Data-Driven Intention Inference and Its Application to Human-Robot Coordination

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Yongming Qin

August 2023

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Yongming Qin

This Dissertation has been read and approved by the examing committee:

Advisor: Tomonari Furukawa

Advisor:

Committee Member: John A. Stankovic

Committee Member: Zongli Lin

Committee Member: Gregory J. Gerling

Committee Member: Lu Feng

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

August 2023

ABSTRACT As robots become increasingly integrated into daily life, their ability to effectively coordinate with human-maneuvered agents, including humans, cars, and drones, becomes crucial. However, existing techniques for human-robot coordination have treated all human agents uniformly, without considering the unique characteristics of each agent. This PhD dissertation proposes a research framework that incorporates the specific attributes of individuals involved in the coordination process and aims to design adaptive coordination abilities for robots.

Inspired by how humans interact and collaborate with familiar individuals, this research investigates how robots can learn from human behavior to enhance their coordination capabilities. Humans intuitively infer each other's intentions and take into account the historical behavior of others to improve collaboration. The proposed research tackles three key aspects: 1) Developing an efficient approach to model human intentions based on historical behavior data. Different types of intentions are classified, and each intention is mapped to a corresponding motion pattern. 2) Inferring human intentions and utilizing the model to determine the current motion pattern for effective coordination. 3) Incorporating inferred intentions and motion patterns into robotic applications. The proposed approaches are applied to two specific applications: state estimation and robotic escorting.

The effectiveness of the proposed approaches is validated through simulations and real experiments. The novel model for intention and motion patterns demonstrates significant advantages in efficiently describing human behavior. In both the state estimation of a human-maneuvered quadrotor and the robotic escorting of a human using a mobile robot, the proposed approaches exhibit benefits such as higher estimation accuracy, enhanced flexibility, and improved user satisfaction.

To my family for their continuous support.

ACKNOWLEDGMENTS

The memories of people who have touched my life have always warmed my heart and nurtured my thoughts. I am immensely grateful to the individuals who have shaped my journey and influenced me in profound ways.

I would like to express my deepest appreciation to my family for their enduring love, unwavering encouragement, and profound understanding throughout this challenging journey. Their patient belief in my decisions and selfless sacrifices have been a constant source of motivation.

I extend my heartfelt gratitude to the faculty and staff of the University of Virginia for fostering a supportive culture and creating a stimulating academic environment. I am particularly grateful to my advisor, Dr. Tomonari Furukawa, whose guidance, support, and mentorship have been instrumental in shaping my doctoral journey. His approachability, unreserved experience sharing, and diligent spirit continue to inspire me and broaden my horizons. I am also indebted to the members of my dissertation committee, Dr. Jack Stankovic, Dr. Zongli Lin, Dr. Gregory Gerling, and Dr. Lu Feng, for their timely feedback, professional guidance, and invaluable insights.

I am deeply appreciative of my friends and colleagues who have provided reliable support, engaging intellectual discussions, and much-needed moments of respite. While countless names flood my mind, spanning across numerous pages, those precious and joyful moments will eternally reside in my memory, even as names may become blurry.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

2D	two-dimensional
3D	three-dimensional
CT	coordinated turn
CV	constant velocity
HRI	human-robot interaction
IMM	interacting MM
KF	Kalman lter
MM	multiple-model
NN	neural network
PF	particle lter
UKF	unscented Kalman lter
MSE	mean squared error

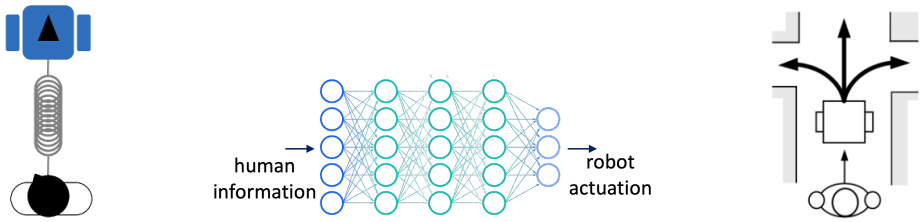
1. Introduction

1.1 Motivation

With the increasing deployment of robots in isolated workspaces, their presence in our daily lives around humans is becoming more prevalent [1, 2]. Small aerial vehicles now assist in photography, three-dimensional (3D) reconstruction, and monitoring, while ground vehicles transport people and goods to various destinations. Public service robots have emerged to interact with humans, providing information and guiding individuals [3]. Assisting robots aid in carrying heavy objects and installing materials [4, 5]. Domestic robots have demonstrated capabilities in performing daily household tasks, including kitchen activities, clothes handling, and sweeping [5]. However, to operate seamlessly in human-populated environments, these robots must effectively interact with various human-maneuvered agents, including humans themselves, as well as agents controlled by humans such as cars and drones. Therefore, the coordination between robots and humans is crucial for their coexistence and cooperation.

Past approaches to human-robot coordination can be categorized into three distinct types, as illustrated in Figure 1.1. The first type, known as physics-based approaches, encompasses the description of human motion through explicit dynamical models based on Newton's laws of motion, which are then utilized to design the coordination strategy [6, 7]. These approaches often employ proportional control strategies for robot tracking and following. The second type, referred to as blackbox-based approaches, leverages deep learning techniques to generate robot actuation [8]. In this approach, the neural network takes human information as input and produces the corresponding robot control commands as output. It essentially learns the mapping between human behavior and robot actions. The third type, pattern-based approaches, describes human motion using predefined patterns [9]. These approaches predict future human motion based on these patterns and subsequently design the coordination strategies for robots to effectively respond to the anticipated human motion.

These three approaches have proven successful in enabling robots to operate in close proximity to humans and have gained widespread application. However, these approaches were designed without considering the unique characteristics exhibited by individuals. It is



(a) Physics based (b) Blackbox based (c) Pattern based

Figure 1.1. Past work on human-robot motion coordination can be classified into three types: physics based, blackbox based, and pattern based approaches.

essential to acknowledge that each person is distinct, displaying specific habits and motions. Therefore, it becomes crucial to customize the coordination of robots for each specific human, accounting for their individuality. None of the three approaches addressed this need for tailored coordination between robots and humans.

1.2 Development

The proposed research draws inspiration from human interaction and collaboration, aiming to replicate and enhance these abilities in robots. Humans possess the capacity to infer each other's intentions and incorporate knowledge of others' historical behavior. To imbue robots with similar capabilities, several key challenges must be addressed: extracting motion characteristics from historical data, inferring human intention, utilizing human intention effectively, and developing the coordination abilities of robots.

In light of these challenges, the proposed research seeks to achieve the following objectives: 1) efficient modeling of human intention using historical data, 2) accurate inference of human intention, and 3) the proposal and validation of effective human-robot coordination through practical implementations.

By accomplishing these objectives, the proposed research makes contributions in the following areas: 1) the efficient extraction of a human's unique characteristics from historical data, and 2) demonstrating the advantages of inferring human intention through two specific robotic applications: state estimation and robotic escorting.

Figure 1.2 shows a diagram of robotic applications with humans involved. The human behaves to influence the robot while the robot observes the human and acts correspondingly.

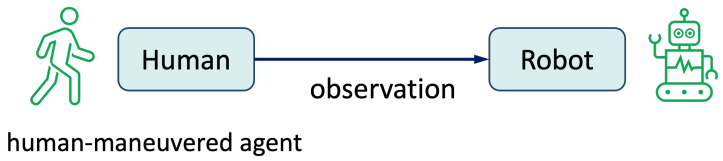


Figure 1.2. A diagram of robotic applications with humans involved

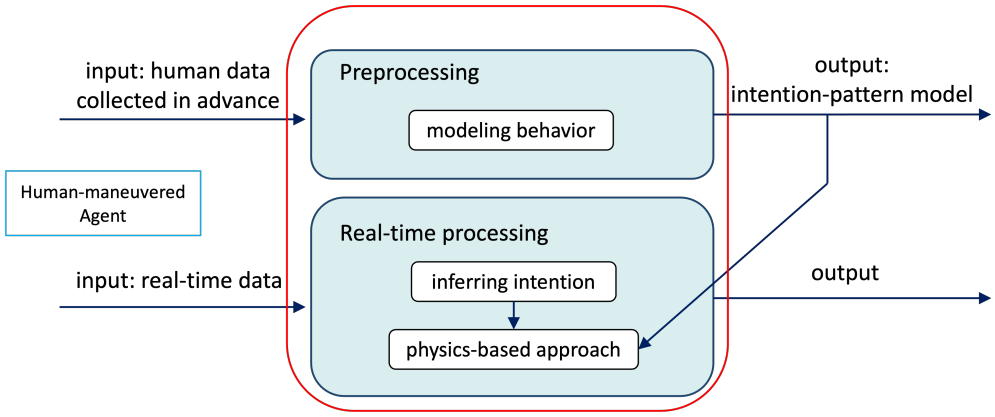


Figure 1.3. The general framework of the proposed research

Figure 1.3 shows the general framework of the proposed research, which takes advantage of both physics-based approaches and pattern-based approaches. There are two processes: preprocessing and real-time processing. The preprocessing captures the special characteristics of a human from the historical data. Human behavior is described by human intentions and motion patterns. The human intention is a discrete value (such as intention 1, 2, 3, ...) that presents what the human plan to do. Each intention is mapped to one motion pattern where the pattern is extracted from the historical data. Human behavior is finally represented as an intention-pattern model.

The real-time processing is for the robotic applications which are state estimation and robotic escorting in this research. The human intention is first inferred based on real-time data. Then by checking the intention-pattern model with the current intention, the

current motion pattern is known. Physics-based approach and the current motion pattern are combined to serve each robotic application.

Figure 1.4 show the specific framework of state estimation. Human-maneuvered agents such as quadrotors teleoperted by people and cars driven by people are treated on the input side. The input of preprocessing is the observation data collected in advance and the input of real-time processing is the real-time observation. By combining the physics-based approach of state estimaiton and the pattern-based approach, the output of estimated state is improved.

Figure 1.5 shows the specific framework of robotic escorting. For robotic escorting, a robot moves in front of a human and complies with the human aim without direct communication. The input of preprocessing is the human demonstration data and the input of real-time processing is the human head direction indicating the human intention. By combining the physics-based approach of common robotic escorting and the pattern-based approach, the robot can better coordinate the human actions in the environment.

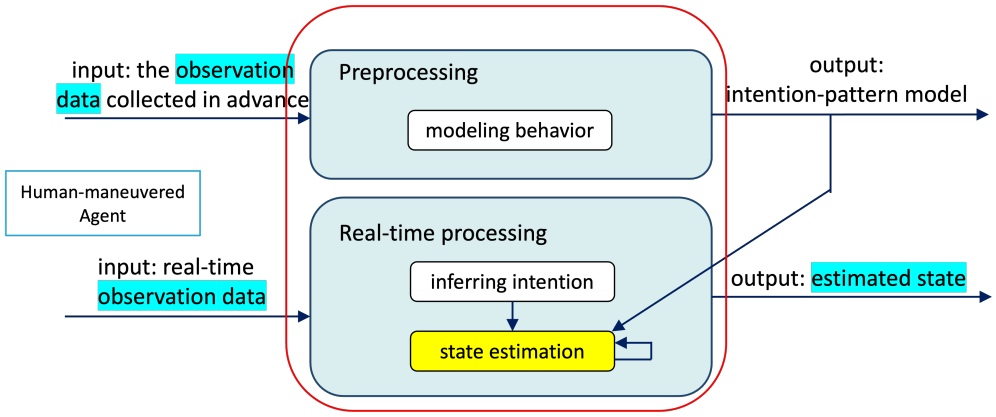


Figure 1.4. The specific framework for state estimation

1.3 Summary of Contributions

This research proposes a framework that learns the unique characteristics of a human from historical data and incorporates the specific attributes of individuals for human-robot coordination. The main contributions are summarized as below.

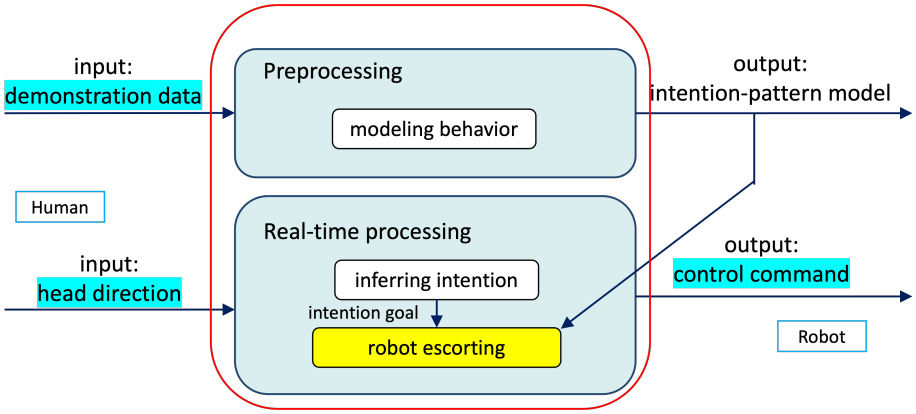


Figure 1.5. The specific framework for robotic escorting

- Efficientlly model human intentions based on historical behavior data of small datasets.
- Infer the human intention with high accuracy.
- Propose and validate the approaches through two applications. For the first application of estimating states of human-maneuvered agents, the proposed research achieves higher accuracy of state with 3 times smaller maximum error and 8.7 times smaller mean squared error (MSE) compared to the based approach. For the second application of robotic escorting, the proposed research leads to smaller prediction error and better user experience.

1.4 Dissertation Structure

In the rest of this dissertation, Chapter 2 proposes the approach for inferring the human intention. Chapter 3 and Chapter 4 respectively present the incorporation of human intention in benefiting state estimation of of human-maneuvered agents and robotic escorting. Finally, Chapter 5 concludes the proposed research and discuss some future directions based on the research.

2. Data-Driven Intention Inferrence

Understanding the aim of a target that is either a human or a robot is a key skill for intelligent systems to coexist with humans or with each other. Related studies play an important role in many areas, such as motion prediction, human-robot interaction (HRI), surveillance, and autonomous driving [10, 11]. Approaches of intention estimation have been developed to estimate such aims based on observations and have attracted increased interest in the last decades [12, 1, 2].

The approaches proposed in the past for intention estimation can be classified into two types. In the first, physics-based approaches describe the target motion with explicit dynamical models based on Newton's law of motion and associate the intentions with the models [13, 14]. Some approaches interpret different intentions with the motion model of different parameters, such as velocities, turning direction, and goals. Kawase et al. [15] considered different turning directions to predict the circular motion for target tracking. Conte et al. [16] predicted the future motion incorporating the human heading direction as an indicator of the target goal for an HRI task. Multiple model (MM) approaches directly utilize several models and relate models to intentions. Liu et al. [17] predicted the trajectory of an aircraft based on the interacting multiple model (IMM) approach and assigned a larger weight to the dynamic model with a heading angle closer to the intended direction. Qin et al. [18] associated intentions with recurring motion patterns which are then described by motion models with different control inputs.

In the second, pattern-based approaches learn the motion behavior from data without explicitly defining the parameterized functions. Bennewitz et al. [19] learned the motion patterns of people from collections of trajectories and derived a hidden Markov model to estimate the current and future positions. Ravichandar et al. [20] trained a neural network (NN) that models the target motion and converges to the goal position. Then different goal positions representing different intentions are coupled with the NN model. Elfring et al. [21] learned the typical motion patterns of humans from data and utilized a person's intended position for improved motion prediction. Dermy et al. [22] learned a set of motion primitives from HRI demonstrations and inferred the intention of the human partner for collaboration.

Lasota et al. [23] determined the most favorable parameters for the prediction methods from task data and improved the motion prediction performance.

Physics-based approaches can be easily interpreted and adapted to new cases. Pattern-based approaches infer related information from data and save the effort to specify parameters [24]. This chapter presents a data-driven multiple model framework for estimating the intention to take the benefit of both. In contrast to conventional approaches which assume the one-to-one relations between intentions and models, the proposed framework infers the multiple-to-multiple relations from labeled observations. Both the multiple-to-multiple relations and the model probabilities of an IMM state estimation approach are incorporated in a recursive Bayesian framework for intention estimation. The strength of the proposed approach lies in the incorporation of the multiple-to-multiple relations which are more common in real cases and also incorporate the dependency information on observations. The proposed framework improves the accuracy of estimation and increases flexibility without specifying the strict one-to-one relations.

The chapter is organized as follows. The next section (2.1) describes the intention estimation problem, its solution using a state estimation approach assuming one-to-one relations, and the resulting limitations. Section 2.2 presents the proposed framework including the recursive Bayesian intention and the inference of multiple-to-multiple relations from labeled observations. Experimental validation investigating the effectiveness of both the inference and the intention estimation is presented in Section 2.3. Conclusions are summarized in the final section (2.4).

2.1 Intention and State Estimation

2.1.1 Intention Estimation

For a target, the discrete motion model and the observation model are generically given by

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_k, \boldsymbol{w}_k) \tag{2.1}$$

$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{v}_k) \tag{2.2}$$

where $\boldsymbol{f}$ and $\boldsymbol{h}$ are the motion and the observation models respectively, $\boldsymbol{x}_k$ is the state of the target at step k, $\boldsymbol{u}_k$ is the input, $\boldsymbol{z}_k$ is the observation, and $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ are the motion and observation noises respectively. A viewer observes the target and estimates the intention. Mostly there is no cooperation between the target and the viewer, thus $\boldsymbol{f}$ and $\boldsymbol{w}_k$ may not be well known while $\boldsymbol{u_k}$ is fully unknown. On the other hand, $\boldsymbol{h}$ and $\boldsymbol{v}_k$ are fully known since they are with the sensor(s) of the viewer. With a short time interval, it is valid to assume that $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ are Gaussian.

The intention is an expression describing an aim or a plan, such as "turning right" and "picking a cup". This chapter denotes intentions discretely as a set $\{\eta^{(a)}\}, a \in \mathbb{N}$ where $\mathbb{N}$ represents all natural numbers. The intention sources from the target, but the understanding of intention is subject to humans. Without knowing the exact intention of the target side, this chapter deals with the intention from the viewer side. The intention at step k is defined as a function of the past observations from step 1 to step k:

$$\eta_k = \boldsymbol{\pi}(\boldsymbol{z}_{1:k}). \tag{2.3}$$

which depends on the viewer and is not explicitly known. Instead, segments of observations that are labeled with intentions η^a are considered to be given. The labeled observations are denoted as a set $\{\boldsymbol{z}^{(a)}_{l_s:l_\mathrm{e}}\}$, which indicates the observations from start step l_s to end step l_e are labeled with intention $\eta^{(a)}$. $l \in \mathbb{N}$ represents the lth segment. Two segments do not overlap, i.e, the intention at each step is unique. The problem of intention estimation is resultantly defined as the estimation of η_k with no knowledge on $\boldsymbol{u}_k$ and $\boldsymbol{\pi}$ and some knowledge on $\boldsymbol{f}$ and $\boldsymbol{w}_k$, given $\boldsymbol{h}$, $\boldsymbol{z}_k$, $\boldsymbol{v}_k$ and $\{\boldsymbol{z}^{(a)}_{l_s:l_\mathrm{e}}\}$.

