Interacting Multiple Model Hand Pose Filtering for Human-Robot Collaboration

Andrew Pasco
Department of Mechanical Engineering
Stanford University
Stanford, CA, USA
apasco@stanford.edu

***Abstract*—In human-robot collaboration contexts, intent prediction is important both for safety and productivity. Humans more effectively collaborate with robots that clearly indicate their intentions, so we explore how robots can be endowed with similar knowledge. In particular, we will apply the interacting multiple model (IMM) filter to estimate human wrist pose during routine tasks that could occur in HRC contexts. With such a filter, we can both distinguish between different motion modes and use these mode probabilities as an indicator of intent, enabling a safer and more effective robotic collaborator. We compared this filter on a dataset of standard and abrupt industrial gestures (DASIG) against single-model baselines like quaternion-based UKF and multiplicative EKF, revealing that the IMM achieves comparable 6-DoF tracking accuracy to the MEKF (∼13.0 mm position and 3.0 degree attitude RMSE) while remaining computationally tractable for real-time applications, and that it drastically reduces false-positive detections of abrupt movements (by over 80%) compared to a baseline velocity threshold, albeit with a slower average detection latency of 0.85 seconds. Ultimately, while the IMM architecture is successfully implemented as a semantic intent estimator without sacrificing tracking fidelity, its detection latency and high sensitivity to hyper-parameter tuning limit its immediate deployment human-robot collaborative environments.**

I. Introduction

Increased prevalence of cobots in industrial environments requires different safety and collaborative considerations than standard caged-off automation cells. Specifically, it is important to account for the multi-modal nature of human manipulation dynamics; a collaborator may frequently switch between static holding/manipulating, smooth reaching, and reactionary responses to external stimuli [1], [2]. Thus, accurate estimation of these regimes is critical both for intent estimation for more effective collaboration and for avoidance-related safety considerations.

To motivate the use of a multiple-model filter, consider a situation where an operator suddenly changes the trajectory of their hand, for example if they drop a tool. A single model filter will predict that the hand continues according to the same dynamics before correction occurs, where a multiple-model filter could adjust its relative prediction probabilities quicker [3]. The overshoot of the former could lead to poor tracking, or worse, collision. To motivate the inclusion of a full 6-DoF wrist pose rather than simply tracking position, certain biomechanical studies suggest that the intent to move can manifest through re-orientation before movement actually begins [4].

This project applies the Interacting Multiple Model filter [5] to 6-DoF wrist pose estimation in industrial robot collaboration contexts. By running parallel filters corresponding to hypothesized modes of motion and soft-switching based on measurement likelihoods, we can both maintain a filtered pose estimate of the hand and utilize the mode probabilities as a proxy for intent. The system was validated on movement data from DASIG, which includes abrupt motions in response to environmental stimuli [6], and implementations of the multiplicative EKF and quaternion UKF were also included for comparison purposes on metrics like RMSE, NEES, and latency to warn an abrupt movement.

II. Related Work

A. Multi-Modal Filtering

While particle filters [7] are commonly used for representing the possibility of multi-modality, the high dimensionality of a 6-DoF maneuvering wrist is prohibitive. The IMM offers a tractable alternative, approximating the posterior as a mixture of Gaussians with fixed computational cost. Often, dynamics models like ”constant velocity” or ”constant acceleration” are used to quantify the different modes [5]. Lee et al. [8] successfully applied IMM filtering to companion robots, though the application was limited to 3D position.

B. Quaternion Filtering

Critically, estimating 3D orientation requires handling the unit-norm constraint of quaternions. Bar-Itzhack and Oshman established the fundamental vector observation models required for this estimation [9]. Building on this, [10] and [11] demonstrated that the QUKF is superior to the EKF for this task, avoiding singularities associated with Euler angles. While effective for continuous motion, these works do not address the ”switching” dynamics inherent to safety incidents. Multiple models using Delta Quaternions were explored in [12], but the focus was latency compensation rather than the explicit categorization of intent for safety.

A core challenge in the application of an IMM to this 6-DoF state space is the mixing step, where estimates from different models are averaged. [13] and [14] argue that rigorous averaging requires mapping errors to the tangent space to

respect the manifold structure, and [15] explicitly formulated an IMM on "Boxplus" manifolds which is directly applicable for quaternions. However, [16] suggested that for small error angles, simplified representations may suffice, which was indeed supported by experimental results in [15].

