Multimodal Target Prediction for Rapid Human-Robot Interaction

Mukund Mitra[*]
Ameya patil
GVS Mothish[*]
Gyanig Kumar[*]
Abhishek Mukhopadhyay[*]
LRD Murthy[*]
Partha Pratim Chakrabarti[†]
Pradipta Biswas[*]

ABSTRACT

Intent prediction finds widespread applications in user interface (UI/UX) design to predict target icons, in automotive industry to anticipate driver's intent, and in understanding human motion during human-robot interactions (HRI). Predicting human intent involves analyzing factors such as hand motion, eye gaze movement, and gestures. This paper introduces a multimodal intent prediction algorithm involving hand and eye gaze using Bayesian fusion. Inverse reinforcement learning was leveraged to learn human preferences for the human-robot handover task. Results demonstrate that the proposed approach achieves the highest prediction accuracy of 99.9% at 60% task completion as compared to state-of-the-art (SOTA) methods.

CCS CONCEPTS

• **Mathematics of computing** → **Bayesian computation**; • **Computing methodologies** → *Inverse reinforcement learning*; • **Computer systems organization** → Robotic autonomy.

KEYWORDS

Human-Robot Interaction, Intent Prediction, Inverse Reinforcement Learning, Multimodal Target Prediction

ACM Reference Format:
Mukund Mitra, Ameya patil, GVS Mothish, Gyanig Kumar, Abhishek Mukhopadhyay, LRD Murthy, Partha Pratim Chakrabarti, and Pradipta Biswas. 2024. Multimodal Target Prediction for Rapid Human-Robot Interaction. In *29th International Conference on Intelligent User Interfaces - Companion (IUI Companion '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640544.3645229

1 INTRODUCTION

Target prediction or intent recognition has been studied for various applications ranging from user interface design [7], automotive [2], human-machine interaction [4], clinical environment [14], domestic settings, and so on. Leveraging multiple input modalities for target prediction enhances the performance and robustness of Human-Robot Interactive (HRI) systems [36]. Human-robot handover is an important aspect of HRI which refers to actions initiated either by the human or their robotic counterpart, to deliver objects to each other [31]. These handovers require the robot to anticipate human hand movements and intentions to effectively plan its path, take control of the object from human, and deliver it to the intended destination, as shown in Fig. 1a. Trajectory prediction algorithms with multiple input modalities are frequently employed to forecast human movements and intended targets [5, 9, 23] during HRI. These algorithms for trajectory prediction can be categorized into four groups: (a) Physics-based (Kalman filter [10], minimum jerk model [6, 22]), (b) Probabilistic graphical models (Gaussian Mixture Models [21, 23], Hidden Markov Models [32]), (c) Recurrent Neural Network (RNN) models [11, 16, 19, 26], and (d) Inverse Reinforcement Learning (IRL) models [15, 24, 25]. Physics-based models cannot provide uncertainty in predicted human motion whereas graphical and RNN-based models lack generalization to unseen environments and adjustment of model parameters is complicated [18]. Most of the IRL formulations assume linear reward-feature dependency, which may not capture complex non-linear rewards during human-robot collaboration (HRC). Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL) [13, 33] leverages the representational capacity of neural networks to capture complex reward distribution. *Wang et al.* proposed a Maximum Entropy IRL (MEIRL) [35] based teaching-learning-collaborative (TLC) framework [30] to predict human intent. The model used speech and data from wearable sensors as input modalities to predict intent using gesture recognition. In contrast, this work predicts human intent with hand and gaze data using MEDIRL. The main contributions of this paper are:

(1) A multimodal target prediction algorithm with hand and gaze movement using Bayesian fusion was proposed for accurate anticipation of the intended target based on partial demonstration. An IRL-based model with a comprehensive

[*] Indian Institute of Science, Bangalore
[†] Indian Institute of Technology, Kharagpur