2.1.2 Interacting Multiple Model State Estimation

For state estimation, the goal is to estimate the target state $\boldsymbol{x}_k$ from observations $\boldsymbol{z}_{1:k}$. Lacking information of $\boldsymbol{f}$ and $\boldsymbol{u}_k$ results in high motion uncertainty. The MM state estimation methods deal with motion uncertainty by describing the motion with several motion models, which are denoted as $\{m^{(\mathrm{i})}\}, \mathrm{i} \in \mathbb{N}$. Most MM methods utilize models of known motion behaviors with different parameters such as variants of the constant velocity (CV) and

constant acceleration (CA) models. Suppose that $\boldsymbol{f}$ of Eq. (2.1) at step k is approximated by a single linear Gaussian model. The motion model $m^{(\mathrm{i})}$ is given by

$$\boldsymbol{x}_k = \boldsymbol{A}_k^{(\mathrm{i})} \boldsymbol{x}_{k-1} + \boldsymbol{B}_k^{(\mathrm{i})} \boldsymbol{u}_k^{(\mathrm{i})} + \boldsymbol{w}_k^{(\mathrm{i})}. \tag{2.4}$$

where $\boldsymbol{A}_k^{(\mathrm{i})}$ is a system matrix, $\boldsymbol{B}_k^{(\mathrm{i})}$ is a control matrix, and $\boldsymbol{w}_k^{(\mathrm{i})}$ is Gaussian with mean $\boldsymbol{0}$ and covariance $\boldsymbol{Q}_k^{(\mathrm{i})}$. The symbol $^{(\mathrm{i})}$ indicates the model $m^{(\mathrm{i})}$. The observation model (2.2) is also supposed to be linear Gaussian:

$$\boldsymbol{z}_k = \boldsymbol{C}_k \boldsymbol{x}_k + \boldsymbol{v}_k \tag{2.5}$$

where $\boldsymbol{C}_k$ is the observation matrix, and $\boldsymbol{v}_k$ is Gaussian with mean $\boldsymbol{0}$ and covariance $\boldsymbol{R}_k$.

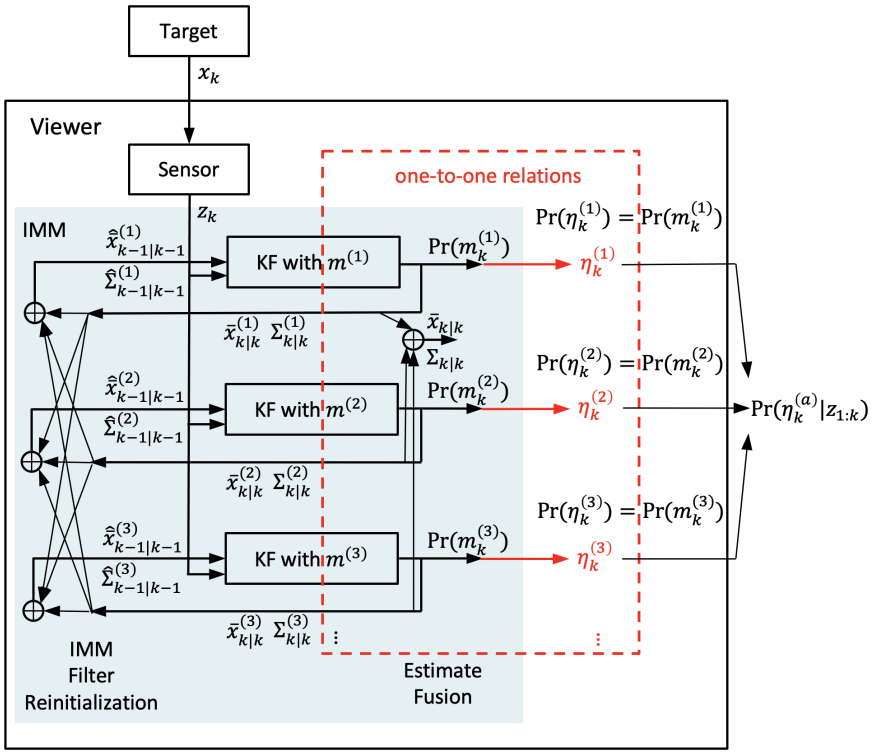


Figure 2.1. The IMM state estimation approach and its application to intention estimation assuming one-to-one relations between intentions and motion models

The shaded area of figure 2.1 shows the IMM approach to estimate the target state. The motion behavior is described with a set of models $\{m^{(1)}, m^{(2)}, m^{(3)}, ...\}$. Having the observation $\boldsymbol{z}_k$ of the state $\boldsymbol{x}_k$, Kalman filter (KF) is applied for each model (2.4). Each KF updates the target state of mean $\hat{\bar{\boldsymbol{x}}}^{(\mathrm{i})}_{k-1|k-1}$ and covariance $\hat{\boldsymbol{\Sigma}}^{(\mathrm{i})}_{k-1|k-1}$ to mean $\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k}$ and covariance $\boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k}$. Each $\hat{\bar{\boldsymbol{x}}}^{(\mathrm{i})}_{k-1|k-1}$ and $\hat{\boldsymbol{\Sigma}}^{(\mathrm{i})}_{k-1|k-1}$ are derived from the IMM filter reinitialization incorporating all of $\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k-1|k-1}$ and $\boldsymbol{\Sigma}^{(\mathrm{i})}_{k-1|k-1}$. The output $\bar{\boldsymbol{x}}_{k|k}$ and covariance $\boldsymbol{\Sigma}_{k|k}$ are calculated by the estimate fusion of all $\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k}$ and $\boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k}$[25].

2.1.3 Conventional Multiple Model Intention Estimation

Denote $\eta^{(a)}_k$ as the event that at step k the intention is $\eta^{(a)}$, i.e, $\eta_k = \eta^{(a)}$. Conventional multiple model framework assumes one-to-one relations between intentions and motion models. The probability of an intention is assumed to be the probability of the corresponding model:

$$Pr(\eta^{(a)}_k) = Pr(m^{\mathrm{i}}_k), \mathrm{i} = a,$$

where $Pr(\cdot)$ indicates the probability of an event. As shown on the right side of figure 2.1, the intention estimation is performed based on this assumption.

There are two main limitations to estimate intention assuming one-to-one relations. First, one intention may not map to one single model exactly. The target behavior of one intention can be too complex to be described by one model. In addition, the model probability does not incorporate the information that an intention is likely to happen with specific observations. Second, multiple model methods are motivated to improve the performance of state estimation. The models designed for satisfactory intention estimation may conflict with state estimation, which reduces the flexibility.

2.2 Proposed Data-Driven Multiple Model Framework

Figure 2.2 shows the proposed framework for intention estimation. Recursive Bayesian intention estimation is designed including modules of intention prediction and intention correction. The intention prediction module predicts the probabilities of all intentions at the

current step based on the probabilities at the previous step. The intention correction module corrects the intention probabilities by incorporating the multiple-to-multiple relations inferred from labeled observations and the model probabilities from IMM. Section 2.2.1 presents the recursive Bayesian intention estimation, and section 2.2.2 presents the inference of the multiple-to-multiple relations from labeled observations.

2.2.1 Recursive Bayesian Intention Estimation

The probability of one intention $\eta^{(a)}$ at step k is derived as

$$\begin{aligned} Pr(\eta_k^{(a)}|\boldsymbol{z}_{1:k}) &= \sum_{\mathrm{i}} Pr(\eta_k^{(a)}, m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k}) \\ &= \sum_{\mathrm{i}} Pr(\eta_k^{(a)}|m_k^{(\mathrm{i})}, \boldsymbol{z}_{1:k}) Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k}) \\ &\propto \sum_{\mathrm{i}} Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k|\eta_k^{(a)}, \boldsymbol{z}_{1:k-1}) Pr(\eta_k^{(a)}|\boldsymbol{z}_{1:k-1}) Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k}) \\ &\propto \sum_{\mathrm{i}} \underbrace{Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k|\eta_k^{(a)}, \boldsymbol{z}_{1:k-1}) Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k})}_{\text{likelihood}} \underbrace{Pr(\eta_k^{(a)}|\boldsymbol{z}_{1:k-1})}_{\text{from prediction}}. \end{aligned} \tag{2.6}$$

Equation (2.6) corresponds to the correction module of the estimation. In the likelihood term, $Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k|\eta_k^{(a)}, \boldsymbol{z}_{1:k-1})$ represents the multiple-to-multiple relations between intentions and motion models. And $Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k})$ is the model probabilities, which is derived from the IMM state estimation.

$Pr(\eta_k^{(a)}|\boldsymbol{z}_{1:k-1})$ is calculated based on $Pr(\eta_{k-1}^{(a)}|\boldsymbol{z}_{1:k-1})$ which is the intention probability at previous step. This process is performed by the prediction module. The proposed framework assumes the probability of transitioning from an intention $\eta^{(a)}$ at step $k-1$ to an intention $\eta^{(b)}$ at step k, $Pr(\eta_k^{(b)})|\eta_{k-1}^{(a)}), b \in \mathbb{N}$, is constant and known, which results in a Markovian process. Given the previous probabilities of all intentions, the prediction probability of each intention is derived as

$$Pr(\eta_k^{(a)}|\boldsymbol{z}_{1:k-1}) = \sum_{b} Pr(\eta_k^{(a)}|\eta_{k-1}^{(b)}) Pr(\eta_{k-1}^{(b)}|\boldsymbol{z}_{1:k-1}). \tag{2.7}$$

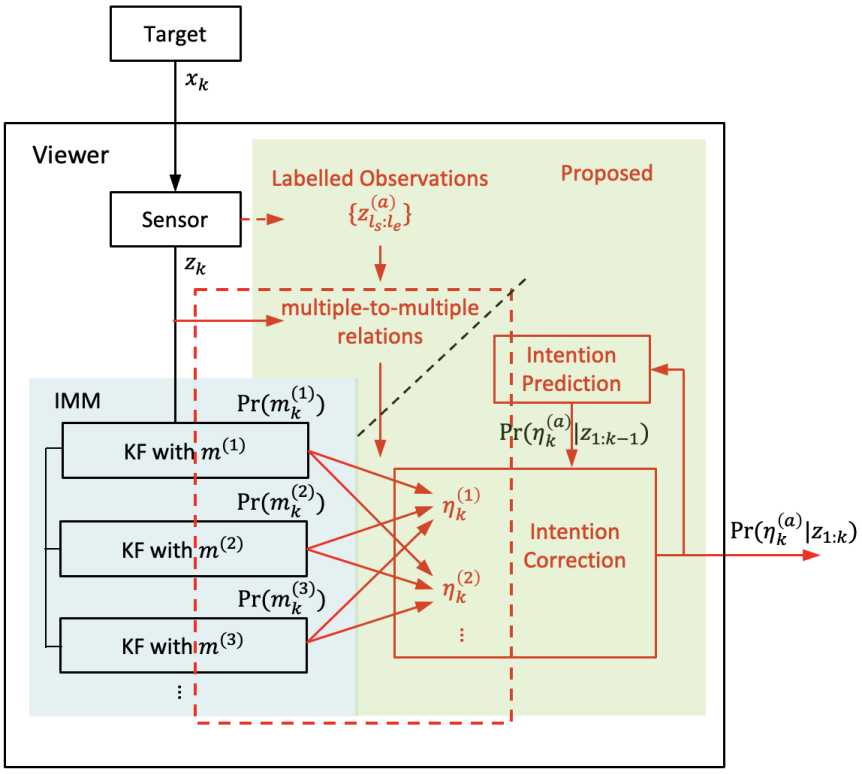


Figure 2.2. The proposed intention estimation framework that infers and utilizes the multiple-to-multiple relations between intentions and motion models from labeled observations.

Similarly, the IMM state estimation assumes the probability of transitioning from a model $m^{(\mathrm{i})}$ at step $k-1$ to a model $m^{(\mathrm{j})}$ at step k, $Pr(m_k^{(\mathrm{j})}|m_{k-1}^{(\mathrm{i})}), \mathrm{j} \in \mathbb{N}$, is also constant and known, which is another Markovian process. The predicted model probability is given by

$$Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k-1}) = \sum_{\mathrm{j}} Pr(m_k^{(\mathrm{i})}|m_{k-1}^{(\mathrm{j})}) Pr(m_{k-1}^{(\mathrm{j})}|\boldsymbol{z}_{1:k-1}).$$

For each KF in IMM state estimation, the observation residual is given by

$$\tilde{\boldsymbol{z}}_k^{(\mathrm{i})} = \boldsymbol{z}_k - \boldsymbol{C}_k \bar{\boldsymbol{x}}_{k|k-1}^{(\mathrm{i})}, \tag{2.8}$$

As indicated by IMM approach, the model likelihood is assumed as

$$\begin{aligned} & Pr(\tilde{\boldsymbol{z}}_k^{(\mathrm{i})}|m_k^{(\mathrm{i})}, \boldsymbol{z}_{1:k-1}) \\ \overset{\text{assume}}{=} & |2\boldsymbol{\pi}\boldsymbol{S}_k^{(\mathrm{i})}|^{-\frac{1}{2}} \exp\big[-\frac{1}{2}(\tilde{\boldsymbol{z}}_k^{(\mathrm{i})})^\top (\boldsymbol{S}_k^{(\mathrm{i})})^{-1}\tilde{\boldsymbol{z}}_k^{(\mathrm{i})}\big]. \end{aligned} \tag{2.9}$$

where $\boldsymbol{S}_k$ is the residual covariance of KF. The model probability is given by

$$\begin{aligned} & Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k}) \\ = & \frac{Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k-1}) Pr(\tilde{\boldsymbol{z}}_k^{(\mathrm{i})}|m_k^{(\mathrm{i})}, \boldsymbol{z}_{1:k-1})}{\sum_{\mathrm{j}} Pr(m_k^{(\mathrm{j})}|\boldsymbol{z}_{1:k-1}) Pr(\tilde{\boldsymbol{z}}_k^{(\mathrm{j})}|m_k^{(\mathrm{j})}, \boldsymbol{z}_{1:k-1})}, \end{aligned} \tag{2.10}$$

which serves as the second part of the likelihood term in Eq. (2.6). It is noted that $Pr(\eta_k^{(a)}|\eta_{k-1}^{(b)})$ and $Pr(m_k^{(\mathrm{i})}|m_k^{(\mathrm{j})})$ can be described by diagonal matrices with few parameters. Refer Li et al. [25] or Qin et al. [18] for the mathematical derivation of IMM and its application to intention estimation.

In summary, the derivation of $Pr(\eta_k^{(a)}|\boldsymbol{z}_{1:k-1})$ and $Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_{1:k})$ has been described by Eq. (2.7) and Eq. (2.10). For calculating $Pr(\eta_k^{(a)}|\boldsymbol{z}_{1:k})$ of Eq. 2.6, the derivation of the remaining term $Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k|\eta_k^{(a)}, \boldsymbol{z}_{1:k-1})$ is proposed in the next subsection.

2.2.2 Multiple-to-Multiple Relations

The multiple-to-multiple relations represented by $Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k|\eta_k^{(a)}, \boldsymbol{z}_{1:k-1})$ include not only $\eta_k^{(a)}$ of intention and $m_k^{(\mathrm{i})}$ of model but also the observations. This means the relation of $\eta_k^{(a)}$ and $m_k^{(\mathrm{i})}$ can be different for distinct observations, which is common in real cases. This chapter considers only the impact of the recent observation $\boldsymbol{z}_k$. Thus, the multiple-to-multiple relations

$$Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k|\eta_k^{(a)}, \boldsymbol{z}_{1:k-1}) = Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k|\eta_k^{(a)})$$

Assuming the relations are independent from the time,

$$\begin{aligned} Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k | \eta_k^{(a)}) &= Pr(m^{(\mathrm{i})}, \boldsymbol{z} | \eta^{(a)}) \\ &= Pr(m^{(\mathrm{i})} | \boldsymbol{z}, \eta^{(a)}) Pr(\boldsymbol{z} | \eta^{(a)}). \end{aligned} \tag{2.11}$$

The raw labeled observations $\{\boldsymbol{z}_{l_s:l_e}^{(a)}\}$ do not contain information of $Pr(m^{(\mathrm{i})})$ directly. The proposed framework first processes $\{\boldsymbol{z}_{l_s:l_e}^{(a)}\}$ using the IMM approach to generate the information of $Pr(m^{(\mathrm{i})})$. Then the data of a tuple $(Pr(m^{(\mathrm{i})}), \boldsymbol{z}, a)$ is derived where i and a are discrete and $\boldsymbol{z}$ are multiple dimensional continuous variables.

The proposed framework takes the grid-based idea to process $\boldsymbol{z}$ [13]. The space of one dimension of $\boldsymbol{z}$ is evenly divided into partitions. Then each $\boldsymbol{z}$ of $\{\boldsymbol{z}_{l_s:l_e}^{(a)}\}$ can be assigned to one of the partitions of all dimensions. To derive $Pr(m^{(\mathrm{i})} | \boldsymbol{z}, \eta^{(a)})$, the average of all $Pr(m^{(\mathrm{i})})$ belonging to each partition is calculated for each pair of i, a. To derive $Pr(\boldsymbol{z} | \eta^{(a)})$, the number of $\boldsymbol{z}$ belonging to this partition is calculated for each a.

By finding the partition that a new observation $\boldsymbol{z}_k$ belongs, $Pr(m^{(\mathrm{i})} | \boldsymbol{z}_k, \eta^{(a)})$ and $Pr(\boldsymbol{z}_k | \eta^{(a)})$ can be given by the values of this partition. Thus, the multiple-to-multiple relations $Pr(m_k^{(\mathrm{i})}, \boldsymbol{z}_k | \eta_k^{(a)}, \boldsymbol{z}_{1:k-1})$ have been derived from the labeled observations, which is a data-driven solution.

2.3 Experimental Validation

The proposed framework was evaluated by estimating the intention of a maneuvered quadrotor, which is a robotic application of this class with high demand. Three steps were performed. The first two steps compared the proposed framework with the conventional framework of Figure 2.1 in a software-in-the-loop (SITL) simulation environment. In the first step, the number of models is equal to the number of intentions where the conventional framework is applied generally. The accuracy of the intention estimation is evaluated through the parametric study using the metric of F1 score [26]. Second, model sets of different numbers of models are tested to show the accuracy and flexibility of the proposed framework. The probabilistic representation of multiple-to-multiple relations is showed to verify the

capability of inferring the relations. Third, the framework was implemented on the real-world data of a quadrotor.

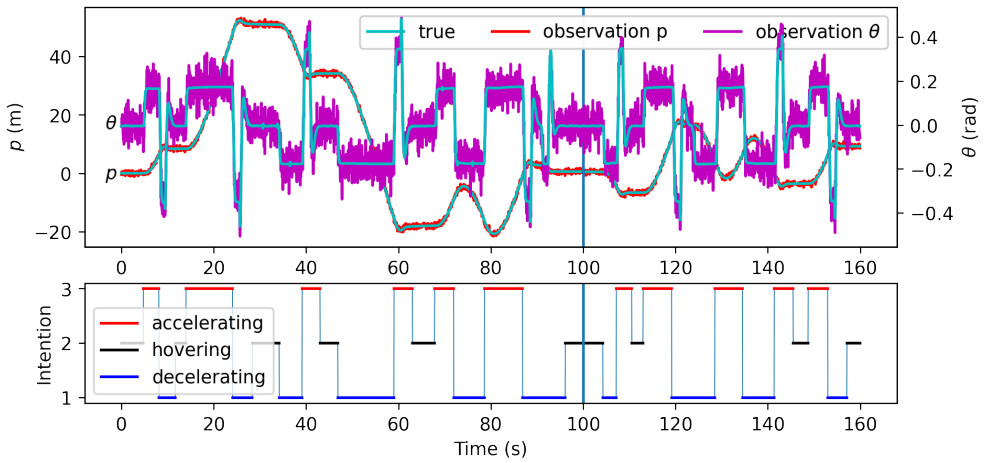


Figure 2.3. The labeled observations with intentions. The first 100 sec was used to infer the multiple-to-multiple relations, and the remaining 60 sec was used to conduct the intention estimation.

