Multiple User Intent Prediction Using Interacting Multiple Model Joint Probabilistic Data Association Filter

Tyler Taplin * **Alexander E. Lyall** ** **Ashwin P. Dani** ***

* *University of Connecticut, Storrs, CT 06269 USA (e-mail: tyler.taplin@uconn.edu)*
** *University of Connecticut, Storrs, CT 06269 USA (e-mail: alexander.lyall@uconn.edu)*
*** *University of Connecticut, Storrs, CT 06269 USA (e-mail: ashwin.dani@uconn.edu)*

Abstract: This paper presents a novel method for multi-user motion intent estimation when the motion is observed by a single sensor. A motion model is associated with each of the activities carried out by the operator and the end location of which is termed as a motion intent. Such modeling of intent is useful in human-robot collaborative tasks. The appropriate model selection is achieved via an interacting multiple model (IMM) filter. When the position measurements of multiple users originating from one sensor are close to each other, then the measurement to operator association becomes challenging. A joint probabilistic data association (JPDA) filter is employed to address this issue. The combined IMM and JPDA filter provides a way to infer the motion intent of each operator. Simulation results show that the IMM-JPDA filter tracks two target states reaching toward goal intent in the presence of clutter measurements originating from the Kinect sensor.

1. INTRODUCTION

Intent inference and/or prediction is a central problem in human-robot collaboration and human-machine (HM) teaming (Dani et al., 2020; Villani et al., 2018; Modares et al., 2017). In manufacturing environment where the product mix is diverse, a complete automation can be cost-inefficient and HM teaming is an appropriate level of automation. Human safety and efficiency of the joint operation are typically studied in HM teaming (Mumm and Mutlu, 2011). It is observed in Li and Ge (2014); Liu et al. (2016); Warrier and Devasia (2016) that incorporating predicted intent in HM teaming improves the efficiency and safety of collaborative tasks for manufacturing robotics and automation and other applications.

Various sensing modalities are used to infer the intent in different application domains, for example, human movement measured using a camera sensor in Mainprice et al. (2015) or a network of camera sensors in Morato et al. (2014) for manufacturing applications, heart rate measurement in Kulic and Croft (2007), and ultrasound measurement for muscle activity in Zhang et al. (2019) for biomedical applications. Based on the application domain, the interpretation of "intent" can be different. In this paper, intent inference is defined as a motion/action intent estimation by using measurements obtained from a Kinect camera sensor. Methods have been developed to infer user motion intent using measurements from a camera sensor and fusion of motion and gaze cues in Ravichandar and Dani (2016); Ravichandar et al. (2018); Trombetta et al. (2020). In Wang et al. (2021), multi-modal data is used to predict the action intent in human-robot handover task. In Ravichandar and Dani (2016), the action/motion intent inference is formulated as a parameter estimation problem, which is solved by using the maximum likelihood estimation. The methods in Ravichandar et al. (2018); Trombetta et al. (2021) formulate the problem in an interacting multiple model (IMM) and variable-structure IMM framework Bar-Shalom et al. (2001), respectively. The method in Trombetta et al. (2021) can handle a larger number of reaching motion intent using a model set augmentation (MSA) algorithm. The MSA selects fewer number of models out of many using constraints from the reachable workspace and gaze cues. A hyperface convolution neural network (CNN) method is used to estimate the gaze direction from the camera images. Information from the two IMMs is fused, one from the gaze IMM and other from the reaching motion IMM. These methods are able to generate promising results when the reaching motion intent analysis is performed for one user.

In this paper, a novel method to infer motion intent of multiple users is developed using a combination of a JPDA filter with an IMM filter. The IMM-JPDA filter (Blom and Bloem, 2002; Chen and Tugnait, 2001) has been used for multi-target tracking applications in Chen and Tugnait (2001) and multi-sensor, multi-target tracking in Tugnait (2004). In this paper, the IMM-JPDA filter from Blom and Bloem (2002) is used for goal reaching motion intent prediction, where each reaching goal motion is characterised as an intent model. The user or target

⋆ This work was supported in part by NSF grant no. SMA-2134367 and in part by a Space Technology Research Institutes grant (number 80NSSC19K1076) from NASA Space Technology Research Grants Program.