IUI Companion '24, March 18–21, 2024, Greenville, SC, USA

ACM ISBN 979-8-4007-0509-0/24/03
https://doi.org/10.1145/3640544.3645229

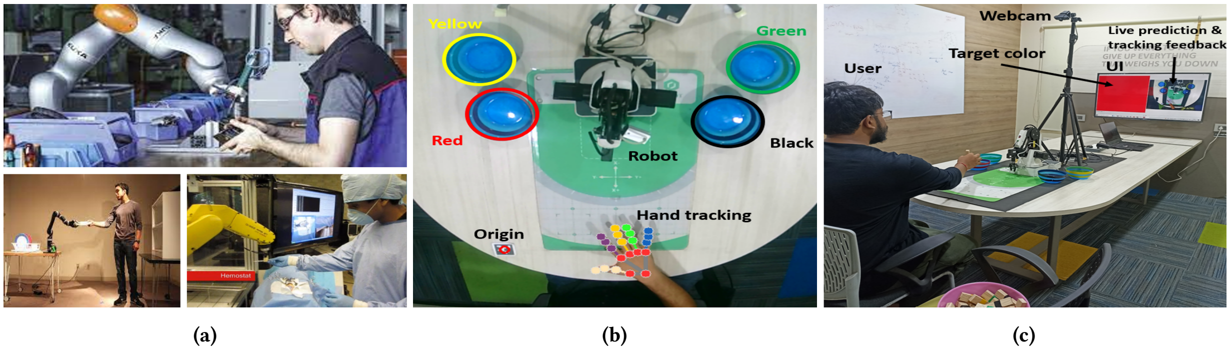


Figure 1: (a) Application of target prediction during human-robot handover in industrial, domestic, and clinical environments. Human hands over assembly parts, domestic appliances, and surgical equipment, respectively to the robot for completing subsequent intended action (b) Top-view of the setup captured by the hand-tracking webcam with four targets and robot symmetrically placed with respect to the user (c) Experimental setup with user interface

set of features was proposed to explicitly capture hand motion during handovers.

(2) User study was performed to analyze the effect of hand movement and eye gaze for target prediction. Results demonstrate that the proposed multimodal target prediction method achieved the highest accuracy of 99.9% at 60% task completion as compared to SOTA Kernel ELM [31] with 99.7% accuracy. Using hand movement, the method outperforms other SOTA methods by reporting an accuracy of 93% at 50% task completion.

2 MULTIMODAL TARGET PREDICTION

The task involved users reaching and placing objects (colored blocks) at distant targets or goals G. Given a partial hand movement ψ, future hand trajectory and the intended target was predicted for handover to the robot. Hand and eye gaze of users were tracked. *Average Fixation Duration (AFD)* which is the total duration of all fixations by the total fixation count [28] for each target, obtained from gaze data was used to estimate prior goal probability $p(G)$, given by:

$$p(G) = \frac{AFD_G}{\sum_{G\epsilon\mathbb{G}} AFD_G} \tag{1}$$

where $\mathbb{G}$ is the set of targets. Probability of a goal given partial demonstration was obtained using Bayes' theorem:

$$p\left(G\epsilon\mathbb{G}|\psi\right) = \frac{p\left(\psi|G\right)p\left(G\right)}{\sum_{G\epsilon\mathbb{G}} p\left(\psi|G\right)p\left(G\right)} \tag{2}$$

where $p\left(\psi|G\right)$ is the probability of the partial hand trajectory obtained from the learned reward distribution using IRL (Section 2.1). Equation (2) assigns higher probabilities to targets towards which the demonstrated partial trajectory approaches. The above approach can be implemented to any prediction pipeline involving multiple input modalities. Probability $p\left(G\epsilon\mathbb{G}|\psi\right)$ were calculated for all possible targets and the target corresponding to the maximum value was predicted G_{pred}. Future hand trajectory to G_{pred} was obtained from the corresponding reward distribution. The robot moves to the predicted hand trajectory for takeover and delivers the block to the predicted target by the user.

