2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

# A 3D Vector Field and Gaze Data Fusion Framework for Hand Motion Intention Prediction in Human-Robot Collaboration

Maleen Jayasuriya[1], Gibson Hu[1], Dinh Dang Khoa Le[1], Karyne Ang[1], Shankar Sankaran[1], Dikai Liu[1]

*Abstract*— **In human-robot collaboration (HRC) settings, hand motion intention prediction (HMIP) plays a pivotal role in ensuring prompt decision-making, safety, and an intuitive collaboration experience. Precise and robust HMIP with low computational resources remains a challenge due to the stochastic nature of hand motion and the diversity of HRC tasks. This paper proposes a framework that combines hand trajectories and gaze data to foster robust, real-time HMIP with minimal to no training. A novel 3D vector field method is introduced for hand trajectory representation, leveraging minimum jerk trajectory predictions to discern potential hand motion endpoints. This is statistically combined with gaze fixation data using a weighted Naive Bayes Classifier (NBC). Acknowledging the potential variances in saccadic eye motion due to factors like fatigue or inattentiveness, we incorporate stationary gaze entropy to gauge visual concentration, thereby adjusting the contribution of gaze fixation to the HMIP. Empirical experiments substantiate that the proposed framework robustly predicts intended endpoints of hand motion before at least $50\%$ of the trajectory is completed. It also successfully exploits gaze fixations when the human operator is attentive and mitigates its influence when the operator loses focus. A real-time implementation in a construction HRC scenario (collaborative tiling) showcases the intuitive nature and potential efficiency gains to be leveraged by introducing the proposed HMIP into HRC contexts. The open-source implementation of the framework is made available at `https://github.com/maleenj/hmip_ros.git`.**

## I. INTRODUCTION

During general human collaboration, predicting the intention behind human hand motion is a fundamental skill that we as humans carry out at an implicit, subconscious, and reactive level based on social/physical cues such as gaze and hand motion trajectories. Such an implicit predictive capability of the intent of hand motion can substantially aid with both the efficiency and safety of human-robot collaboration (HRC) tasks as it provides valuable information necessary for high-level decision-making and planning. Thus, a robot with fast hand motion intention prediction can ensure a more intuitive collaborative experience for all parties involved.

The problem of hand motion intention prediction (HMIP) in the context of HRC is typically framed in terms of matching an intended action of a human participant with the robot's model of a known set of actions or intended end points (see Section II for more details). State-of-the-art approaches in this regard have been made by utilising deep learning-based approaches such as RNNs. However, accurate and robust HMIP, which is generalisable beyond repetitive tasks, remains a difficult challenge due to the stochastic nature of human hand motion. Furthermore, HMIP is a low-level function that typically informs more resource-intensive high-level tasks such as decision-making, planning and execution. Thus, an ideal HMIP mechanism has to be fast and require fewer computational resources.

Our previous work [1] proposed a vector field-based representation of hand motion for endpoint intent prediction in 2D workspaces. The primary utility of this method was its low computational cost and scalability. However, it is clear that for more complex interactions in 3D, a multimodal approach from a variety of human signals can improve the robustness and predictive capabilities.

Eye tracking, in particular, can provide valuable information regarding human intention and focus. Related work incorporating gaze information in the context of HMIP primarily utilises gaze fixation under the assumption that fixating on a particular endpoint is predictive of the intention to move to that endpoint (see Section II). This assumption may be challenged when a human operator is unfocused or under high cognitive load, causing random or unfocused saccadic eye movements. To counteract this effect, the proposed framework utilises stationary gaze entropy (SGE), which has been shown to be a measure of human visual focus, to weigh the contribution of gaze fixation before fusing it with the proposed hand-tracking approach. Thus, the contributions of this paper are as follows:

- A 3D vector field-based method to represent hand motion and predict endpoint intentions based on minimum jerk trajectory (MJT) model-based predictions.
- A stationary gaze entropy (SGE) based weighted gaze fixation approach for HMIP.

The above modalities, which can operate independently, are combined using a Naive Bayes Classifier (NBC) to provide a statistically consistent HMIP framework. The remainder of this paper is structured as follows: Section II discusses the related work relevant to HMIP. Section III presents the 3D vector fields approach. Section IV describes the SGE-based weighting of gaze fixation and NBC-based framework. An evaluation of the proposed framework and a real-time implementation in an HRC construction scenario, i.e: collaborative tiling, is presented in V. Finally, Section VI concludes the paper with remarks on the experimental results and thoughts on future work.

## II. RELATED WORK

Approaches to human motion intention prediction can broadly be categorised into explicit model-based approaches and learning-based approaches.

[1] Associated to the Robotics Institute (RI), University of Technology Sydney, Australia.

