Gaze-based Human Intent Estimation for Intelligent Prosthetic Arm

Zhongchun Yu
Department of Mechanical Engineering
Stanford University
Stanford, United States
zcyu@stanford.edu

Yiyang Mu
Department of Mechanical Engineering
Stanford University
Stanford, United States
yiyangmu@stanford.edu

***Abstract*—This paper presents a novel approach to human intent estimation using gaze-based input for controlling a semi-autonomous prosthetic arm in simulated environments. Our method employs a Gaussian Hidden Markov Model (HMM) to interpret the temporal patterns of gaze behavior as proxies for the user's proximal intentions, such as picking and placing objects during a color sorting task. The system uses AR headsets to enhance interaction and intuitive control, aligning visual focus with the prosthetic's operations. The effectiveness of gaze-based intent estimation is evaluated through a task that tests the prosthetic's precision and spatial awareness, highlighting the potential of gaze as a robust control modality in assistive robotics. The proposed approach aims to provide intelligent prosthetic devices capable of inferring human intent, thereby making them more accessible and effective for individuals with disabilities.**

***Index Terms*—human–robot collaboration, intention inference, probabilistic methods, human–robot interaction**

I. Introduction

The increasing interest in human-robot collaboration and shared autonomy arises from the growing need for advanced assistive devices that can compensate for the limitations of human users and significantly aid in various applications [1], [2], [3], [4]. In the specific field of assistive robotics, such as prosthetic arms, human users depend on robotic assistance for daily tasks. However, given the limited actuation input channels from human users and the need for advanced control mechanisms to perform dexterous tasks, it is crucial to enable the robot to manage lower-level control while accurately inferring the human's intent behind their high-level goals.

In this work, we focus on gaze-based human intent estimation in an intelligent (semi-autonomous) robotic prosthetic arm within a simulated scenario to investigate the effectiveness in task execution by bridging the gap between the degrees of freedom (DOF) in robotic actions and user inputs. This will be achieved through an integrated robot control interface and a proposed discrete Gaussian Hidden Markov Model (HMM) based on gaze tracking for goal inferring. The system leverages AR headsets, which have shown significant potential as an input modality [5] and testing setup [6], providing users with a visual representation of the robotic arm and task space while simultaneously facilitating gaze tracking to infer user goals.

II. Related Works

Intent Estimation

Intention inference with discrete states and actions has been extensively studied through various probabilistic approaches. Common methods include Bayesian inference, Sequential Decision Making, and Hierarchical Planning within Probabilistic Models. Bayesian methods employ Bayes' Theorem to determine the probability of a future event based on past observations, integrating prior knowledge with new data to form a posterior distribution for more accurate state estimation and decision-making. This method follows a structured workflow: starting from the prior distribution, updating it with the likelihood from new measurements, and combining these to form a posterior distribution [7] [8]. The Sequential Decision Making is a process in which decisions are made in stages over time, with each stage capturing new information from observations that could influence subsequent decisions. Typical model formulations for the Sequential Decision Making include Markov Decision Process (MDP), Partially Observable Markov Decision Process (POMDP), and Hidden Markov Models (HMM). The MDP serves as a fundamental model for sequential decision making, which requires fully observable states, a completely known transition model, and a stationary (i.e., time-independent) environment. However, these prerequisites render MDPs less suitable for robust decision making and planning for complex robotic systems in stochastic and unknown environments. The POMDP is an extension of the MDP that addresses situations where the agent does not have complete information about the current state of the system. Since the true state of the system is not fully observable, a POMDP would maintain an explicit probability distribution to estimate the current state. This probability distribution will be updated based on the actions taken and the observations received at each stage [9] [10]. The HMM is a more generalized model for sequential decision making, wherein the system is presumed to follow a Markov process with non-observable (i.e., hidden) states. In an HMM, the underlying states are not directly observed but can be inferred from past observations. This attribute makes HMM particularly useful in real-world scenarios where the state of a system cannot be directly measured [11]. The Hierarchical Planning within probabilistic models refers to a structured approach that breaks down a complex decision-making problem into a hierarchy of smaller, more manageable subproblems. It could incorporate with MDPs, POMDPs, HMMs, and other probabilistic models to handle uncertainty in state transition and observations [12].