Table 2.1 lists the parameters used for simulation. The quadrotor dynamics was calculated in Gazebo, which also created motion noise artificially. As the most fundamental and typical motion, the linear horizontal motion of the quadrotor was considered. The quadrotor's state, $\boldsymbol{x}$, is given by:

$$\boldsymbol{x} = [p, \dot{p}, \theta, \dot{\theta}]^\top$$

where p is the position in the moving direction, $\dot{p}$ is the linear velocity, θ is the attitude (pitch angle), and $\dot{\theta}$ is the angular velocity. p and θ of the quadrotor were observed. The variances of the observation noise were specified as $[0.25, v_o^2]$ where v_o was varied in the parametric study. Figure 2.3 shows the noisy observations and the corresponding three intentions: $\eta^{(a)}$ of decelerating ($a = 1$), hovering ($a = 2$), and accelerating ($a = 3$). The labeling was based on the viewer's perception of the quadrotor acceleration. As the acceleration is independent from the position p, the validation considers the multiple-to-multiple relations $Pr(m^{(\mathrm{i})}, \boldsymbol{z}|\eta^{(a)})$ based on θ instead of the whole $\boldsymbol{z}$, which makes it easier to depict the probabilistic representations.

Table 2.1. Parameters for simulation

Parameter	Value
labeled intentions	Decelerating, Hovering, Accelerating
Motion noise	Specified in Gazebo
Observation noise variances	$[0.25, v_o^2]$
Duration of observations for inferring relations [s]	100
Duration of estimation [s]	60

Table 2.2. Parameters of the proposed framework

Parameter	Value
Time step [s]	$\Delta t = 0.05$
Variances of noise $\boldsymbol{Q}$	$[0, 1, 0, 0.36]$
Variances of noise $\boldsymbol{R}$	$[1, (1.5v_o)^2]$
Model Set 1	$(v_{m1} = 0, v_{m2} = [0, -0.2, 0.2])$
Model Set 2	Set 1 $\cup$ $(v_{m1} = 1, v_{m2} = 0)$
Model Set 3	Set 2 $\cup$ $(v_{m1} = 0, v_{m2} = [-0.1, 0.1])$
Model Set 4	Set 3 $\cup$ $(v_{m1} = 0, v_{m2} = [-0.3, 0.3])$

Through the analysis of the quadrotor, each motion model is given by

$$\boldsymbol{A}^{(\mathrm{i})} = \begin{bmatrix} 1 & dt & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & v_{m1} & dt \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ \boldsymbol{B}^{(\mathrm{i})}\boldsymbol{u}^{(\mathrm{i})} = \begin{bmatrix} 0 \\ 0 \\ v_{m2} \\ 0 \end{bmatrix}. \tag{2.12}$$

and the observation matrix $\boldsymbol{C}_k$ is a four-dimensional identity matrix. v_{m1} and v_{m2} are varied parameters for below description. Since a quadrotor flies forward or backward by adjusting the pitch angle, $v_{m1} = 0$ and v_{m2} of different values can represent different intentions, for instance, $v_{m1} = 0$ and $v_{m2} > 0$ for the accelerating intention. Besides, $v_{m1} = 1$ and $v_{m2} = 0$ result in a CV model. Table 2.2 lists the parameters of the proposed framework and the conventional framework.

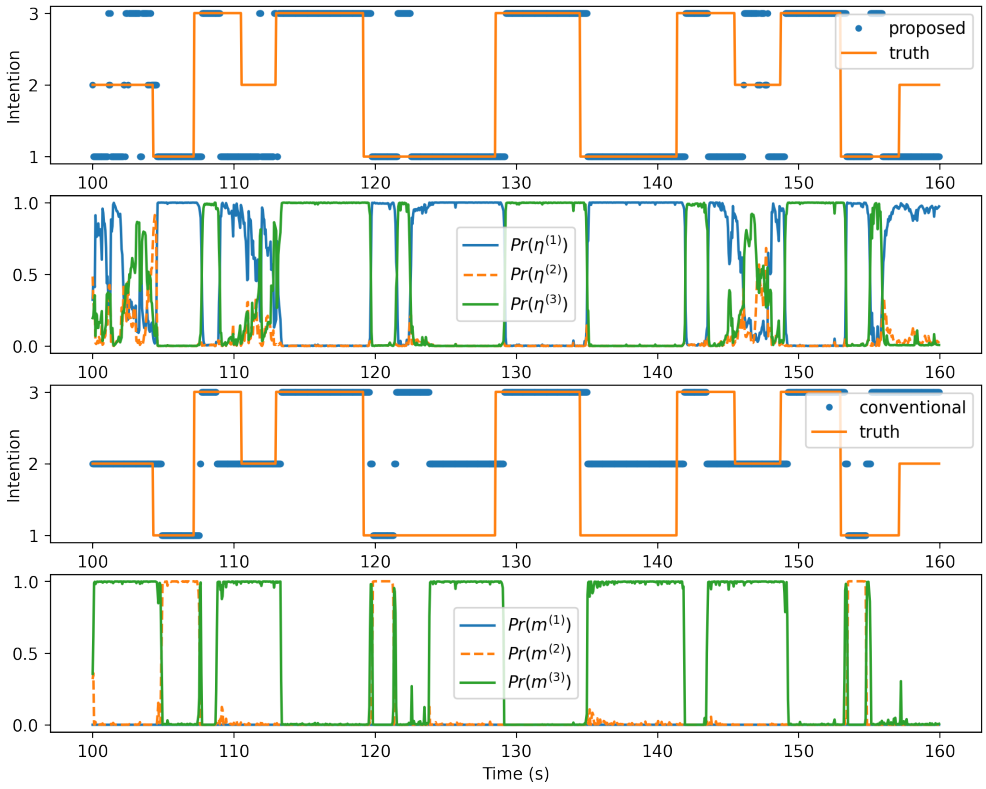


Figure 2.4. The most probable intentions of both frameworks, the intention probabilities $Pr(\eta^{(a)}$ of the proposed framework, and the model probabilities $Pr(m^{(\mathrm{i})}$ of both frameworks

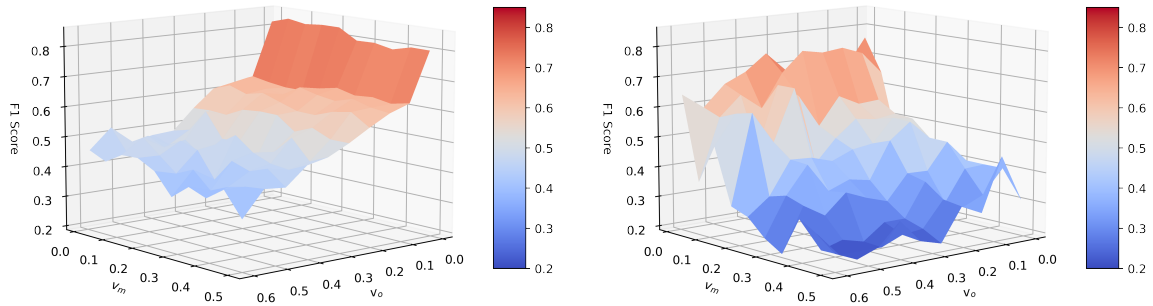
2.3.1 Equal Numbers of Intentions and Models

This section considers the case that three models of Model Set 1 in Table 2.2 are used for estimating the three intentions. Figure 2.4 shows the most probable intentions based on $Pr(\eta_k^{(a)})$ and $Pr(m_k^{(\mathrm{i})})$. The most probable intention is derived as

$$\eta_k = \begin{cases} \arg\max_a Pr(\eta_k^{(a)}) & \text{proposed} \\ \arg\max_{\mathrm{i}} Pr(m_k^{(\mathrm{i})}) & \text{conventional.} \end{cases} \tag{2.13}$$

Since $Pr(m_k^{(\mathrm{i})})$ is generated by the IMM approach, it is same for the proposed framework and the conventional framework; $Pr(\eta_k^{(a)})$ is of the proposed framework. The dynamic change of $Pr(\eta_k^{(a)})$ indicates the uncertainty of the estimated intention. The F1 scores are 0.65 and

0.55 respectively. The accuracy of the proposed framework is higher than the conventional framework even though the conventional framework is designed for the cases that the numbers of intentions and models are equal. This is because perfect one-to-one relations rarely exist.



(a) Proposed framework

(b) Conventional framework

Figure 2.5. F1 score evaluating the intention estimation accuracy with respect to v_o of observation noise and the parameter v_m of three models that $v_{m1} = 0, v_{m2} = [0, -v_m, v_m]$. Higher score indicates higher accuracy.

Figure 2.5 shows the F1 score over v_o of observation noise and the parameter v_m of three models that $v_{m1} = 0, v_{m2} = [0, -v_m, v_m]$. It is first seen the accuracy of the proposed framework is higher on average. The accuracy becomes lower when the observation noise gets larger for both frameworks, which is reasonable since the frameworks obtain little information to estimate the intention if the observation is considerably noisy. However, while the model parameter v_m affects the accuracy of the conventional framework significantly, the influence on the proposed framework is small. The proposed framework is resilient to the change of model parameters by incorporating the relative relations of all models instead of the sole relation to one single model.

2.3.2 Different Numbers of Intentions and Models

This section compares the proposed framework and the conventional framework using all of the four model sets. Figure 2.6 gives an example of the probabilistic representations of $Pr(m_k^{(\mathrm{i})}|\boldsymbol{z}_k, \eta_k^{(a)})$ when Model Set 2 is used. The multiple-to-multiple relations of three

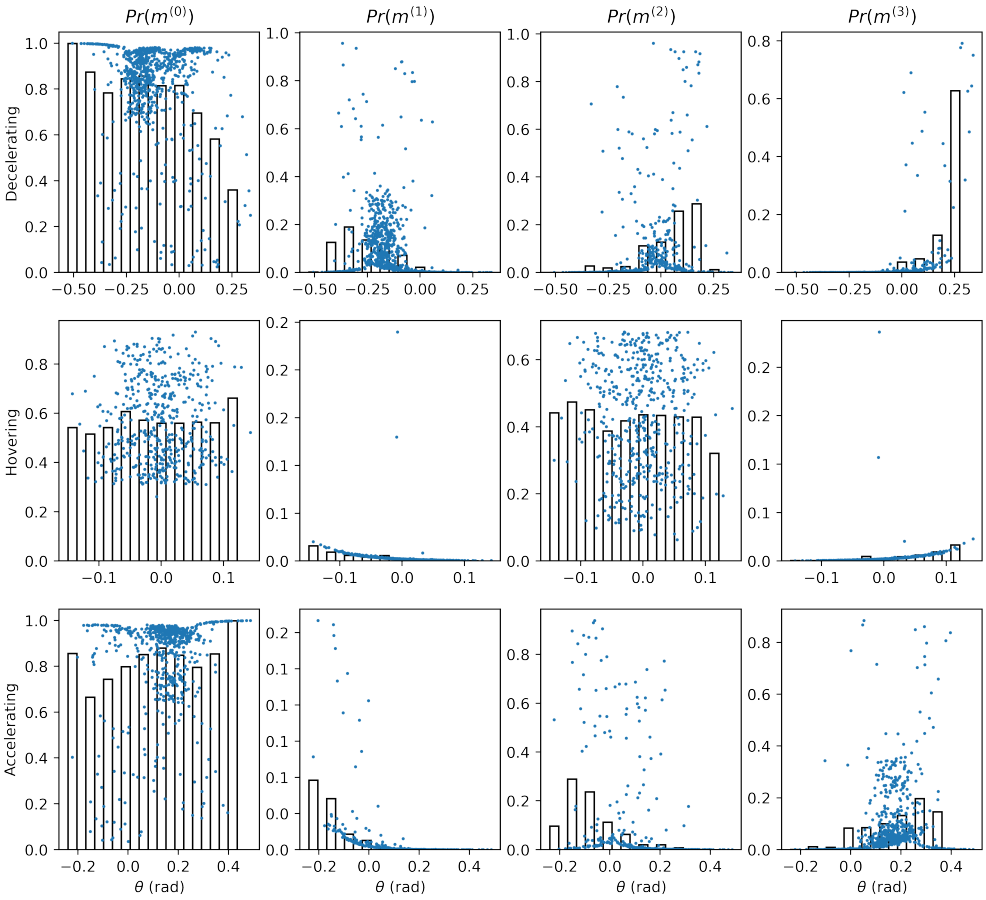


Figure 2.6. Multiple-to-multiple relations of $Pr(m^{(\mathrm{i})}|\boldsymbol{z}, \eta^{(a)})$ between the three intentions and the four models of model set 2

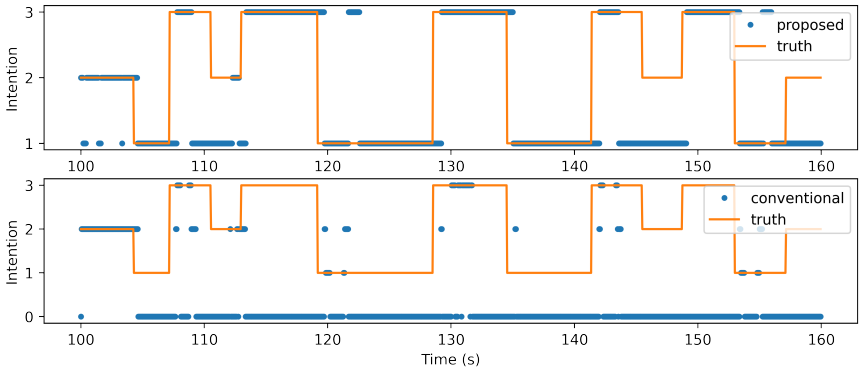


Figure 2.7. The most probable intentions of both frameworks when the four models of model set 2 are used

intentions and four models result in 12 histograms. As seen from the distributions, no apparent relation between one intention and one model appears, which is common in real cases. The memory space is proportional to the multiplication of the number of intentions, the number of models, and the number of partitions of $\boldsymbol{z}$.

Figure 2.7 shows the estimated intentions of each framework when model set 2 is used. The F1 scores are 0.70 and 0.13 respectively. Since one-to-one relations between intentions and models no longer hold, the accuracy of the conventional framework becomes significantly low compared with the case of Model Set 1.

Table 2.3 shows the F1 scores and mean squared errors (MSEs) of estimated θ for all model sets. The intention accuracy of the proposed framework stays high for all model sets. For both frameworks, the MSE of estimated θ is larger for Model Set 1 compared to other sets since Model Set 1 needs to consider both state estimation and intention estimation satisfying the strict one-to-one relations. Since state estimation is usually designed for or along with intention estimation, the proposed framework features increased flexibility when designing models for state and intention estimation.

Table 2.3. F1 score of each framework and the MSE of the estimated θ for both frameworks

Model Set	1	2	3	4
F1 score (proposed)	0.65	0.70	0.72	0.72
F1 score (conventional)	0.55	0.13	0.10	0.10
MSE (e-3 rad)	2.74	1.28	1.04	1.03

2.3.3 Test on Real-World Data

The proposed framework was further verified on real-world data. Figure 2.8(a) shows the experimental environment and a quadrotor as the target. Optitrack motion capture system was used to provide the observations. Figure 2.8(b) 2.8(c) 2.8(d) show examples of the estimation results. Real-time estimation is showed in the supplemental material.

(a) Experimental environment and the target

(b) Decelerating

(c) Hovering

(d) Accelerating

Figure 2.8. The experimental environment, the target, and examples of the estimation results

2.4 Summary

This chapter has presented a data-driven multiple model framework for estimating the intention of a target from observations. The framework infers the multiple-to-multiple relations between intentions and motion models from labeled observations. Both the multiple-to-multiple relations and model likelihoods are incorporated for intention estimation. Experimental analysis shows that the proposed framework features higher accuracy compared with the conventional framework especially when one-to-one relations do not hold. The estimation performance is resilient to parameter changes of models, which results in superior flexibility when designing models for state and intention estimation.

3. State Estimation Benefiting from Intention Inference

Most dynamic targets to track or engage are either human-maneuvered or humans themselves. Estimating the state of such a human-maneuvered target is essential and important, and has attracted tremendous interest in the last decades [27, 28, 29, 30]. Despite the importance, difficulty in the estimation of the human-maneuvered target lies in the motion uncertainty. Even though the motion model of the target may be well or precisely known, the control of the human is often unknown [31]. The motion, as a result, becomes considerably different from the expectation. This gives rise to need for the ability to handle motion uncertainty [13].

For a human-maneuvered target, estimation techniques proposed in the past to handle motion uncertainty can be classified into two types. In the first, a single accurate motion model is developed and used to describe the motion behavior. Due to their robust estimation upon past observations, various Bayesian methods, including the parametric Kalman filters (KFs) and the nonparametric particle filters, have been applied by characterizing the estimation problem and identifying the best estimation technique for the problem [15, 32, 33, 34, 35, 36, 37]. Steckenrider et al. [38] proposed to introduce higher-order terms to the motion model through Taylor series expansion and adaptively estimated the target state. Gindele et al. [39] improved the motion model by incorporating the situational context and extending the state space. Since human control is unknown most of the time, conservative motion behaviors such as constant velocity (CV) and constant acceleration (CA) have been incorporated as the most probable human controls [31]. Instead of the motion model, Mehra [40] estimated the covariances of motion noise and observation noise when the filter is detected not working optimally. Almagbile et al. [41] evaluated three adaptation methods of noise covariances and showed improvements over the conventional Kalman filter. It is effective to control uncertainty when the deterministic motion accuracy can no longer be improved. In addition to the model and its uncertainty, other work has dealt with unknown human control and its uncertainty from the motion noise [31]. The human control dominates the motion behavior when the target has a large unconstrained workspace. Bogler [42] represented the time-varying human control deterministically by piecewise constants and es-

timated the control in addition to the state. Chakrabarty et al. [43] assumed the exogenous input and its derivative to be bounded for a class of nonlinear systems in state estimation. Conte et al. [16] used head motion as an additional indicator when the target is a human and improved the estimation accuracy. While they are more detailed and more adaptively represented, these motion models cannot keep capturing the target motion and estimating its state well particularly if the motion is drastically changed by a human. This is due to the limited representation of a single model.

In the second, multiple models, which are either superposed or switched, have been used to estimate more varying motion behavior [44, 45, 46, 47, 48]. The multiple-model (MM) estimation methods extend existing techniques to handle multiple models and cover a wider range of motion behavior [25]. Blom et al. [44] proposed the interacting MM (IMM) method that uses a fixed set of motion models with Markovian switching coefficients. The transition probability and model likelihood were introduced to recursively adapt the model probabilities. Li et al. [49] proposed the variable-structure MM (VSMM) method to overcome the limitations of using a fixed set of models in describing the motion. The VSMM method introduces model set adaption besides the model adaptation and thus can describe and estimate even a broader range of motion behavior. Recently, Xu et al. [50] has engaged with estimating varying motion behaviors by adapting parameters where a fixed coarse grid and an adaptive fine grid of the parameters were combined to determine the models that best match the target motion behavior. Despite the wider covering, it is still insufficient to capture and estimate the target if the human control changes considerably. The MM methods are rather formulated to cover a larger state space given the most probable human control. Since the drastic control change may magnify changes in state space, the resulting target state could be beyond the permissible space of the MM estimation. In addition, the use of the deterministic control makes the estimation underestimated as the human control is most uncertain.

This chapter presents an approach for estimating the state of a human-maneuvered target by associating the recurring motion behaviors with human intentions. The proposed approach consists of a pre-process, which constructs the so-called intention-pattern model to encapsulate the human intention, and the main process, which allows state estimation using

the intention-pattern model. In the pre-process, the intention-pattern model is constructed from the prior observations by running a revised IMM estimation, extracting motion behaviors of each human intention, aligning them and probabilistically representing its behavior. The main process, then, uses standard state estimation such as KF extensively using the probabilistically represented intention-pattern model. The strength of the proposed approach lies in the incorporation of the intention-pattern model as the incorporation can make the estimation not only accurate in mean but also precise in covariance.

The chapter is organized as follows. The next section describes the estimation problem and its solution using the IMM estimation method, which is not only a generalized formulation but also the technique used in the pre-process of the proposed approach. Section 3.2 presents the proposed estimation approach including the pre-process and the main process. Numerical validation investigating the effectiveness of both the intention-pattern model and the state estimation is presented in Section 3.3. Conclusions are summarized in the final section.