Explicit model-based approaches attempt to predict human motion and intention by utilising explicitly defined dynamic equations based on the physics of human motion. For instance, the minimum torque change model proposed by Uno et al. [2] attempts to model human movement by minimizing the time derivative of joint torques. This requires explicitly defining the dynamics equations of the musculoskeletal system. An alternative modelling approach known as the minimum jerk trajectory (MJT) model proposed by Flash et al. [3] proposes that human arms tend to follow a path between two points that minimizes the third-order derivative of position (i.e jerk). This approach does not require an explicit definition of the musculoskeletal system and has been successful in describing limb motion based on central nervous system (CNS) data from both human and animal studies. The MJT model was used by Landi et al. [4] to predict targets of human reaching motion, and by Dinh et al. [5] for local obstacle avoidance in HRC scenarios. Due to the difficulty of mathematically modelling the erratic behaviour of human arm motions, these approaches are limited to short-term trajectory predictions in repetitive tasks in highly controlled settings that include motion capture or VR systems.

In contrast, learning-based approaches are gaining popularity due to their ability to capture complex non-linear relationships. Machine learning approaches such as Hidden Markov Models (HMMs) have been used in predicting human intentions in hierarchical human-robot collaborative assembly tasks [6]–[8]. Gaussian Mixture Models (GMMs) have also been utilised to classify trajectories in repetitive tasks as a means to predict endpoints of human hand motion [9]–[11].

Deep learning approaches such as Recurrent Neural Networks (RNNs) have recently emerged as powerful tools capable of predicting sequences based on time series data and have been applied to human motion prediction [12]. Formica et al [13] utilised RNNs to predict the destination of human hand motion in a pick-and-place case study. Improving on basic RNNs, Long Short-Term Memory (LSTM) based RNN networks have also been utilised to predict human hand motion [14], [15]. However, learning-based techniques as a whole are computationally expensive and require large training data sets to generalise enough to capture previously unseen trajectories or sudden changes in hand motion. Furthermore, deep learning based approaches lack uncertainty information about the predictions vital for higher-level decision-making and planning purposes. The transparency of the prediction, which these techniques do not offer, is also a critical issue concerning safety in HRC applications where humans and robots collaborate with physical contact.

Eye tracking and gaze data is another modality that is gaining popularity due to the multitude of information it provides about human intention and behaviour. Zhou et al. [16] combined hand motion trajectories and gaze focus trajectories to cluster and predict hand motion patterns based on a deep learning technique in a VR simulation. Choi et al. [17] used a similar gaze-hand relationship to predict human hand motion for object handover tasks. However, gaze-related work in this domain has been limited to gaze metrics such as fixation and focus trajectories. More informative metrics, such as stationary gaze entropy (SGE) [18] and transitionary gaze entropy (TGE) [19], have been largely ignored. SGE and TGE have been shown to be very powerful metrics in understanding human attention and visual focus [20]. The utility of these entropy metrics has been demonstrated in applications such as multiple object tracking in air traffic control tasks [21] and capturing human trust in HRC applications [22]. We hypothesize that these metrics can provide valuable information for human hand intention prediction.

Our previous work [1] utilised vector fields to represent human hand motion trajectories in 2D workspaces for HMIP. Although theoretically, a trajectory is a continuous moving path that can be represented by a sequence of vectors, sensor noise and discretisation errors can make such a vector path unrepresentative of the nature of the motion. Representing trajectory data using a vector field has been shown to be more robust as it represents the likeness of where a trajectory appears, capturing the overall flow, especially in the context of similarity calculations [23]. Furthermore, the computational scalability of vector field data and the ability to represent different trajectories with the same dimensionality makes vector field representations ideal for tasks such as clustering, analysis, and similarity query [24]. Such vector field-based trajectory analyses have been utilised widely in predicting pedestrian and vehicle motion paths, primarily in a 2D setting [25], [26]. To the best of our knowledge, 3D vector field representations have not been utilised in the context of analysing and predicting human hand motion in HRC applications.

The proposed framework (See Figure 1) combines a vector field-based 3D representation of hand motion trajectories with gaze metrics to provide a robust framework that predicts the intended endpoint target of hand motion in an HRC setting.

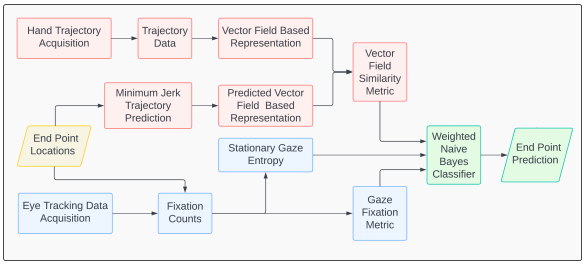


Fig. 1: Overview of Proposed Framework

## III. 3D VECTOR FIELD METHOD FOR HMIP

This section describes the 3D vector field-based representation of the hand trajectory data. Hand trajectory data can be acquired by any state-of-the-art sensors and tracking methodologies, such as the Ultra Leap hand tracker, Kinect

sensors or motion capture systems, and is not the focus of this paper.