2.1 Hand Movement Prediction

As evident from Equation (2), the multimodal target prediction algorithm involves likelihood of the partial hand demonstration $p\left(\psi|G\right)$. This was obtained using Maximum Entropy Deep IRL (MEDIRL). Human hand motion was modeled as an agent following a Markov Decision Process (MDP). An MDP is defined as $\{S, A, T, \gamma, r\}$ consisting of of states $s\epsilon S$, actions $a\epsilon A$, probability of transition T, discount factor γ, and reward function $r : S \longrightarrow \mathbb{R}$. Let $\mathfrak{D} = \{\tau_i\}_{i=1}^{M}$ be the expert dataset consisting of M hand motion trajectories τ given by $\tau = [s_1, a_1, s_2, a_2, ..., a_{N-1}s_N]$. IRL aims to learn a reward function r, under which the expert demonstration is optimal. MEDIRL [33] approximates the reward function using a deep neural network parameterized by ω, i.e., $r_\omega\left(s\right) = g\left(f\left(s\right), \omega\right)$ where $f(s)$ are the features of state s and ω are the weights of the network or reward parameters. Function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ maps the feature space of dimension d to a real-valued reward r. The IRL problem was framed by maximizing the joint posterior distribution of observing the expert dataset and reward parameters under a given reward structure.

$$L\left(\omega\right) = logP\left(\mathfrak{D}, \omega|r\right) = logP\left(\mathfrak{D}|r\right) + logP\left(\omega\right) \tag{3}$$

The optimal reward parameters can be obtained by backpropagating the gradient with a regularization technique [33] as given below:

$$\frac{\partial L}{\partial\omega} = \left(\mu_{\mathfrak{D}} - \mathbb{E}\left[\mu\right]\right)\frac{\partial g\left(f, \omega\right)}{\partial\omega} \tag{4}$$

where $\mu_{\mathfrak{D}}$ is the State Visitation Frequency (SVF) from the expert data and $\mathbb{E}\left[\mu\right]$ is the expected SVF [35] given the learned reward at each iteration. Approximate Value Iteration and Policy Propagation algorithm [8] were used to estimate SVF. A feedforward neural network with 4 hidden layers with 128 neurons each and a ReLU activation function was employed for feature parameterized reward network. The state space was defined by discretizing the collaborative workspace of $60 \times 60cm$ into a 30×30 uniform grid. The action space consists of five actions *(up, left, right, up-left, up-right)* considering only forward motion of hand.

A goal-conditioned reward distribution was obtained by assigning higher rewards to goal states. This was fused with the reward distribution r learned using MEDIRL. For each goal G, a reward distribution $r_G\epsilon R$ was learned such that the expectation of features of the expert with respect to the probability distribution π_G over the actions (i.e., hand movements) at a given state equals that of the learner. Using the maximum entropy formulation of IRL [34, 35], π_G is recursively defined by:

$$\pi_G\left(a_i|s_i\right) \propto e^{Q_G(s_i,a_i)} \tag{5}$$

$$V_G\left(s_i\right) = \underset{a_i}{\mathrm{softmax}}\left\{r_G\left(s_i\right) + Q_G\left(s_i, a\right)\right\} \tag{6}$$

$$Q_G\left(s_i, a_i\right) = \mathbb{E}_{T(s_{i+1}|s_i,a)}\left[V_G\left(s_{i+1}\right)|s_i, a_i\right] \tag{7}$$

where V_G and Q_G are the future expected and cumulative rewards respectively. They are recursively calculated using equations (6), (7) and the MDP given in Section 2.1. The above equations have a closed form solution when the state transition dynamics are linear. The probability distribution over a partial hand trajectory $\psi = \{s_1, ..., s_m\}$, $m < N$ is given by:

$$p\left(\psi|G\right) = \prod_{i=1}^{m} \pi_G\left(a_i|s_i\right) = e^{\left\{\sum_{i=2}^{m} r_G(s_i)\right\}+V_G(s_m)-V_G(s_1)} \tag{8}$$

The above equation assigns higher probabilities to trajectories which maximizes future expected rewards. A summary of target prediction is given in Algorithm 1.

