MiXR-Intent: Intention Prediction for Immersive Remote Encountered-Type Haptics

N. B. Takele[1,2], D. Delehelle[1,2], Y. Kim[1], S. Anastasi[4], N. Deshpande[3], J. Ortiz[1], D. G. Caldwell[1], C. Recchiuto[2], and Y. T. Tefera[1]

***Abstract*— This paper introduces MiXR-Intent, a deep learning-based system for intention estimation in immersive telepresence with haptic feedback. MiXR-Intent uses gaze direction, upper body posture, and hand movement patterns to accurately predict user intentions and contact points within virtual reality (VR) environments. By interpreting these cues, the system aims to minimize interaction noise and latency—challenges that are especially critical when haptic feedback is employed for remote telepresence and teleoperation applications. In this paper, we propose a Long Short-Term Memory (LSTM) neural network as a method for predicting user intentions and contact points during interaction. Our approach focuses on using three interaction types: pushing, pointing, and grasping, which could be used as a fundamental interaction types for enabling natural and intuitive user experiences. To improve robustness and generalization in all directions and positions, each interaction type data is collected across six distinct directional movements, to make sure diverse trajectory variations relative to the user's body. This data variability allows the proposed LSTM neural network to capture user interaction under different spatial positions. The model is trained on a dataset of 6120 samples collected from 20 subjects, using 17 contact points and capturing gaze and upper body movements across six distinct directions. We evaluated the proposed system using various parameters of the LSTM architecture and conducted a performance analysis across multiple metrics, achieving up to 95% accuracy in classification.**

***Index Terms*— Mixed Reality, Body Tracking,Haptics, Intention Prediction, Teleoperation, Telepresence**

I. INTRODUCTION

Robots have become invaluable assets to the human being across various applications, from household cleaning assistants to executing high-risk missions in environments too hazardous for humans, driven by recent advancements in consumer-grade high-performance computing devices. However, despite the hardware and software advancements, enabling robots to operate effectively in semi-structured or completely unstructured environments without direct human supervision remains a significant challenge [2]. This has led to a growing focus on teleoperation, which uses human capabilities alongside robots to provide adaptable and precise control in complex scenarios [3], [4].

This research is supported by and in collaboration with the Italian National Institute for Insurance against Accidents at Work (INAIL), under the project "Sistemi Cibernetici Collaborativi - Robot Teleoperativo 3".

[1]Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163 Genova, Italy

[2]University of Genova, via all'Opera Pia 13, 16145, Genova, Italy

[3]School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, UK

[4]Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro (INAIL), P.le Pastore 6, 00144 Rome, Italy

Yet, teleoperation itself introduces a unique set of challenges that must be addressed to have a seamless and efficient operation. Primarily, it requires an intuitive and real-time interaction between the human operator and the robotic system by maintaining high fidelity in control actions, where every movement of the robot mirrors the operator's intent [2]. Additionally, the system must provide essential feedback to the user in real-time to allow operators to sense remote environments [5]. Numerous techniques have been proposed to provide this senses using several types of teleoperation interfaces. Virtual Reality (VR) has been used to provide immersive visualization [5], and encountered-type haptic (ETH) feedback systems have been explored to deliver a realistic sense of touch and force perception [1] by using the user's hand pose and gaze to generate force feedback through a robotic manipulator, seen in Fig. 1. However, the position accuracy of the robotic manipulator can be impacted by measurement noise and latency [1]. To address these challenges, this paper proposes a data-driven approach that predicts the operator's intended interaction type and contact points using body tracking. Specifically, it uses upper-body, hand, and gaze tracking to anticipate user interactions and compensate for measurement errors, thereby enhancing the accuracy and responsiveness of the system.

II. RELATED WORK AND CONTRIBUTION

Our proposed approach draws on multiple fields, such as haptics, virtual reality, and deep learning for intention prediction. In the following, we briefly highlight the most relevant recent related works.

