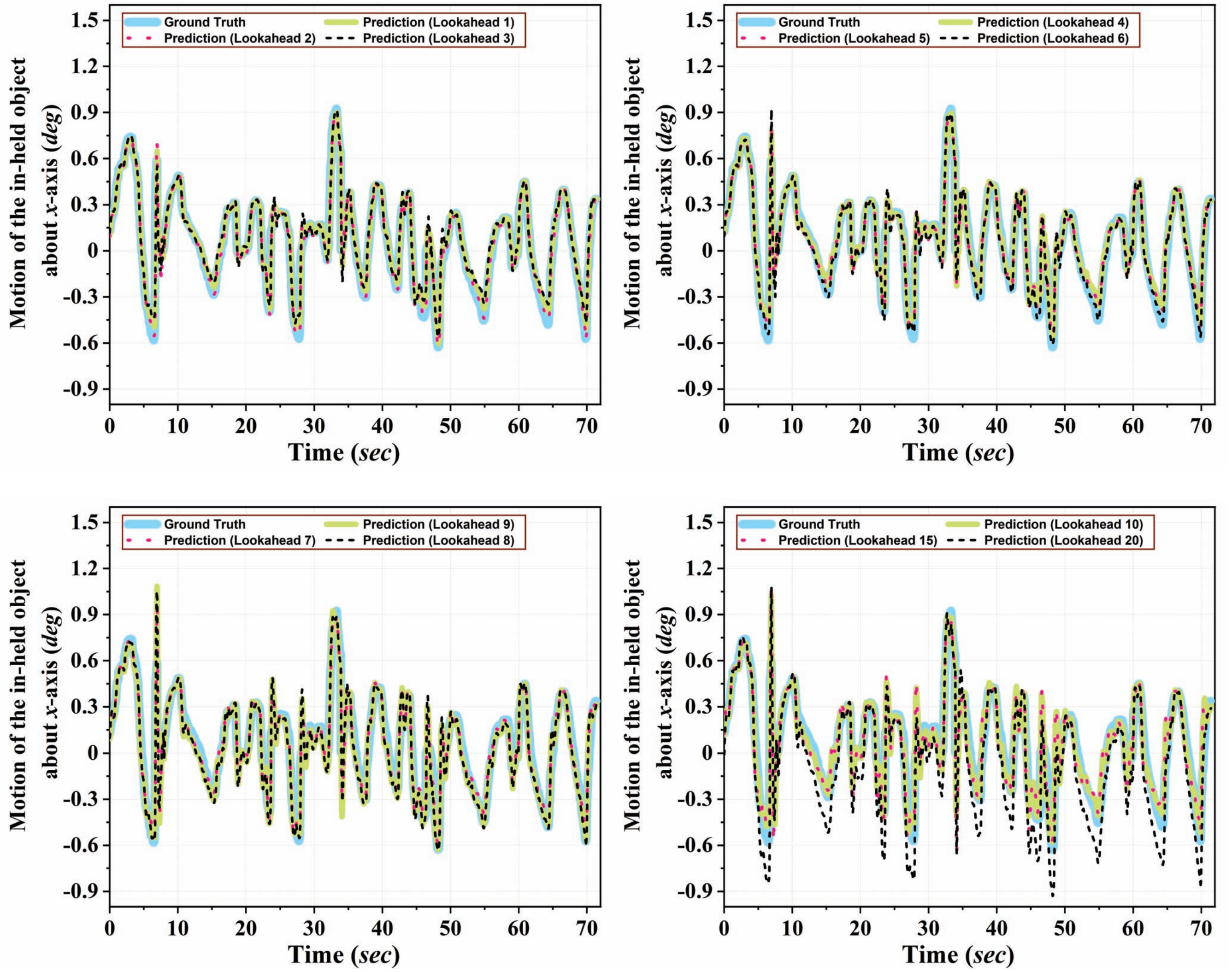

Figure 8. Showing the trajectory of true motion of the object and the predicted intent of the motion of the object using various lookahead values.

3. Conclusion

This paper summarizes the contribution of utilizing visual, kinaesthetic modalities and their hybrid version of estimating the human intent in manipulating various objects of interest. We propose an ensemble stacking approach to discuss the improvements made in accurate predictions of the human intent template that can help in mitigating the delays occurring in bilateral move-and-wait strategy of teleoperation. A total MSE of 0.0016 is observed while using kinaesthetic data individually while we observe that a total MSE of 0.006 is observed when using visual data independently. A considerable performance in accuracy is observed by realizing lesser error when combining both the modalities. Our proposed system outperforms benchmark approach in the literature in the respect of modelling the prediction of human intent as a regression problem with significant results.