While Bayesian methods and others provide robust frameworks for intent estimation, the Hidden Markov Model (HMM) has been chosen for its specific advantages in handling temporal sequences where the state is not directly observable. HMMs allow for the modeling of hidden states through observable sequences, which is particularly advantageous in scenarios like gaze tracking where the user's intent must be inferred from indirect cues. The HMM's ability to manage the temporal dependencies and uncertainties inherent in such data makes it a preferred choice over other probabilistic models, which might not as effectively handle the sequential and latent aspects of the data.

III. Proposed Methods

A. Setup

Framework: Our platform (Fig. 1), is built on the Robot Operating System (ROS) Noetic, operating on a Linux computer. This setup communicates over WiFi with an 8-camera OptiTrack PrimeX 13 motion-capture system, connected to a Windows computer running the Motive Tracker software. Additionally, a custom mixed-reality application developed in Unity runs on a Microsoft HoloLens 2 AR headset using the open-source ROS framework.

In our system, the prosthetic arm is simulated in Gazebo along with the task workspace. The user visualizes both the arm and the objects to be manipulated through the AR headset. Motion capture markers placed near the shoulder serve as reference points, enabling the virtual arm to accurately follow the participants' body movements, making it appear as if it is worn on their body at all times.

Control Interface: For the baseline experiment, the user directly controls the end-effector pose of the arm to pick and place the selected object using a joystick. In the gaze-based experiment, object selection is determined by the intent estimation model, and the pick-and-place motion planning is executed using the MoveIt framework. EMG electrodes placed on the ipsilateral forearm are used to send confirmation signals for object selection, ensuring precise and reliable control of the prosthetic arm.

B. Gaze-based Intent Estimation Model

Our method is designed to infer human intent using early gaze cues that indicate the proximal intention [13] [14] for a pick-and-place task.

The temporal patterns of gaze are effectively modeled using a Gaussian Hidden Markov Model (GHMM), where the hidden states, denoted as $X(t)$, capture the underlying intentions of the user. These hidden states may reflect the user's perceptual focus, such as attention directed toward objects to be picked up or specific target locations for placement. The predictions from the Gaussian Hidden Markov Model (GHMM) are enhanced by the model's understanding of context, which significantly relies on prior knowledge. This context-awareness, encapsulating specific sequences of block movements or color preferences in tasks like color sorting, allows the GHMM to make more accurate state transitions and infer user intent with greater precision.

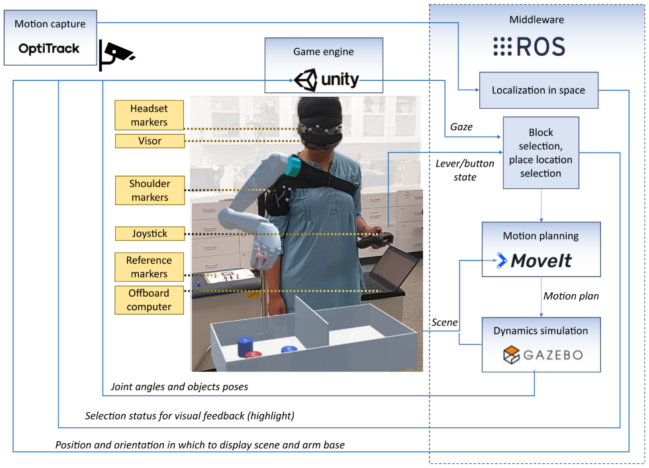

Fig. 1. System architecture and components: a non-amputee participant operates the arm using a joystick. The block in the hand was highlighted and the reach motion planned using MoveIt

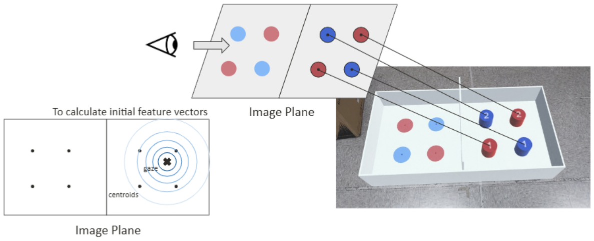

Fig. 2. Diagram illustrating the process of calculating initial feature vectors for AoI-likelihood estimation in gaze-based intent detection. The image plane displays how 3D scene elements (blocks of different colors) are projected as 2D gaze points. Each block's centroid is marked to compute the distance from the gaze points.