III. Methods

The following section details the problem formulation, including the state and measurement models.

A. State Space & Dynamics

The state vector $x_t = [p_t, v_t, q_t, \omega_t]^T$ includes the pose and velocity. In the context of the pick-and-place tasks with disruptions in DASIG, hand movement is modeled with three different modes. With inspiration from [5], the filter bank will include a static model, a constant velocity (CV) model, and a "maneuvering" (M) model. The static model has an identity state transition matrix and low process noise covariance, representing static holds or in-place tool manipulations. Because we do not include the accelerations in the state, the CV and M models have the same state transitions:

$$p_{t+1} = p_t + v_t * \Delta t$$

$$v_{t+1} = v_t + W_t$$

$$q_{t+1} = q_t \otimes \exp\left(\frac{1}{2}\omega_t \Delta t\right)$$

$$\omega_{t+1} = \omega_t + W_t$$

$$W_t \sim GWN(0, Q)$$

However, CV has low process noise covariance while M has high process noise covariance. These models represent smooth movements like reaching towards a bin and jerky movements like a sudden response to an alarm, respectively. Given these state dynamics, we then have the 13×13 Jacobian:

$$\begin{bmatrix} I_{3\times3} & \Delta t I_{3\times3} & \mathbf{0}_{3\times4} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & I_{3\times3} & \mathbf{0}_{3\times4} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{4\times3} & \mathbf{0}_{4\times3} & R(\delta q)_{4\times4} & \frac{1}{2}\Delta t L(q)_{4\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times4} & I_{3\times3} \end{bmatrix}$$

where $R(\delta q)$ is the 4×4 right-quaternion-multiplication matrix for the incremental rotation $\delta q = \exp(\frac{1}{2}\omega_t \Delta t)$, $L(q)$ is the 4×3 matrix representing the first three columns of the left-quaternion-multiplication matrix (coupling 3D angular velocity components to the 4D quaternion space with small-angle approximation), and $\Delta t I_{3\times3}$ is the position-velocity term [17].

B. Measurement Model

The selected measurement model is the 7-dimensional position and orientation of the hand at each timestep. While this does not exactly match the measurement method of the original dataset, it is unlikely that IMU-based measurements would be used in a real industrial environment. We assume that some other means (perhaps vision-based) exists to provide this 7-d measurement at each timestep, which is taken as the ground-truth injected with Gaussian white noise. For the 3D position, the measurement model follows standard additive noise:

$$z_{p,t} = p_t + V_{p,t}$$

where $V_{p,t} \sim GWN(0, R_p)$ is the 3D position measurement noise. For the 4D orientation quaternion, we employ a simplified additive noise model followed by explicit normalization. While rigorous quaternion estimation typically models noise as a 3D rotational perturbation to strictly maintain the unit-norm constraint of S^3, directly adding 4D Gaussian noise and renormalizing is practically sufficient for small angular displacements and low noise variances. The quaternion measurement is modeled as:

$$z_{q,t} = \frac{q_t + V_{q,t}}{\|q_t + V_{q,t}\|}$$

where $V_{q,t} \sim GWN(0, R_q)$ is the 4D additive measurement noise. Combining these, the full 7-dimensional measurement $z_t = [z_{p,t}^T, z_{q,t}^T]^T$ is extracted directly from the corresponding position and orientation of the state vector x_t. We take $R_p = I_3 * 10^{-4}$ and $R_q = I_4 * 10^{-4}$.

C. Filters

a) IMM with "Naive Mixing": The IMM filter (Figure 1) combines multiple state hypotheses from component filter models [5]. Before each state update, estimates from the previous iteration ($\hat{x}^j$) are mixed using conditional model probabilities (μ) indicating the likelihood of the system's current mode. These probabilities are computed at each step using the prior step's probabilities, the measurement likelihood (Λ_t^j) derived from the current measurement innovation, and a state switching matrix p_{ij}. The updated probability for model j at time t is computed as:

$$\mu_t^j = \frac{1}{c}\Lambda_t^j \sum_i p_{ij}\mu_{t-1}^i$$

where c is a normalization constant ensuring the probabilities sum to one. Estimates from each separate filter ($\hat{x}^j$) after the current step dynamics update and measurement innovation are then combined as a weighted sum using these conditional model probabilities to yield a more accurate overall state prediction. While this weighted average step breaks the manifold structure of quaternions, previous works like [15] recommended that a naive alternative with quaternions averaged in parameter space and normalized after is suitable, so we use this method to avoid the higher computational cost of manifold optimization. We will implement this naive mixing by first aligning all estimates in the same hemisphere to avoid averaging near-antipodal quaternions (since $q = -q$, a simple weighted sum can remove information). Then, we compute the overall combined quaternion estimate $\hat{q}_t$ by weighting the individual filter estimates $\hat{q}_t^j$ and normalizing the result:

$$\hat{q}_t = \frac{\sum_j \mu_t^j \hat{q}_t^j}{\left\|\sum_j \mu_t^j \hat{q}_t^j\right\|}$$