Researchers have explored the advantages of integrating immersive VR and haptic interfaces into remote teleoperation and telepresence applications. Studies such as [6], [7], [8] have investigated immersive visualization systems that enable real-time 3D scene rendering with dynamic viewpoint control using head-tracked stereo 3D displays. Together with immersive visualization, haptic technologies such as ETH has been proposed [9], [10], [11], [12] to provide intuitive haptic feedback. Although, traditional desktop-type haptic devices can provide advanced force feedback [13], but they are limited by the requirement for the operator to hold the device and limited workspace. In contrast, VR based Encountered-type haptics enable bare-hand interaction, which has been shown to improve operator immersion in teleoperation scenarios [12]. However, the effectiveness of ETH systems depends on accurately interpreting the operator's hand and body movements, as different positions and

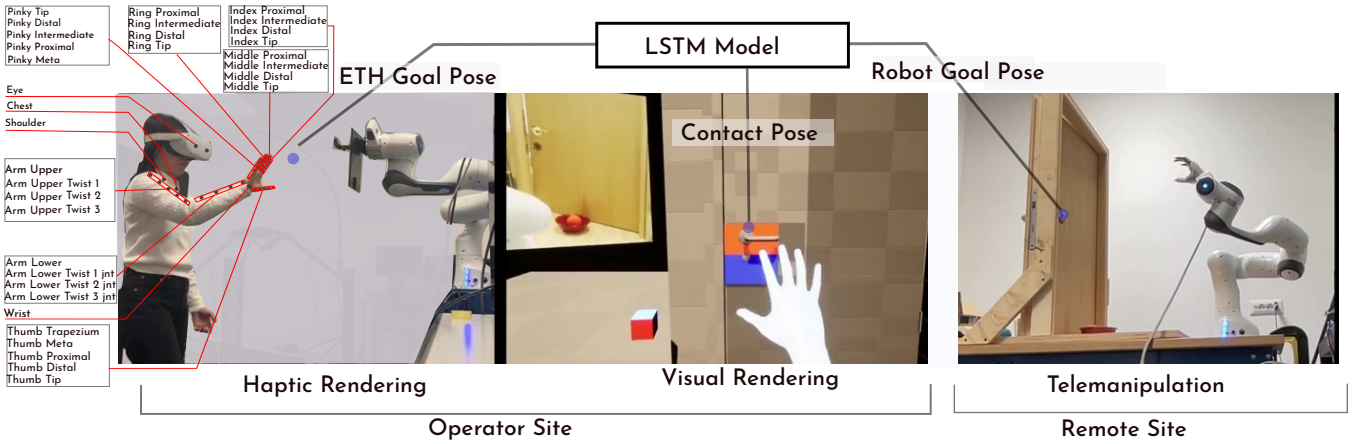

Fig. 1. The figure illustrates a typical bilateral encountered-type haptic teleoperation system - Figure adapted from [1]. In this setup, the human operator's upper body and gaze are tracked while they interact with a remote environment using their bare hands. The proposed system adds an LSTM model to predict both the operator's intended actions and contact poses.

gestures convey distinct interaction intentions.

Here, we present a brief overview of various techniques used in human action understanding and intention prediction within human-computer and human-robot interaction fields. Several approaches have been proposed to predict and interpret human actions. One example is Festor et al. [14], who introduced a fully data-driven approach to predict human actions during object manipulation tasks. This method relies only on gaze cues, using eye movement patterns to infer user intentions. Salvato et al. [15] tackled the issue of VR latency by using a self-attention mechanism to predict user touch interactions, to create smoother and more responsive haptic feedback. Similarly, Du et al. [16] integrated Electromyography (EMG) and Inertial Measurement Unit (IMU) sensors with a Long Short-Term Memory (LSTM) neural network to achieve real-time predictions of hand movements and grasp forces, which is particularly relevant for robotic teleoperation requiring precise force control. Another important contribution comes from Belardinelli et al. [17], who leveraged gaze and hand motion data in combination with Generalized Hidden Markov Models (GHMMs) to predict user intentions. This approach improved the accuracy of action prediction in interactive scenarios, demonstrating the importance of multimodal data fusion in understanding human intention.

Building on advancements in the research works mentioned above, this article presents our work aimed at improving the use of hand and body position for interactions. To the best of our knowledge, this is the first attempt to apply intention prediction to immersive interactions and the main contributions are listed bellow:

1) We proposed a predictive model for intention prediction using LSTM for VR-based encountered-type haptics.
2) A novel dataset that provides gaze, upper body movements, hand movements, and object poses, providing researchers carefully evaluated dataset resource.