Hand trajectory data, once obtained, is represented as a vector field to better capture its overall flow amidst discretisation errors and sensor-tracking noise. A predicted set of vector fields is also constructed for each possible known endpoint location in the human-robot collaborative space based on the minimum jerk trajectory (MJT) model. A similarity metric is then calculated between each of the predicted vector fields and the observed vector fields, which predicts the intended endpoint of the human hand motion. Sections III-A to III-C describe this process in more detail.

### A. Vector Field Based Representation

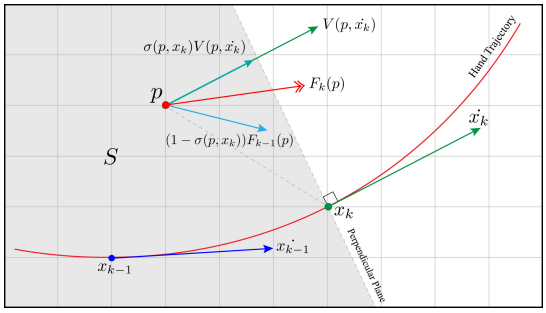


Fig. 2: 2D Projection of the Derivation of Proposed 3D Vector Field Representation

Consider an observed hand motion trajectory $\gamma$ which takes discrete time steps $n = f$ to complete, characterised by:

$$\gamma = \{x_0, x_1, ...x_k..., x_f\} = \{x_n\}_{n=0:f} \tag{1}$$

where $x_k \in \mathbb{R}^3$ represents a trajectory point in space at time step $n = k$ and $\dot{x_k}$ the corresponding velocity at that point. Thus, let $\dot{\gamma} = \{\dot{x_n}\}_{n=0:f}$ be the velocities at each point along the trajectory. Consider the workspace domain of the human-robot interaction to be discretised by a grid resolution of $g$. Let $S_k$ be the subdomain of the workspace that is behind a plane which is perpendicular to the trajectory vector $x_k$. A vector field $F_k$ at time step $k$ assigns a vector at each point $p \in S_k$, within the given workspace such that:

$$F_k(p) = \sigma(p, x_k)V(p, \dot{x_k}) + (1 - \sigma(p, x_k))F_{k-1}(p) \tag{2}$$

where $V(p, \dot{x_k})$ represents the velocity vector $\dot{x_k}$ translated to point $p$ (See Figure 2). The parameter $\sigma$ is used to smoothly scale the vector field based on the Euclidean distance to the original trajectory points to better capture the overall flow and likeness of the hand motion trajectory (see figure 3). A sigmoid function $\sigma(p, x_k)$ is used to achieve this smooth scaling given by:

$$\sigma(p, x_k) = \frac{1}{1 + e^{(\frac{\|p - x_k\|}{\beta})}} \tag{3}$$

The parameter $\beta$ controls the impact of this scaling at each point on the vector field. This representation is a 3D extension and refinement of work presented in [1].

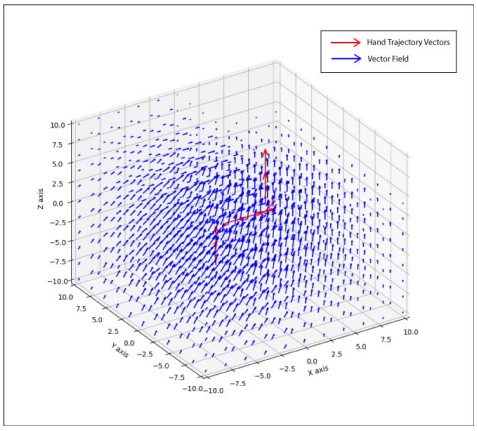


Fig. 3: Example of 3D Vector Field Generated for a Simple Hand Trajectory

### B. Trajectory Prediction Based on MJT Model

We utilise this MJT model to make predicted trajectories to each of the possible end point locations in the collaborative human-robot workspace. Assume a set of $m$ possible endpoint locations given by:

$$E = [{}^1x_e, {}^2x_e.... {}^mx_e] = \{{}^ax_e\}_{a=1:m} \tag{4}$$

Given an initial time step 0, the final time for trajectory completion $t_f$, and model-based predicted hand position vector $\overline{x}$, the MJT model (in continuous time) can be formally presented as minimising the following cost function:

$$C = \frac{1}{2}\int_0^{t_f} \left(\frac{d^3\overline{x}}{dt^3}\right)^T \left(\frac{d^3\overline{x}}{dt^3}\right) dt \tag{5}$$

Considering the point-to-point movement of a human hand, it can be assumed that the boundary conditions for velocity and acceleration would be zero. Furthermore, the boundary conditions for the position will be $\overline{x(t_0)} = x_0$ and $\overline{x(t_f)} =^a x_e$ for a given end point location $a$. Utilising these boundary conditions, a unique solution for the optimisation problem given in equation 5 can be solved for position, as the following polynomial [4]:

$$\overline{x_t} = x_0 + ({}^ax_e - x_0)(10\tau^3 + 15\tau^4 + 6\tau^5) \tag{6}$$

Where $\tau = \frac{t}{t_f}$, which is the normalised time relative to the time taken to complete the trajectory. Thus, differentiating Equation 6 gives the velocity of the predicted trajectory:

$$\overline{\dot{x_t}} = \frac{d\tau}{dt}\frac{\overline{dx_t}}{d\tau} \tag{7}$$

This MJT model is utilised to calculate a set of predicted trajectories from the current observed hand position $x_k$ to all possible endpoint locations $E$ in the collaborative workspace.