Algorithm 1 Multimodal Target Prediction

```
Input: Partial hand trajectory $\psi = \{s_1, ..., s_m\}$, Goal $G \epsilon \mathbb{G}$, MDP = $\{S, A, T, \gamma, r_G\}$, $r_G \epsilon R$, Average Fixation Duration corresponding to each goal ($AFD_G$).
Output: Predicted goal $G_{pred}$
for Goal G in set of goals $\mathbb{G}$ do
    Initialize $V_G = 0$ {%% Initialise value function to zero}
    Update $V_G \leftarrow (r_G, S, A, T, \gamma)$, [Equation (5) - (7)] {%% Perform Approximate Value Iteration until convergence}
    $p\left(\psi|G\right) = e^{\left\{\sum_{i=2}^{m} r_G(s_i)\right\}+V_G(s_m)-V_G(s_1)}$ {%% Probability of partial hand trajectory}
    $p\left(G\right) \leftarrow \frac{AFD_G}{\sum_{G\epsilon\mathbb{G}} AFD_G}$ {%% Prior of a goal using gaze data}
    $p\left(G|\psi\right) \leftarrow \frac{p(\psi|G)p(G)}{\sum_{G\epsilon\mathbb{G}} p(\psi|G)p(G)}$ {%% Combined probability of a goal using Bayes' theorem}
end for
$G_{pred} \leftarrow \ max\ p\left(G|\psi\right)$ {%% Predicted goal is the one with highest combined probability}
```

2.2 Features Identification

From Equation (4), it may be noted that reward distribution is a function of feature space. The following features were used to model human preferences during the task:

- Distance feature (f_d): To minimize effort, humans prefer the shortest path. The distance feature captures deviation from this path:

$$f_d\left(s\right) = e^{-\left\{d(s)-d_{shortest}\left(s'\right)\right\}} \tag{9}$$

where $d_{shortest}$ is the straight line path from start to end of the trajectory with states s'.

- Velocity feature (f_v): Given the close proximity of targets, relying solely on the distance feature is inadequate for distinguishing between targets. To address this, the velocity feature was introduced, which captures deviations from the desired hand velocity v_{des}. Desired velocity was determined based on previous work [17, 34], which reported velocity profiles concerning the distance to the endpoint. Linear, quadratic, and cubic velocity profiles were fit with respect to the distance to the endpoint. Table 1, indicates that quadratic velocity profile gives the lowest error and highest R^2. Therefore, the desired velocity at each state is given by:

$$v_{des}(s) = aX(s)^2 + bX(s) + c \tag{10}$$

where $X(s)$ represents the distance of state s from the end state of the given trajectory, and a,b,c are constants derived from quadratic polynomial fitting. Deviations from this desired velocity is the velocity feature:

$$f_v\left(s\right) = -\left(v_{des}\left(s\right) - v\left(s\right)\right)^2 \tag{11}$$

Table 1: Velocity Curve Fit

	Average R^2	Average error (cm/s)
Linear	0.80	10.33
Quadratic	**0.97**	**7.21**
Cubic	0.79	11.45

Considering hand dynamics and user comfort, acceleration $f_a(s)$ and jerk $f_j(s)$ features were introduced. These features were computed as the sum of squared acceleration and jerk at each state, respectively. Feature of a trajectory was expressed as the sum of features of each individual state of the trajectory. The above features could be implemented to model any task involving rapid aiming movement [1] through hand motion. All the features were normalized to (0,1) to have equal contributions to the learned reward function.

3 USER STUDY

User study was conducted to evaluate the performance of the target prediction model with the following input modalities:

(1) Hand: This examines the effectiveness of using MEDIRL for target prediction. The IRL model learns a reward distribution trained from hand motion data. This was used to predict the target.
(2) Eye: This examines reliability of eye gaze for target prediction. Average fixation duration was used to predict the target.
(3) Hand and eye: This examines the use of multiple input modalities for target prediction.

Participants: 10 participants (8 males and 2 females) were recruited from our university, averaging 27.9 years of age (SD: 4.2). Their average arm length was 57.83 cm (SD: 2.5), with 7 being right-handed and 3 left-handed individuals. None of the participants had color blindness. All participants provided necessary permissions and consent for the trials.