III. SYSTEM OVERVIEW

As seen in Fig. 1, the proposed approach builds on immersive teleoperation interface system adapted from [12], consisting of an operator site and a remote robot site. These sites are linked via an immersive VR interface, managing interaction and the implementation of the proposed tracking and prediction system. At the operator site, an encountered-type haptic interface is used. The 7-degree-of-freedom (DoF) Franka Emika Panda, was used as a haptic manipulator (ETH robot), providing physical feedback to the operator. The core contribution of this work lies in the integration of an LSTM neural network to predict user intentions and contact poses. By continuously tracking the operator's gaze, hand position, and upper body movement, the system anticipates the operator's intended actions and the corresponding contact points.

A. Interaction Dataset

Gaze, upper body, and hand interactions plays a very important role in improving in human computer or human robot interaction. For example, gaze direction could provide important information about the user's focus and intent [7], upper body movements could show posture and spatial orientation, and hand interactions can provide cues about motor skills essential for object manipulation [18]. Using these body postures and movements separately neglects the broader context of user behavior and could leading to less adaptive and intuitive systems. But, by combining them, we can develop a more comprehensive user interactions system that can respond in a more natural manner. The first important task to achieve this is to collect body-tracked dataset for various types of interactions. Inspired by authors in [19], we defined three key interaction types: grasping, pushing, and pointing (Fig. 3). These interactions are chosen because they represent a wide range of common human-object or human-environment interactions. Each of these three interactions was performed in six distinct directions to capture a wide variety of movement patterns: forward,

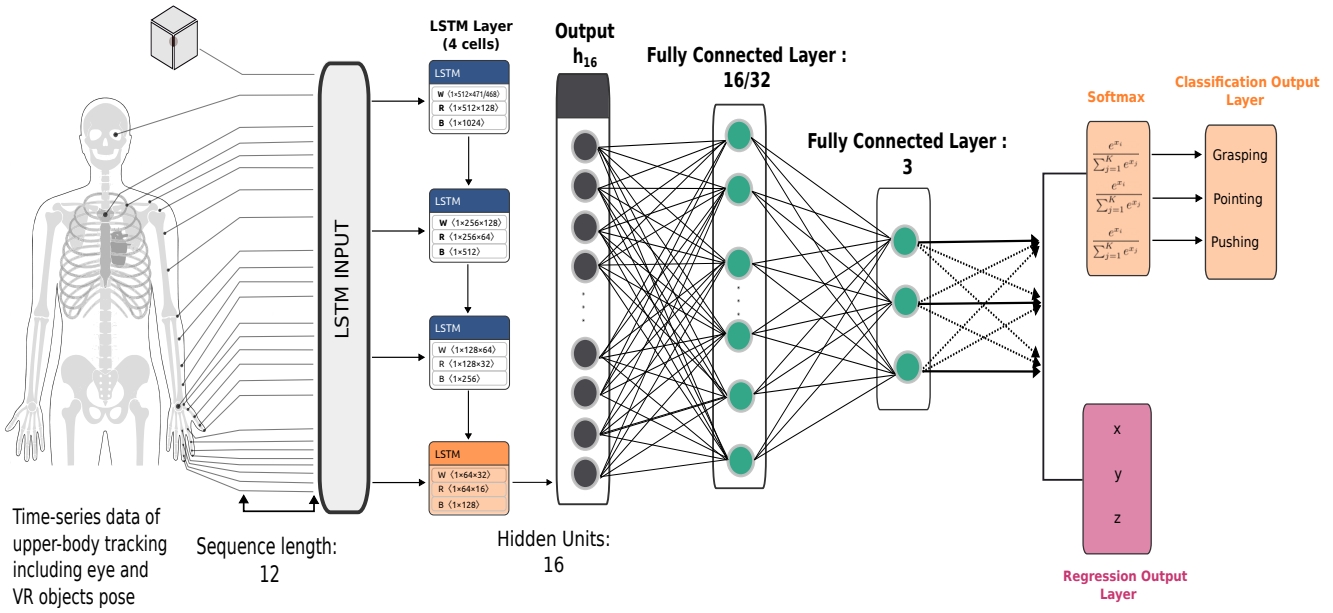

Fig. 2. LSTM architecture for intention classification and regression. For brevity, we used orange and pink colors to highlight layers for each task. Classification task uses all four LSTM layers (including orange) with Softmax and the classification output layer. Regression task uses the first three LSTM layers, excluding the orange one, with X, Y, and Z outputs.

backward, left-to-right, right-to-left, upward, and downward. For each direction, 17 contact points were collected, ensuring sufficient data coverage for training robust models, resulting in 306 samples per subject (3 interaction types $\times$ 6 directions $\times$ 17 contact points). With data from 20 subjects, the dataset comprises a total of 6120 samples. This novel dataset incorporates gaze tracking, upper body movements, hand motions, and virtual object poses, providing a rich foundation for improving interaction prediction and enhancing the responsiveness of teleoperation and haptic feedback systems.