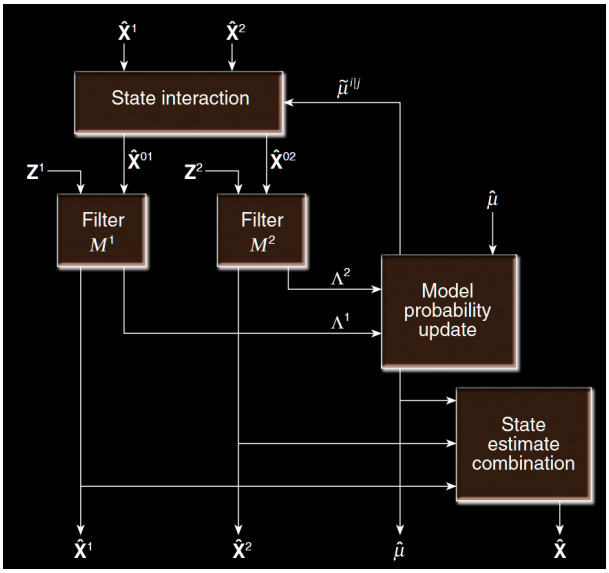

Figure 1. Block diagram of the IMM filter from [5]

b) Multiplicative EKF: The MEKF [16] addresses the singularity and constraint issues of standard EKFs by parameterizing the global attitude with a four-component unit quaternion while employing a three-component representation for the attitude errors. The true attitude is represented as the product of a reference unit quaternion, $q_{ref}(t)$, and an error quaternion parameterized by a three-vector, $a(t)$:

$$q(t) = \delta q(a(t)) \otimes q_{ref}(t)$$

The filter performs an unconstrained estimation of the three-component error, bypassing the need for constrained quaternion estimation. To remove redundancy between the two representations, the expectation of the attitude error is forced to be identically zero, making the reference quaternion the optimal attitude estimate. During a discrete measurement update, the filter calculates a non-zero update for the error vector, $\hat{a}(+)$. To prepare for the next propagation step, a nonlinear "reset" operation folds this updated error into the reference quaternion via quaternion multiplication and resets the error state back to zero:

$$q_{ref}(+) = \delta q(\hat{a}(+)) \otimes q_{ref}(-)$$

c) Quaternion EKF: The QUKF [10] adapts the standard UKF to handle the same fact that unit quaternions do not belong to a standard vector space. Because of the mismatch between the six-dimensional noise vector and the seven-dimensional state vector, simple arithmetic addition for sigma point generation is again invalid. As in to the MEKF approach, the QUKF converts 3D rotation noise vectors into noise quaternions and applies them to the state via quaternion multiplication. It is also important to note that computing the a priori mean of the projected sigma points using a standard average fails to produce a valid unit quaternion. To compute a valid geometric mean, the QUKF employs an iterative intrinsic gradient descent algorithm: For an estimated mean $\bar{q}_t$ at iteration t, it calculates an error quaternion e_i representing the relative rotation required to turn the estimated mean into each sigma point q_i as $e_i = q_i \bar{q}_t^{-1}$. The corresponding 3D rotation error vectors $\vec{e}_i$ are then averaged to find a barycentric adjustment vector:

$$\vec{e} = \frac{1}{2n} \sum_{i=1}^{2n} \vec{e}_i$$

The quaternion representation of this adjustment vector, e, is then applied to refine the mean orientation for the next iteration:

$$\bar{q}_{t+1} = e \bar{q}_t$$

This process iterates until the adjustment vector reaches zero, indicating the true geometric mean has been found. Finally, the state covariance is computed using these same 3D rotation error vectors rather than the arithmetic difference of the quaternions. This is notably more computationally expensive than either of the prior approaches.