Since a prediction is being made for each endpoint, only a portion of the trajectory $\gamma = [x_0 ... x_k]$ is observed at timestep $k$. Furthermore, considering the equations 6 and 7, the parameters $x_0$,$^a x_e$ and $t$ are known for a given endpoint location and partial observation. Thus in order to calculate the predicted trajectory and its velocities, only a reasonable estimate of $t_f$ is necessary. For this purpose, the partially observed trajectory points are fitted to the MJT model in order to calculate $t_f$ by solving the following optimisation problem:

$$C(t_f) = \frac{1}{2} \sum_{n=0}^{k} \|x_k - \overline{x}_k\|^2 \tag{8}$$

This estimate of $t_f$ is then utilised in equations 6 and 7 to obtain predicted trajectories to each endpoint in $E = \{^a x_e\}_{a=1:m}$ at each time step $k$. This results in a set of predicted trajectories $\Gamma_k = \{\overline{^a\gamma_k}\}_{a=1:m}$ and their corresponding predicted velocities $\dot{\Gamma}_k = \{\overline{^a\dot{\gamma}_k}\}_{a=1:m}$ where each element relates to a given endpoint $a$.

Each predicted trajectory is then used to calculate a predicted vector field for each endpoint location utilising the equations outlined in Section III-A. Thus a set of predicted vector fields $\overline{\Phi} = \{\overline{^a F_k}\}_{a=1:m}$ at time $k$ is obtained whose elements relate to each endpoint location.

### C. Vector Field Similarity Metric

Based on sections III-A and III-B; at each time step $k$ we obtain an observed vector field $F_k$ and a set of $m$ predicted vector fields $\overline{\Phi} = \{\overline{^a F_k}\}_{a=1:m}$.

A similarity metric $^a\alpha_k$ between the observed vector field $F_k$ and a particular predicted vector field $\overline{^a F_k}$ is calculated based on vector cosine similarity. This metric quantifies the alignment of vectors at each corresponding point $p$ of the vector fields and is given using equation 9:

$$^a\alpha_k = \frac{1}{n(S_k)} \sum^{S_k} F_k \cdot \overline{^a F_k} \tag{9}$$

This similarity metric quantifies how close an observed vector field is to a given predicted vector field and is normalised by the number of points $n(S_k)$ in the workspace domain $S_k$. This, in turn, quantifies how close an observed hand trajectory is to a predicted MJT trajectory for a specific endpoint. This vector field similarity metric, $^a\alpha_k$, is the first feature attribute used in the NBC when fusing with Gaze data. The value $^a\alpha_k$ ranges from $-1$ to 1, indicating unaligned to perfectly aligned vector fields, respectively.

## IV. FUSION OF SGE WEIGHTED GAZE FIXATIONS

As with hand trajectory data, raw gaze data can be acquired through a multitude of state-of-the-art sensors and approaches such as dedicated eye-tracking glasses and headsets, including or not limited to the Hololens, Pupilcore and Tobii eye tracker and is not the focus of this paper.

The gaze data extracted is primarily in the form of fixation counts associated with each possible endpoint. For a given moment in time, the fixation counts are utilised to calculate Stationary Gaze Entropy (SGE), which quantifies the visual focus of a human. This is utilised to weigh the contribution of the gaze fixation metric $^a\beta_k$, which is fused with the vector field similarity metric $^a\alpha_k$ utilising an NBC. This process is outlined from Section IV-A to IV-C.

### A. Gaze Fixation Counts

Fixation counts associated with each potential endpoint in $E = \{^a x_e\}_{a=1:m}$ are counted from when hand motion is detected up to the current time step $k$. The counting is reset when the hand motion stops. Thus a set of gaze fixation counts $G = [^1 g_k, ^2 g_k .... ^m g_k] = \{^a g_k\}_{a=1:m}$ associated with each end point at time $k$ is obtained. Fixation counts to a given endpoint $a$ as a proportion of the total fixations is considered as the second metric utilised as a feature for the NBC.

$$^a\beta_k = \frac{^a g_k}{^T g_k} \tag{10}$$

Where $^T g_k$ is the total fixation count at time step $k$. Here, it is assumed that a higher gaze fixation metric $^a\beta_k$ is associated with the intended endpoint of the human hand motion. However, since the robustness of this assumption varies based on the focus of the user, the contribution made by the gaze data is weighted based on the visual focus of a particular human operator using SGE.