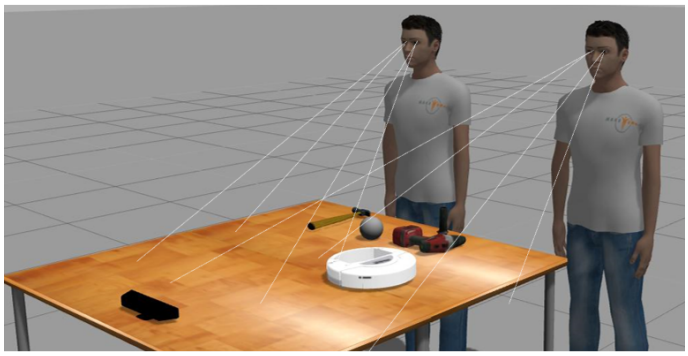

Fig. 1. Modeling of motion intent for multiple users.

motions are observed using a 3D camera sensor as shown in Fig. 1. The problem is formulated as a model selection problem from various number of possible motion models per target in the sense of an IMM filter. Since only one sensor is available for the measurement and many possible measurements can originate from skeletal tracking of Kinect camera sensor at a time instance, the association of sensor measurement to its correct target for state estimation of user's motion intent becomes challenging. To overcome this challenge, a JPDA filter is used, which computes associations of the measurements of the target's previous state to the current motion state before it is used for updating the state in the IMM filter. Similar to JPDA, it is assumed that at-most one measurement originates from the target within the validation region. The rest of the validated measurements are assumed to be from clutter. A dynamic neural network (DNN) is trained on human reaching motion data to produce trajectories from hand position to a goal location in the single target case, see Ravichandar et al. (2018); Trombetta et al. (2020). A combined model is constructed by stacking the states and evolving the two motion profiles independent of one another. A contraction constraint developed in Lohmiller and Slotine (1998) is used such that the motion converges at the goal location specified in the model. The proposed method is implemented in simulation using skeletal joint data obtained from the Kinect sensor. The results show the ability of the proposed method to interpret correct reaching motion intent for two targets to two different objects placed on a workbench.

2. MULTIPLE USER INTENT ESTIMATION PROBLEM

Consider a human-robot collaborative setup where the human operators or targets are carrying out a sequence of tasks to support the robot with its own function, e.g., picking up an object from different locations and handing them to the robots for manipulation or assembly, or carrying out a joint task with the robot, e.g., moving a large object together. A motion model is associated with each of the tasks carried out by the targets and the tasks being considered are goal oriented tasks, where the end location is known in 3D space. The motion models are learned a-priori from data and are stored as a library of models. At each time instance, the motion of the operators is observed by a Kinect sensor mounted on the robot. After image processing and skeletal tracking, a noisy position measurement of the targets is obtained in the senor frame. Based on the position information, the multi-user intent estimation problem is formulated as a selection of the current model under which each of the target's motion is currently evolving. Since each of the motion model is associated with a goal location, the selected model is used to infer the goal motion intents of the operators from one sensor. Since multiple operators are observed in one sensor frame, multiple measurements are extracted from the frame. In addition, there might be spurious measurements that originate from the sensor, for instance, skeletal tracking detects additional joints or some other object in the scene produces a measurement. To tackle this scenario, sensor measurement to operator state data association problem is solved using the JPDA with the IMM filter. In the next section, motion and measurement models are presented, which will be used in the combined IMM-JPDA filter.

3. MOTION AND MEASUREMENT MODELS

Given a set of $n \in \mathbb{N}$ targets $\mathcal{T} = \{1, 2, \ldots, n\}$ and $m \in \mathbb{N}$ models $\mathcal{M} = \{1, 2, \ldots, m\}$ corresponding to a set of m goal locations $\mathcal{G} = \{g(1), g(2), \ldots, g(m)\}$ s.t. $\mathcal{G} \subset \mathbb{R}^3$, let $\theta_k^i \in \mathcal{M}$ indicate the mode of target i at time k. Consider the nonlinear discrete time system for each target

$$x_{k+1}^i = f^i(x_k^i, \theta_k^i) + \nu_k^i, \tag{1}$$

where $x_k^i \in \mathbb{R}^3$ is the state and ν_k^i is i.i.d. multivariate zero-mean Gaussian noise with the covariance $Q_k^i \in \mathbb{R}^{3\times3}$. The nonlinear function $f^i(\cdot)$ is modeled by a DNN of the form

$$f^i(x_k^i, \theta_k^i) = x_k^i + W^T \sigma\left(U^T s_k^i(\theta_k^i)\right) + \epsilon\left(s_k^i(\theta_k^i)\right), \tag{2}$$

where $s_k^i(\theta_k^i) = \left[\left(x_k^i - g\left(\theta_k^i\right)\right)^T, 1\right]^T$ is the input vector to the DNN, $W \in \mathbb{R}^{n_h \times 3}$ and $U \in \mathbb{R}^{4\times n_h}$ are the constant weight matrices with n_h hidden layers. The reconstruction error function $\epsilon(\cdot)$ approaches zero as the DNN is fully trained and the vector Sigmoid activation function

$$\sigma(\cdot) = \left[\frac{1}{1+e^{-(\cdot)_1}}, \frac{1}{1+e^{-(\cdot)_2}}, \cdots, \frac{1}{1+e^{-(\cdot)_d}}\right]^T, \tag{3}$$