3.1 Estimation of a Human-maneuvered Target Using IMM Estimation

3.1.1 Estimation Problem Formulation

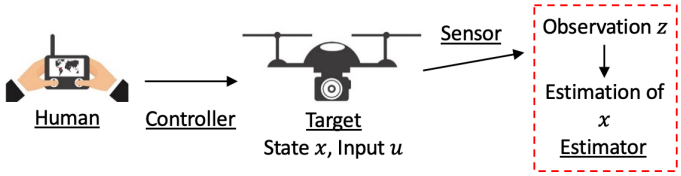


Figure 3.1. The problem of estimating a human-maneuvered target from observations.

Figure 3.1 shows a schematic diagram of the problem of estimating the state of a human-maneuvered target in case the target is a multirotor. When maneuvering a target, a human operator interacts with a controller using an interface device such as a vehicle panel or a joystick. The controller may be implemented in the interface device, in the target, or both. Some parameters of the controller, such as the maximum speed of the target, are usually configurable to realize different motion behavior. The information of human operation and

configurable parameters are not known as no communication with the target is available. The estimator does not affect the human operator and target as well. Having the target observed in the field of view (FOV) of a fixed sensor such as a stereo camera, the goal of the problem is to design the estimator to estimate the target state from observations. The discrete motion model of the target and the observation model are generically given by

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_k, \boldsymbol{w}_k) \tag{3.1}$$

$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{v}_k) \tag{3.2}$$

where $\boldsymbol{f}$ and $\boldsymbol{h}$ are the motion and the observation models respectively, $\boldsymbol{x}_k$ is the state of the target at step k to estimate, $\boldsymbol{u}_k$ is the input, $\boldsymbol{z}_k$ is the observation, and $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ are the motion and observation noises respectively. Because it is a target maneuvered by a human, $\boldsymbol{f}$ and $\boldsymbol{w}_k$ may not be well known while $\boldsymbol{u_k}$ is fully unknown. On the other hand, $\boldsymbol{h}$ and $\boldsymbol{v}_k$ are fully known since they are with the sensor(s) of the estimator. With short time interval, it is valid to assume that $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ are Gaussian. The problem is resultantly defined as the estimation of $\boldsymbol{x}_k$ with no knowledge on $\boldsymbol{u}_k$ and some knowledge on $\boldsymbol{f}$ and $\boldsymbol{w}_k$, given $\boldsymbol{h}$, $\boldsymbol{z}_k$ and $\boldsymbol{v}_k$.

3.1.2 IMM Estimation

Lacking information of $\boldsymbol{f}$ and $\boldsymbol{u}_k$ results in high motion uncertainty. The MM estimation methods deal with motion uncertainty by describing the motion with several motion behaviors called modes, which are denoted by $\mathbb{S} = \{s^{\mathrm{j}}\}, \mathrm{j} \in \mathbb{N}$ where $\mathbb{N}$ represents all natural integers. With the definition, a mode s^{j} is used to represent the motion at step k when it approximates the motion behavior well, i.e., $s_k = s^{\mathrm{j}}$. To describe the behavior with minimum complexity, a mode s^{j} is most commonly described by a mathematical model m^{i}, which is collectively represented by $\mathbb{M} = \{m^{\mathrm{i}}\}, \mathrm{i} \in \mathbb{N}$. A model m^{i} thus represents the motion behavior at a step, i.e., $s_k = s^{\mathrm{j}} = m^{\mathrm{i}}$. Li et al. Li et al. [51] is referred for more description. Most MM estimation methods utilize models of known motion behaviors with different parameters

such as variants of the CV and CA models. As an example, suppose that $\boldsymbol{f}$ of Eq. (3.1) at step k is approximated by a single linear Gaussian model. The motion model m^{i} is given by

$$\boldsymbol{x}_k = \boldsymbol{A}_k^{(\mathrm{i})}\boldsymbol{x}_{k-1} + \boldsymbol{B}_k^{(\mathrm{i})}\boldsymbol{u}_k^{(\mathrm{i})} + \boldsymbol{w}_k^{(\mathrm{i})}. \tag{3.3}$$

where $\boldsymbol{A}_k^{(\mathrm{i})}$ is a system matrix, $\boldsymbol{B}_k^{(\mathrm{i})}$ is a control matrix, and $\boldsymbol{w}_k^{(\mathrm{i})}$ is Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{Q}_k^{(\mathrm{i})}$. The symbol $^{(\mathrm{i})}$ indicates that the model m^{i} is used. The observation model (3.2) is also supposed to be linear Gaussian:

$$\boldsymbol{z}_k = \boldsymbol{C}_k\boldsymbol{x}_k + \boldsymbol{v}_k \tag{3.4}$$

where $\boldsymbol{C}_k$ is the observation matrix, and $\boldsymbol{v}_k$ is Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{R}_k$.

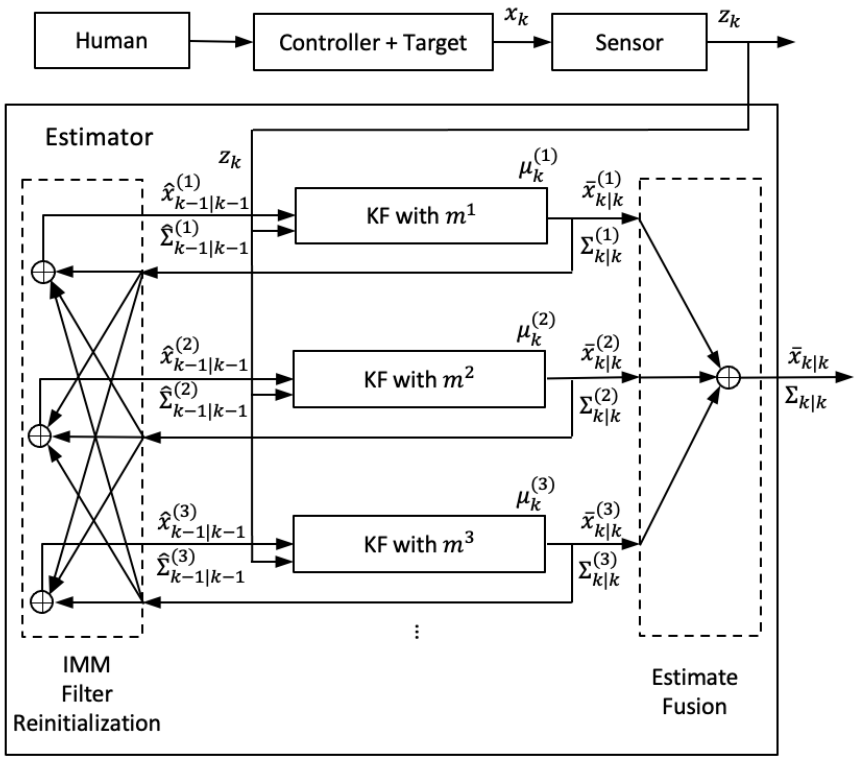


Figure 3.2. The IMM estimation method.

Figure 3.2 shows the framework of the IMM estimation method to estimate the target state. The motion behavior is described with a set of models $\{m^1, m^2, m^3, ...\}$. Having the

observation $\boldsymbol{z}_k$ of the state $\boldsymbol{x}_k$, KF is applied for each model (3.3) with the linear Gaussian assumption. Each KF updates the target state of mean $\hat{\bar{\boldsymbol{x}}}^{(\mathrm{i})}_{k-1|k-1}$ and covariance $\hat{\boldsymbol{\Sigma}}^{(\mathrm{i})}_{k-1|k-1}$ to mean $\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k}$ and covariance $\boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k}$. Each $\hat{\bar{\boldsymbol{x}}}^{(\mathrm{i})}_{k-1|k-1}$ and $\hat{\boldsymbol{\Sigma}}^{(\mathrm{i})}_{k-1|k-1}$ are derived from the IMM filter reinitialization incorporating all of $\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k-1|k-1}$ and $\boldsymbol{\Sigma}^{(\mathrm{i})}_{k-1|k-1}$. The output $\bar{\boldsymbol{x}}_{k|k}$ and covariance $\boldsymbol{\Sigma}_{k|k}$ are calculated by the estimate fusion of all of $\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k}$ and $\boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k}$.

The mathematical derivation is as follows. The event that the model m^{i} matches the mode at step k is denoted as $m^{(\mathrm{i})}_k \triangleq \{s_k = m^{\mathrm{i}}\}$. The probability of $m^{(\mathrm{i})}_k$ is denoted as $\mu^{(\mathrm{i})}_k \triangleq Pr\{m^{(\mathrm{i})}_k\}$ where $Pr\{\cdot\}$ indicates the probability of an event. The IMM estimator assumes the probability of transitioning from a model m^{i} at step k to a model m^{j} at step $k+1$ is constant and known as $\boldsymbol{\pi}_{\mathrm{ij}}$:

$$Pr\{m^{(\mathrm{j})}_{k+1}|m^{(\mathrm{i})}_k\} = Pr\{s_{k+1} = m^{\mathrm{j}}|s_k = m^{\mathrm{i}}\} = \boldsymbol{\pi}_{\mathrm{ij}}, \tag{3.5}$$

where $\mathrm{i}, \mathrm{j} \in \mathbb{N}$. For one cycle, the predicted model probability, $\hat{\mu}^{(\mathrm{i})}_{k|k-1}$, is given by

$$\hat{\mu}^{(\mathrm{i})}_{k|k-1} \triangleq Pr\{m^{(\mathrm{i})}_k|\boldsymbol{z}_{1:k-1}\} = \sum_{\mathrm{j}} \mu^{(\mathrm{j})}_{k-1}\boldsymbol{\pi}_{\mathrm{ji}} \tag{3.6}$$

where $\boldsymbol{z}_{1:k-1}$ are the observations from step 1 to step $k-1$. The weight that $m^{(\mathrm{j})}_{k-1}$ contributes to $m^{(\mathrm{i})}_k$ is derived as

$$\mu^{\mathrm{j|i}}_{k-1} \triangleq Pr\{m^{(\mathrm{j})}_{k-1}|m^{(\mathrm{i})}_k, \boldsymbol{z}_{1:k-1}\} = \mu^{(\mathrm{j})}_{k-1}\boldsymbol{\pi}_{\mathrm{ji}}/\hat{\mu}^{(\mathrm{i})}_{k|k-1} \tag{3.7}$$

The KF of each model starts with the derivation of input:

$$\hat{\bar{\boldsymbol{x}}}^{(\mathrm{i})}_{k-1|k-1} \triangleq E[\boldsymbol{x}_{k-1}|m^{(\mathrm{i})}_k, \boldsymbol{z}_{1:k-1}] = \sum_{\mathrm{j}} \bar{\boldsymbol{x}}^{(\mathrm{j})}_{k-1|k-1}\mu^{\mathrm{j|i}}_{k-1}, \tag{3.8a}$$

$$\hat{\boldsymbol{\Sigma}}^{(\mathrm{i})}_{k-1|k-1} = \sum_{\mathrm{j}} \left[\boldsymbol{\Sigma}^{(\mathrm{j})}_{k-1|k-1} + (\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k-1|k-1} - \hat{\boldsymbol{x}}^{(\mathrm{j})}_{k-1|k-1})(\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k-1|k-1} - \hat{\boldsymbol{x}}^{(\mathrm{j})}_{k-1|k-1})^{\top}\right] \mu^{(\mathrm{j|i})}_{k-1}. \tag{3.8b}$$

According to the KF formulation, the predicted mean and covariance are derived as

$$\bar{\boldsymbol{x}}_{k|k-1}^{(\mathrm{i})} = \boldsymbol{A}_k^{(\mathrm{i})} \hat{\boldsymbol{x}}_{k-1|k-1}^{(\mathrm{i})} + \boldsymbol{B}_k^{(\mathrm{i})} \boldsymbol{u}_k^{(\mathrm{i})}, \tag{3.9a}$$

$$\boldsymbol{\Sigma}_{k|k-1}^{(\mathrm{i})} = \boldsymbol{A}_k^{(\mathrm{i})} \hat{\boldsymbol{\Sigma}}_{k-1|k-1}^{(\mathrm{i})} (\boldsymbol{A}_k^{(\mathrm{i})})^\top + \boldsymbol{Q}_k^{(\mathrm{i})}. \tag{3.9b}$$

For correction, the KF gain is first computed through

$$\boldsymbol{K}_k^{(\mathrm{i})} = \boldsymbol{\Sigma}_{k|k-1}^{(\mathrm{i})} (\boldsymbol{C}_k)^\top (\boldsymbol{S}_k^{(\mathrm{i})})^{-1} \tag{3.10}$$

where the residual covariance is given by

$$\boldsymbol{S}_k^{(\mathrm{i})} = \boldsymbol{C}_k \boldsymbol{\Sigma}_{k|k-1}^{(\mathrm{i})} (\boldsymbol{C}_k)^\top + \boldsymbol{R}_k \tag{3.11}$$

The corrected mean and covariance are derived as

$$\bar{\boldsymbol{x}}_{k|k}^{(\mathrm{i})} = \bar{\boldsymbol{x}}_{k|k-1}^{(\mathrm{i})} + \boldsymbol{K}_k^{(\mathrm{i})} \tilde{\boldsymbol{z}}_k^{(\mathrm{i})}, \tag{3.12a}$$

$$\boldsymbol{\Sigma}_{k|k}^{(\mathrm{i})} = (\boldsymbol{I} - \boldsymbol{K}_k^{(\mathrm{i})} \boldsymbol{C}_k) \boldsymbol{\Sigma}_{k|k-1}^{(\mathrm{i})}, \tag{3.12b}$$

where the observation residual is given by

$$\tilde{\boldsymbol{z}}_k^{(\mathrm{i})} = \boldsymbol{z}_k - \boldsymbol{C}_k \bar{\boldsymbol{x}}_{k|k-1}^{(\mathrm{i})} \tag{3.13}$$

For IMM estimation, the model likelihood $L_k^{(\mathrm{i})}$ is assumed as

$$\begin{aligned} L_k^{(\mathrm{i})} &\triangleq Pr\{\tilde{\boldsymbol{z}}_k^{(\mathrm{i})} | m_k^{(\mathrm{i})}, \boldsymbol{z}_{1:k-1}\} \\ &\overset{\text{assume}}{=} |2\pi \boldsymbol{S}_k^{(\mathrm{i})}|^{-\frac{1}{2}} \exp\big[-\frac{1}{2} (\tilde{\boldsymbol{z}}_k^{(\mathrm{i})})^\top (\boldsymbol{S}_k^{(\mathrm{i})})^{-1} \tilde{\boldsymbol{z}}_k^{(\mathrm{i})} \big] \end{aligned} \tag{3.14}$$

and the model probability $\mu_k^{(\mathrm{i})}$ is given by

$$\mu_k^{(\mathrm{i})} \triangleq Pr\{m_k^{(\mathrm{i})} | \boldsymbol{z}_{1:k}\} = \frac{\hat{\mu}_{k|k-1}^{(\mathrm{i})} L_k^{(\mathrm{i})}}{\sum_{\mathrm{j}} \hat{\mu}_{k|k-1}^{(\mathrm{j})} L_k^{(\mathrm{j})}}. \tag{3.15}$$

The overall mean and covariance are derived as

$$\bar{\boldsymbol{x}}_{k|k} \triangleq E[\boldsymbol{x}_k|\boldsymbol{z}_{1:k}] = \sum_{\mathrm{i}} \bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k}\mu^{(\mathrm{i})}_k \tag{3.16a}$$

$$\boldsymbol{\Sigma}_{k|k} = \sum_{\mathrm{i}} [\boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k} + (\boldsymbol{x}^{(\mathrm{i})}_{k|k} - \bar{\boldsymbol{x}}_{k|k})(\boldsymbol{x}^{(\mathrm{i})}_{k|k} - \bar{\boldsymbol{x}}_{k|k})^\top]\mu^{(\mathrm{i})}_k. \tag{3.16b}$$

Owing to the introducing of transition probability $Pr\{m^{(\mathrm{j})}_{k+1}|m^{(\mathrm{i})}_k\}$ of Eq. (3.5) and the likelihood $L^{(\mathrm{i})}_k$ of Eq. (3.14), the model probabilities $\mu^{(\mathrm{i})}_k$ adapt to match the current motion. Suppose the model $m^{(\mathrm{i})}$ matches the current mode better, the filter of $m^{(\mathrm{i})}$ contributes more on $\bar{\boldsymbol{x}}_{k|k}$ and $\boldsymbol{\Sigma}_{k|k}$ by having a higher model probability $\mu^{(\mathrm{i})}_k$.

For estimation with one motion model, the single motion model (3.1) cannot impair its inconsistency from the actual motion when the human has changed the target motion considerably. The IMM method estimates in a larger state space due to the usage of multiple models, but it still uses the most probable deterministic human control such as the CV and CA. If the control is different, the multiple models of the IMM method may not be able to cover the space of estimation and could lead to a wrong estimation. The uncertainty could also be underestimated since the unknown human control, which is most uncertain, is handled deterministically. This limitation of the conventional techniques affects the quality of estimation when the target is human-maneuvered.

3.2 Proposed Approach Using Intention-Pattern Model

Since the contributions of this chapter are the construction of a intention-pattern model and the state estimation using the intention-pattern model, this section describes each contribution in a subsection. Section 3.2.1 presents the overview of the construction of an intention-pattern model, followed by the details of the two major components, which are the intention inference and the intention-pattern modeling. The implementation of the constructed intention-pattern models into the state estimation is then detailed in Section 3.2.2.

3.2.1 Construction of Intention-Pattern Model

Overview

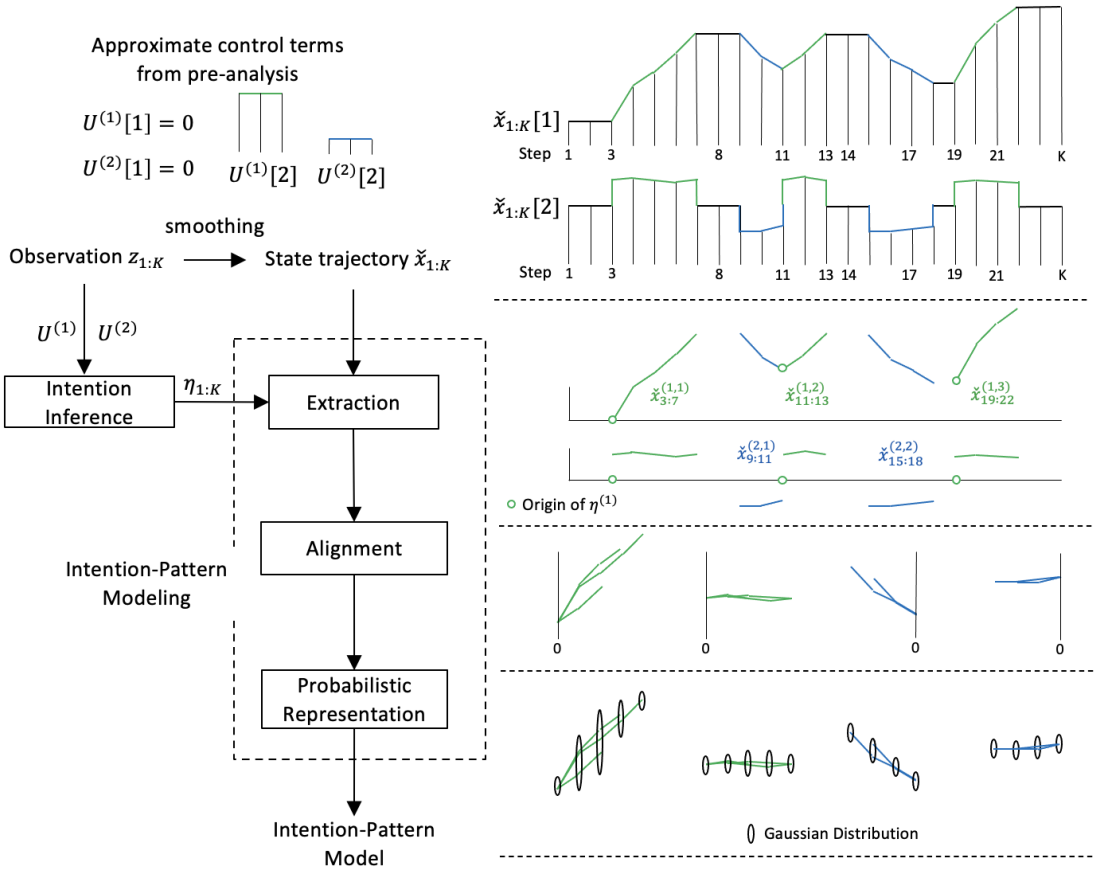


Figure 3.3. Construction of intention-pattern model. $(\cdot)[\mathrm{i}]$ represents the ith dimension of $(\cdot)$.