### B. Stationary Gaze Entropy

A common practice in studying saccadic gaze data associated with fixations to multiple areas of interest (AOIs), such as with the available endpoints in a collaborative workspace, is to model the gaze transitions as a first-order Markov Chain [18]. This allows for the calculation of two entropy metrics based on Shannon's Entropy: the Gaze Transition Entropy (GTE) and Stationary Gaze Entropy (SGE). In the proposed solution, we utilise the SGE metric, as many studies have established that it provides valuable information that quantifies visual focus in a particular AOI/endpoint [19]. SGE for a given time step $k$ is calculated as:

$$H_k = -\sum_{a=1}^{m} {}^a\beta_k \log_2({}^a\beta_k) \tag{11}$$

When the SGE value $H_k$ is higher, it indicates that the subject is distributing their visual attention across the available AOIs/endpoints. Conversely, a lower value suggests that fixations are focused on a few particular AOIs. Thus, SGE quantifies visual attention and focus and can be utilised as an appropriate weighting factor for the gaze fixation metric in the NBC.

### C. Weighted Naive Bayes Classification

For a given hand trajectory motion at time step $k$ the vector field similarity metric $\alpha_k$ and gaze fixation metric $\beta_k$ is utilised as features for the NBC classifier. For this purpose, we assume that $\alpha_k$ and $\beta_k$ are independent features.

We utilise the standard formulation of the NBC for this purpose, with SGE-related weights modulating the contribution provided by the gaze fixation counts.

For the purpose of our application, the NBC is utilised as a binary classifier to identify if a given hand trajectory is intended for a given endpoint or not. Thus given a specific endpoint, the classifier will query between the following two classes:

- Y: Hand motion is intended for the specified endpoint
- N: Hand motion is not intended for the specified endpoint

Based on Baye's theorem, the probability that a hand motion is intended for a specified endpoint $a$ given a particular vector field similarity metric and gaze fixation metric is:

$$p(Y|{}^a\alpha_k, {}^a\beta_k) \propto p(Y)p({}^a\alpha_k|Y)p({}^a\beta_k|Y) \tag{12}$$

For a given prediction all endpoint priors are considered to be equally likely, thus $p(Y) = 1/m$ and can be factored out as a constant. Based on the log probability form of equation 12, and the standard NBC formulation, we obtain a probability score associated with each possible endpoint at time $k$.

$${}^aP_k = ln[p({}^a\alpha_k|Y)] + (1 - H_k)ln[p({}^a\beta_k|Y)] \tag{13}$$

here $(1 - H_s)$ weights the contribution made by gaze data based on SGE. Since gaze fixation counts as a proportion of the total fixations behave as direct likelihoods, we can consider:

$$p({}^a\beta_k|Y) = {}^a\beta_k \tag{14}$$

The vector field similarity metric $\alpha_k$ is a continuous variable bounded between $-1$ to 1. Thus considering class $Y$, $p({}^a\alpha_k|Y)$ can be assumed to be a negative half-normal distribution with a mean $\mu_y = 1$ which indicates perfect alignment between predicted and observed vector fields (i.e class $Y$). The variance $\sigma_y$ associated with class Y is obtained utilising a short training process or can be tuned heuristically. Formulating the problem as a binary classification problem generalises a simple training process to any scenario with different endpoints.

Finally, the endpoint with the highest log probability score is considered as the most plausible predicted intended endpoint $\hat{P}_k$, based on the observed hand trajectory data and gaze data at a time instant $k$:

$$\hat{P}_k = \underset{a}{\text{argmax}}({}^aP_k) \tag{15}$$

A confidence gate-based threshold can be heuristically utilised to make decisions based on this prediction for higher-level decision-making.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

Two sets of experiments were carried out with the objective of first evaluating the robustness of the proposed framework and secondly verifying its validity in terms of efficiency gains in a real-world HRC application in the construction industry. As previously outlined, hand and eye tracking data can be obtained from any state-of-the-art source. For the purpose of these experiments, hand and eye tracking data were obtained using a monocular camera, OpenCV and Google Mediapipe deep learning libraries.

### A. Experiment 1: Robustness Evaluation

Experiment 1 consisted of four objects placed on a desk at varying locations in 3D space, emulating a generic collaborative HRC workspace consisting of reaching tasks (See Figure 4). The proposed framework was utilised to predict which of the four objects (endpoints) a particular human participant intended to reach.

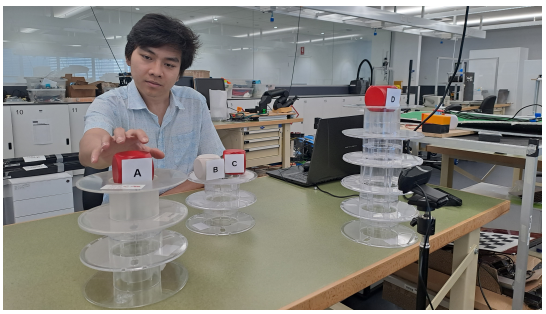


Fig. 4: Experiment 1: Robustness Evaluation

Five different trajectories were tested, as described in Table I. Each trajectory was executed by four participants five times for a total of a hundred trajectories. To evaluate the prediction accuracy of the proposed HMIP framework at different portions of a given trajectory, the prediction accuracy at $25\%$ time intervals of the total trajectory time is presented in Table I. The results presented in Table I are the mean accuracy of all trajectories collected among the four participants. Prediction results are presented based solely on the vector field similarity metric and the gaze fixation metric separately, along with the combined weighted NBC result.