D. Experimental Validation

DASIG [6] contains 6-DoF IMU data of subjects performing industrial tasks like assembly and pick-and-place while periodically interrupted by simulated alarms. This is appropriate for developing filters for more responsive collaborative robots. The data was acquired at 200Hz with measurements of accelerations along the three sensor axes, angular velocities around the three sensor axes, and orientations of the sensors with respect to a global reference frame. There are also records of the occurrences of events like LED illumination and alarm buzzer activation. Ground-truth pose trajectories can be generated by utilizing the orientation and angular velocity as given and using the provided participant kinematic data along with orientations to derive positions. Velocities are then obtained by single differences.

The IMM filter will be compared against alternative single-modal models like quaternion UKF [10] and multiplicative EKF [16]. Since the ground truth data from the dataset is higher-frequency than is likely available for real-time hand pose detection in an industrial setting, measurements are down-sampled to 50Hz (with noise as mentioned previously) to better represent the processing speed of a possible vision-based detection system.

E. Metrics

Tracking performance (trajectory estimate RMSE) will be used as a primary evaluation metric, with normalized estimation error squared (NEES) utilized to check whether the filters are over- or under-confident. A NEES above the state dimension indicates overconfidence (the estimation is off but covariance is not high enough) while a NEES below the state dimension indicates underconfidence (the estimation is close to ground truth but covariance is too high) [18]. Additionally,

we will consider the evolution of the IMM's conditional model probabilities during "abrupt" operator movements to see whether they could be utilized as a lower-latency intent estimation signal than the evolution of the state estimation itself (comparing against predicted velocity spikes beyond a certain threshold). Because these mode probabilities are subject to high-frequency shifts, we will utilize an exponential moving average (EMA) filter to produce a more "legible" representation of the probabilities for semantic interpretation.

Certain specific metrics are relevant for analyzing the quality of abrupt movement detection. High recall (% of actual positives successfully predicted) is desired to ensure abrupt events are not missed, given the possibility of injury to operator by a robot which fails to classify an abrupt movement. However, high precision (% of positive predictions which were actually positive) is also important to ensure that false positives do not make the system less efficient. Thus, F1 score, the harmonic mean of R and P ($2\frac{RP}{R+P}$) is an appropriate metric for balancing both.

IV. Results & Discussion

A. Parameter Tuning

Before evaluating the filters against each other, it was important to identify appropriate values for the various hyperparameters, most importantly the mode transition probability matrix and the process noise covariances Q_{CV} and Q_M. Additionally, the span of the EMA and the mode probability threshold above which should be considered a detection of an abrupt movement needed to be selected to ensure, if possible, both high recall and high precision. For the purpose of brevity, assume all selected Q are scalings of the identity matrix of the appropriate size.

After identifying reasonable ranges for these hyperparameters with some initial exploration, a grid search was used to identify optimal values. First, EMA spans of 20, 30, and 40 samples, probability thresholds of 0.4, 0.5, 0.6, and 0.7, values for diag(Q_{CV}) of $5 * 10^{-7}$, 10^{-6}, and $5 * 10^{-6}$, and diagonal values for the mode transition matrix of 0.95, 0.98, and 0.99 were considered. Each combination was evaluated on a subset of 10 trajectories. Results from this initial parameter sweep revealed that there is a tradeoff between filter latency and prediction precision. Lower EMA spans or probability thresholds yielded more false positive detections, as did decreasing Q_{CV} too far from the selected Q_M of 10^{-3}. In general, using the EMA to smooth the probabilities naturally causes detection latency, but without it the mode probabilities are too "jumpy" to indicate anything actionable. The highest performing combinations in terms of F1 score are presented in Table I.

Another interesting result from this initial parameter exploration was that the audio buzzer events (requiring the user to abruptly raise their hand in the air) were more easily detected than the visual interruption events (requiring the user to redirect from one delivery station to another). This is evident in Fig. 2, the mode probabilities over the course of a trial.