Figure 3.3 shows the construction of the intention-pattern model where an example illustration is given on the right side. The prior analysis of the target behavior leads to the extraction of a set of human intentions $H = \{\eta^{(\mathrm{i})} | \forall \mathrm{i}\}$ and the corresponding approximate control terms, $\boldsymbol{U}^{(\mathrm{i})}$. Each human intention $\eta^{(\mathrm{i})}$ is an expression describing an aim or a plan, such as "moving forward" and "turning right".

Each human intention at step k is defined as a function of the recent states of N_h steps:

$$\eta_k = \alpha(\boldsymbol{x}_{k-N_h+1:k}). \tag{3.17}$$

Compared with the human actions which are of one step and are associated with the control input $\boldsymbol{u}_k$, the human intentions are labels of continuous states of multiple steps.

The corresponding control term could vary, but let it be constant for simplicity. Because the intention is defined for a period, the figure illustratively shows each control term with two steps. Given a sequence of observations $\boldsymbol{z}_{1:K}$, the human intention at step k, $\eta_k = \eta^{(\mathrm{i})} \in H$, is first inferred for all steps, i.e., $\eta_{1:K} = \{\eta_1, ..., \eta_K\}$. We assume that the observations are fully observable for simplicity. After smoothing the observations and deriving the state trajectory $\check{\boldsymbol{x}}_{1:K}$, the proposed construction technique identifies segments in state space exhibiting the extracted intention, $\check{\boldsymbol{x}}_{k_{s(\mathrm{i},\mathrm{j})}:k_{\mathrm{e}(\mathrm{i},\mathrm{j})}}, \mathrm{j} \in \mathbb{N}$, where $k_{s(\mathrm{i},\mathrm{j})}$ and $k_{\mathrm{e}(\mathrm{i},\mathrm{j})}$ are the starting and the ending steps. Three segments of intention $\eta^{(1)}$ and two segments of intention $\eta^{(2)}$ are shown in the example illustration. The segments of the same intention are aligned to characterize the pattern of motion probabilistically. The intention-pattern model, describing the relationship between the input intention and the output motion pattern, is finally represented by a set of Gaussian distributions.

Intention Inference

Since states $\boldsymbol{x}_{k-N_h+1:k}$ are not directly available, this section proposes the approach to infer the human intention η_k based on observations. Given observations $\boldsymbol{z}_{1:k}$ in addition to the control terms corresponding to the extracted intentions, $\boldsymbol{U}^{(\mathrm{i})}, \forall \mathrm{i}$, the first process of the intention inference is to run the IMM estimation. There is only one motion model, but the motion is simulated for each intention $\boldsymbol{U}^{(\mathrm{i})}$:

$$\boldsymbol{x}_k = \boldsymbol{A}_k \boldsymbol{x}_{k-1} + \boldsymbol{U}^{(\mathrm{i})} + \boldsymbol{w}_k, \tag{3.18}$$

where $\boldsymbol{A}_k$ and $\boldsymbol{w}_k$ are determined from the analysis of target motion. Having Eq. (3.4) as an observation model, the KF updates the mean $\boldsymbol{x}_{k|k}$ and the covariance $\boldsymbol{\Sigma}_{k|k}$ similarly to Eqs. (3.9)-(3.13) for each intention:

$$\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k-1} = \boldsymbol{A}_k \bar{\boldsymbol{x}}^{(\mathrm{i})}_{k-1|k-1} + \boldsymbol{U}^{(\mathrm{i})} \tag{3.19a}$$

$$\boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k-1} = \boldsymbol{A}_k \boldsymbol{\Sigma}^{(\mathrm{i})}_{k-1|k-1} (\boldsymbol{A}_k)^\top + \boldsymbol{Q}_k. \tag{3.19b}$$

$$\boldsymbol{K}^{(\mathrm{i})}_k = \boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k-1} (\boldsymbol{C}_k)^\top (\boldsymbol{S}^{(\mathrm{i})}_k)^{-1} \tag{3.19c}$$

$$\boldsymbol{S}^{(\mathrm{i})}_k = \boldsymbol{C}_k \boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k-1} (\boldsymbol{C}_k)^\top + \boldsymbol{R}_k \tag{3.19d}$$

$$\tilde{\boldsymbol{z}}^{(\mathrm{i})}_k = \boldsymbol{z}_k - \boldsymbol{C}_k \bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k-1}, \tag{3.19e}$$

$$\bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k} = \bar{\boldsymbol{x}}^{(\mathrm{i})}_{k|k-1} + \boldsymbol{K}^{(\mathrm{i})}_k \tilde{\boldsymbol{z}}^{(\mathrm{i})}_k, \tag{3.19f}$$

$$\boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}^{(\mathrm{i})}_k \boldsymbol{C}_k) \boldsymbol{\Sigma}^{(\mathrm{i})}_{k|k-1}. \tag{3.19g}$$

The likelihood of the motion model at step k is given by Eq. (3.14). Since the intention is determined for a period, let the number of steps that defines an intention be N_h steps. The intention likelihood is defined and derived as the joint likelihood of the model likelihoods $L^{(\mathrm{i})}_k$:

$$\mathcal{L}_k(\boldsymbol{U}^{(\mathrm{i})}) = \prod_{\kappa=k-N_h+1}^{k} L^{(\mathrm{i})}_k(\boldsymbol{U}^{(\mathrm{i})}) = \prod_{\kappa=k-N_h+1}^{k} |2\pi \boldsymbol{S}^{(\mathrm{i})}_\kappa|^{-\frac{1}{2}} \exp\big[-\frac{1}{2} (\tilde{\boldsymbol{z}}^{(\mathrm{i})}_\kappa)^\top (\boldsymbol{S}^{(\mathrm{i})}_\kappa)^{-1} \tilde{\boldsymbol{z}}^{(\mathrm{i})}_\kappa \big]. \tag{3.20}$$

The control term that maximizes the intention likelihood is then selected:

$$\mathrm{i}_k = \arg\max_{\mathrm{i}} \{ \mathcal{L}_k(\boldsymbol{U}^{(\mathrm{i})}) | \forall \mathrm{i} \}, \tag{3.21}$$

if the intention likelihood is above the threshold:

$$\mathcal{L}_k(\boldsymbol{U}^{(\mathrm{i}_k)}) \geq \mathcal{L}^*(\boldsymbol{U}^{(\mathrm{i}_k)}).$$

The corresponding intention η_k is given by

$$\eta_k = \begin{cases} \eta^{(\mathrm{i}_k)} & \mathcal{L}_k(\boldsymbol{U}^{(\mathrm{i}_k)}) \geq \mathcal{L}^*(\boldsymbol{U}^{(\mathrm{i}_k)}) \\ \text{ø} & \text{Otherwise} \end{cases} \tag{3.22}$$

where ø indicating an empty element means that there is no matching intention. The recursive operation infers intention for all steps, $\eta_{1:K}$.

Intention-Pattern Modeling

The first process of the intention-pattern modeling, the extraction of the intended motions, checks the intention η_k and identifies its period. Let the jth segment of the ith intention extracted from the smoothed state trajectory $\check{\boldsymbol{x}}_{1:K}$ be $\check{\boldsymbol{x}}_{k_s(\mathrm{i},\mathrm{j}):k_\mathrm{e}(\mathrm{i},\mathrm{j})}$. The second process of alignment aligns the extracted segments by co-locating their origins $\check{\boldsymbol{x}}_{k_0(\mathrm{i},\mathrm{j})}$:

$$k_0(\mathrm{i},\mathrm{j}) = 0,$$
$$\check{\boldsymbol{x}}_{k_0(\mathrm{i},\mathrm{j})} = \check{\boldsymbol{x}}_{k_0(\mathrm{i})} = \text{const.},$$

where the step of the origin $k_0(\mathrm{i},\mathrm{j}) \in \{k_s(\mathrm{i},\mathrm{j}), ..., k_\mathrm{e}(\mathrm{i},\mathrm{j})\}$.

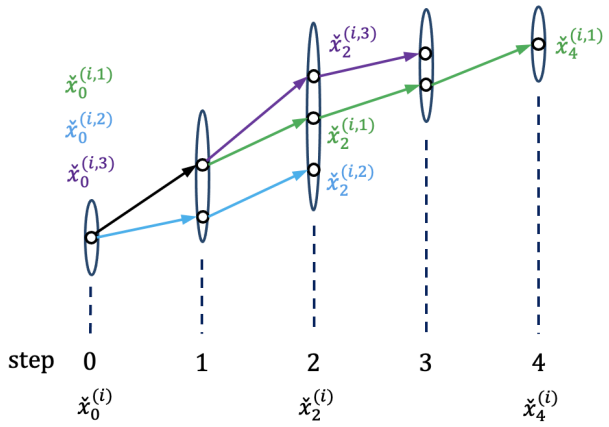


Figure 3.4. A set of Gaussian distributions representing the motion pattern which is the output of intention-pattern model.

The final process of motion pattern characterization derives the intention-pattern model by probabilistically characterizing the aligned segments. Figure 3.4 shows the characterization after three segments, green, blue and purple, are aligned. As the number of segments increases, it is valid to assume that the variation of the motion follows a Gaussian distribution:

$$\check{\boldsymbol{x}}_{\kappa}^{(\mathrm{i})} \sim \mathcal{N}\Big(\bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i})}, \check{\boldsymbol{\Sigma}}_{\kappa}^{(\mathrm{i})}\Big),$$

where κ is a step of the intention-pattern model after alignment, and the mean and the covariance are

$$\bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i})} = \frac{1}{n_{\kappa}^{(\mathrm{i})}} \sum_{\mathrm{j}} \check{\boldsymbol{x}}_{\kappa}^{(\mathrm{i,j})},$$
$$\check{\boldsymbol{\Sigma}}_{\kappa}^{(\mathrm{i})} = \frac{1}{n_{\kappa}^{(\mathrm{i})}} \sum_{\mathrm{j}} (\check{\boldsymbol{x}}_{\kappa}^{(\mathrm{i,j})} - \bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i})})(\check{\boldsymbol{x}}_{\kappa}^{(\mathrm{i,j})} - \bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i})})^{\top}.$$

$n_{\kappa}^{(\mathrm{i})}$ is the number of segments at step κ for the ith intention [24]. The intention-pattern model is finally derived as

$$\check{\boldsymbol{x}}_{k}^{(\mathrm{i})} \sim \mathcal{N}\Big(\bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i})}, \check{\boldsymbol{\Sigma}}_{\kappa}^{(\mathrm{i})}\Big)\delta(k-\kappa),$$

or

$$\bar{\check{\boldsymbol{x}}}_{k}^{(\mathrm{i})} = \bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i})}\delta(k-\kappa), \tag{3.24a}$$
$$\check{\boldsymbol{\Sigma}}_{k}^{(\mathrm{i})} = \check{\boldsymbol{\Sigma}}_{\kappa}^{(\mathrm{i})}\delta(k-\kappa), \tag{3.24b}$$

where $\delta(\cdot)$ is a Dirac delta function. This means that the intention-pattern model is defined by a set of Gaussian distributions:

$$\mathcal{N}^{(\mathrm{i})} = \{\mathcal{N}\Big(\bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i})}, \check{\boldsymbol{\Sigma}}_{\kappa}^{(\mathrm{i})}\Big) | \forall\kappa\}. \tag{3.25}$$

3.2.2 Estimation Using Intention-Pattern Model

Figure 3.5 shows the schematics of the proposed state estimator using the intention-pattern model. Given a new observation $\boldsymbol{z}_k$, Eqs. (3.19)-(3.22) outputs the intention of the current step, $\eta_k = \eta^{(\mathrm{i}_k)}$. The estimator then checks the corresponding Gaussian distributions $\mathcal{N}^{(\mathrm{i}_k)}$ to find the matching step κ_k with respect to the recent estimate state $\boldsymbol{x}_{k-1|k-1}$:

$$\mathcal{L}(\boldsymbol{x}_{k-1|k-1}|\mathcal{N}^{(\mathrm{i}_k)}) = \max_{\kappa} Pr\{\boldsymbol{x}_{k-1|k-1}|\mathcal{N}\Big(\bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i}_k)}, \check{\boldsymbol{\Sigma}}_{\kappa}^{(\mathrm{i}_k)}\Big)\} \tag{3.26}$$

$$\kappa_k = \arg\max_{\kappa} Pr\{\boldsymbol{x}_{k-1|k-1}|\mathcal{N}\Big(\bar{\check{\boldsymbol{x}}}_{\kappa}^{(\mathrm{i}_k)}, \check{\boldsymbol{\Sigma}}_{\kappa}^{(\mathrm{i}_k)}\Big)\}. \tag{3.27}$$

The step κ_k of $\mathcal{N}^{(\mathrm{i}_k)}$ is chosen if the intention-pattern model of the i_kth intention is satisfactory:

$$\mathcal{L}(\boldsymbol{x}_{k-1|k-1}|\mathcal{N}^{(\mathrm{i}_k)}) \geq \mathcal{L}^*(\mathcal{N}^{(\mathrm{i}_k)}), \tag{3.28}$$

and the step κ_k matches the current step k.

Having the step κ_k of the intention i_k identified, the state is predicted by

$$\bar{\boldsymbol{x}}_{k|k-1} = \boldsymbol{A}_k \bar{\boldsymbol{x}}_{k-1|k-1} + \bar{\boldsymbol{U}}_k^{(\mathrm{i}_k)}, \tag{3.29a}$$

$$\boldsymbol{\Sigma}_{k|k-1} = \boldsymbol{A}_k \boldsymbol{\Sigma}_{k-1|k-1} (\boldsymbol{A}_k)^\top + \boldsymbol{T}_k^{(\mathrm{i}_k)} + \boldsymbol{Q}_k. \tag{3.29b}$$

where the control term $\boldsymbol{U}_k^{(\mathrm{i}_k)}$ and the covariance of the control uncertainty, $\boldsymbol{T}_k^{(\mathrm{i}_k)}$, are derived from the intention-pattern model as

$$\bar{\boldsymbol{U}}_k^{(\mathrm{i}_k)} = \bar{\check{\boldsymbol{x}}}_{\kappa_k+1}^{(\mathrm{i}_k)} - \boldsymbol{A}_k \bar{\check{\boldsymbol{x}}}_{\kappa_k}^{(\mathrm{i}_k)}, \tag{3.30a}$$

$$\begin{aligned} \boldsymbol{T}_k^{(\mathrm{i}_k)} &= \mathbf{E}[(\boldsymbol{U}_k^{(\mathrm{i}_k)} - \bar{\boldsymbol{U}}_k^{(\mathrm{i}_k)})(\boldsymbol{U}_k^{(\mathrm{i}_k)} - \bar{\boldsymbol{U}}_k^{(\mathrm{i}_k)})^\top] \\ &= \mathbf{E}\Big\{[\check{\boldsymbol{x}}_{\kappa_k+1}^{(\mathrm{i}_k)} - \bar{\check{\boldsymbol{x}}}_{\kappa_k+1}^{(\mathrm{i}_k)} - \boldsymbol{A}_k(\check{\boldsymbol{x}}_{\kappa_k}^{(\mathrm{i}_k)} - \bar{\check{\boldsymbol{x}}}_{\kappa_k}^{(\mathrm{i}_k)})][\check{\boldsymbol{x}}_{\kappa_k+1}^{(\mathrm{i}_k)} - \bar{\check{\boldsymbol{x}}}_{\kappa_k+1}^{(\mathrm{i}_k)} - \boldsymbol{A}_k(\check{\boldsymbol{x}}_{\kappa_k}^{(\mathrm{i}_k)} - \bar{\check{\boldsymbol{x}}}_{\kappa_k}^{(\mathrm{i}_k)})]^\top\Big\} \\ &= \check{\boldsymbol{\Sigma}}_{\kappa_k+1}^{(\mathrm{i}_k)} + \boldsymbol{A}_k \check{\boldsymbol{\Sigma}}_{\kappa_k}^{(\mathrm{i}_k)} \boldsymbol{A}_k^\top, \end{aligned} \tag{3.30b}$$

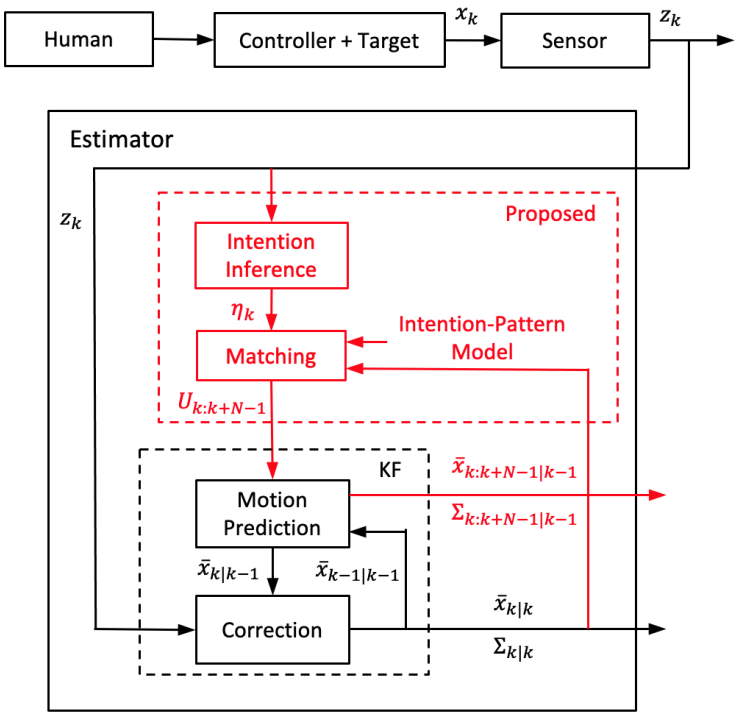


Figure 3.5. Estimation taking advantage of the proposed intention-pattern model. The red indicates the proposed parts compared with the conventional KF.

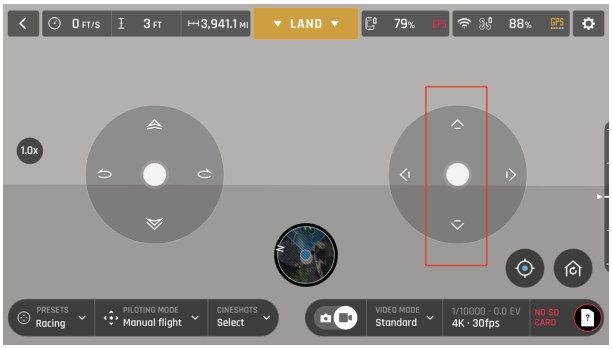
Equation (3.29) shows that the covariance propagation is more than that of the conventional KF-based estimation by the addition of $\boldsymbol{T}_k^{(\mathrm{i}_k)}$. The correction is conducted by Eqs. (3.19c)-(3.19g) of KF with i_k on behalf of i.