The dataset was collected with clear objectives to ensure systematic and meaningful data collection. The primary goal was to identify the most likely contact points where a user intends to interact and enable their rapid localization. A secondary objective was to provide insights into the type of interaction, to clearly identify actions such as pushing, grasping, or pointing. To achieve these objectives, maintaining consistency during data collection was essential to ensure reliability, reproducibility, and accuracy. To support this objectives, the dataset collection was structured around the following three key characteristics:

1) **Diversity in interaction**: the dataset includes grasping, pushing, and pointing actions performed in six distinct directions to capture a full range of motion.
2) **Consistency in collection**: a structured experimental protocol was followed to maintain uniformity in data flow and direction.
3) **Accuracy in contact point interaction**: contact points in MR were designed to be large enough for easy interaction yet small enough to require precision.

Due to space limitations and to maintain focus on the main contribution, the detailed data analysis is kept concise. For further details, readers are encouraged to refer to the supplementary materials published in a workshop [20].

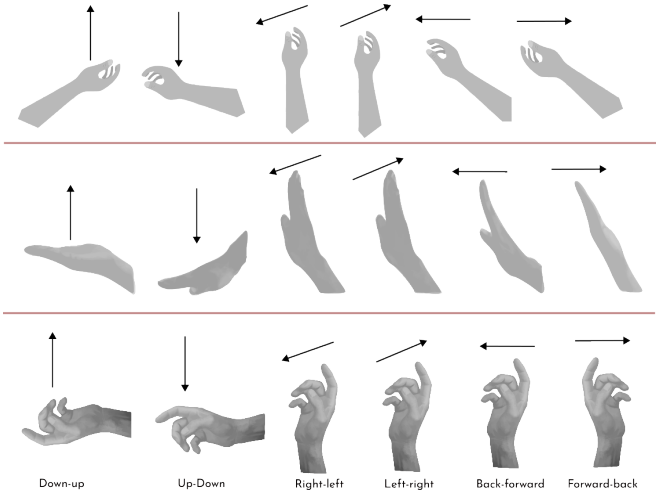

Fig. 3. The six directional grasping, pushing, and pointing gestures

B. Interaction Type Prediction and Contact Pose Estimation

To estimate the contact pose and understand the interaction type, We formulate the problem as two complementary tasks: (1) a classification task: predicting the user's interaction type based on upper body motion and gaze cues, and (2) a regression task: estimating the 3D contact pose ${}^{VR}\mathbf{P}_t$. Let $\mathbf{Y_C} = \{y_1, y_2, \ldots, y_C\}$ represent the set of interaction type classes, where C corresponds to the three distinct interaction categories: grasping, pointing, and pushing. The

objective is to train a model that can map each interaction, represented as a sequence $\mathbf{x_t}$ for $t_0 < t < t_N$, with N being the length of the sequence, to a predicted interaction type class $\hat{\mathbf{Y}}_c$, ensuring alignment with the true interaction type label $\mathbf{y}$. The classification task is defined as learning a function $f_c : \mathbb{R}^n \rightarrow Y_c$, where f_c maps the input features to a specific interaction type class, Y_c. This function is controlled by θ_c, which represents the model's weights. During training, the model updates these weights iteratively through an optimization process, typically by minimizing a loss function that quantifies the difference between predicted and actual labels. Through this process, the model gradually improves its ability to accurately predict the interaction type for each input.