1) AoIs-Likelihood Feature Computation: For the gaze-based intent estimation, we calculate the likelihoods of gaze points, g_t, relative to predefined Areas of Interest (AoIs) using a multivariate Gaussian distribution (see Fig. 2). Initially, we project the 3D scene onto a 2D image plane and label the gaze points according to the AoIs. In this setup, the centroid of each block is designated as an AoI.. The gaze points are modeled as:

$$g_t \sim \mathcal{N}(\mu_t, \Sigma) \tag{1}$$

where μ_t represents the (u, v) coordinates of the gaze at time t and $\Sigma = \sigma^2 I_2$ signifies an isotropic covariance, indicating equal variance in both u and v directions, symbolizing a circular uncertainty region around the gaze point.

The likelihood of each AoI given the gaze point g_t is computed as:

$$P(\mathrm{AoI}_i \mid g_t) = \frac{\exp(-\frac{1}{2}(g_t - \mathrm{AoI}_i)^T \Sigma^{-1}(g_t - \mathrm{AoI}_i))}{\sum_j \exp(-\frac{1}{2}(g_t - \mathrm{AoI}_j)^T \Sigma^{-1}(g_t - \mathrm{AoI}_j))} \tag{2}$$

The likelihoods are normalized across all AoIs to ensure that the sum of likelihoods for each gaze point equals one, thus forming a valid probability distribution.

This yields a feature vector F_t for each gaze point:

$$F_t = [P(\mathrm{AoI}_1 \mid g_t), \ldots, P(\mathrm{AoI}_n \mid g_t)] \tag{3}$$

where each component represents the normalized likelihood of the gaze being directed towards each AoI. This probabilistic framework quantifies attention distribution across the AoIs, facilitating accurate intent estimation from gaze data.

2) Model Framework: The model framework is shown in Fig. 3.

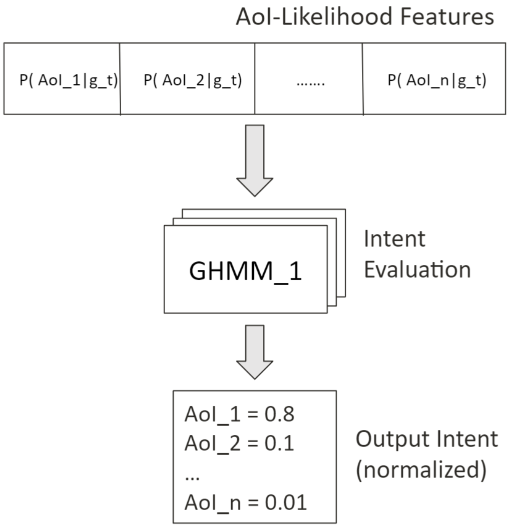

Fig. 3. Schematic of the Gaussian Hidden Markov Model (GHMM) workflow for intent estimation from gaze data

The observable gaze sequence in our model is governed by a Hidden Markov Process, where the observation vector comprises the AoI likelihoods of the target blocks and potential placement positions. These likelihoods are derived from a multivariate Gaussian distribution and are assembled into observation vectors representing the probability distributions of objects under gaze at different time steps. Governed by the emission probabilities of the Hidden Markov Model (HMM), these distributions reflect the current state of the hidden variables, allowing the model to output the likelihood of user intents by analyzing these sequences, thereby providing insights into intended actions based on gaze patterns.

Each intent within the model correlates with an Area of Interest (AoI) as defined by the task configuration. A dedicated Gaussian Hidden Markov Model (GHMM) is employed for each intent, incorporating a number of internal states that depend on the AoIs. Transition and emission priors are learned and iteratively refined throughout the training process.

For this study, both the training and testing data are simulated offline. It is assumed that all picking and placing actions within these simulations are executed flawlessly, ensuring a controlled evaluation environment for the GHMMs.

IV. Experiment Design

This experiment serves as a functional test to measure the gross manual dexterity of individuals using our semi-autonomous prosthetic arm based on the proposed gaze-based human intent estimation model. .

A. Task Description

Our approach modifies the Box and Blocks Test to reduce the influence of compensatory body movements, thereby providing a more precise assessment of the prosthetic arm's capabilities

In our new task design (Fig. 4), blocks of mixed colors are placed arbitrarily within a drawer. The task subject, using a simulated prosthetic arm, must sort and transport the blocks by color. Each block must be placed in its designated color-specific drawer as quickly and accurately as possible. The drawers are divided into grid-like sections to better analyze space occupancy and the arm's performance.