Table I
Top 5 Filter Configurations (Ranked by F1-Score)

Hyperparameters				Detection				Tracking
Span	**Thr.**	Q_{CV}	**PI**	**F1**	**Rec.**	**Prec.**	**Lat. (s)**	**RMSE (mm)**
40	0.7	5e-7	0.99	0.68	0.63	0.74	0.91	13.97
30	0.7	5e-7	0.95	0.67	0.58	0.79	0.85	13.61
40	0.6	1e-6	0.95	0.67	0.60	0.75	0.80	13.57
40	0.6	5e-7	0.95	0.66	0.60	0.73	0.80	13.61
40	0.7	1e-6	0.99	0.66	0.60	0.73	0.92	13.93

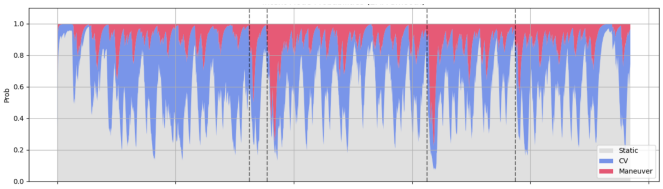

Figure 2. Large spikes are seen in the probability of the "Maneuvering" mode (red) after the interruption events (dashed vertical lines).

As a result, it was relevant to split up the recall / precision / F1 score outcomes according to the type of alarm. Another parameter sweep was conducted with this consideration, yielding the final values of $Q_{CV} = 10^{-6}$, diagonal entries for p_{ij} of 0.95, EMA span of 40, and probability threshold for detection of 0.6, which were associated with the second highest overall F1 score (including 100% recall for the audio alarms) and the lowest detection latency for the audio alarms of 0.81s.

Table II
Top 5 Filter Configurations with Modality Breakdown

Hyperparameters				Comb.	Visual (V)		Audio (A)		Tracking
Span	**Thr.**	Q_{CV}	**PI**	**F1**	**Rec.**	**Lat. (s)**	**Rec.**	**Lat. (s)**	**RMSE (mm)**
40	0.7	5e-7	0.99	0.68	0.25	0.92	1.00	0.91	13.97
30	0.7	5e-7	0.95	0.67	0.15	0.68	1.00	0.88	13.61
40	**0.6**	**1e-6**	**0.95**	**0.67**	**0.20**	**0.75**	**1.00**	**0.81**	**13.57**
40	0.6	5e-7	0.95	0.66	0.20	0.75	1.00	0.81	13.61
40	0.7	1e-6	0.99	0.66	0.20	0.98	1.00	0.91	13.93

In addition to tuning these parameters for the IMM, it was also important to determine appropriate values for the process noise Q used in the constant velocity models for the MEKF and the QUKF to ensure there was a fair comparison between the models. A sweep through Q values on a sample trajectory yielded the following results:

Table III
QUKF/MEKF Process Noise Sweep Results

	QUKF		MEKF	
Q	**Pos. RMSE (mm)**	**NEES**	**Pos. RMSE (mm)**	**NEES**
1e-6	44.01	9672.93	43.64	9667.00
1e-5	17.70	699.66	17.34	697.73
5e-5	13.77	115.89	13.61	115.57
1e-4	**14.40**	**56.45**	**14.30**	**56.32**
5e-4	17.10	11.80	17.03	11.79

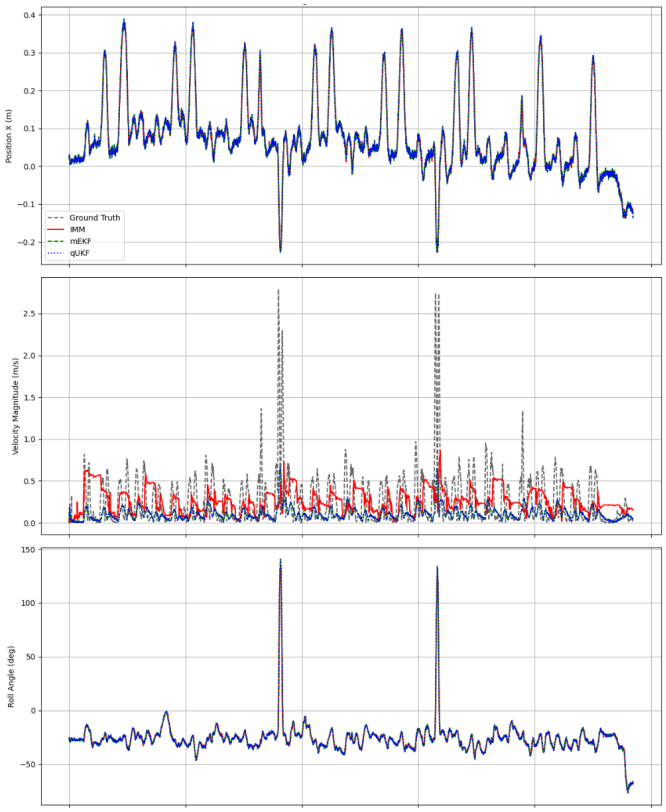

Figure 3. Tracking performance on example trajectory for IMM, MEKF, and QUKF.