The regression model is designed to learn a function $f_r : \mathbb{R}^n \rightarrow \mathbb{R}^3$ that takes an input feature vector $\mathbf{x}_t$ and predicts the coordinates ${}^{VR}\hat{\mathbf{P}}_t \in \mathbb{R}^3$, which should closely approximate the true contact point $\mathbf{p}_t$. The goal is to minimize the discrepancy between the predicted and actual coordinates. To achieve this, the model iteratively updates its parameters θ_r through an optimization process that adjusts the weights to reduce the prediction error. Over successive iterations, the model refines its parameters to better capture the relationship between the input features and the contact point, gradually improving the accuracy of the predicted coordinates.

1) Interaction Type Prediction: For classification, the model processes input sequences of 12 time steps. Each time step includes 471 features extracted from gaze, upper body, and hand pose. The classification network is structured with four stacked LSTM layers. The first LSTM layer has 128 hidden units, generating a sequence of hidden states that capture the temporal dependencies in user interactions. The second LSTM layer, with 64 hidden units, processes these hidden states and passes its output to the next stage. The third LSTM layer, with 32 hidden units, continues processing the temporal features, while the final LSTM layer, with 16 hidden units, produces the final output, which is then passed on to the next stage. Dropout layers with a rate of 0.4 are applied after each LSTM layer to prevent overfitting. A fully connected dense layer with 16 neurons and ReLU activation follows, leading to the final output layer with 3 neurons, corresponding to the three interaction types. The ReLU activation function is selected for its ability to introduce non-linearity [21], enabling the model to learn complex interaction patterns while mitigating the vanishing gradient problem, thus enhancing training stability and efficiency. A softmax activation function is then used to generate class probabilities, converting the output into a probability distribution over the three classes, which is essential for multiclass classification tasks.

2) Contact Pose Estimation: The contact pose estimation model follows a similar sequential structure but is designed to do regression for continuous values. It takes input sequences of 12 time steps, each with 468 features, and consists of three LSTM layers. The first layer has 128 hidden units and passes the full sequence to the next LSTM layer, which has 64 hidden units. The second LSTM layer also passes the full sequence to the third LSTM layer, which has 32 hidden units. Unlike the previous layers, the third LSTM layer only passes its final output to the next stage. To reduce overfitting, dropout layers with a rate of 0.3 are applied after each LSTM layer. Following this, a fully connected dense layer with 32 neurons and ReLU activation refines the learned features. The final output layer consists of 3 neurons representing the x, y, and z coordinates for the contact points, with a linear activation function for continuous value prediction.

As shown in Fig. 2, the proposed LSTM model comprises two architectures within a single figure for conciseness and clarity: one dedicated to interaction type prediction and the other to contact pose estimation.

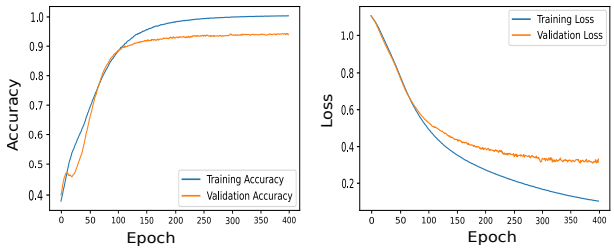

Fig. 4. The left plot illustrates the model's accuracy over epochs, while the right plot depicts the loss values

IV. EXPERIMENTAL EVALUATION

A. Evaluation Metrics

We evaluate the proposed intention prediction algorithm using standard metrics. For classification, we report Accuracy, Precision, Recall, and F1-Score to assess overall performance, along with Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC), and confusion matrices for per-class analysis. For regression, we use Mean Absolute Error (MAE) to measure prediction accuracy.

B. Evaluation and Analysis

The models use a window size of 12 time steps, determined through experimental analysis. A summary of the model's performance across different time steps is provided in Table I.

TABLE I
MODEL PERFORMANCE METRICS ACROSS VARIOUS TIME STEPS

Time Steps	8	10	**12**	14	16	18
Accuracy	0.7501	0.7229	**0.9507**	0.7395	0.7404	0.6559
MAE	0.0959	0.0944	**0.0911**	0.0976	0.1207	0.1012

1) Classification: The final model achieved an overall test accuracy of 0.9507, correctly predicting approximately 95.1% of the test samples, as shown in Fig. 4. This reflects a good level of classification performance. To further evaluate the model's performance across individual classes, Fig. 5