In the initial phase of the task design, we will employ two drawers, each divided into four empty sections, and provide a total of four cylindrical blocks (two blocks per color for placement). Progressing to subsequent stages, we will introduce additional areas of interest (AoIs) to incrementally increase the task complexity.

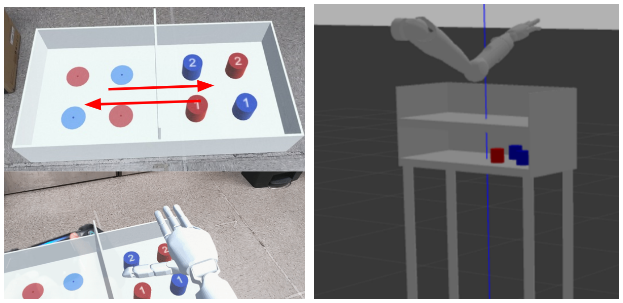

Fig. 4. Illustration of the task design: original BBT (left) and the new color sorting task (right)

B. Rationale for Task Design

The color sorting task design serves two primary purposes: it tests the functionality of the prosthetic arm by requiring precise movements that reduce compensatory body motions, and it incorporates occlusions and prior knowledge of drawer configurations to rigorously assess our intent estimation model.

V. Data Preparation

In this experiment, we utilized simulated gaze data projected onto an image plane to emulate a human user interacting with our system. This simulation approach was carefully selected to expedite the research process by bypassing the extended approval periods typically required for human subject research. Additionally, this method allows us to focus more intently on refining the gaze-based human intent estimation component within a controlled environment.

It is important to note that the simulated gaze points are generated based on previously collected data, ensuring a realistic approximation of human gaze behavior. However, there may still be discrepancies, particularly in the dynamics of state transitions, which do not fully capture the natural variability present in actual human gaze patterns. This acknowledgment is crucial as it frames the limitations of our simulated data and its impact on the Gaussian Hidden Markov Model (GHMM) testing and system development phases.

A. BBT - Configuration 1

Initially, we replicate the traditional Block Box Test (BBT) as a baseline study, which is depicted in Fig. 5. In this initial setup, each block is distinctly positioned without overlap, ensuring clear separation and accessibility. The layout consists of four designated blocks for picking on the right side and four corresponding placement targets on the left side. Regarding the temporal dynamics of data generation, we model the gaze movements to sequentially transition through the blocks—starting from block 1 to block 4. The gaze path for each block is structured to progress from the picking phase $(pick_i)$ to the placement phase $(place_i)$. Additionally, we assign a normalized period for each target, with this configuration utilizing 100 timestamps per intent.

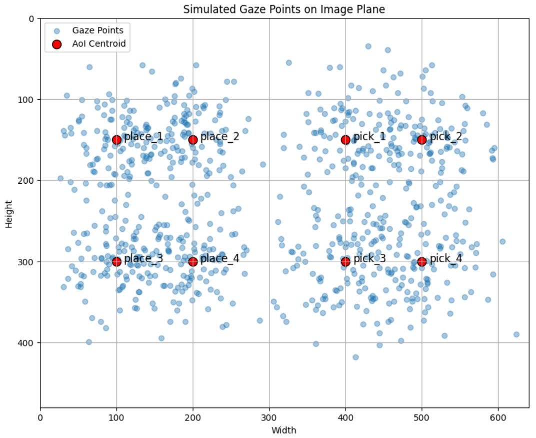

Fig. 5. Simulated gaze points for BBT - configuration 1

B. BBT - Configuration 2

This configuration modifies the original Block Box Test (BBT) setup (see Fig. 6), closely aligning with Configuration 1, but with an intentional adjustment: the pick targets are positioned closer together. This alteration is designed to test the robustness of the intent estimation model under conditions that may challenge or confuse the baseline predictions. By reducing the distance between pick targets, we aim to evaluate the model's accuracy in distinguishing between closely situated intents, thereby providing insights into its effectiveness in more complex scenarios.