While a choice of $Q = 5 * 10^{-4}$ would yield the most consistent filter in both cases, we instead choose the slightly lower $Q = 10^{-4}$ for better tracking performance for a fairer comparison to the IMM, which, due to the way it was constructed to give the opportunity of extracting intent from the changing filter modes, also has a higher NEES indicating overconfidence. We can also note that this is between $Q_M = 10^{-3}$ and $Q_{CV} = 10^{-6}$.

Another important outcome from this initial parameter sweep and performance comparison of the MEKF and QUKF is that while the performance of both seem to be very similar, the runtime of the QUKF is much higher. The average time to compute one step of the MEKF is 0.23ms while the average time for one step of the QUKF is 8.89ms.

B. Initial Filter Comparison

After identifying these appropriate model parameters, an initial comparison of the models was performed. Results from running all three filters on one sample trajectory can be seen in Fig. 3.

Relevant outcomes from this initial test were as follows: all filters performed similarly, with position RMSE of 12.98, 13.15, and 13.22mm for the IMM, MEKF, and QUKF, respectively. The ANEES of 245, 216, and 441 are all far above the state dimensions, but this was to be expected with the choice of parameters above. Finally, we again see the large runtime difference, with execution time of 0.45ms/step for the IMM, 0.21ms/step for the MEKF, and 7.70ms/step for the QUKF. Given this infeasible runtime difference and considering that the prediction quality is also lacking in this sample of results, we choose to exclude the QUKF from the full dataset filter comparison.

Another important shortcoming evident in the performance of position/orientation tracking compared with velocity/angular velocity is it seems that with the chosen model parameters, we are "trusting" the measurements quite heavily. For the full dataset, we will test using measurements downsampled to 50Hz and downsampled further at 10Hz to see if any further differences between the filters emerge in terms of predictive quality. It is important to note that even though we may be relying heavily on the measurements, the shifting mode probabilities for the IMM due to these measurement innovations may still give some semantic indication with useful qualities.

C. Full Dataset Evaluation

Finally, an evaluation was carried out on the full dataset (176 trajectories where four alarm events were triggered) to compare the IMM against the single-model MEKF alternative. Comparisons were carried out both with 50Hz updating (the standard, realistic rate) and at a further down-sampled 10Hz to test the limits of our selected parameters. The results of the 50Hz run are as follows:

Table IV
50HZ PERFORMANCE COMPARISON: IMM VS. MEKF

Metric	**IMM**	**MEKF**
Tracking & Consistency		
Pos. RMSE (mm)	13.00	13.51
Att. RMSE (deg)	2.96	3.38
Mean NEES	306.32	384.54
Combined Detection		
F1-Score	0.58	0.38
Precision	0.63	0.26
Recall	0.54	0.67
False Positives	224	1318
Visual Alarms		
Recall	0.19	0.34
Latency (s)	0.89	0.55
Audio Alarms		
Recall	0.89	0.99
Latency (s)	0.80	0.55

We can see from these results that the tracking performance is similar for both filters, and both filters are inconsistent (to about the same degree). However, important differences exist for the detection of abrupt movements. We consider this signal to be triggered for the IMM when the "M" mode probability crosses the threshold of 0.6, while this signal is triggered for the MEKF when the predicted velocity exceeds 0.3m/s. While the MEKF has lower latency and slightly higher recall for both types of alarms, it raises too many false positive alerts, yielding

a lower F1 score of 0.38 compared to 0.58 for the IMM. As an interesting comparison, we also present the results from an alternative run through the entire dataset with downsampling to 10Hz instead (and the EMA window adjusted to 8 samples to still account for 0.8s):

Table V
OVERALL PERFORMANCE COMPARISON: IMM VS. MEKF (10.0 HZ)

Metric	**IMM**	**MEKF**
Tracking & Consistency		
Pos. RMSE (mm)	15.37	29.52
Att. RMSE (deg)	3.19	7.84
Mean NEES	213.68	2026.24
Combined Detection		
F1-Score	0.12	0.48
Precision	0.06	0.40
Recall	1.00	0.60
False Positives	10261	629
Visual Alarms		
Recall	1.00	0.22
Latency (s)	0.16	0.70
Audio Alarms		
Recall	1.00	0.98
Latency (s)	0.17	0.64

Now, neither filter is particularly usable. The tracking of the MEKF has completely deteriorated, while the IMM has become overly sensitive. Effectively, every measurement innovation is so large that it is considered a "manuever." Clearly, the usability of the IMM's mode probabilities for intent/semantic estimation is very sensitive to the choice of hyper-parameters. We also likely owe some of this breakdown to the deterioration of the small-angle approximation assumptions.