Because the proposed approach estimates the state incorporating the human intention, the mean of the estimated state is potentially more accurate than the conventional state estimation if observations are not available or reliable. The covariance of the estimated state is also more precise since it is updated adding the uncertainty of the human intention and prevents underestimation. Finally, it is to be noted that the proposed state estimation allows future prediction with human intention in addition to the current estimation by recursively predicting the state with Eq. (3.29).

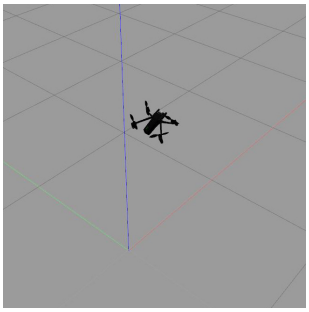
3.3 Numerical Validation

Having the strength of the intention-pattern model identified, it is essential to test the proposed approach numerically and identify the capability and limitations. The approach was evaluated by applying to the state estimation of a human-maneuvered multirotor, which is one of the applications of this class with high demand. To identify the capability and limitations in depth, a simulated environment was created and used.

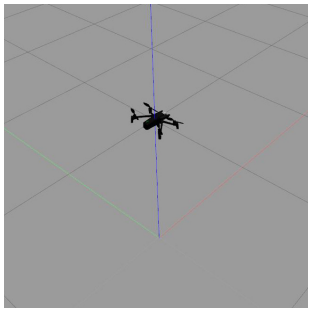
Figure 3.6 shows the controller interface used to create the multirotor motion and the resulting hovering, accelerating and decelerating motions in the software-in-the-loop (SITL) simulation environment whereas Table 3.1 lists the parameters used for simulation. With the right joystick of the controller interface, the human issues void command for hovering and forward or backward command for accelerating or decelerating. The multirotor dynamics was



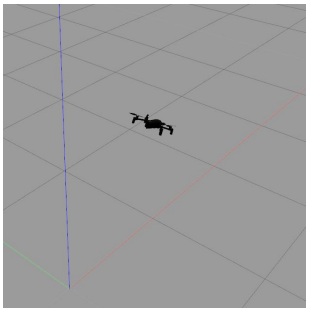
(a) Controller interface



(b) Decelerating



(c) Hovering



(d) Accelerating

Figure 3.6. The controller interface of the SITL simulation environment and the motion examples of the multirotor for three intentions.

calculated in Gazebo, which also created motion noise artificially. As the most fundamental

Table 3.1. Parameters for simulation.

Parameter	Value
Human commands	Void, Forward, Backward
Simulation motion noise	Specified in Gazebo
Simulation observation noise variances	$[o_{s1}^2, o_{s1}^2, o_{s2}^2, o_{s2}^2]$
Cruising inclination limit [rad]	0.17

and typical motion, the linear horizontal motion of the multirotor was considered. The multirotor's state, $\boldsymbol{x}$, is given by:

$$\boldsymbol{x} = [p, \dot{p}, \theta, \dot{\theta}]^\top$$

where p is the position in the moving direction, $\dot{p}$ is the linear velocity, θ is the attitude (pitch angle) and $\dot{\theta}$ is the angular velocity. The estimator was assumed to observe all the state variables of the multirotor, i.e., $\boldsymbol{z} = [z^p, z^{\dot{p}}, z^\theta, z^{\dot{\theta}}]^\top$. The observations were created by adding noise to the true state where the noise variances are indicated as $[o_{s1}^2, o_{s1}^2, o_{s2}^2, o_{s2}^2]$ since the variances are varied in the parametric study. Figure 3.7 shows the time-varying human command, true state and observation. The observation was created with $[o_{s1}, o_{s2}] = [1, 0.05]$. The observation noise was set high as the proposed approach is effective when the observation is uncertain or unavailable. The first 100 sec was used to construct the intention-pattern model, and the state estimation using the proposed approach was conducted with the observation of the remaining 60 sec. The command varies dynamically, and the multirotor motion is seen to reflect the commands of forward, void and backward.

Through the analysis of the multirotor state estimation problem, the motion and observation models used by the proposed approach were linear. The motion matrix $\boldsymbol{A}_k$ is given by

$$\boldsymbol{A}_k = \begin{bmatrix} 1 & dt & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & dt \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.31}$$

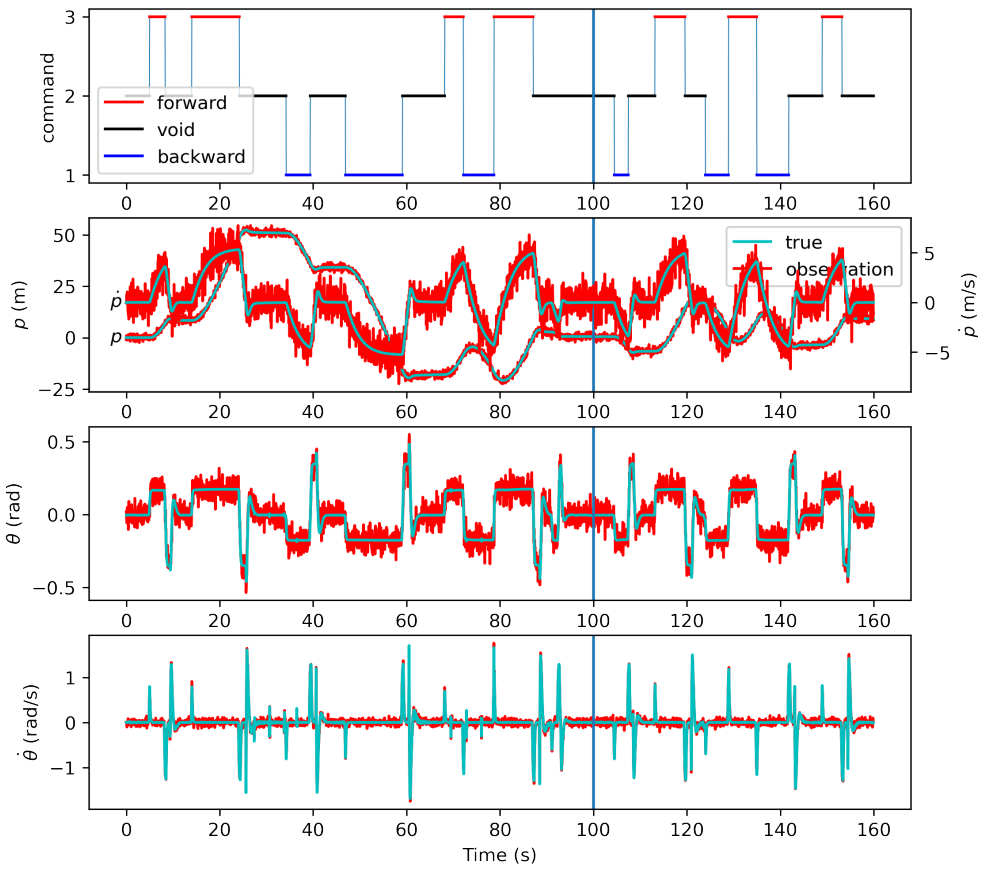


Figure 3.7. The human command and the true state and observation trajectories.

whereas the observation matrix $\boldsymbol{C}_k$ is a four-dimensional identity matrix. Table 3.2 lists the parameters of the proposed approach for both the intention-pattern model construction and the state estimation. The number of prediction steps between observations is denoted as n_p since it takes a different value for each process/study. While the variances of the observation noise is known, those of the motion noise were determined from the theoretical and experimental analyses. $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$ and $\boldsymbol{U}^{(3)}$ were chosen to infer the decelerating intention $\eta^{(1)}$, the hovering intention $\eta^{(2)}$ and the accelerating intention $\eta^{(3)}$ respectively. θ_U is a parameter to control the value of $\boldsymbol{U}^{(\mathrm{i})}$ for parametric study.

Section 3.3.1 investigates the validity of the construction process of intention-pattern model through the parametric study. Section 3.3.2 then validates the estimation performance using the intention-pattern model.

Table 3.2. Parameters of the proposed approach.

Parameter	Value
Time step [s]	$\Delta t = 0.05$
Number of predictions between two observations	n_p
Variances of motion noise $\boldsymbol{Q}$	$[0, (1 - o_{s1}/4)^2 \Delta t^2, 0, 2^2 \Delta t^2]$
Variances of observation noise $\boldsymbol{R}$	$[o_{s1}^2, o_{s1}^2, o_{s2}^2, o_{s2}^2]$
Human intention	Decelerating, Hovering, Accelerating
$\boldsymbol{U}^{(1)}$	$[0, 0, -\bar{x}^{(\mathrm{i})}_{k-1\|k-1}[3] - \theta_U, 0]^\top$
$\boldsymbol{U}^{(2)}$	$[0, 0, -\bar{x}^{(\mathrm{i})}_{k-1\|k-1}[3], 0]^\top$,
$\boldsymbol{U}^{(3)}$	$[0, 0, -\bar{x}^{(\mathrm{i})}_{k-1\|k-1}[3] + \theta_U, 0]^\top$
Smoothing technique	Triangular moving average
Duration of construction [s]	100
Duration of estimation [s]	60

3.3.1 Construction of Intention-Pattern Model

Figure 3.8 shows the inferred intentions and those in the corresponding smoothed trajectories when θ_U was 0.2. The smoothed trajectories are segmented based on the inferred intentions. The position is seen to appropriately increase and decrease when the human intention is with accelerating and decelerating respectively. Since $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$ and $\boldsymbol{U}^{(3)}$ differ from each other in the pitch angle θ, the pitch angle plot also shows intentions clearly: θ near 0 indicates hovering; positive θ with large magnitude indicates accelerating; negative θ with large magnitude indicates decelerating. Figure 3.9 shows the aligned segments and the variances of each resulting intention-pattern model when $n_p = 1$. It is first seen that the aligned segments are consistent, which indicates that the proposed intention inference is valid. More consistency is shown in position than in pitch angle partly because the Gaussian assumption is not flexible enough to describe the pitch angle. The derived variances show that the intention-pattern models are modelled probabilistically from observations and could be used to perform state estimation more precisely.

To analyse the dependency of the intention inference, the F1 score [26] evaluating the inference performance was derived with different levels of observation noises and control terms. The parameters varied were o_{s2} for the observation noise and θ_U for the control term since the pitch angle θ characterizes the intention. The ground truth intention was defined

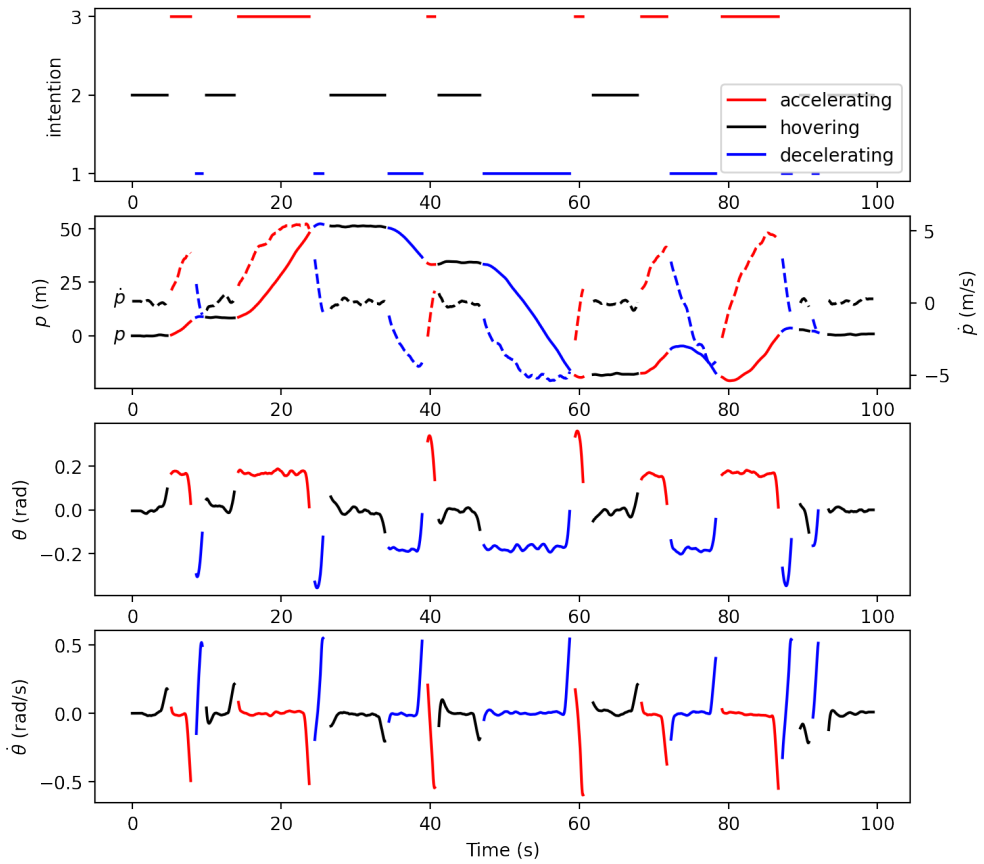
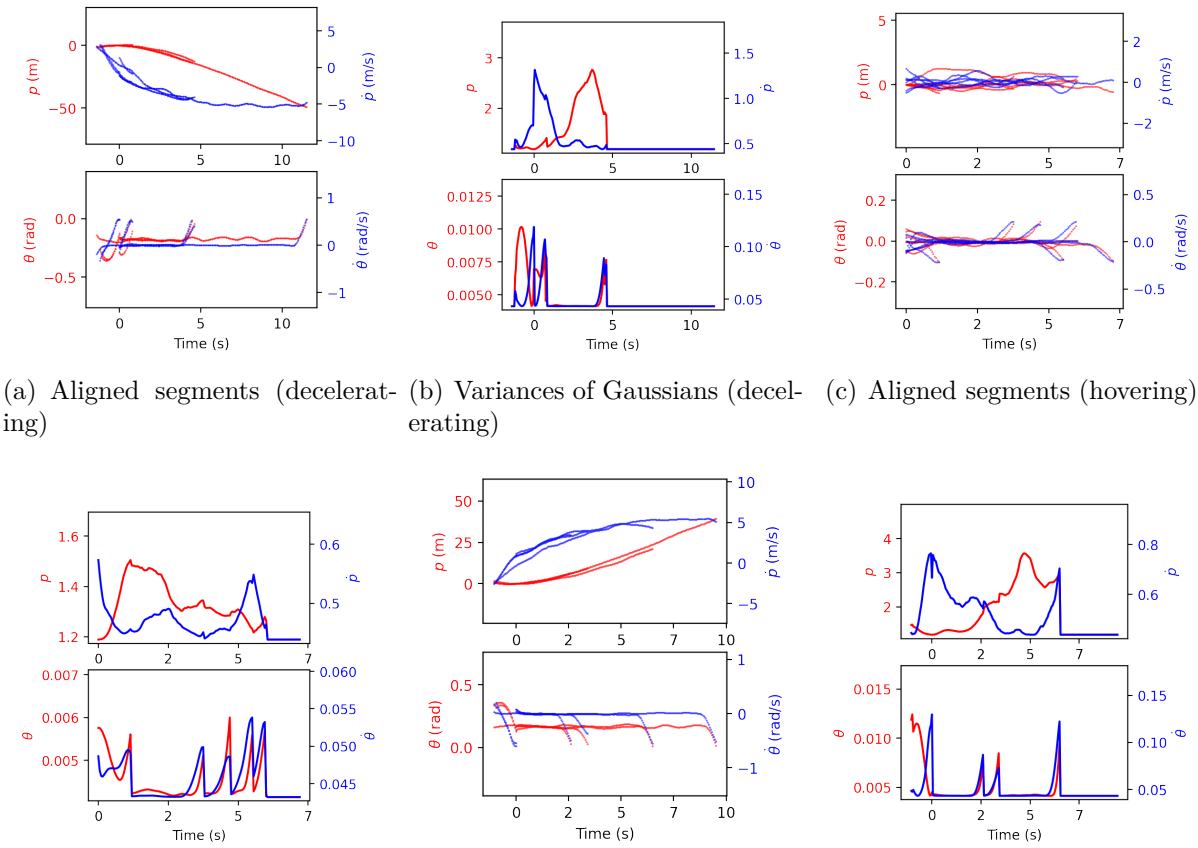


Figure 3.8. The inferred intention and the corresponding smoothed trajectory.

based on the real θ value: hovering when $|\theta| \leq 0.05$; accelerating when $\theta > 0.05$; decelerating when $\theta < -0.05$. The F1 score is calculated as

$$\frac{2}{\frac{\text{TP+FP}}{\text{TP}} + \frac{\text{TP+FN}}{\text{TP}}}$$

where TP, FP, and FN each represents the number of steps of true positive, false positive and false negative [26]. The F1 score which is closer to 1 indicates better inference. Figure 3.10 shows the distribution of the F1 score over o_{s2} and θ_U. As seen from the figure, the smaller the noise, the better the inference. For θ_U, there is a best value in the middle; either too large or too small will result in poor inference.



(a) Aligned segments (decelerating)

(b) Variances of Gaussians (decelerating)

(c) Aligned segments (hovering)

(d) Variances of Gaussians (hovering)

(e) Aligned segments (accelerating)

(f) Variances of Gaussians (accelerating)

Figure 3.9. Construction of the intention-pattern model for three intentions.

Figure 3.11 shows the resulting performance of each intention-pattern model. The two red broken lines show the range of motion pattern defined by the variance of the intention-pattern model constructed from the first 100 sec whereas the solid black lines the motions of the same intention-pattern model identified in the next 60 sec. It is seen that motions extracted in the 60 sec are well along with the range of the intention-pattern model. This verifies the validity of the probabilistically represented intention-pattern model.

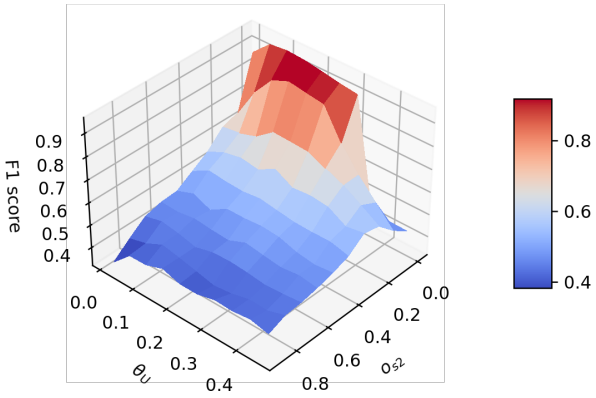
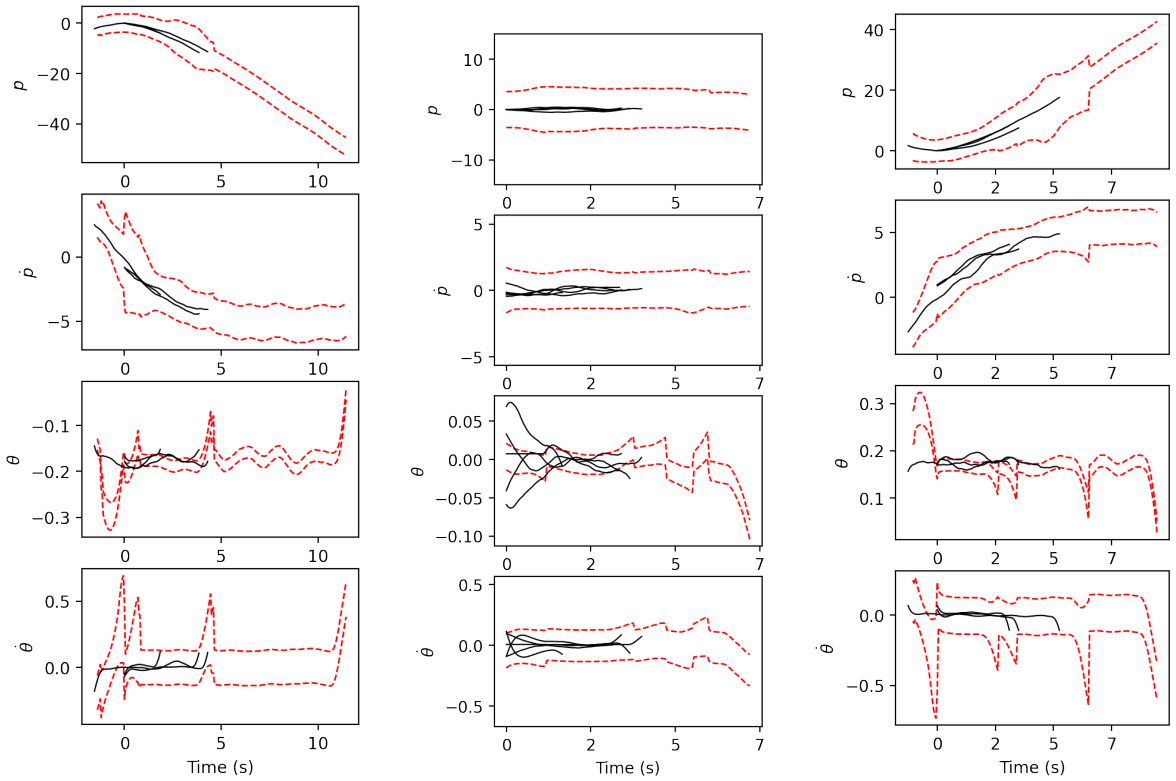


Figure 3.10. F1 score evaluating the intention inference accuracy with respect to simulation observation noise and the control term $\boldsymbol{U}^{(\mathrm{i})}$.