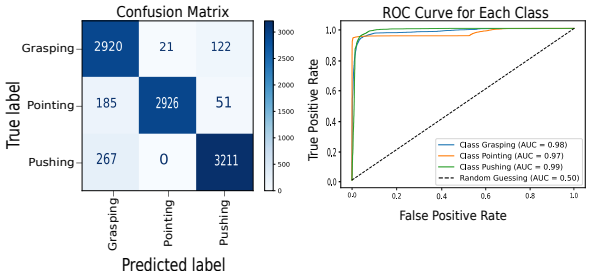

Fig. 5. The left plot displays the model's confusion matrix, while the right plot presents the ROC curves

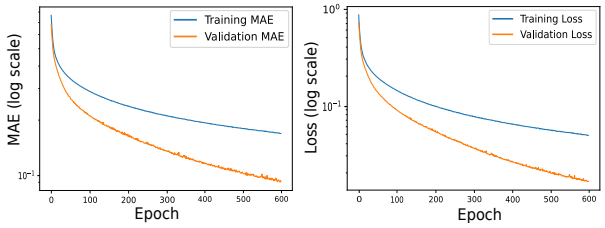

Fig. 6. MAE (left) and Loss (right) per Training Iteration

presents both the confusion matrix and ROC curves. For instance, in the confusion matrix, the first row shows that the model correctly classified 2920 instances of "Grasping," while misclassifying 21 as "Pointing" and 122 as "Pushing." The second and third rows detail the model's performance on the "Pointing" and "Pushing" classes, respectively. The ROC curve (Fig. 5, right) offers additional insight into the model's ability to distinguish between the three interaction classes. The AUC for "Grasping" was 0.98, while "Pointing" and "Pushing" achieved 0.97 and 0.99, respectively. These high values demonstrate the model's strong ability to differentiate between user actions, further validating its classification effectiveness.

To analyze the impact of different model configurations, we conducted three ablation studies (Table II). In Ablation Study 1, with two LSTM layers, the model performed reasonably well, achieving an AUC of 0.98 for "Grasping" and 0.99 for "Pointing" and "Pushing," respectively, though the precision for "Grasping" was slightly lower at 0.8981. In Ablation Study 2, increasing to three LSTM layers led to a slight improvement, with an overall accuracy of 92.46%. Precision for "Grasping" (0.8708), "Pointing" (0.9958), and "Pushing" (0.9180) showed minor variations. Finally, in Ablation Study 3, using four LSTM layers further refined the model, improving precision across most classes, particularly for "Pointing," which reached 0.9822, while maintaining a high AUC for all classes: 0.98 for "Grasping," 0.97 for "Pointing," and 0.99 for "Pushing." The results indicate that adding more LSTM layers enhances classification performance, though the improvements become less significant beyond three layers.

2) Regression: The regression model's ability to predict 3D contact points, which are essential for providing realistic

TABLE II
CLASSIFICATION METRICS FOR EACH ABLATION STUDY

Class	Accuracy	Precision	Recall	F1-Score	AUC	LSTM Layer
Ablation Study 1						
Grasping	0.9262	0.8981	0.9262	0.9119	0.98	
Pointing	0.9118	0.9924	0.9118	0.9504	0.99	2
Pushing	0.9741	0.9310	0.9741	0.9521	0.99	
Ablation Study 2						
Grasping	0.9507	0.8708	0.9507	0.9090	0.98	
Pointing	0.8969	0.9958	0.8969	0.9438	0.98	3
Pushing	0.9267	0.9180	0.9267	0.9223	0.97	
Ablation Study 3						
Grasping	0.9301	0.9332	0.9301	0.9317	0.98	
Pointing	0.9428	0.9822	0.9428	0.9621	0.97	4
Pushing	0.9761	0.9391	0.9761	0.9573	0.99	

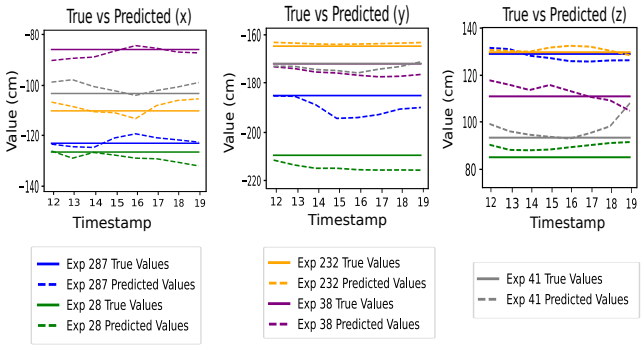

Fig. 7. True vs. Predicted Contact Points for Selected Experiments

haptic feedback, was evaluated under various configurations. Adjustments were made to the learning rate (controlling the speed of learning), the number of training epochs, and the structure of the LSTM network (number of layers). By experimenting with these factors, we aimed to determine the optimal combination that offers high accuracy without unnecessary complexity. To evaluate the model's performance, we used the MAE metric, which measures the average difference between the model's predictions and the actual contact points. A lower MAE indicates more precise predictions.