D. Discussion

The results from the full dataset evaluation highlight a fundamental tradeoff to this IMM design. By separating the process noise covariances (Q_{CV} and Q_M) by several orders of magnitude, we forced the filter to transition between modes to account for sudden movements. However, this parameter selection yields a statistically "overconfident" filter, as evidenced by the high Mean NEES values for both the 50Hz and 10Hz evaluations. This allows us to use the shifting mode probabilities as an embedded intent estimator, classifying abrupt movements with a higher F1-score and far fewer false positives than a simple velocity threshold applied to the MEKF.

Despite the advantage in precision, the prediction latency of the IMM remains a significant hurdle. Detection latency of 0.80s to 0.89s is likely insufficient for realistic usage in a fast-paced industrial or human-robot collaborative environment, where safety systems must react to abrupt human movements in milliseconds. The required EMA smoothing window (since the model is currently effectively classifying each new measurement innovation as matching one of the three models), is the primary contributor to this delay.

Furthermore, the alternative evaluation at 10Hz demonstrates that the IMM's utility as a precise intent estimator is highly sensitive to hyper-parameter selection and sensor rate. At a lower update frequency, the increased gap between measurements causes the constant-velocity prediction to naturally diverge further from the true human motion. Then, every measurement innovation becomes large enough to trigger the "maneuvering" mode hypothesis, yielding an overly sensitive filter. Clearly, the hyper-parameters cannot be generalized across different setups without rigorous re-tuning.

V. LIMITATIONS, FUTURE WORK & CONCLUSION

A primary limitation of this study is the complexity of the high-dimensional hyper-parameter space. Optimizing the process noise matrices (Q), the mode transition probability matrix (Π), the measurement noise covariance (R), the smoothing window span, and the detection threshold simultaneously is a non-convex problem, and the grid search used here likely settled on a local optimum. A specific shortcoming of these chosen hyper-parameters were their leading to a filter with a very high Kalman gain. Because the filter trusts the measurement innovations too closely, it exhibits overconfidence (high NEES) and struggles to gracefully reject noise, which especially contributed to the degraded performance at 10Hz.

Additionally, the methodology for extracting the "ground truth" state variables from the dataset was slightly flawed. Because the original DASIG dataset does not provide true continuous velocity, these variables were estimated numerically from the pose data (which itself was derived from kinematics). This approach is vulnerable to noise amplification, evidenced by the large spikes in ground truth velocity during abrupt movements. These spikes artificially inflate the RMSE during transitions and make the evaluation of the MEKF's velocity-based detection threshold overly sensitive.

Future work should address the latency and hyper-parameter sensitivity issues by exploring data-driven sequence models. Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks could be applied to the signal analysis and detection task, replacing the EMA and static thresholding of the IMM's mode probabilities with a model that learns the temporal signatures of true alarms versus normal operational noise. Furthermore, an RNN could potentially be applied to the filtering task itself, learning the to predict end-to-end rather than relying on handcrafted process noise matrices.

Overall, this work demonstrated that by utilizing "naive" additive EKFs within an IMM architecture, we successfully achieved comparable tracking accuracy to a MEKF while gaining the ability to semantically estimate operator intent. While, with appropriately chosen model parameters, the IMM excels at suppressing false positive detections compared to a baseline velocity threshold, future work to refine detection latency and state extraction are required to deploy such a system in a live, human-robot collaborative environment.

Find code at my Github.

References

[1] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland Oregon USA: ACM, Mar. 2, 2015, pp. 51–58, ISBN: 978-1-4503-2883-8. DOI: 10.1145/2696454.2696473.

[2] M. Rubagotti, I. Tusseyeva, S. Baltabayeva, D. Summers, and A. Sandygulova, "Perceived safety in physical human–robot interaction—a survey," *Robotics and Autonomous Systems*, vol. 151, p. 104 047, May 2022, ISSN: 09218890. DOI: 10.1016/j.robot.2022.104047.

[3] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, "Interacting multiple model methods in target tracking: A survey," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 103–123, Jan. 1998, ISSN: 00189251. DOI: 10.1109/7.640267.