is the element-wise application of the sigmoid function where $d = \dim(\cdot)$. The states of each target are stacked in a column vector to form the global state vector $x_k = \left[x_k^1; x_k^2; \cdots; x_k^n\right] \in \mathbb{R}^{3n}$. Similarly, the global noise term $\nu_k = \left[\nu_k^1; \nu_k^2; \cdots; \nu_k^n\right]$ with covariance Q_k, global mode $\theta_k = \left[\theta_k^1; \theta_k^2; \cdots; \theta_k^n\right] \in \mathbb{R}^{nm}$, and global input vector $s_k(\theta_k) = \left[s_k^1(\theta_k^1); s_k^2(\theta_k^2); \cdots; s_k^n(\theta_k^n)\right]$ are defined. The combined weights, $\mathbb{W} = \mathbb{I}_{n\times n} \otimes W$ and $\mathbb{U} = \mathbb{I}_{n\times n} \otimes U$, are block diagonal matrices with n blocks and $\otimes$ is the Kronecker product. Let the global reconstruction error function be $\epsilon\left(s_k(\theta_k)\right) = \left[\epsilon\left(s_k^1(\theta_k^1)\right); \epsilon\left(s_k^2(\theta_k^2)\right); \cdots; \epsilon\left(s_k^n(\theta_k^n)\right)\right]$. Then the global model of all n targets is defined as

$$x_{k+1} = f(x_k, \theta_k) + \nu_k, \tag{4}$$

where the global nonlinear function $f(\cdot)$ is modeled by

$$f(x_k, \theta_k) = x_k + \mathbb{W}^T \sigma\left(\mathbb{U}^T s_k(\theta_k)\right) + \epsilon\left(s_k(\theta_k)\right). \tag{5}$$

Calculating the Jacobian of (5) yields

$$\frac{\partial f}{\partial x_k} = \mathbb{I}_{3\times3} + \mathbb{W}^T \left.\frac{\partial \sigma}{\partial x_k}\right|_{\mathbb{U}^T s_k(\theta_k)} \mathbb{U}^T \frac{\partial s_k}{\partial x_k}, \tag{6}$$

where $\mathbb{U}^T \frac{\partial s_k}{\partial x_k} = \mathbb{I}_{n\times n} \otimes U^T \left[\mathbb{I}_{3\times3} | \mathbf{0}_3\right]^T$ and $\frac{\partial \sigma}{\partial x_k}(\cdot) = \mathrm{diag}(\sigma([\cdot]_1)(1-\sigma([\cdot]_1)), \sigma([\cdot]_2)(1-\sigma([\cdot]_2)), \cdots, \sigma([\cdot]_d)(1-$

$\sigma([\cdot]_d)))$, the state transition matrix can be approximated by

$$F_k = \left.\frac{\partial f}{\partial x_k}\right|_{x_{k-1}} \tag{7}$$

A contraction analysis is used to verify that the DNN exponentially converges to the goal location. Similar to Trombetta et al. (2020), the discrete time contraction analysis is performed in accordance with Lohmiller and Slotine (1998). The optimization problem used to train the DNN is given by

$$[\hat{\mathbb{W}}, \hat{\mathbb{U}}] = \arg\min_{\mathbb{W},\mathbb{U}} \{E_D + \kappa E_W\} \tag{8}$$

where $E_D = \sum_{i=1}^{N_D} ||y_i - a_i||^2$, $y_i \in \mathbb{R}^{3T\times 1}$ represents the state data x_{k+1} and a_i is the output of 5 and E_W is the sum of the squares of the DNN weights, $\kappa \in \mathbb{R}^+$ is a scalar parameter of regularizations. The contraction matrix is determined by using the Jacobian from 7 in the following inequality

$$F_k^T M_{k+1} F_k - M_k \leq -\gamma_c M_k, \qquad M_k > 0 \tag{9}$$

where $\gamma_c \in \mathbb{R}_+$ is a strictly positive constant, $M_k \in \mathbb{R}^{3\times 3}$ represents the uniformly positive definite (PD) contraction metric which is a symmetric matrix as given in Lohmiller and Slotine (1998). The measurement model for target i is given by