Objects B and C were intentionally placed close to each other (about $4cm$ apart) to gauge the framework's response. Considering the results for trajectories 1 to 4, when reliable gaze focus is maintained, the combined prediction accuracy is superior compared to each modality in isolation. This is to be expected since a multimodal framework would be more robust to occlusions or tracking failures that could plague a particular single modality. It should also be noted that all results are generally consistently accurate by around $50\%$ of the total trajectory, and each modality can still successfully carry out HMIP as an individual system. However, results indicate that the vector field-based method drops in accuracy, especially when objects are close to each other.

Although the gaze fixation metric appears to outperform the vector field metric, it may be prone to false positives if

TABLE I: Experiment 01: Prediction Accuracy at 25% Increments of Total Trajectory Time

| Traj No | Description | Metric | 0-25% | 25-50% | 50-75% | 75-100% |
|---|---|---|---|---|---|---|
| 1 | To A with focused gaze fixations | VF Similarity Metric | 0.85 | 0.75 | 0.88 | 1.00 |
| | | Gaze Fixation Metric | 0.94 | 1.00 | 0.75 | 0.63 |
| | | Combined | 0.91 | 1.00 | 1.00 | 1.00 |
| 2 | To B with focused gaze fixations | VF Similarity Metric | 0.36 | 0.53 | 0.63 | 0.75 |
| | | Gaze Fixation Metric | 1.00 | 1.00 | 1.00 | 1.00 |
| | | Combined | 0.67 | 0.80 | 0.91 | 1.00 |
| 3 | To C with focused gaze fixations | VF Similarity Metric | 0.39 | 0.41 | 0.95 | 0.88 |
| | | Gaze Fixation Metric | 0.88 | 0.75 | 0.75 | 1.00 |
| | | Combined | 0.97 | 0.83 | 0.80 | 1.00 |
| 4 | To D with focused gaze fixations | VF Similarity Metric | 0.50 | 0.34 | 0.38 | 0.75 |
| | | Gaze Fixation Metric | 1.00 | 1.00 | 1.00 | 1.00 |
| | | Combined | 0.86 | 0.78 | 0.92 | 1.00 |
| 5 | To A with unreliable gaze fixations | VF Similarity Metric | 0.66 | 0.76 | 0.65 | 1.00 |
| | | Gaze Fixation Metric | 0.40 | 0.50 | 0.25 | 0.25 |
| | | Combined | 0.45 | 0.75 | 0.58 | 1.00 |

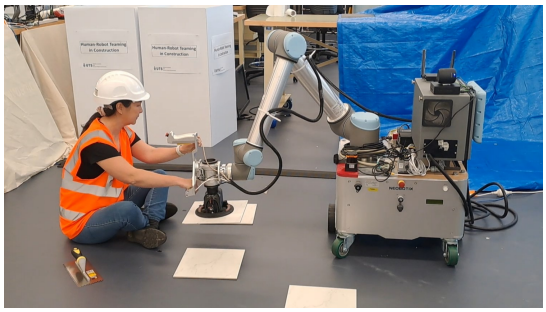

Fig. 5: Experiment 2: Collaborative Tiling with HMIP

the human operator is unfocused or inattentive. To evaluate such scenarios and the SGE-based weighting, trajectory 5 was carried out with random saccadic eye movements simulating an unfocused or tired participant. The low prediction accuracy given by the gaze metric is indicative of these random saccades. However, the combined result is minimally impacted by this low accuracy as the SGE lowers the contribution of the gaze fixation metric when unreliable random and unfocused gaze data is fed into the framework. Thus, reliable, focused gaze data adds value and robustness to the predictive capabilities, while unreliable, unfocused gaze data does not drastically impact the predictive capabilities of the proposed HMIP framework.

### *B. Experiment 2: Real-time Efficiency Verification*

Experiment 2 consisted of a collaborative tiling scenario between a human and a robot (see figure 5). Two scenarios are juxtaposed to showcase the intuitive nature of incorporating the proposed HMIP into an HRC context. Scenario 1 is a conventional approach to HRC-based tiling with no HMIP, where the robot is tasked with bringing a tile from a pile of tiles to a fixed location close to the human. The human then uses the arm and places it in the correct tiling location decided by the user.

In Scenario 2, users employed an intuitive hand motion to instruct the robot on the desired destination for the tile. Leveraging the proposed real-time HMIP framework, the robot brings the tile directly to the user's desired tile location based on the HMIP prediction made before the hand motion is fully completed. The mode of the final predictions, with a confidence level exceeding a 2-sigma confidence gate, is utilized to ascertain the predicted endpoint.