Setup: The experimental setup consists of four target bowls colored yellow, black, red and green as shown in Fig. 1b. $\{yellow, green\}$ were placed at a distance of $115cm$ and $\{black, red\}$ at a distance of $95cm$ from the user, symmetrically at proximity to each other. The user interface displays a random color from the target set, continuous hand tracking, target positions, and the robot's configurations. The target bowl size increases upon prediction, providing visual feedback. Additionally, it shows the number of iterations and the predicted target. The interface ran on a TV positioned at $3.2m$ away from the participant. A fixed-base robotic manipulator, Dobot Magician, was situated at a distance of $100cm$ from the user. The user sat on a chair positioned $40cm$ from the nearest edge of the robot's workspace as shown in Fig. 1c. Hand coordinates were tracked with Google Mediapipe [20] using a webcam. These coordinates were transformed into Cartesian coordinates relative to the April tag located at the bottom-left corner of the workspace using linear regression. User's eye gaze was captured using Tobii Glasses 2 eye tracker.

Design: The user's task involves reaching and placing a color block at the corresponding colored target. However, before the user's hand completely reaches the target, based on the partial motion, the intended target and the corresponding path to the target were predicted. The robot advances along the predicted trajectory to a fixed point for takeover. The user hands over the block to the robot, which subsequently places the block at the predicted target. This completes one iteration of the task. Each participant is required to carry out a total of 20 iterations, with 5 randomly designated for each target and varying input partial motion. For details please refer to the supplementary video (Link).

Procedure: Participants were briefed on distinct prediction scenarios — one involving solely hand movement, only eye gaze, and both. Participants were then asked to take trials. Following each trial, the robot was set to its home position and accuracy was calculated based on the number of blocks in the same colored target.

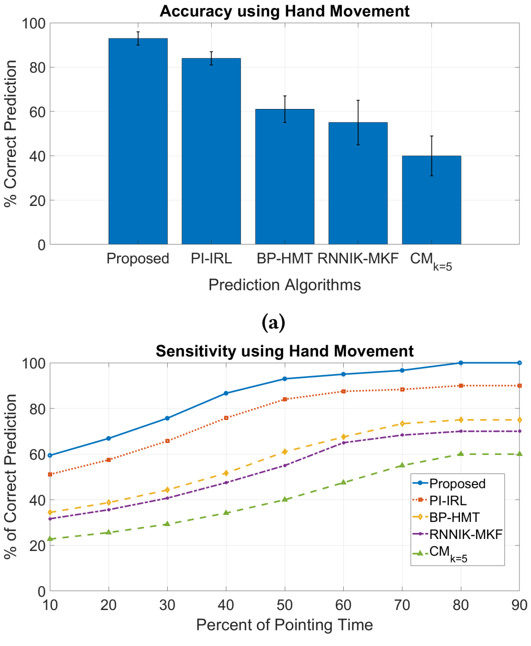
4 RESULT AND DISCUSSION

The expert dataset consists of 200 hand trajectories with velocity, acceleration, and jerk data at 30Hz. Corresponding 3D eye gaze data were recorded at 120Hz. The gaze data was downsampled to 30Hz by taking an average of 4 frames. A total of 180 trajectories were used for training and remaining 20 for testing. For prediction using hand motion, the proposed approach was validated with SOTA methods: $CM_{k=5}$ [12], Bayesian Predictor for Human Motion Trajectory (BP-HMT) [18], Recurrent Neural Network-Inverse Kinematics-Modified Kalman Filtering (RNNIK-MKF) [19], and Path Integral-Inverse Reinforcement Learning (PI-IRL) [29]. To evaluate target prediction, accuracy and sensitivity metrics [3] were used:

Accuracy: It is the percentage of correct target prediction (colored block in same color target) among all predictions. Average accuracy at a particular instant is the mean of accuracies obtained from that time to the end of the task.

Sensitivity: It signifies how quickly the intended target could be predicted. It is the accuracy obtained for different fractions of total pointing time.