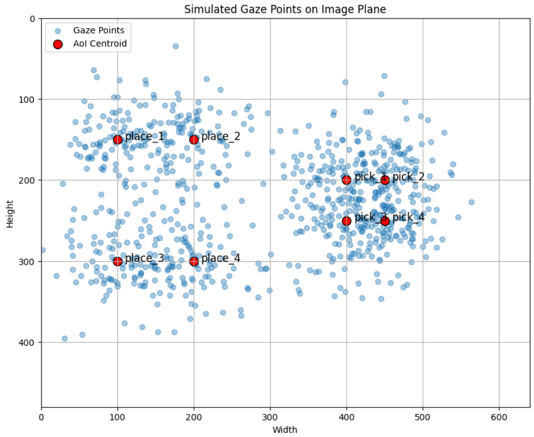

Fig. 6. Simulated gaze points for BBT - configuration 2 (Closer AoIs)

C. Color Sorting Task - Configuration 3

In this configuration, we execute the color sorting task, focusing on a single layer within a drawer, assumed to be designated for blue items. As illustrated in Figure 7, the scenario includes a simplified setup with three blocks: two blue and one partially occluded red block. We utilize this arrangement to simulate potential real-world conditions where block visibility may be compromised. In the corresponding simulated data depicted in Fig. 8, we define five Areas of Interest (AoIs) labeled as $empty_1$, $blue_1$, red_1, $blue_2$, and $empty_2$, respectively. This configuration helps us explore the gaze-based model's effectiveness in identifying and differentiating between occupied and vacant spaces within a cluttered environment.

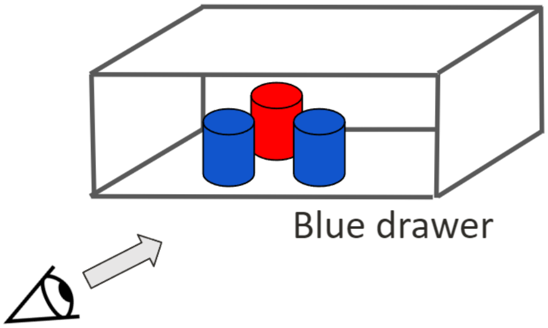

Fig. 7. Illustration of the drawer setup with color-specific blocks and designated AoIs for the gaze simulation

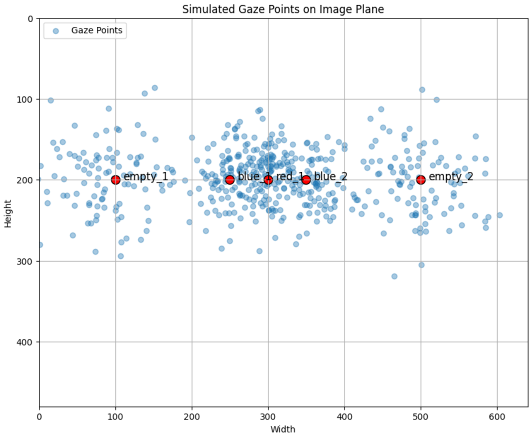

Fig. 8. Simulated gaze points for Color Sorting Task - configuration 3

VI. Results

Baseline Model: To compare the intent estimation results, we establish a baseline model using the initial likelihoods of Areas of Interest (AoIs) calculated directly from the raw gaze data. These initial calculations serve as the foundation for our comparison, representing the simplest form of intent inference based solely on unprocessed gaze observations.

Prior Knowledge: For Configurations 1 and 2, prior knowledge about the sequence of block movements for pick-and-place tasks was utilized. In Configuration 3, we assumed blue drawer which increases the likelihood that the intent is focused on the red block, possibly to remove it and make space. Additionally, we operated under the assumption of a grasping state, suggesting that the blocks are more likely targets of intent rather than the empty slots.

A. Regular BBT - Configuration 1

The initial task setup includes four pick-up positions with clear visibility and no occlusion. Fig. 5 displays the simulated gaze points and the centroids of the AoIs for each pick and place operation. Fig. 9 shows the likelihoods of each AoI for each pick, derived from the temporal gaze data.

The performance of our model is quantified using a confusion matrix, presented in Fig. 10. This matrix compares the accuracy of predictions from our approach against true labels. In an ideal scenario, the main diagonal of the confusion matrix, representing correct predictions, would prominently feature higher values indicating accurate recognition, while non-diagonal elements, indicating errors, would register lower values.

However, for the baseline method where no intent inference is employed, the confusion matrix shows a spread of values not confined to the diagonal, suggesting frequent misidentification of blocks as their adjacent counterparts (e.g., confusing Block 1 with Block 2, or Block 3 with Block 4). In contrast, the confusion matrix from the GHMM shows higher values along the diagonal and minimal values elsewhere, indicating more accurate predictions and a clear distinction between the intended picks. This demonstrates the effectiveness of our model in accurately inferring user intent.