The input features were standardized using their mean and standard deviation, ensuring that all features were on a similar scale, which helps the model learn more effectively. Likewise, the target values (contact points) were also standardized in the same way to prevent large variations in output values that could negatively impact the training process. The MAE and loss per iteration during training are shown in Fig. 6. The loss curve illustrates how the model's error decreases over time, indicating its learning progress. The MAE plot, presented in normalized form, reflects how accurately the model predicts the target values throughout training.

To ensure consistent evaluation, the analysis focused on samples with identical sequence lengths. Specifically, Experiments 287, 28, 232, 38, and 41 were selected. This approach allowed for a fair comparison of model performance across different instances without variability in sequence length affecting the results. The MAE values varied across these experiments, reflecting differences in how well the model predicted contact points for each case. The lowest error was observed in Experiment 232, with an MAE of 1.67 cm, indicating relatively accurate predictions. In contrast, Experiment 28 showed the highest error, with an MAE of 4.00 cm, suggesting a slight deviation between predicted and actual contact points. The remaining experiments yielded MAE values of 3.03 cm (Experiment 287), 3.14 cm (Experiment 38), and 2.91 cm (Experiment 41).

Fig. 7 compares the true and predicted values for the x, y, and z coordinates across the selected experiments. The visualizations highlight varying levels of accuracy in the model's predictions. In some cases, the predicted values closely align with the actual values, while others show slight deviations. Notably, Experiment 232 yielded the most accurate predictions, with minimal differences between the predicted and true values. In contrast, Experiment 28 exhibited moderate discrepancies. Overall, the plot shows that the model predicted the contact points well, with predicted values closely matching the actual values in most experiments.

V. Conclusions and Future Work

In this paper, we introduced MiXR-Intent, an LSTM-based system, to predict user intentions in immersive teleoperation with encountered-type haptics (ETH). It uses gaze direction, hand gestures, and upper body movements to predict pushing, pointing, and grasping interactions. This allowed us to classify different interaction types and to estimate contact points, achieving an overall test accuracy of 95.1% and a low mean absolute error (MAE) in predicting 3D contact points. We believe the proposed method will improve the intuitiveness and responsiveness of ETH systems in teleoperation. Future work will involve selecting a subset of input features to identify which ones—such as gaze, hand movements, or upper body posture—contribute most significantly to accurate intention prediction. Additionally, we will further investigate this through subjective experiments using the ETH teleoperation setup.

References

[1] Y. Kim, S. Anastasi, and N. Deshpande, "Towards encountered-type haptic interaction for immersive bilateral telemanipulation," 2023.

[2] Y. T. Tefera, I. Sarakoglou, S. N. Deore, Y. Kim, V. Barasuol, M. Villa, S. Anastasi, D. G. Caldwell, N. Tsagarakis, C. Semini, and N. Deshpande, "Robot teleoperativo: Collaborative cybernetic systems for immersive remote teleoperation," in *2024 IEEE Conference on Telepresence*, pp. 1–4, 2024.

[3] A. Acemoglu, G. Peretti, M. Trimarchi, J. Hysenbelli, J. Krieglstein, A. Geraldes, N. Deshpande, P. M. V. Ceysens, D. G. Caldwell, M. Delsanto, *et al.*, "Operating from a distance: robotic vocal cord 5g telesurgery on a cadaver," *Annals of internal medicine*, vol. 173, no. 11, pp. 940–941, 2020.

[4] O. Khatib, X. Yeh, G. Brantner, B. Soe, B. Kim, S. Ganguly, H. Stuart, S. Wang, M. Cutkosky, A. Edsinger, P. Mullins, M. Barham, C. R. Voolstra, K. N. Salama, M. L'Hour, and V. Creuze, "Ocean one: A robotic avatar for oceanic discovery," *IEEE Robotics and Automation Magazine*, vol. 23, no. 4, pp. 20–29, 2016.