$$z_k^i = h^i(x_k^i) + w_k^i \tag{10}$$

where $z_k^i \in \mathbb{R}^3$ and w_k^i are i.i.d. multivariate zero-mean Gaussian measurement noise with covariance $R_k \in \mathbb{R}^{3\times 3}$. A global measurement model is constructed in a similar fashion by stacking the vector such that $z_k = [z_k^1; z_k^2; \cdots; z_k^n] \in \mathbb{R}^{3n}$. To include false measurements and missed detections in the measurement model, a detection vector $\phi_k = [\phi_k^1, \phi_k^2, \cdots, \phi_k^n]^T$ is defined such that $\phi_k^i \in \{0, 1\}$ is a Boolean variable that indicates whether target i is detected, $\phi_k^i = 1$, or not, $\phi_k^i = 0$. Then $D_k = \sum_{i=1}^n \phi_k^i$ is the number of detected targets at time step k. Define the operator Φ, which produces the $D_k \times n$ matrix resulting from removing all zero rows from $\text{diag}\{\phi_k\}$ and $\tilde{\Phi} = \Phi \otimes \mathbb{I}_{3\times 3}$. Furthermore, let χ be a $D_k \times D_k$ permutation matrix with $\bar{\chi} = \chi \otimes \mathbb{I}_{3\times 3}$, then the set of measurements originating from detected targets, $\tilde{z}_k$ is given by

$$\tilde{z}_k = \bar{\chi}\tilde{\Phi}(\phi_k) z_k. \tag{11}$$

Suppose that from L_k measurements there are $D_k \leq L_k$ detected target measurements and $L_k = D_k + C_k$, where C_k is the number of clutter measurements. Then let $C_k > 0$ define the set of measurements for the augmentation of $\tilde{y}_k = [\tilde{z}_k^T; z_k^{*T}]^T \in \mathbb{R}^{3L_k}$, where z_k^* are measurements from clutter. Finally, let $\psi_k = [\psi_k^1, \psi_k^2, \cdots, \psi_k^L]^T$ be an indicator vector introduced to generalize the measurement model such that a varying number of detected measurements from the targets and clutter can be incorporated at each time instance, ψ_k indicates whether a measurement corresponds to a detected target and $\psi_k^* = \mathbb{I}_{L\times 1} - \psi_k$ indicates if measurement comes from clutter. Based on the indicator vectors ψ_k and ψ_k^*, the full measurement model including target and clutter measurements can be expressed as

$$y_k = \left[\tilde{\Phi}(\psi_k), \tilde{\Phi}(\psi_k^*)\right] \tilde{y}_k. \tag{12}$$

Let $\tilde{\chi}_k^T = \chi^T \Phi(\psi_k) \otimes \mathbb{I}_{3\times 3}$ so that the exact measurement model can be written

$$\tilde{\chi}_k^T y_k = \tilde{\Phi}(\phi_k) x_k + \tilde{\Phi}(\phi_k) w_k. \tag{13}$$

where the set of association events is covered by the hypotheses (χ_k, ϕ_k).

4. MULTI-USER INTENT ESTIMATION USING IMM AND JPDA FILTER

A multi-user motion intent filter is presented in this section, using a combined IMM-JPDA filter. The IMM filter is used to select the appropriate model for the ith operator and JPDA is used to associate the measurement coming from the sensor frame to the operator for state estimation. The intent estimation filter is presented in following steps.

4.1 Interaction/Mixing

In this step the initial state and covariance for each filter in the IMM of each target i are computed by mixing the state and covariance from the previous step. Let $\pi_{\theta^i,\theta^j} = Pr\left\{\theta_k^i = \theta^i \middle| \theta_{k-1}^i = \theta^j\right\}$ be the model transition probability and $\mu_{k-1}(\theta^i|\theta^j)$ be the mixing probability computed by

$$\mu_{k-1}(\theta^i|\theta^j) = \frac{\pi_{\theta^i,\theta^j}\mu_{k-1}(\theta^j)}{\bar{\mu}_{k-1}^i(\theta^j)} \tag{14}$$

where $\bar{\mu}_{k-1}^i(\theta^j) = \sum_{\theta^j=1}^m \pi_{\theta^i,\theta^j}\mu_{k-1}^i(\theta^j)$, $\mu_{k-1}(\theta^j)$ is the mode probability of the IMM, the initial state and covariance for ith user and θ^ith mode is computed by

$$\hat{x}_{k-1|\theta_k^i}^i(\theta^i) = \sum_{\theta^j=1}^m \mu_{k-1}(\theta^i|\theta^j)\hat{x}_{k-1}^i(\theta^j), \tag{15}$$

$$\begin{aligned} P_{k-1|\theta_k^i}^i\left(\theta^i\right) = \sum_{\theta^j=1}^m \mu_{k-1}(\theta^i|\theta^j)\times \\ \left(P_{k-1}^i\left(\theta^j\right) + \left[\hat{x}_{k-1}^i\left(\theta^j\right) - \hat{x}_{k-1|\theta_k^i}^i\left(\theta^i\right)\right]\right. \\ \left.\left[\hat{x}_{k-1}^i\left(\theta^j\right) - \hat{x}_{k-1|\theta_k^i}^i\left(\theta^i\right)\right]^T\right) \end{aligned} \tag{16}$$

4.2 Predicted State and Covariance

The predicted state $\bar{x}_k^i$ computed using the θ^ith dynamic model in (4) and the corresponding predicted covariance are given by

$$\begin{aligned} \bar{x}_k^i\left(\theta^i\right) &= f(\hat{x}_{k-1|\theta_k^i}^i(\theta^i), \theta^i) \\ \bar{P}_k^i\left(\theta^i\right) &= F_k P_{k-1|\theta_k^i}^i\left(\theta^i\right) F_k^T + Q_k \end{aligned} \tag{17}$$

where F_k is the Jacobian of (5).