References

[1] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1

[2] Sanika Doolani, Callen Wessels, Varun Kanal, Christos Sevastopoulos, Ashish Jaiswal, Harish Nambiappan, and Fillia Makedon. A review of extended reality (xr) technologies for manufacturing training. *Technologies*, 8(4):77, 2020. 1

[3] Claudia D'Ettorre, Andrea Mariani, Agostino Stilli, Ferdinando Rodriguez y Baena, Pietro Valdastri, Anton Deguet, Peter Kazanzides, Russell H Taylor, Gregory S Fischer, Simon P DiMaio, 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. 1

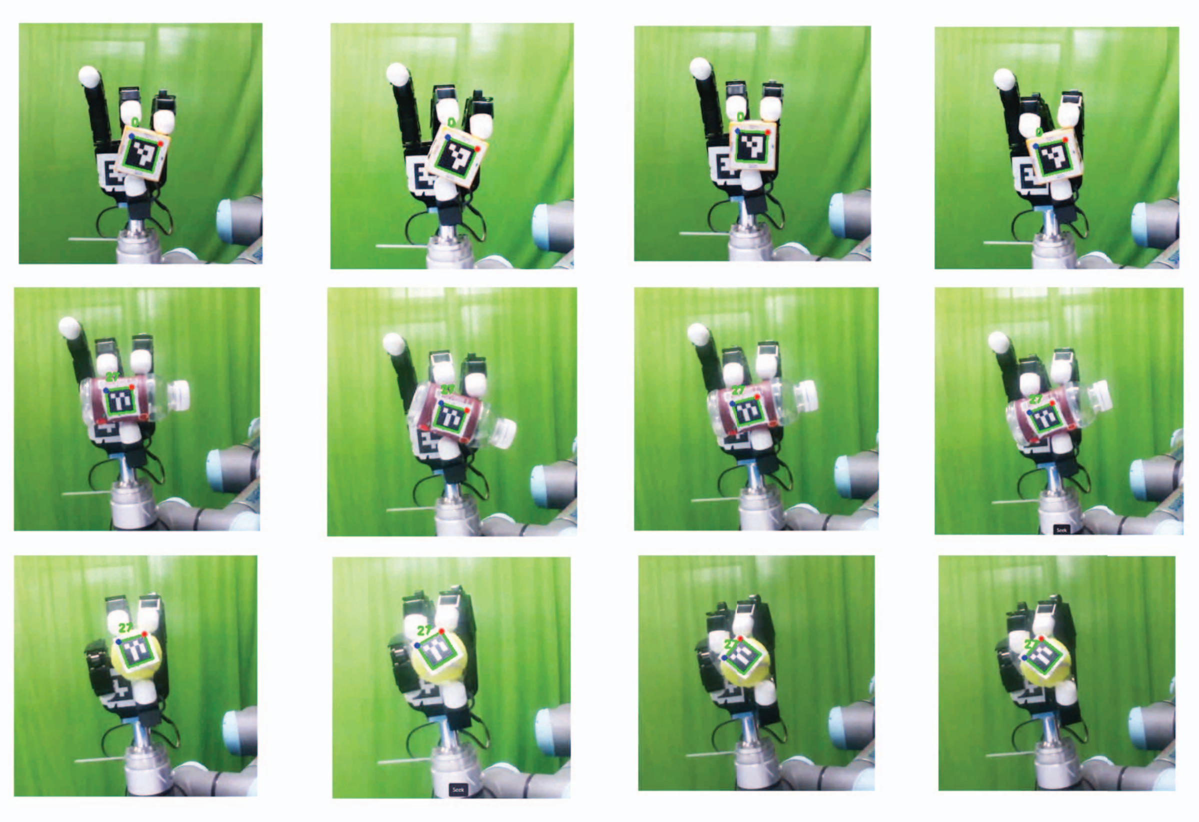

Figure 9. The estimation/prediction of intent of motion using different objects from initial pose to goal pose (left to right).

[4] Guillaume Gamelin, Amine Chellali, Samia Cheikh, Aylen Ricca, Cedric Dumas, and Samir Otmane. Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments. *Personal and Ubiquitous Computing*, 25:467–484, 2021. 1

[5] Vicent Girbes-Juan, Vinicius Schettino, Yiannis Demiris, and Josep Tornero. Haptic and visual feedback assistance for dual-arm robot teleoperation in surface conditioning tasks. *IEEE Transactions on Haptics*, 14(1):44–56, 2020. 1

[6] Peter F Hokayem and Mark W Spong. Bilateral teleoperation: An historical survey. *Automatica*, 42(12):2035–2057, 2006. 1

[7] Jing Luo, Weibin Liu, Wen Qi, Jianwen Hu, Junming Chen, and Chenguang Yang. A vision-based virtual fixture with robot learning for teleoperation. *Robotics and Autonomous Systems*, 164:104414, 2023. 1

[8] Catharine LR McGhan, Ali Nasir, and Ella M Atkins. Human intent prediction using markov decision processes. *Journal of Aerospace Information Systems*, 12(5):393–397, 2015. 5

[9] Leonid Prokhorenko, Daniil Klimov, Denis Mishchenkov, and Yuri Poduraev. Modular robot interface for a smart operating theater. *Journal of Robotic Surgery*, pages 1–13, 2023. 1

[10] Luca Scimeca, Josie Hughes, Perla Maiolino, Liang He, Thrishantha Nanayakkara, and Fumiya Iida. Action augmentation of tactile perception for soft-body palpation. *Soft robotics*, 9(2):280–292, 2022. 1

[11] Geng Yang, Honghao Lv, Zhiyu Zhang, Liu Yang, Jia Deng, Siqi You, Juan Du, and Huayong Yang. 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. 1

[12] Yi Zhang, Susan Finger, and Stephannie Behrens. *Introduction to mechanisms*. Carnegie Mellon University, 2003. 2

[13] Chengzhi Zhu, Chenguang Yang, Yiming Jiang, and Hui Zhang. Fixed-time fuzzy control of uncertain robots with guaranteed transient performance. *IEEE Transactions on Fuzzy Systems*, 2022. 1