A temporal efficiency comparison was conducted between the two scenarios by four participants. Scenario 2 with HMIP showed a $24\%$ efficiency gain overall, underscoring the enhanced efficiency derived from the predictive and intuitive nature of the proposed HMIP-based interaction.

## VI. CONCLUSIONS

The proposed 3D vector field-based representation utilising predicted MJT trajectories, which extends our previous work [1], provides a robust and scalable solution to HMIP. Results indicate that the fusion of gaze fixation adds value to this method. The proposed utilisation of SGE as a weighting factor in the NBC classifier ensures that the impact of potential false positives caused by fatigue or lack of focus is mitigated. Preliminary results indicate that this weighting successfully exploits focused gazed fixations and discounts unfocused gaze fixations. Thus, it can be concluded that the vector field-based method and SGE weighted gaze fixation act synergistically to provide a robust HMIP framework.

The implementation of the proposed framework in a real-time HRC scenario in a construction context showcases how the proposed HMIP improves the efficiency of collaboration due to its intuitive nature. Due to the lack of open-source HMIP algorithms for benchmarking, the real-time implementation of the framework has been presented as an open-source ROS package in an attempt to welcome future comparative analysis and benchmarking in this area.

In terms of future work, we hope to investigate the proposed HMIP framework's impact on cognitive and physical load in HRC contexts. Furthermore, we hope to exploit other multimodal sources, such as skeletal tracking and EEG, to improve the robustness of the HMIP framework. We also plan on exploring more robust data fusion and classification approaches.

## REFERENCES


[1] M. Manitta, M. Jayasuriya, and D. Liu, "A Vector Field-Based Method for Human Action Representation and Recognition During Human-Robot Collaboration," in *IEEE International Conference on Automation Science and Engineering (CASE)*, 2023.

[2] Y. Uno, M. Kawato, and R. Suzuki, "Formation and Control of Optimal Trajectory in Human Multijoint Arm Movement," *Biological Cybernetics*, vol. 61, no. 2, pp. 89–101, Jun. 1989. [Online]. Available: https://link.springer.com/10.1007/BF00204593

[3] T. Flash and N. Hogan, "The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model," *The Journal of Neuroscience*, vol. 5, no. 7, pp. 1688–1703, Jul. 1985. [Online]. Available: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.05-07-01688.1985

[4] C. T. Landi, Y. Cheng, F. Ferraguti, M. Bonfe, C. Secchi, and M. Tomizuka, "Prediction of Human Arm Target for Robot Reaching Movements," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Macau, China: IEEE, Nov. 2019, pp. 5950–5957. [Online]. Available: https://ieeexplore.ieee.org/document/8968559/

[5] K. H. Dinh, O. Oguz, G. Huber, V. Gabler, and D. Wollherr, "An Approach To Integrate Human Motion Prediction Into Local Obstacle Avoidance in Close Human-Robot Collaboration," in *IEEE International Workshop on Advanced Robotics and its Social Impacts (ARSO)*. Lyon: IEEE, Jun. 2015, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7428221/

[6] A. M. Zanchettin and P. Rocco, "Probabilistic Inference of Human Arm Reaching Target for Effective Human-Robot Collaboration," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, BC: IEEE, Sep. 2017, pp. 6595–6600. [Online]. Available: http://ieeexplore.ieee.org/document/8206572/

[7] C. Lenz, A. Sotzek, T. Roder, H. Radrich, A. Knoll, M. Huber, and S. Glasauer, "Human Workflow Analysis Using 3D Occupancy Grid Hand Tracking in a Human-Robot Collaboration Scenario," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. San Francisco, CA: IEEE, Sep. 2011, pp. 3375–3380. [Online]. Available: http://ieeexplore.ieee.org/document/6094570/

[8] H. Ding, G. Reissig, K. Wijaya, D. Bortot, K. Bengler, and O. Stursberg, "Human Arm Motion Modeling and Long-Term Prediction for Safe and Efficient Human-Robot-Interaction," in *IEEE International Conference on Robotics and Automation (ICRA), 2018*. Shanghai, China: IEEE, May 2011, pp. 5875–5880. [Online]. Available: http://ieeexplore.ieee.org/document/5980248/

[9] J. Lyu, P. Ruppel, N. Hendrich, S. Li, M. Gorner, and J. Zhang, "Efficient and Collision-Free Human-Robot Collaboration Based on Intention and Trajectory Prediction," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–1, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9920012/

[10] R. Luo, R. Hayne, and D. Berenson, "Unsupervised Early Prediction of Human Reaching for Human–Robot Collaboration in Shared Workspaces," *Autonomous Robots*, vol. 42, no. 3, pp. 631–648, Mar. 2018. [Online]. Available: http://link.springer.com/10.1007/s10514-017-9655-8

[11] R. Luo and D. Berenson, "A Framework for Unsupervised Online Human Reaching Motion Recognition and Early Prediction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Hamburg, Germany: IEEE, Sep. 2015, pp. 2426–2433. [Online]. Available: http://ieeexplore.ieee.org/document/7353706/