Figures 2a and 2b depict the accuracy at 50% task completion and sensitivity, respectively, when relying on hand movement. As illustrated in Fig. 2a, the proposed approach employing MEDIRL achieves the highest target prediction accuracy of 93% at 50% task completion. In comparison, PI-IRL reaches a prediction accuracy of 84% while BP-HMT, RNNIK-MKF, and $CM_{k=5}$ report accuracies of 61%, 55% and 40%, respectively. Highest prediction accuracy using the proposed approach is due to the ability of the neural network in MEDIRL to learn complex reward functions during handover. BP-HMT and RNNIK-MKF exhibit lower accuracy due to their dependency on specific set of demonstrations, which hampers generalization to unseen hand movements. These methods require a large dataset for training compared to IRL. From Fig. 2b, the sensitivity remains below 80% during the initial 30% of the task but increases as the task progresses. This is due to the precise modeling of the handover task through the proposed set of features as more input data is available for prediction. Accuracy approaches 99.9% when less than 20% of the task remains.

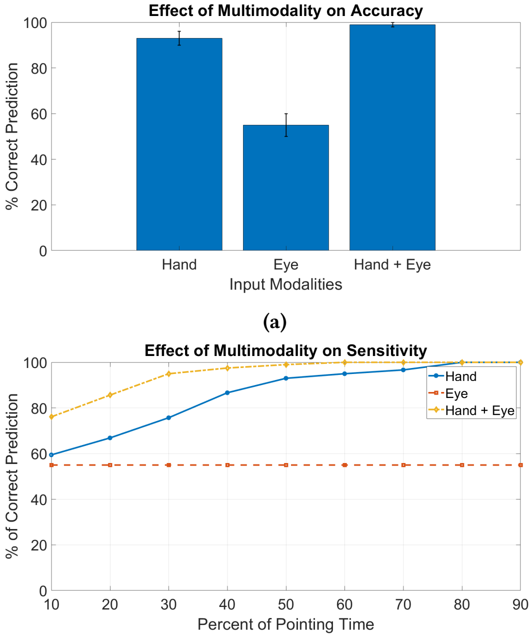


(a)

(b)

Figure 2: (a) Average target prediction accuracy at 50% task completion (b) Sensitivity as the task progresses, using hand movement for target prediction

Target prediction using only eye gaze reports an accuracy of 55% as shown in Fig. 3a. The low accuracy is attributed to the vergence gaze movement where black and red targets were visited, whereas the user was looking at yellow and green, respectively. The impact of multimodal target prediction using the proposed method on accuracy and sensitivity is illustrated in Fig. 3a and 3b. It can be noted that prediction accuracy of 99.1% is achieved using gaze and hand as input modality, at 50% task completion. Fig. 3b shows that the proposed method's sensitivity with hand and gaze inputs is comparatively higher than using only hand motion or gaze data. The multimodal target prediction algorithm achieves an accuracy of 99.9%, compared to 95% with only hand and 55% with gaze as input, when 60% of the task time has elapsed. This is attributed to the



(a)

(b)

Figure 3: Effect of multimodal target prediction on (a) Average target prediction accuracy at 50% task completion (b) Sensitivity, as the task progresses. Note that target was predicted at 2.5 seconds using eye gaze after the task commenced and does not depend on the pointing time.

use of gaze for providing additional information about the user's intent by assigning prior probabilities to targets in addition to the likelihood obtained from the partial hand motion. Hand motion intuitively conveys details about the user's intention, while eye gaze rapidly signals the intended target. Note that a uniform prior probability of 0.25 was assigned to the targets when only a single input modality was used.

The effectiveness of the proposed method was evaluated by comparing its results with those of relevant previous works across diverse HRI applications, as detailed in Table 2. Each algorithm was trained on application-specific data and was tested for the identical task. Despite the varied applications, a meaningful comparison of algorithms was facilitated by considering the average accuracy. The results indicate that the proposed method demonstrates a competitive average accuracy of 99.9% at 60% task completion when compared to other established methods in the field.