[4] A. Roby-Brami, N. Bennis, M. Mokhtari, and P. Baraduc, "Hand orientation for grasping depends on the direction of the reaching movement," *Brain Research*, vol. 869, no. 1, pp. 121–129, 2000, ISSN: 0006-8993. DOI: https://doi.org/10.1016/S0006-8993(00)02378-7.

[5] A. Genovese, "The interacting multiple model algorithm for accurate state estimation of maneuvering targets," *Johns Hopkins APL Technical Digest (Applied Physics Laboratory)*, vol. 22, pp. 614–623, Oct. 2001.

[6] E. Digo, M. Polito, E. Caselli, L. Gastaldi, and S. Pastorelli, "A dataset of standard and abrupt industrial gestures recorded through MIMUs," *Robotics*, vol. 14, no. 12, p. 176, Nov. 28, 2025, ISSN: 2218-6581. DOI: 10.3390/robotics14120176.

[7] P. Del Moral, "Nonlinear filtering: Interacting particle resolution," *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, vol. 325, no. 6, pp. 653–658, 1997, ISSN: 0764-4442. DOI: https://doi.org/10.1016/S0764-4442(97)84778-7.

[8] D. Lee, C. Liu, and J. K. Hedrick, "Interacting multiple model-based human motion prediction for motion planning of companion robots," in *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, West Lafayette, IN, USA: IEEE, Oct. 2015, pp. 1–7, ISBN: 978-1-5090-1959-5. DOI: 10.1109/SSRR.2015.7443013.

[9] I. Bar-Itzhack and Y. Oshman, "Attitude determination from vector observations: Quaternion estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-21, no. 1, pp. 128–136, Jan. 1985, ISSN: 0018-9251. DOI: 10.1109/TAES.1985.310546.

[10] E. Kraft, "A quaternion-based unscented kalman filter for orientation tracking," in *Sixth International Conference of Information Fusion, 2003. Proceedings of the*, Cairns, Queensland, Australia: IEEE, 2003, pp. 47–54, ISBN: 978-0-9721844-4-1. DOI: 10.1109/ICIF.2003.177425.

[11] N. Enayati, E. De Momi, and G. Ferrigno, "A quaternion-based unscented kalman filter for robust optical/inertial motion tracking in computer-assisted surgery," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2291–2301, Aug. 2015, ISSN: 0018-9456, 1557-9662. DOI: 10.1109/TIM.2015.2390832.

[12] H. Himberg, Y. Motai, and A. Bradley, "A multiple model approach to track head orientation with delta quaternions," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 90–101, Feb. 2013, ISSN: 2168-2267, 2168-2275. DOI: 10.1109/TSMCB.2012.2199311.

[13] C. Hertzberg, R. Wagner, U. Frese, and L. Schröder, "Integrating generic sensor fusion algorithms with sound state representations through encapsulation of manifolds," *Information Fusion*, vol. 14, no. 1, pp. 57–77, Jan. 2013, ISSN: 15662535. DOI: 10.1016/j.inffus.2011.08.003.

[14] J. Solà, "Quaternion kinematics for the error-state KF," 2015.

[15] T. L. Koller and U. Frese, "The interacting multiple model filter on boxplus-manifolds," in *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Karlsruhe, Germany: IEEE, Sep. 14, 2020, pp. 88–93, ISBN: 978-1-7281-6422-9. DOI: 10.1109/MFI49285.2020.9235232.

[16] F. L. Markley, "Attitude error representations for kalman filtering," *Journal of Guidance, Control, and Dynamics*, vol. 26, no. 2, pp. 311–317, Mar. 2003, ISSN: 0731-5090, 1533-3884. DOI: 10.2514/2.5048.

[17] I. Bar-Itzhack, J. Deutschmann, and F. Markley, "Quaternion normalization in additive EKF for spacecraft attitude determination," en, in *Navigation and Control Conference*, New Orleans,LA,U.S.A.: American Institute of Aeronautics and Astronautics, Aug. 1991. DOI: 10.2514/6.1991-2706. Accessed: Mar. 16, 2026. [Online]. Available: https://arc.aiaa.org/doi/10.2514/6.1991-2706.

[18] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, "State estimation in discrete-time linear dynamic systems," in *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Ltd, ch. 5, pp. 199–266, ISBN: 9780471221272. DOI: https://doi.org/10.1002/0471221279.ch5. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471221279.ch5. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221279.ch5.