4.3 Measurement Validation using Gating

To validate the measurements at time instance k, a gate $\mathcal{V}_k^i$ is defined for each target i, which is given by

$$\mathcal{V}_k^i = \{z^i : [z^i - \bar{x}_k^i(\hat{\theta}_k^{i*})]^T S_k^i(\hat{\theta}_k^{i*})^{-1}[z^i - \bar{x}_k^i(\hat{\theta}_k^{i*})] \leq \gamma\} \tag{18}$$

where $\mathcal{V}_k^i$ is an elliptical region centered around $\bar{x}_k^i(\theta^i)$ with a gate size $\gamma \in \mathbb{R}^+$ and $\hat{\theta}_k^{i*}$ is obtained by solving the following maximization problem. Find $\hat{\theta}_k^i$ such that

$$\hat{\theta}_k^i = \text{argmax}_{\theta^i}\left\{\det S_k^i(\theta^i)\right\} \tag{19}$$

where the cross covariance $S_k^i(\theta^i)$ for the target i is computed as follows

$$S_k^i\left(\theta^i\right) = H_k P_{k-1|\theta_k^i}^i\left(\theta^i\right) H_k^T + R_k. \tag{20}$$

with $H_k = \mathbb{I}$ as the Jacobian of the measurement model in (10). The validated measurements are obtained by evaluating (18) for each measurement y_k^j. If $y_k^j \in \mathcal{V}_k^i$ then it is considered as a validated measurement and those lying outside the region, $y_k^j \notin \mathcal{V}_k^i$, are considered as false. The reduced set of possible detection and permutation hypothesis at kth time instance for various ϕ is given by $\tilde{\mathcal{X}}_k(\phi)$.

4.4 Evaluation of Measurement to Target Association and Measurement Update

The target state $x_k(\theta^i)$ and the covariance $P_k(\theta^i)$ are updated by incorporating the measurements appropriately weighted by β_k^{ij}. The filter update equations are given below.

$$\hat{x}_k^i(\theta^i) = \bar{x}_k^i(\theta^i) + K_k^i(\theta^i)\Big(\sum_{j=1}^{L_t} \beta_k^{ij}(\theta^i)\nu_k^{ij}(\theta^i)\Big) \tag{21}$$

$$\begin{aligned} P_k^i(\theta^i) = \bar{P}_k^i(\theta^i) &- K_k^i(\theta^i) H_k^i \bar{P}_k^i(\theta^i)\Big(\sum_{j=1}^{L_t} \beta_k^{ij}(\theta^i)\Big) \\ &+ K_k^i(\theta^i)\left(\sum_{j=1}^{L_t} \beta_k^{ij}(\theta^i)\nu_k^{ij}(\theta^i)\nu_k^{ij}(\theta^i)^T\right) K_k^i(\theta^i) \\ &- K_k^i(\theta^i)\left(\sum_{j=1}^{L_t} \beta_k^{ij}(\theta^i)\nu_k^{ij}(\theta^i)\right) \\ &\times \left(\sum_{j=1}^{L_t} \beta_k^{ij}(\theta^i)\nu_k^{ij}(\theta^i)\right)^T K_k^i(\theta^i)^T \end{aligned} \tag{22}$$

where the innovation term is $\nu_k^{ij}(\theta^i) = y_k^j - h^i(\bar{x}_k^i(\theta^i))$, the filter gain $K_k^i(\theta^i)$ and $\beta_k^{ij}(\theta^i)$ term are computed as follows.

$$K_k^i(\theta^i) = \bar{P}_k^i(\theta^i) H_k^T S_k^i(\theta^i)^{-1} \tag{23}$$

$$\beta_k^{ij}(\theta^i) = \sum_{\phi,\tilde{\chi},\eta} \frac{[\Phi(\phi)]_{*i}^T \tilde{\chi}_{*j} \beta_k(\phi,\tilde{\chi},\eta)}{\sum_{\tilde{\chi},\phi} \beta_k(\phi,\tilde{\chi},\theta)} \tag{24}$$

with $\Phi(\phi)_{*i}$ and $\tilde{\chi}_{*j}$ are the ith and jth columns of $\Phi(\phi)$ and $\tilde{\chi}$, and