5 CONCLUSION

This work introduced a multimodal target prediction algorithm using Bayesian fusion to predict human hand movements and intended targets during handover. Maximum-Entropy Deep IRL (MEDIRL) was employed to learn the reward distribution, accompanied by a set of task-specific feature functions designed for capturing hand motion. User study was conducted to evaluate the proposed method involving a combination of hand and eye gaze. The proposed method demonstrated highest prediction accuracy compared to other methods. Subsequent efforts will utilize computer vision to directly estimate 3D gaze fixation within the fixed workspace of the robot. A more refined metric will be developed to effectively model eye gaze behavior for improved target prediction.

REFERENCES

[1] Bashar I Ahmad, Patrick M Langdon, Simon J Godsill, Robert Hardy, Lee Skrypchuk, and Richard Donkor. 2015. Touchscreen usability and input performance in vehicles under different road conditions: an evaluative study. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 47–54.

[2] Bashar I Ahmad, James K Murphy, Patrick M Langdon, Simon J Godsill, Robert Hardy, and Lee Skrypchuk. 2015. Intent inference for hand pointing gesture-based interactions in vehicles. *IEEE transactions on cybernetics* 46, 4 (2015), 878–889.

[3] Pradipta Biswas and Patrick Langdon. 2014. Multimodal target prediction model. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 1543–1548.

[4] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. 2018. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 4563–4570.

[5] Laura Cohen, Sinan Haliyo, Mohamed Chetouani, and Stéphane Régnier. 2014. Intention prediction approach to interact naturally with the microworld. In *2014 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. IEEE, 396–401.

[6] Brecht Corteville, Erwin Aertbeliën, Herman Bruyninckx, Joris De Schutter, and Hendrik Van Brussel. 2007. Human-inspired robot assistant for fast point-to-point movements. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, 3639–3644.

[7] Tor-Salve Dalsgaard, Jarrod Knibbe, and Joanna Bergström. 2021. Modeling Pointing for 3D Target Selection in VR. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. 1–10.

[8] Nachiket Deo and Mohan M Trivedi. 2020. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735* (2020).

[9] Jos Elfring, René Van De Molengraft, and Maarten Steinbuch. 2014. Learning intentions for improved human motion prediction. *Robotics and Autonomous Systems* 62, 4 (2014), 591–602.

[10] Ashraf Elnagar. 2001. Prediction of moving objects in dynamic environments using Kalman filters. In *Proceedings 2001 IEEE International Symposium on Computational Intelligence in Robotics and Automation (Cat. No. 01EX515)*. IEEE, 414–419.

[11] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*. 4346–4354.

[12] Nisal Menuka Gamage, Deepana Ishtaweera, Martin Weigel, and Anusha Withana. 2021. So predictable! continuous 3d hand trajectory prediction in virtual reality. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 332–343.

[13] Lu Gan, Jessy W Grizzle, Ryan M Eustice, and Maani Ghaffari. 2022. Energy-based legged robots terrain traversability modeling via deep inverse reinforcement learning. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8807–8814.

[14] Mithun Jacob, Yu-Ting Li, George Akingba, and Juan P Wachs. 2012. Gestonurse: a robotic surgical nurse for handling surgical instruments in the operating room. *Journal of Robotic Surgery* 6 (2012), 53–63.

[15] Mrinal Kalakrishnan, Peter Pastor, Ludovic Righetti, and Stefan Schaal. 2013. Learning objective functions for manipulation. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 1331–1336.

[16] Philipp Kratzer, Marc Toussaint, and Jim Mainprice. 2020. Prediction of human full-body movements with motion optimization and recurrent neural networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1792–1798.

[17] Edward Lank, Yi-Chun Nikko Cheng, and Jaime Ruiz. 2007. Endpoint prediction using motion kinematics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 637–646.

[18] Qinghua Li, Zhao Zhang, Yue You, Yaqi Mu, and Chao Feng. 2020. Data driven models for human motion prediction in human-robot collaboration. *IEEE Access* 8 (2020), 227690–227702.

[19] Ruixuan Liu and Changliu Liu. 2020. Human motion prediction using adaptable recurrent neural networks and inverse kinematics. *IEEE Control Systems Letters* 5, 5 (2020), 1651–1656.