[12] J. Martinez, M. J. Black, and J. Romero, "On Human Motion Prediction Using Recurrent Neural Networks." arXiv, May 2017, arXiv:1705.02445 [cs]. [Online]. Available: http://arxiv.org/abs/1705.02445

[13] F. Formica, S. Vaghi, N. Lucci, and A. M. Zanchettin, "Neural Networks based Human Intent Prediction for Collaborative Robotics Applications," in *International Conference on Advanced Robotics (ICAR)*. Ljubljana, Slovenia: IEEE, Dec. 2021, pp. 1018–1023. [Online]. Available: https://ieeexplore.ieee.org/document/9659328/

[14] Y. Cheng and M. Tomizuka, "Long-Term Trajectory Prediction of the Human Hand and Duration Estimation of the Human Action," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 247–254, Jan. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9599389/

[15] M. Mavsar, M. Denisa, B. Nemec, and A. Ude, "Intention Recognition with Recurrent Neural Networks for Dynamic Human-Robot Collaboration," in *International Conference on Advanced Robotics (ICAR)*. Ljubljana, Slovenia: IEEE, Dec. 2021, pp. 208–215. [Online]. Available: https://ieeexplore.ieee.org/document/9659473/

[16] T. Zhou, Y. Wang, Q. Zhu, and J. Du, "Human Hand Motion Prediction Based on Feature Grouping and Deep Learning: Pipe Skid Maintenance Example," *Automation in Construction*, vol. 138, p. 104232, Jun. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0926580522001054

[17] A. Choi, M. K. Jawed, and J. Joo, "Preemptive Motion Planning for Human-to-Robot Indirect Placement Handovers," in *IEEE International Conference on Robotics and Automation (ICRA)*. Philadelphia, PA, USA: IEEE, May 2022, pp. 4743–4749. [Online]. Available: https://ieeexplore.ieee.org/document/9811558/

[18] K. Krejtz, T. Szmidt, A. T. Duchowski, and I. Krejtz, "Entropy-Based Statistical Analysis of Eye Movement Transitions," in *Proceedings of the Symposium on Eye Tracking Research and Applications*. Safety Harbor Florida: ACM, Mar. 2014, pp. 159–166. [Online]. Available: https://dl.acm.org/doi/10.1145/2578153.2578176

[19] K. Krejtz, A. Duchowski, T. Szmidt, I. Krejtz, F. González Perilli, A. Pires, A. Vilaro, and N. Villalobos, "Gaze Transition Entropy," *ACM Transactions on Applied Perception*, vol. 13, no. 1, pp. 1–20, Dec. 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2834121

[20] B. Shiferaw, L. Downey, and D. Crewther, "A Review of Gaze Entropy as a Measure of Visual Scanning Efficiency," *Neuroscience & Biobehavioral Reviews*, vol. 96, pp. 353–366, Jan. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0149763418303075

[21] S. Lanini-Maggi, I. T. Ruginski, T. F. Shipley, C. Hurter, A. T. Duchowski, B. B. Briesemeister, J. Lee, and S. I. Fabrikant, "Assessing How Visual Search Entropy and Engagement Predict Performance in a Multiple-Objects Tracking Air Traffic Control Task," *Computers in Human Behavior Reports*, vol. 4, p. 100127, Aug. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2451958821000750

[22] Y. Zhang, S. Hopko, A. Y, and R. K. Mehta, "Capturing Dynamic Trust Metrics during Shared Space Human Robot Collaboration: An eye-tracking approach," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 66, no. 1, pp. 536–536, Sep. 2022. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1071181322661296

[23] Y. Jia and C.-R. Lee, "Vector Field Model for Trajectory Data and Its Application in Similarity Query," in *IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. Yanuca Island, Cuvu, Fiji: IEEE, Dec. 2020, pp. 1180–1187. [Online]. Available: https://ieeexplore.ieee.org/document/9407909/

[24] N. Ferreira, J. T. Klosowski, C. Scheidegger, and C. Silva, "Vector Field k-Means: Clustering Trajectories by Fitting Multiple Vector Fields," Aug. 2012, arXiv:1208.5801 [cs]. [Online]. Available: http://arxiv.org/abs/1208.5801

[25] J. Frederico Carvalho, M. Vejdemo-Johansson, F. T. Pokorny, and D. Kragic, "Long-term Prediction of Motion Trajectories Using Path Homology Clusters," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Macau, China: IEEE, Nov. 2019, pp. 765–772. [Online]. Available: https://ieeexplore.ieee.org/document/8968125/

[26] C. Barata, J. C. Nascimento, and J. S. Marques, "A Sparse Approach to Pedestrian Trajectory Modeling Using Multiple Motion Fields," in *IEEE International Conference on Image Processing (ICIP)*. Beijing: IEEE, Sep. 2017, pp. 2538–2542. [Online]. Available: http://ieeexplore.ieee.org/document/8296740/