$$\beta_k(\phi,\tilde{\chi},\theta) = \begin{cases} \dfrac{1}{c_k}\lambda^{L_k - D(\phi)} \displaystyle\prod_{i=1}^{m} [b_k^i(\phi,\tilde{\chi},\theta^i) \\ \times (1-P_d^i)^{1-\phi^i}(P_d^i)^{\phi^i}\bar{\mu}_k^i(\theta^i)], & \text{for} \quad \tilde{\chi} \in \tilde{\mathcal{X}}_k(\phi) \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

where c_k is the normalizing constant, λ is the spatial density of clutter, P_d^i is the detection probability of the user i

$$\begin{aligned} b_k^i(\phi,\tilde{\chi},\theta^i) &= [(2\pi)^{3L_k}\det(S_k^i(\theta^i))]^{-\frac{\phi^i}{2}} \\ &\times \exp\left(-\frac{1}{2}\sum_{j=1}^{L_k}([\Phi(\phi)]_{*i}^T \tilde{\chi}_{*j} \nu_k^{ij}(\theta^i)^T [S_k^i(\theta^i)]^{-1} \nu_k^{ij}(\theta^i)\right) \end{aligned} \tag{26}$$

The mode probability $\mu_k^i(\theta^i)$ is updated for targets $i = \{1,...,n\}$ and models $\theta^i = \{1,..,m\}$ using

$$\mu_k^i(\theta^i) = \sum_{\phi,\tilde{\chi},\theta^i} \beta_k(\phi,\tilde{\chi},\theta) \tag{27}$$

4.5 Combined State Estimates and Covariance

The combined state and covariance of the IMM-JPDA filter is then obtained as follows.

$$\begin{aligned} \hat{x}_k^i &= \sum_{\theta^i=1}^{m}\sum_{\tilde{\chi},\phi} \beta_k(\phi,\tilde{\chi},\theta)\hat{x}_k^i(\theta^i) \\ P_k^i &= \sum_{\theta^i=1}^{m}\sum_{\tilde{\chi},\phi} \beta_k(\phi,\tilde{\chi},\theta)\Bigg(P_k^i(\theta^i) + [\hat{x}_k^i(\theta^i) - \hat{x}_k^i] \\ &\qquad [\hat{x}_k^i(\theta^i) - \hat{x}_k^i]^T\Bigg) \end{aligned} \tag{28}$$

which gives the estimated position of the target i and associated covariance obtained after processing all the measurements in one sensor frame coming from multiple targets and background clutter, i.e., false measurement. The IMM-JPDA filter is implemented for the intent estimation validation as shown in the following section.

5. EXPERIMENTAL RESULTS

5.1 Experimental Setup

Experiments are conducted to validate combined IMM-JPDA filtering for multi-user intent prediction. A Microsoft Kinect motion sensor is used to collect skeletal tracking motion data, and provides the measurements in Euclidean position coordinates. A prefiltering step is utilized for smoothing the measurement signal prior to applying IMM-JPDA filter. Arm motion data of two participants reaching for distinct objects on a workbench is used. In order to highlight the advantages of JPDA filtering, data includes reaching across different objects on the workbench in such a manner as to occlude measurements from the sensor. When hand motions cross in the camera focal plane the Kinect skeletal tracking software is unable to distinguish hand positions and results in erroneous measurements. Another issue arises as targets are mobile around the workstation. Identifiers for associating measurements to specific targets are lost when crossing from one side of the camera view to the other. For these reasons, experiments were performed in an attempt to push the limitations of the measurement technique and provide an ideal testing environment for the application of IMM-JPDA tracking method to reaching intent inference problem.

JPDA filtering allows for loosening assumptions that the identifiers attached to measurements are trustworthy. In addition, multiple target tracking through combining JPDA with IMM filtering is insensitive to scenarios involving persistent false alarms and clutter. The identifiable information of the measurements is removed and the filter relies on validation regions to narrow down the measurements used in the filter update step. The event space for two targets reaching toward two goal locations leads to four models to consider. The IMM distinguishes models

based upon the likelihood that the motion is approaching a goal location. The contraction constraint provides information about which goal is the intended outcome. The IMM-JPDA filter for intent inference is coded and run in Matlab R2022b on a laptop with a 2.8GHz Intel Xeon CPU with 6 cores and 64 GB of RAM.

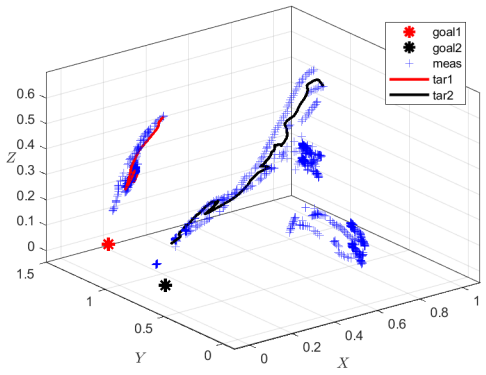

Fig. 2. Trajectories for multiple targets using IMM-JPDA filter for scenario 1.