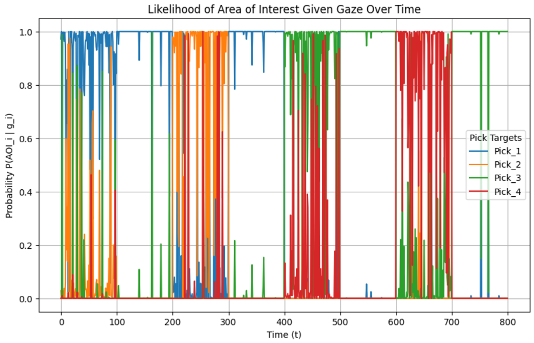

Fig. 9. Likelihood for basic BBT

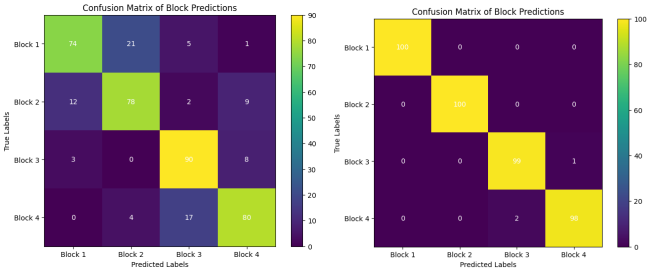

Fig. 10. Normalized confusion matrices for the baseline (left) and GHMM (right) models for configuration 1. The baseline model tends to make errors when classifying AoIs with closer locations (e.g. pick_1 and pick_2, pick_3, and pick_4.

B. BBT with Closer AoIs - Configuration 2

This task also features four pick-up positions, but with the blocks positioned closer together, increasing the challenge of accurate intent discrimination. Fig. 6 illustrates the simulated gaze points and the centroids of the AoIs for each operation. Fig. 11 presents the likelihood of each AoI calculated from the gaze data over time. The effectiveness of our intent estimation model compared to the baseline is shown in Fig. 12. Our model demonstrates significantly clearer results under the more challenging close proximity of blocks. Notably, the performance of the baseline model decreases significantly as the distance between the blocks decreases, highlighting its limitations in handling closely spaced AoIs.

C. Color Sorting Task - Case 3

The color sorting task is similar to BBT, but it sorts the blocks by color. This task allows occlusion of blocks and prior knowledge about the likelihood of the color of the blocks that have been moved, which is more close to real life scenario. Fig. 8 shows the simulated gaze points for this task.

Fig. 13 shows the likelihood of the AoI for each pick given the gaze input derived from the gaze data over time. Fig. 14 presents the comparative results between the baseline model and our intent prediction approach. The findings indicate that

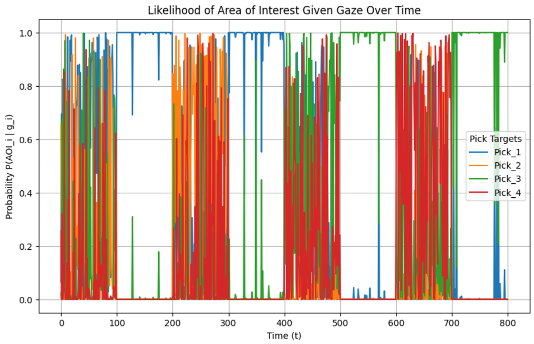

Fig. 11. Likelihood for BBT with close AoIs

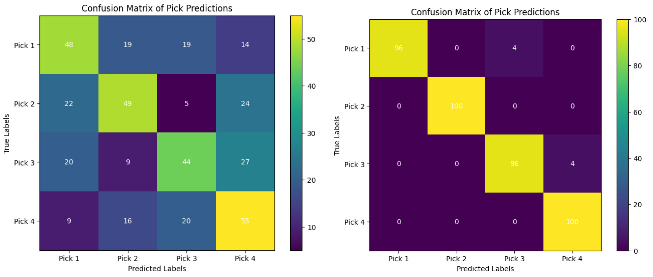

Fig. 12. Normalized confusion matrices for the baseline (left) and GHMM (right) models for configuration 2

our model significantly enhances the accuracy of selecting occluded objects by using prior knowledge about the task, demonstrating the practical benefits of integrating contextual understanding into intent prediction models.