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

Table 2: Comparison to Various Previous Works

Works	Algorithm	Mode of Intent Prediction	Interface	Average accuracy
Ours	Deep MEIRL	Hand and eye motion	Google Mediapipe and eye tracker	99.1% at 50%
				99.9% at 60%
Wang et. al [30]	MEIRL	Speech and gesture	Google cloud speech recognition	95% at 90%
Mitra et. al [27]	MEIRL	Hand motion	Mixed Reality	95.58% at 70%
Elfring et. al [9]	GHMMs	Body Pose	Vision system	90% at 90%
Wang et. al [31]	Kernel ELM	Speech and gesture	Natural language and wearable sensing	99.7% at 90%
Gamage et. al [12]	Classical kinematics	Hand motion	Virtual Reality	88% at 90%
Mainprice et. al [23]	GMM	Body Pose	Simulation (MATLAB)	92% at 90%

[21] Ruikun Luo and Dmitry Berenson. 2015. A framework for unsupervised online human reaching motion recognition and early prediction. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2426–2433.

[22] Yusuke Maeda, Takayuki Hara, and Tamio Arai. 2001. Human-robot cooperative manipulation with motion estimation. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, Vol. 4. Ieee, 2240–2245.

[23] Jim Mainprice and Dmitry Berenson. 2013. Human-robot collaborative manipulation planning using early prediction of human motion. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 299–306.

[24] Jim Mainprice, Rafi Hayne, and Dmitry Berenson. 2016. Goal set inverse optimal control and iterative replanning for predicting human reaching motions in shared workspaces. *IEEE Transactions on Robotics* 32, 4 (2016), 897–908.

[25] Omey M Manyar, Zachary McNulty, Stefanos Nikolaidis, and Satyandra K Gupta. 2023. Inverse Reinforcement Learning Framework for Transferring Task Sequencing Policies from Humans to Robots in Manufacturing Applications. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 849–856.

[26] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2891–2900.

[27] Mukund Mitra, Preetam Pati, Vinay Krishna Sharma, Subin Raj, Partha Pratim Chakrabarti, and Pradipta Biswas. 2023. Comparison of Target Prediction in VR and MR using Inverse Reinforcement Learning. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 55–58.

[28] Anneli Olsen. 2012. The Tobii I-VT fixation filter. *Tobii Technology* 21 (2012), 4–19.

[29] Sibo Tian, Xiao Liang, and Minghui Zheng. 2023. An Optimization-Based Human Behavior Modeling and Prediction for Human-Robot Collaborative Disassembly. In *2023 American Control Conference (ACC)*. IEEE, 3356–3361.

[30] Weitian Wang, Rui Li, Yi Chen, Z Max Diekel, and Yunyi Jia. 2018. Facilitating human–robot collaborative tasks by teaching-learning-collaboration from human demonstrations. *IEEE Transactions on Automation Science and Engineering* 16, 2 (2018), 640–653.

[31] Weitian Wang, Rui Li, Yi Chen, Yi Sun, and Yunyi Jia. 2021. Predicting human intentions in human–robot hand-over tasks through multimodal learning. *IEEE Transactions on Automation Science and Engineering* 19, 3 (2021), 2339–2353.

[32] Zhan Wang, Patric Jensfelt, and John Folkesson. 2015. Modeling spatial-temporal dynamics of human movements for predicting future trajectories. In *Workshop at the Twenty-Ninth AAAI Conference on Artificial Intelligence," Knowledge, Skill, and Behavior Transfer in Autonomous Robots", AAAI Conference on Artificial Intelligence, Austin, USA, January 25, 2015*. Association for the advancement of Artificial Intelligence.

[33] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. 2015. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888* (2015).

[34] Brian Ziebart, Anind Dey, and J Andrew Bagnell. 2012. Probabilistic pointing target prediction via inverse optimal control. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. 1–10.

[35] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.

[36] Athanasia Zlatintsi, Isidoros Rodomagoulakis, Petros Koutras, AC Dometios, Vassilis Pitsikalis, Costas S Tzafestas, and Petros Maragos. 2018. Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3171–3175.