(a) Decelerating (b) Hovering (c) Accelerating

Figure 3.11. Validation of the constructed intention-pattern model.

3.3.2 Estimation Using Intention-Pattern Model

Having the intention-pattern model constructed using the first 100 sec, Figure 3.12 shows the result of state estimation incorporating the constructed intention-pattern model in the subsequent 60 sec. Unlike the intention-pattern model, the state estimation uses $n_p = 5$ since the effect of the proposed approach can be seen with the motion prediction. The ground truth and the result of the conventional KF estimation without intention incorporation are also shown for comparison. The estimation result of the proposed approach is seen to be closer to the ground truth than that of the conventional approach. The estimation of p and $\dot{p}$ particularly shows the responsive estimation of the proposed approach when the target motion is changed by the human while the conventional estimation exhibits notable delay. The faster response is due to the use of the intention-pattern model. The conventional approach could improve estimation by frequent accurate observation, but observations are often uncertain or unavailable.

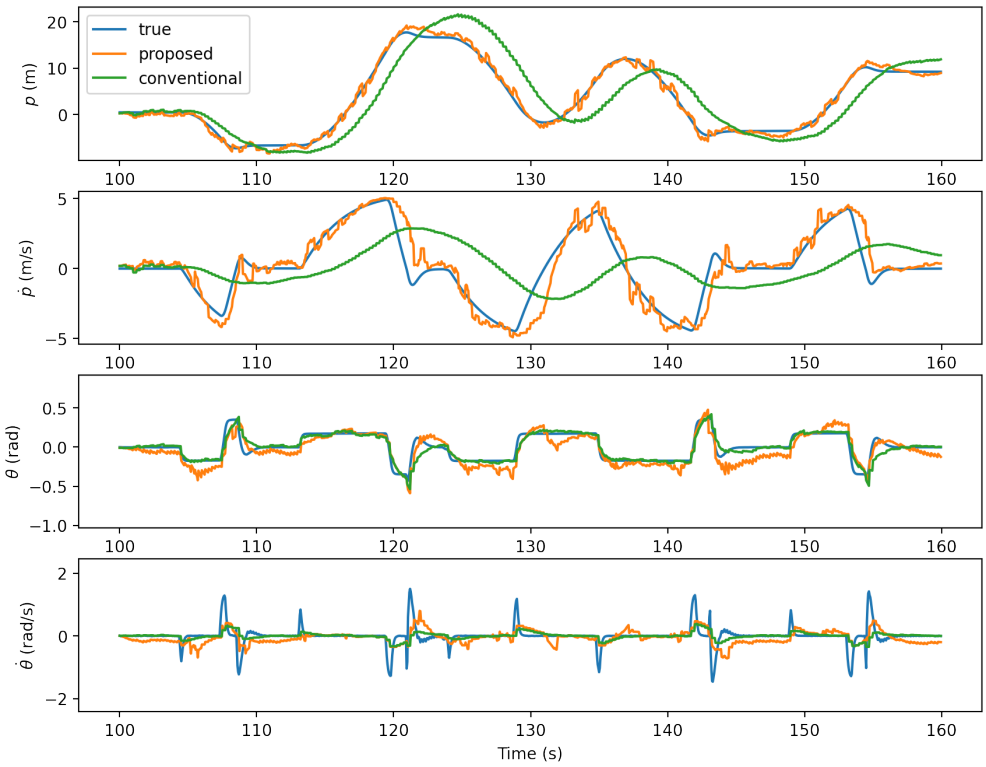


Figure 3.12. State estimation by the proposed and the conventional approaches.

Figure 3.13 shows the absolute error of estimated mean of each state variable with respect to time. While seeing less difference in θ and $\dot{\theta}$, the error of the proposed approach in p and $\dot{p}$ consistently and significantly stays low compared to the conventional approach. The difference is particularly large when the human changes the target motion since the conventional approach does not take the human intention into account. The maximum error and the mean squared error (MSE), integrating the absolute errors to a single quantity, are improved by almost three times and 8.7 times, respectively, when the proposed approach was deployed.

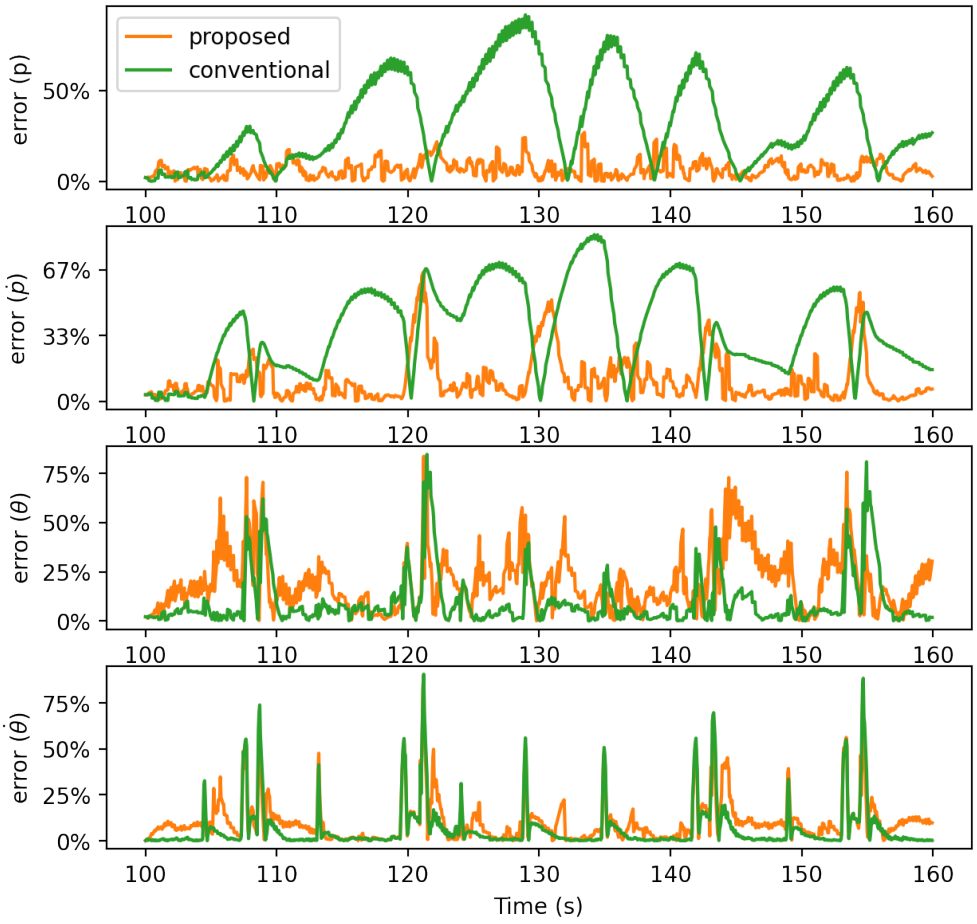


Figure 3.13. Absolute error of mean estimated by the proposed and the conventional approaches.

Figure 3.14 shows the variance of each state variable estimated by the proposed and the conventional approaches. The result shows that the proposed approach exhibits larger

variances than those of the conventional approach when the error is large. Since the proposed approach infers human intentions and adds their uncertainties, its variance is estimated more precisely and adequately. The variance of the conventional approach, on the other hand, is significantly smaller though the mean estimation is wrong. Having the human control deterministically treated without inferring intentions, the uncertainty of the conventional approach is markedly underestimated.

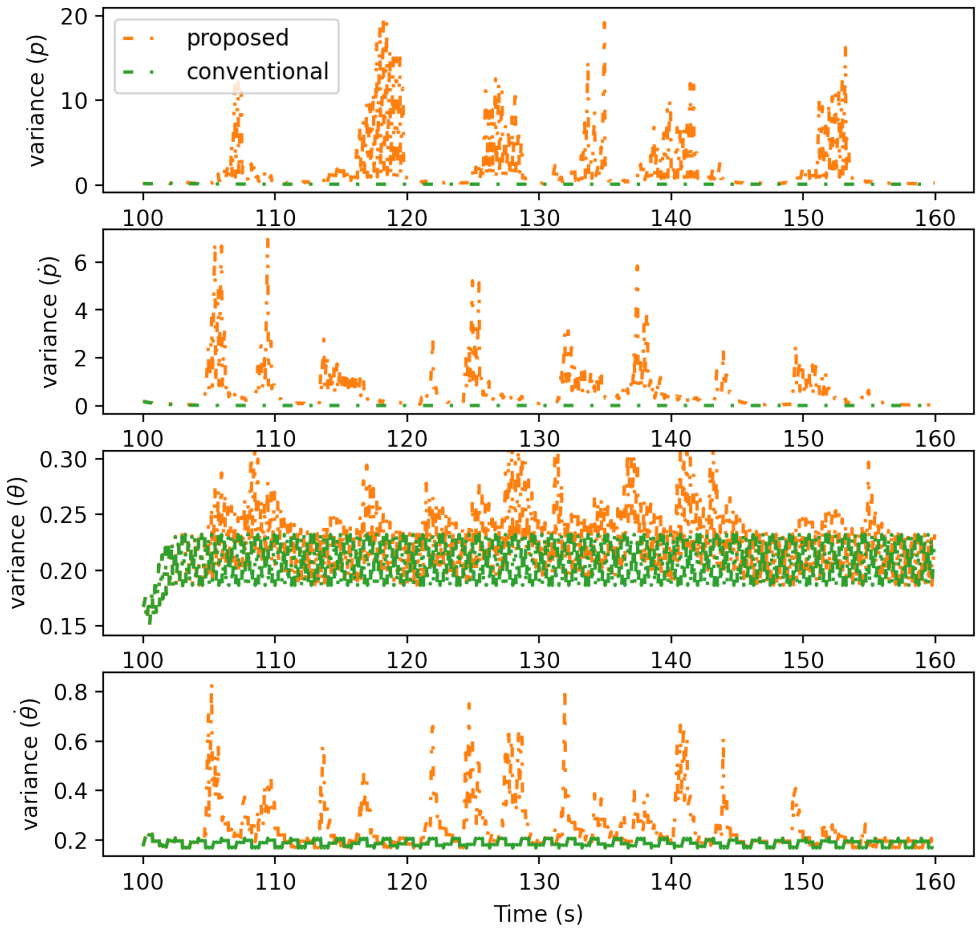
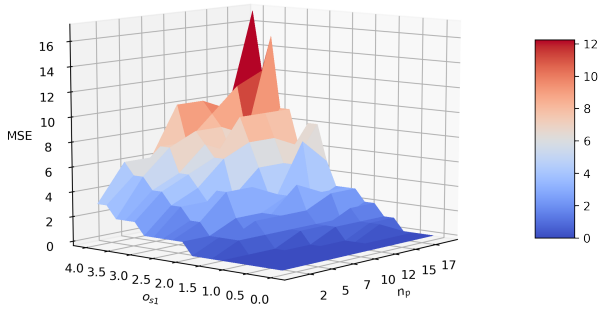


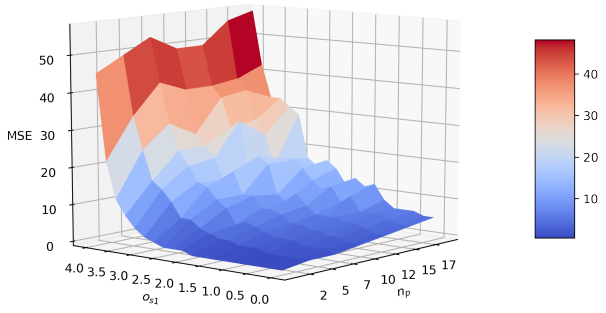
Figure 3.14. Variance estimated by the proposed and the conventional approaches.

The performance of the proposed approach in state estimation was further examined through the parametric study. Figure 3.15 shows the MSE of the proposed approach when o_{s1} and n_p were varied. o_{s1} was varied to examine the effect of the observation noise since it contributed less at the construction of the intention-pattern model. The result of the

conventional approach is also shown for comparison. It is first seen that the MSE of the proposed approach is significantly lower than that of the conventional approach when o_{s1} is large. The large o_{s1} increases the dependency of state estimation onto the prediction. As a result, the proposed approach, incorporating human intention and effective in prediction, can thus keep the MSE low. The result also shows that the MSE of the proposed approach remains low even when n_p is large. n_p also increases the dependency of state estimation onto the prediction, so the proposed approach becomes better than the conventional approach in accuracy. Meanwhile, the proposed and the conventional approaches exhibit a similar MSE when o_{s1} is low and n_p is one. This is because the estimation becomes correction-driven since the frequency of the accurate correction becomes high.



(a) Proposed approach



(b) Conventional approach

Figure 3.15. MSE with different observation noise and the number of prediction steps between observations

3.4 Summary

This chapter has presented an approach that estimates the state of a human-maneuvered target incorporating human intention, which consists of a pre-process constructing a intention-pattern model, and the main process allowing state estimation using the intention-pattern model. The pre-process constructs the intention-pattern model from the prior observations and probabilistically represents the model. The main process, then, uses standard state estimation such as KF extensively leveraging the probabilistically represented intention-pattern model. In the application of the proposed approach to the state estimation of a human-maneuvered multirotor, the numerical result has first shown that the constructed intention-pattern model represents the human intention appropriately. The result of state estimation of the human-maneuvered multirotor then shows that the proposed approach estimates the state more accurately than the conventional approach particularly when observations are uncertain or unavailable. The proposed approach has also demonstrated that it can estimate the covariance more precisely.

4. Robotic Escorting Benefiting from Intention Inference

4.1 Introduction

After the deployment of numerous robots in isolated workspaces, there has been a growing presence of robots in daily life, collaborating with humans [1, 2, 52]. One particular application is robotic following, where a robot moves alongside a human and maintains a relative position [53, 54]. Studies have shown that having a robot in front of a human induces less anxiety compared to having it behind, and it offers convenience when the human needs to physically interact with the robot [55]. The application of having a robot in front is referred as **robotic escorting** in this paper. Related approaches have attracted increasing interest in the last decade, as they offer various potential applications such as exercise companions, touchless shopping carts, and sanitary transportation in hospitals [56].

To successfully escort in front of a human, the robot needs to predict human behavior and move proactively [57]. There are two main categories of prediction techniques, differing in the description of human behavior [58]. The first category employs physics-based approaches, which utilize explicit dynamic models based on Newton's laws of motion to describe human behavior. These dynamic models are commonly used within recursive Bayesian estimation frameworks [13, 34, 35]. As the human input is unknown to the robot, Bogler [42] and Chakrabarty et al. [43] made assumptions about the input of the dynamic model. Other approaches ignore the input and assume conservative motion behaviors, such as constant velocity (CV) and coordinated turn (CT), as the most probable behavior [31]. The second category comprises pattern-based approaches that learn human motion behavior from data without explicitly defining parameterized functions. Bennewitz et al. [19], Elfring et al. [21], and Ravichandar et al. [20] learned motion patterns from collections of trajectories. Dermy et al. [22] learned a set of motion primitives from human-robot interaction demonstrations and inferred the intent of the human partner for collaboration. Qin et al. [18] utilized motion patterns to describe and identify recurring human behavior.

Physics-based approaches of the first category have been extensively utilized in the field of robotic escorting. Early studies conducted by Ho et al. [59], Jung et al. [60], and Cifuentes et al. [61] employed proportional control to assign fixed goal positions for the robot at a

predetermined distance in front of the observed human. These approaches utilized simple kinematic motion models, such as CV, CT, or non-holonomic motion models, incorporated within a Bayesian filter to track the human's movements. Subsequent work by Saiki et al. [53] and Hu et al. [57] extended these approaches to consider more complex motions, such as the combination of walking straight and turning. However, despite their achievements, these approaches heavily rely on the human consistently providing accurate indicators (e.g., head direction) to control the robot at each time step. This requirement poses challenges in accommodating distracted behavior and often leads to increased fatigue.

While prior research on robotic escorting predominantly focused on physics-based approaches, this paper harnesses the potential of pattern-based methods to characterize human behavior and guide robot movements. Human behavior is captured through a model encompassing intent and motion patterns. By leveraging demonstration data, the model is translated into a graph structure with vertices and edges on the environmental map. Combining this information with head direction, the human's intent is inferred by identifying the appropriate vertex, enabling autonomous robot movement towards that vertex. Due to the richer information contained in the model, the proposed approach achieves superior long-term prediction compared to the head direction-based methods alone. The robot's movement towards a vertex is autonomously guided by a path planner, thus reducing the demand for human attention. Notably, changes in human head direction only impact the robot's movement if they lead to a new intent vertex, mitigating the impact of distracted behavior.

Safety concerns is important in robotics [62, 63, 64, 65] and the trust beween human and the robot also play major role [66]. This paper further verifies the proposed approaches with specifically designed experiments and use study.

Safety concerns are of paramount importance in the field of robotics, as highlighted by several studies [62, 63, 64, 65]. In addition, building trust between humans and robots is equally crucial [66]. To address these critical aspects, this paper not only proposes novel approaches but also conducts meticulously designed experiments and user studies to verify their effectiveness. By combining rigorous testing and user feedback, this research aims to contribute to the advancement of both safety and trust in human-robot interactions.

The paper is organized as follows. Section 4.2 introduces the problem of robotic escorting and highlights the limitations of conventional solutions. Section 4.3 presents an innovative escorting approach that characterizes human behavior using a model of intent and motion patterns. To evaluate the effectiveness of the approach, Section 4.4 provides a thorough analysis of the modeling techniques employed and presents the results of extensive experiments conducted to assess the performance of the escorting system. Finally, Section 4.5 summarizes the key findings and draws meaningful conclusions. To simplify the description, this chapter will use the term "intent" to encompass both "intent" and "intention".

4.2 Escorting and Fundamentals

4.2.1 Robotic Escorting Problem

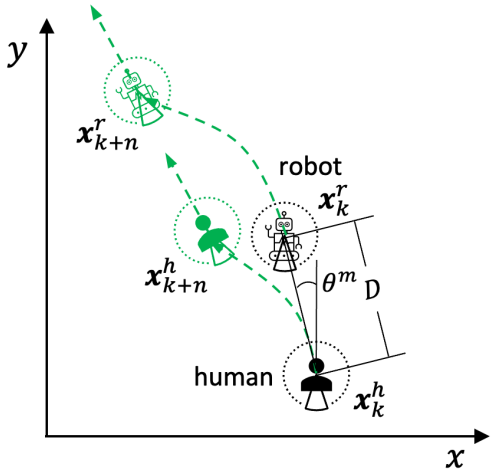


Figure 4.1. The problem of robotic escorting. The robot moves in front of the human while complying with the human's aim to maintain the relative position (the offset distance D and the azimuth θ^m).