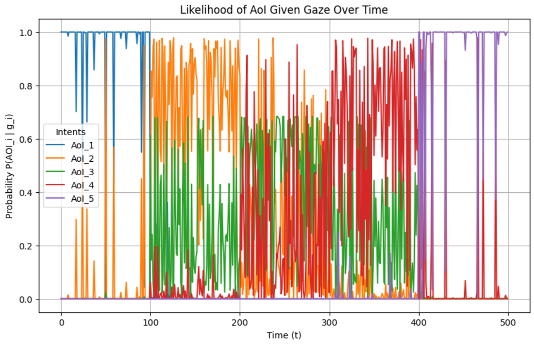

Fig. 13. Likelihood for Color Sorting Task

VII. Conclusion and Future Work

For future developments, we plan to validate the efficacy of our approach in a real-world setup by utilizing actual gaze data, moving beyond the current simulations. Furthermore, we aim to enhance the versatility of the Gaussian Hidden Markov Model (GHMM) by transitioning it to an action-based model (e.g., picking or grasping), which will enable it to adapt to a broader range of potential states. Lastly, we intend to refine the model to handle dynamic changes in the environment, such as replacing blocks with those of different colors, thereby increasing its applicability to varied and evolving tasks.

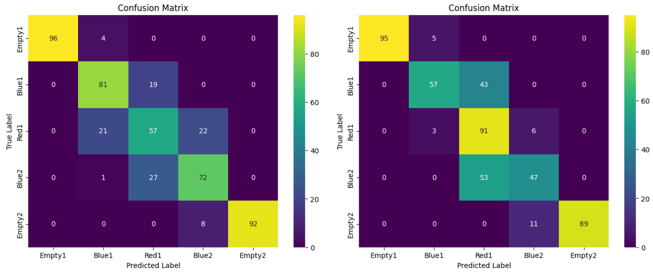

Fig. 14. Normalized confusion matrices for the baseline (left) and GHMM (right) models for configuration 3

References

[1] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. Behal, "How Autonomy Impacts Performance and Satisfaction: Results From a Study With Spinal Cord Injured Subjects Using an Assistive Robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, pp. 2–14, Jan. 2012.

[2] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, pp. 717–742, June 2018.

[3] S. Jain, A. Farshchiansadegh, A. Broad, F. Abdollahi, F. Mussa-Ivaldi, and B. Argall, "Assistive robotic manipulation through shared autonomy and a Body-Machine Interface," in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*, (Singapore, Singapore), pp. 526–531, IEEE, Aug. 2015.

[4] D. Gopinath, S. Jain, and B. D. Argall, "Human-in-the-Loop Optimization of Shared Autonomy in Assistive Robotics," *IEEE Robotics and Automation Letters*, vol. 2, pp. 247–254, Jan. 2017.

[5] M. Markovic, S. Dosen, C. Cipriani, D. Popovic, and D. Farina, "Stereovision and augmented reality for closed-loop control of grasping in hand prostheses," *Journal of Neural Engineering*, vol. 11, p. 046001, Aug. 2014.

[6] A. Sharma, W. Niu, C. L. Hunt, G. Levay, R. Kaliki, and N. V. Thakor, "Augmented Reality Prosthesis Training Setup for Motor Skill Enhancement," 2019.

[7] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and others, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 1, 2021. Publisher: Nature Publishing Group UK London.

[8] Z. Wang, K. Mülling, M. P. Deisenroth, H. Ben Amor, D. Vogt, B. Schölkopf, and J. Peters, "Probabilistic movement modeling for intention inference in human–robot interaction," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 841–858, 2013. Publisher: SAGE Publications Sage UK: London, England.

[9] K. P. Murphy, "A survey of POMDP solution techniques," *environment*, vol. 2, no. 10, 2000.

[10] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online POMDP planning for autonomous driving in a crowd," in *2015 ieee international conference on robotics and automation (icra)*, pp. 454–460, IEEE, 2015.

[11] S. R. Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996. Publisher: Elsevier.

[12] J. K. Lenstra, A. Rinnooy Kan, and L. Stougie, "A framework for the probabilistic analysis of hierarchical planning systems," *Annals of Operations Research*, vol. 1, pp. 23–42, 1984. Publisher: Springer.

[13] J. D. Velleman and M. E. Bratman, "Intention, Plans, and Practical Reason.," *The Philosophical Review*, vol. 100, p. 277, Apr. 1991.

[14] M. Land, N. Mennie, and J. Rusted, "The Roles of Vision and Eye Movements in the Control of Activities of Daily Living," *Perception*, vol. 28, pp. 1311–1328, Nov. 1999.