5.2 Model Parameters and Filter Design

In the experiments, the wrist and hand positions for both the right and left arms of each user are chosen as the input measurements. The benefit of obtaining multiple data points for each target, instead of only considering a single position for each target, is that for situations when a measurement is unreliable or occluded it may be ignored while the alternative measurement may still be processed in the motion model. For our case, this results in eight measurements to be considered prior to measurement validation. The initial target information, i.e., which hand belongs to which target is assumed to be known at the beginning of the simulation with number of targets given by the number of detected skeletons the Kinect sensor tracks. The initial state for the tracking algorithm is determined by averaging the initial position of each hand and wrist pair. Goal locations are known and were determined by selecting pixel locations of objects in the image of the workspace and transforming them from the camera frame to the world frame using known intrinsic parameters of the camera and an estimation of the objects depth from the Kinect sensor.

For the IMM-JPDA filter, the measurement covariance R_k is chosen as $0.5\mathbb{I}_{3\times3}$, while the process covariance Q_k for all the models is selected as $0.1\mathbb{I}_{3\times3}$. Each motion model is produced by considering the combinations of targets and goals. A uniform initial mode probability μ_{k-1} is assigned to the IMM-JPDA filter. The mode transition probability matrix for two modes is set to $\begin{pmatrix} 0.667 & 0.333 \\ 0.333 & 0.667 \end{pmatrix}$. The state prediction from the IMM is used along with a measurement gating threshold to determine which measurements to use for updating each model. A value for gating threshold $\gamma = 0.1$ is chosen to produce a reasonable size gating region for measurement validation. Validation is performed on the set of measurements in the region where the probability of detection for a target is Binomially distributed, with a probability of detection $P_d = 0.99$. False measurements are considered to be Poisson distributed processes with a distribution determined by the expected spatial density of the clutter λ in the validation region. The JPDA filter results in a set of weights used in calculating the update step. The weights calculated in the JPDA filter correspond to the likelihood of measurement to target association and incorporate the clutter.

5.3 Scenario 1

In the first set of experiments, the reaching motion tasks are designed to test the ability of the filter without obscuring the measurements with dense clutter or intermittent sampling rates. For the simple case of two targets reaching toward two goals, the IMM-JPDA filter is able to estimate the state in the presence of clutter. Fig. 2 shows the trajectories for each target's reaching motion produced by the method. As seen from Fig. 2, the estimated target states tend toward the true goal location or intent. In Fig. 4, the Euclidean distance between two target tracks as a function of time is shown in blue. For this scenario, the two target motions do not cross each other, which can be seen from the estimated states of the filter and the track differences show in Fig. 4.

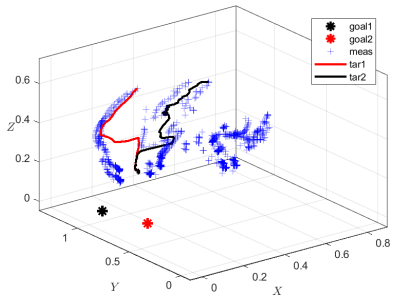

Fig. 3. Track coalescence in the result of the target measurements crossing experiment.

5.4 Scenario 2

In the next set of experiments, the IMM-JPDA filter is tested in a scenario when the operators' reaching motions are close to each other, which provides a challenging scenario for the filter to estimate the states of the targets. Fig. 3 shows the trajectories for each target's reaching motion produced by the method. Initially, when the target measurements are far apart the filter is able to estimate the states of two targets correctly leading to correct estimation of the goal location, but as the measurements from two targets get closer to one another, the estimated target states coalesce leading to difficulties in estimating the reaching goal intent. This happens as a result of complete overlap in the measurements within the validation regions of the targets and can eventually lead to mode collapse. Fig. 4 shows the difference between tracks for the corresponding example scenarios in red. The undesirable behavior is observed due to crossing of targets toward the end of the experiment. Hypothesis pruning has been explored in Blom and Bloem (2002) as a means of avoiding coalescence by disregarding association events which have a lower likelihood of occurrence. The track difference information shown in Fig. 4 can be used to indicate when the targets are crossing. Furthermore, it provides insight into

whether or not incoming and future measurements should be trusted, or if additional information from an alternate sensor is required.

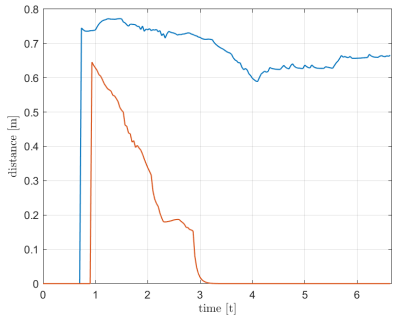

Fig. 4. Distance between tracks for selected examples from twelve experiments. The blue line corresponds to scenario 1 and red line corresponds to scenario 2.