[5] A. Naceri, D. Mazzanti, J. Bimbo, Y. T. Tefera, D. Prattichizzo, D. G. Caldwell, L. S. Mattos, and N. Deshpande, "The vicarios virtual reality interface for remote robotic teleoperation: Teleporting for intuitive tele-manipulation," *Journal of Intelligent & Robotic Systems*, vol. 101, pp. 1–16, 2021.

[6] P. Milgram and J. Ballantyne, "Real world teleoperation via virtual environment modeling," in *International Conference on Artificial Reality & Tele-existence*, Citeseer, 1997.

[7] Y. T. Tefera, D. Mazzanti, S. Anastasi, D. G. Caldwell, P. Fiorini, and N. Deshpande, "Forecast: Real-time foveated rendering and unicasting for immersive remote telepresence.," in *ICAT-EGVE*, pp. 75–84, 2022.

[8] A. Mossel and M. Kroeter, "Streaming and exploration of dynamically changing dense 3d reconstructions in immersive virtual reality," in *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pp. 43–48, IEEE, 2016.

[9] W. A. McNeely, "Robotic graphics: a new approach to force feedback for virtual reality," in *Proceedings of IEEE Virtual Reality Annual International Symposium*, pp. 336–341, IEEE, 1993.

[10] T. Susumu, "A construction method of virtual haptic space," in *International Conference on Artificial Reality and Tele-Exsistence (ICAT'94)*, pp. 131–138, 1994.

[11] Y. Yokokohji, R. L. Hollis, and T. Kanade, "What you can see is what you can feel-development of a visual/haptic interface to virtual environment," in *Proceedings of the IEEE 1996 Virtual Reality Annual International Symposium*, pp. 46–53, IEEE, 1996.

[12] Y. Kim, M. C. Castillo Silva, S. Anastasi, and N. Deshpande, "Towards immersive bilateral teleoperation using encountered-type haptic interface," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1354–1359, 2023.

[13] J. K. Salisbury and M. A. Srinivasan, "Phantom-based haptic interaction with virtual objects," *IEEE Computer Graphics and Applications*, vol. 17, no. 5, pp. 6–10, 1997.

[14] P. Festor, A. Shafti, A. Harston, M. Li, P. Orlov, and A. A. Faisal, "Midas: Deep learning human action intention prediction from natural eye movement patterns," *arXiv preprint arXiv:2201.09135*, 2022.

[15] M. Salvato, N. Heravi, A. M. Okamura, and J. Bohg, "Predicting hand-object interaction for improved haptic feedback in mixed reality," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3851–3857, 2022.

[16] Y. Du, H. B. Amor, J. Jin, Q. Wang, and A. Ajoudani, "Learning-based multimodal control for a supernumerary robotic system in human-robot collaborative sorting," *IEEE Robotics and Automation Letters*, 2024.

[17] A. Belardinelli, A. R. Kondapally, D. Ruiken, D. Tanneberg, and T. Watabe, "Intention estimation from gaze and motion features for human-robot shared-control object manipulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9806–9813, IEEE, 2022.

[18] R. Ren, Z. Wang, C. Yang, J. Liu, R. Jiang, Y. Zhou, S. Jiang, and B. He, "Enhancing robotic skill acquisition with multimodal sensory data: A novel dataset for kitchen tasks," *Scientific Data*, vol. 12, no. 1, p. 476, 2025.

[19] Y. Xiao, K. Miao, and C. Jiang, "Mapping directional mid-air unistroke gestures to interaction commands: A user elicitation and evaluation study," *Symmetry*, vol. 13, no. 10, p. 1926, 2021.

[20] N. B. Takele, D. Delehelle, Y. Kim, Y. T. Tefera, N. Deshpande, D. G. Caldwell, J. Ortiz, and C. T. Recchiuto, "Mixr-interact: Mixed reality interaction dataset for gaze, hand, and body," in *HRI 2025 Workshop VAM-HRI*, 2025.

[21] A. I. Rodríguez and X. D. Buitrago, "How to choose an activation function for deep learning," *Tekhnê*, vol. 19, no. 1, pp. 23–32, 2022.