6. CONCLUSION AND FUTURE WORK

A novel multi-user action intent estimation method using IMM-JPDA filter is presented in this paper. The IMM-JPDA filter is used to estimate the position states of two different users from multiple measurements originating from a Kinect reference frame. The filter first computes a validation region for each target's state and uses measurements within the validation region to update the state estimate. The measurements that are coming from other joints of the skeleton are treated as a clutter and weighted appropriately in the filter to update the state estimate. The methods is validated using a simulation scenario where two user are reaching for two different objects. Future work will focus on evaluating the performance of the proposed method on more use cases.

REFERENCES

Bar-Shalom, Y., Li, X.R., and Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation.* John Wiley and Sons.

Blom, H.A. and Bloem, E.A. (2002). Combining IMM and JPDA for tracking multiple maneuvering targets in clutter. In *International Conference on Information Fusion*, 705–712.

Chen, B. and Tugnait, J.K. (2001). Tracking of multiple maneuvering targets in clutter using IMM/JPDA filtering and fixed-lag smoothing. *Automatica*, 37(2), 239–249.

Dani, A.P., Salehi, I., Rotithor, G., Trombetta, D., and Ravichandar, H. (2020). Human-in-the-loop robot control for human-robot collaboration: Human intention estimation and safe trajectory tracking control for collaborative tasks. *IEEE Control Systems*, 40(6), 29–56.

Kulic, D. and Croft, E.A. (2007). Affective state estimation for human–robot interaction. *IEEE Transactions on Robotics*, 23(5), 991–1000.

Li, Y. and Ge, S. (2014). Human-robot collaboration based on motion intention estimation. *IEEE/ASME Transactions on Mechatronics*, 19(3), 1007–1014.

Liu, C., Hamrick, J.B., Fisac, J.F., Dragan, A.D., Hedrick, J.K., Sastry, S.S., and Griffiths, T.L. (2016). Goal inference improves objective and perceived performance in human-robot collaboration. In *International Conference on Autonomous Agents & Multiagent Systems*, 940–948.

Lohmiller, W. and Slotine, J.J.E. (1998). On contraction analysis for nonlinear systems. *Automatica*, 34(6), 683–696.

Mainprice, J., Hayne, R., and Berenson, D. (2015). Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In *IEEE International Conference on Robotics and Automation*, 885–892.

Modares, H., Ranatunga, I., AlQaudi, B., Lewis, F.L., and Popa, D.O. (2017). Intelligent human–robot interaction systems using reinforcement learning and neural networks. In *Trends in control and decision-making for human–robot collaboration systems*, 153–176.

Morato, C., Kaipa, K.N., Zhao, B., and Gupta, S.K. (2014). Toward safe human robot collaboration by using multiple kinects based real-time human tracking. *Journal of Computing and Information Science in Engineering*, 14(1).

Mumm, J. and Mutlu, B. (2011). Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *International Conference on Human-Robot Interaction*, 331–338.

Ravichandar, H.C. and Dani, A.P. (2016). Human intention inference using Expectation-Maximization algorithm with online model learning. *IEEE Transactions on Automation Science and Engineering*, 14(2), 855–868.

Ravichandar, H.C., Kumar, A., and Dani, A.P. (2018). Gaze and motion information fusion for human intention inference. *International Journal of Intelligent Robotics and Applications*, 2(2), 136–148.

Trombetta, D., Rotithor, G., Salehi, I., and Dani, A.P. (2021). Variable structure human intention estimator with mobility and vision constraints as model selection criteria. *Mechatronics*, 76, 102570.

Trombetta, D., Rotithor, G.S., Salehi, I., and Dani, A.P. (2020). Human intention estimation using fusion of pupil and hand motion. In *IFAC World Congress*, 9535–9540.

Tugnait, J.K. (2004). Tracking of multiple maneuvering targets in clutter using multiple sensors, IMM, and JPDA coupled filtering. *IEEE Transactions on Aerospace and Electronic Systems*, 40(1), 320–330.

Villani, V., Pini, F., Leali, F., and Secchi, C. (2018). Survey on human-robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, 55, 248 – 266.

Wang, W., Li, R., Chen, Y., Sun, Y., and Jia, Y. (2021). Predicting human intentions in human-robot hand-over tasks through multimodal learning. *IEEE Transactions on Automation Science and Engineering*, (2), 2339–2353.

Warrier, R.B. and Devasia, S. (2016). Inferring intent for novice human-in-the-loop iterative learning control. *IEEE Transactions on Control Systems Technology*, 25(5), 1698–1710.

Zhang, Q., Kim, K., and Sharma, N. (2019). Prediction of ankle dorsiflexion moment by combined ultrasound sonography and electromyography. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(1), 318–327.