Figure 4.1 presents a schematic diagram of robotic escorting, where a robot is expected to move in front of the human. The targeted robot's position is offset from the human by a distance D and an azimuth θ^m, with the goal of aligning the robot's absolute orientation

with that of the human. The relation at step k in a two-dimensional (2D) space can be expressed as

$$\begin{bmatrix} x_k^r \\ y_k^r \\ \theta_k^r \end{bmatrix} = \begin{bmatrix} x_k^h \\ y_k^h \\ \theta_k^h \end{bmatrix} + \begin{bmatrix} D\cos(\theta_k^h + \theta^m) \\ D\sin(\theta_k^h + \theta^m) \\ 0 \end{bmatrix}, \tag{4.1}$$

where the superscript $()^r$ and $()^h$ denote the robot and human, respectively. The variables x and y represent the position, while θ represents the orientation [57]. Notably, $\theta^m = 0$ in the case of escorting.

This paper addresses different types of mobile robots. The robot motion model is described generally as:

$$\boldsymbol{x}_k^r = \boldsymbol{f}^r(\boldsymbol{x}_{k-1}^r, \boldsymbol{u}_k^r) + \boldsymbol{w}_k^r, \tag{4.2}$$

where $\boldsymbol{x}_k^r = [x_k^r, y_k^r, \theta_k^r]^\top$ represents the robot state, $\boldsymbol{u}_k^r$ is the control input, and $\boldsymbol{w}_k^r$ represents the motion noise. The robot state is observed in the world coordinate using techniques such as simultaneous localization and mapping (SLAM). The robot's sensors capture human information, including the human's position and head direction, which are then transformed to the world coordinate system.

To enable proactive movement in escorting, the robot predicts the future human states by employing a human motion model. The general form of the motion model is given as

$$\boldsymbol{x}_k^h = \boldsymbol{f}^h(\boldsymbol{x}_{k-1}^h, \boldsymbol{u}_k^h) + \boldsymbol{w}_k^h \tag{4.3}$$

where $\boldsymbol{x}_k^h$ represents the human state, $\boldsymbol{u}_k^k$ denotes the human input that affects the state, and $\boldsymbol{w}_k^h$ represents the motion noise. Prediction, especially for longer periods, heavily relies on the human input $\boldsymbol{u}_k^h$ which is however unknown to the robot.

The equation (4.3) can be formulated in various ways depending on the chosen state $\boldsymbol{x}_k^h$, human input $\boldsymbol{u}_k^h$ and their corresponding $\boldsymbol{f}^h, \boldsymbol{w}^h$. For the purpose of prediction, Eq. (4.3) can be expressed as a velocity motion model

$$\begin{bmatrix} x_k^h \\ y_k^h \\ \theta_k^h \end{bmatrix} = \begin{bmatrix} x_{k-1}^h \\ y_{k-1}^h \\ \theta_{k-1}^h \end{bmatrix} + \begin{bmatrix} \cos(\theta_{k-1}^h) & 0 \\ \sin(\theta_{k-1}^h) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_k^h \\ \omega_k^h \end{bmatrix} \Delta t + \boldsymbol{w}_k^h. \tag{4.4}$$

This model assumes that human motion is governed by two velocities: a rotational velocity ω_k and a translational velocity v_k [13]. Here, Δt represents the time between two steps. The state prediction is achieved by iteratively computing Eq. (4.4). To predict the state over a longer period, assumptions of CV ($v_k^h \approx v_{k-1}^h$) or CT ($\omega_k^h \approx \omega_{k-1}^h$) are introduced.

As stated in Hu et al. [57], the velocity motion model Eq.(4.4) is applicable to normal walking scenarios, where humans primarily move forward without significant sideways motion. However, humans behavior is agile, and their input changes over time. Moreover, the interaction between the robot and the human can influence human motion. Consequently, the dynamic models represented by Eq. (4.4) can describe motion over multiple steps but exhibit limitations in capturing human motion over longer durations, such as periods exceeding 2 seconds.

4.2.2 Escorting with Head Direction

Since the prediction of human states using Eq. (4.4) shows limitations, additional information about the human head direction is employed since it exhibits human intent and allows the prediction of human motion in a longer run [56, 67]. By observing the head yaw ϕ_k^h and utilizing Eq. (4.4), the predicted position of the human after n steps can be calculated as follows:

$$\begin{bmatrix} x_{k+n}^h \\ y_{k+n}^h \\ \theta_{k+n}^h \end{bmatrix} = \begin{bmatrix} x_k^h \\ y_k^h \\ \theta_k^h \end{bmatrix} + \begin{bmatrix} \cos(\phi_k^h)\, v_k^h\, n\Delta t \\ \sin(\phi_k^h)\, v_k^h\, n\Delta t \\ f^{\phi}(\phi_k^h, n\Delta t) \end{bmatrix}. \tag{4.5}$$

The first two rows indicate that the human position after n steps will be aligned with the head yaw ϕ_k^h. The third row indicates that the human orientation will be adjusted by a function f^ϕ, which depends on the head yaw ϕ_k^h and the time period of $n\Delta t$. The function f^ϕ is determined empirically based on the relationship between the head direction and the future orientation of the human. It is to be noted that the uncertainty of the predicted human position primarily stems from the incorrelation of the head motion with the human motion.

To ensure that the robot maintains a desired distance ahead of the human, the goal pose of the robot $\boldsymbol{x}_{k+n}^{r*}$ is calculated as follows:

$$
\begin{aligned}
x_{k+n}^{r*} &= x_{k+n}^{h} + D\cos\theta_{k+n}^{h}, \\
y_{k+n}^{r*} &= y_{k+n}^{h} + D\sin\theta_{k+n}^{h}, \\
\theta_{k+n}^{r*} &= \theta_{k+n}^{h}.
\end{aligned}
$$

Finally, the control inputs for the robot $\boldsymbol{u}_{k+n}^{r}$ are determined to achieve autonomous escorting and can be derived as a function of $|\boldsymbol{f}^r(\boldsymbol{x}_{k+n-1}^{r}, \boldsymbol{u}_{k+n}^{r}) - \boldsymbol{x}_{k+n}^{r*}|$. The sequence of control actions aims to minimize the distance between the predicted future robot poses and the desired robot poses.

The conventional escorting approaches described above utilize the head yaw to compute the control command at each time step. Although these approaches enable the human to control the robot in a teleoperation-like manner, they exhibit several limitations. The head yaw ϕ_k^h does not always indicate the direction in which the human is moving, as the human may get distracted and look in other directions. Moreover, relying on the head yaw for prolonged periods can be mentally and physically exhausting for the human operator.

4.3 Proposed Escorting

Figure 4.2 illustrates the diagram of the proposed approach. Prior to escorting, the human operates the robot within the environment, generating demonstration data through the utilization of a built-in SLAM technique. This data encompasses the robot's trajectory

and the map of the environment. From this data, an intent-pattern model, representing the model of intent and motion patterns, is constructed and represented as a graph denoted by $\mathcal{G}\{V, E\}$. During the escorting phase, the human's intent is inferred by identifying the appropriate vertex in the graph based on the head direction. The discovered vertex is then passed to the navigation stack, which plans the path towards the vertex and controls the robot accordingly.

Section 4.3.1 details the process of constructing the intent-pattern model from the collected demonstrations, while Section 4.3.2 outlines the methodology for inferring the intent using this model.

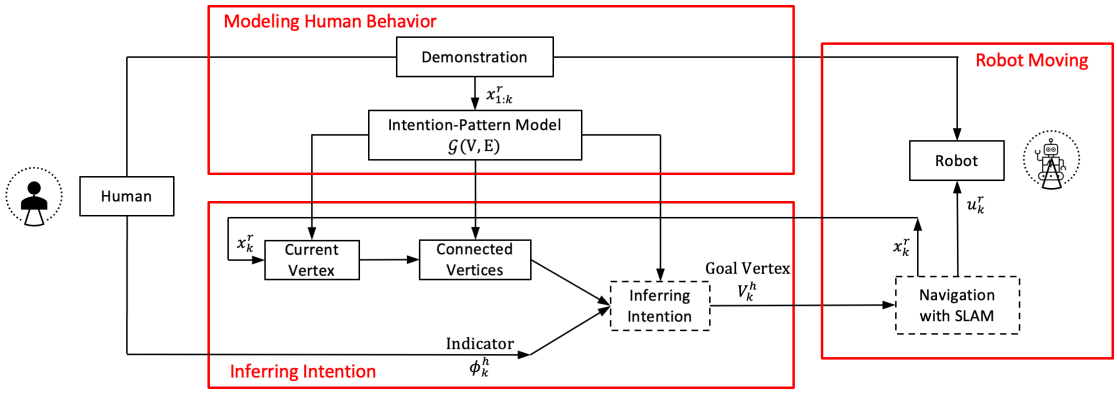


Figure 4.2. The diagram of the proposed approach which describes human behavior with the intent-pattern model and utilizes the model to infer the human intent for escorting.

4.3.1 Modeling Demonstration Data

Human behavior often extends over a certain duration, making it challenging to describe such prolonged behavior using motion models like Eq. (4.3). To address this, the proposed approach introduces the intent-pattern model to represent human behavior within a given period. The intent-pattern model establishes one-to-one relationships between intents and motion patterns. With this model, if the current behavior aligns with a specific motion pattern, the corresponding intent is likely to be occurring. Conversely, if a human intent is present, the future states can be derived using the corresponding motion pattern.

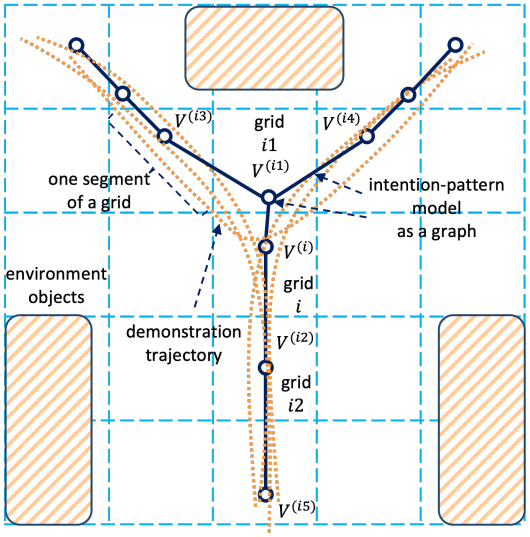


Figure 4.3. Modeling the demonstrations as an intent-pattern model, which is represented as a graph consisting of vertices and edges

To construct the intent-pattern model, this paper applies the Wizard of Oz concept and utilizes demonstrations that the human operates the robot [68]. A SLAM technique is employed to localize the robot and build a map of the environment. The robot's trajectory is represented as $\boldsymbol{x}_{1:k} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ where $\boldsymbol{x}_k = [x_k^r, y_k^r, \phi_k^r]^\top$ denotes the robot's pose.

Figure 4.3 presents an example of constructing the intent-pattern model from demonstrations. The map is initially divided into grids, each indexed with a side length of l. In the figure, the grid indices are represented by natural numbers $\mathrm{i}, \mathrm{i}1, \mathrm{i}2, \ldots$ for explanatory purposes. The trajectory is then segmented within each grid, as illustrated by the orange dotted lines, with each dot representing a trajectory point at a specific step.

The proposed approach assumes that the points within each grid follow a Gaussian distribution, given an appropriate grid side length l. For grid i, a point is sampled as

$$\boldsymbol{x}^{(\mathrm{i})} \sim \mathcal{N}\left(\bar{\boldsymbol{x}}^{(\mathrm{i})}, \boldsymbol{\Sigma}^{(\mathrm{i})}\right). \tag{4.7}$$

The mean and covariance are calculated as follows:

$$\bar{\boldsymbol{x}}^{(\mathrm{i})} = \frac{1}{n^{(\mathrm{i})}} \sum_{k_\mathrm{i}} \boldsymbol{x}^{(\mathrm{i},k_\mathrm{i})}, \tag{4.8a}$$

$$\boldsymbol{\Sigma}^{(\mathrm{i})} = \frac{1}{n^{(\mathrm{i})}} \sum_{k_\mathrm{i}} (\boldsymbol{x}^{(\mathrm{i},k_\mathrm{i})} - \bar{\boldsymbol{x}}^{(\mathrm{i})})(\boldsymbol{x}_\kappa^{(\mathrm{i},k_\mathrm{i})} - \bar{\boldsymbol{x}}^{(\mathrm{i})})^\top, \tag{4.8b}$$

where $n^{(\mathrm{i})}$ denotes the number of points in grid i, and k_i represents the index of each point [24]. The mean is treated as a graph vertex, denoted as $V^{(\mathrm{i})} = \bar{\boldsymbol{x}}^{(\mathrm{i})}$. Connections between grids are represented by edges. For example, the edge $E^{(\mathrm{i},\mathrm{i}1)}$ in the figure connects vertex $V^{(\mathrm{i})}$ and $V^{(\mathrm{i}1)}$.

An **intent** is defined as moving from a start vertex to an end vertex. The start vertex represents the robot's current position, while the end vertex signifies the robot's desired goal position. The **motion pattern** associated with an intent consists of the edges that connect the start and end vertices. Motion patterns are derived as sets of Gaussian distributions connected from one to another. For instance, in Figure 4.3, the motion pattern of the intent from $V^{(\mathrm{i}5)}$ to $V^{(\mathrm{i})}$ comprises the Gaussian distributions of $V^{(\mathrm{i}5)}, V^{(\mathrm{i}2)}$, and $V^{(\mathrm{i})}$, along with

their corresponding connections. The **intent-pattern model** is ultimately represented as a graph $\mathcal{G}\{V, E\}$, with each vertex being a Gaussian distribution as defined in Eq. (4.7).

4.3.2 Inferring Intents as Navigation Goals

The proposed approach infers the intent by determining the vertex that the robot is currently closest to and the vertex that the robot should navigate towards. Figure 4.4 illustrates an example scenario for explanation purposes, where both the robot and the human are moving along an edge. The robot's state is denoted as $\boldsymbol{x}_k^r$, and the human's head yaw is represented by ϕ_k^h. To calculate the nearest vertex to the robot, this paper computes the probability of each vertex being the current position of the robot:

$$\begin{aligned} \mathrm{i}_k &= \arg\max_{\mathrm{i}} Pr\Big\{\boldsymbol{x}_k | \mathcal{N}\Big(\bar{\boldsymbol{x}}^{(\mathrm{i})}, \boldsymbol{\Sigma}^{(\mathrm{i})}\Big)\Big\} \\ &= \arg\max_{\mathrm{i}} (2\boldsymbol{\pi}\boldsymbol{\Sigma}^{(\mathrm{i})})^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}[\boldsymbol{x}_k^r - \bar{\boldsymbol{x}}^{(\mathrm{i})}]^\top (\boldsymbol{\Sigma}^{(\mathrm{i})})^{-1} [\boldsymbol{x}_k^r - \bar{\boldsymbol{x}}^{(\mathrm{i})}]\Big\}. \end{aligned} \tag{4.9}$$

This is achieved by maximizing the likelihood of the robot's state given the Gaussian distribution associated with each vertex. Considering that there may be grids with a small number of points, the nearest vertex is determined based on distances when the number of points surrounding the robot's position $[x_k^r, y_k^r]$ is below a specified threshold:

$$\mathrm{i}_k = \arg\min_{\mathrm{i}} \Big[(x_k^r - \bar{x}^{(\mathrm{i})})^2 + (y_k^r - \bar{y}^{(\mathrm{i})})^2\Big]^{-\frac{1}{2}}. \tag{4.10}$$

where $\bar{x}^{(\mathrm{i})}$ and $\bar{y}^{(\mathrm{i})}$ are the elements of $\bar{\boldsymbol{x}}^{(\mathrm{i})}$.

Next, the goal vertex is selected, taking into account both the intent-pattern model and the head direction. The goal vertex is chosen among the vertices in the graph $\mathcal{G}\{V, E\}$. The BFS algorithm is utilized to find the connected vertices of the current position vertex, V^{i_k}[69]. The number of expanding layers in the BFS algorithm is denoted as n_{BFS}. In figure 4.4, $V^{\mathrm{i}_k,1}$ and $V^{\mathrm{i}_k,2}$ are on the BFS layer 1 while $V^{\mathrm{i}_k,3}, V^{\mathrm{i}_k,4}$ and $V^{\mathrm{i}_k,5}$ are on the BFS layer 2. This paper aims to find more connected vertices, as it increases the likelihood of identifying the goal vertex that aligns with the head yaw, potentially avoiding sharp turns and enabling smooth robot motion.

The connected vertices of $V^{(\mathrm{i}_k)}$ are represented as $V^{(\mathrm{i}_k,\diamond)} \in \{V^{(\mathrm{i}_k,1)}, V^{(\mathrm{i}_k,2)}, ...\}$. The orientation from $V^{(\mathrm{i}_k)}$ to each connected vertex is calculated using

$$\phi^{(\mathrm{i}_k,\diamond)} = \mathrm{atan2}(\mathrm{y}^{(\mathrm{i}_k,\diamond)} - \mathrm{y}^{\mathrm{r}}_{\mathrm{k}}, \mathrm{x}^{(\mathrm{i}_k,\diamond)} - \mathrm{x}^{\mathrm{r}}_{\mathrm{k}}), \tag{4.11}$$

where $x^{(\mathrm{i}_k,\diamond)}, y^{(\mathrm{i}_k,\diamond)}$ denote the elements of $V_k^{(\mathrm{i}_k,\diamond)}$. Taking the head yaw ϕ_k^h as a variable, the probability of selecting each connected vertex is calculated as a normal distribution:

$$\begin{aligned} Pr\{V^{(\mathrm{i}_k,\diamond)}\} &= \mathcal{N}(\phi^{(\mathrm{i}_k,\diamond)} | \phi_k^h, \sigma^{(\mathrm{i}_k,\diamond)}) \\ &= (2\boldsymbol{\pi}\sigma^{(\mathrm{i}_k,\diamond)})^{-\frac{1}{2}} \exp\{-\frac{1}{2}[(\phi^{(\mathrm{i}_k,\diamond)} - \phi_\kappa^h)/\sigma^{(\mathrm{i}_k,\diamond)}]^2\} \end{aligned} \tag{4.12}$$

where the head yaw ϕ_k^h serves as the mean and a manually set deviation parameter, $\sigma^{(\mathrm{i}_k,\diamond)}$, determines the spread of the distribution. The goal vertex of $\boldsymbol{x}_k^r$ and ϕ_k^h is chosen based on the maximum probability:

$$\begin{aligned} V_k &= V^{(\mathrm{i}_k,\diamond^*)}, \\ \diamond^* &= \arg\max_{\diamond} \{Pr\{V^{(\mathrm{i}_k,\diamond)}\} \mid \forall \diamond\}. \end{aligned} \tag{4.13}$$

The proposed approach integrates the selected goal vertex with the navigation stack to leverage existing techniques for practical performance [70]. The navigation stack consists of a global planner which plans the path from the current position to the goal position and a local planner which generates control commands for the robot. The navigation stack also incorporates a SLAM technique to localize the robot and map the environment, enabling real-time sensing of the surroundings and collision avoidance by the local planner.

During ongoing escorting, if the human moves out of the field of view (FOV), the robot will continue towards the last available navigation goal and come to a stop upon reaching it. However, if the navigation goal is obstructed by other objects or becomes unreachable, the navigation stack will halt the robot. In summary, the robot ensures safety by avoiding collisions with environmental objects while in motion and stops if no new goal can be computed based on the human's head direction.

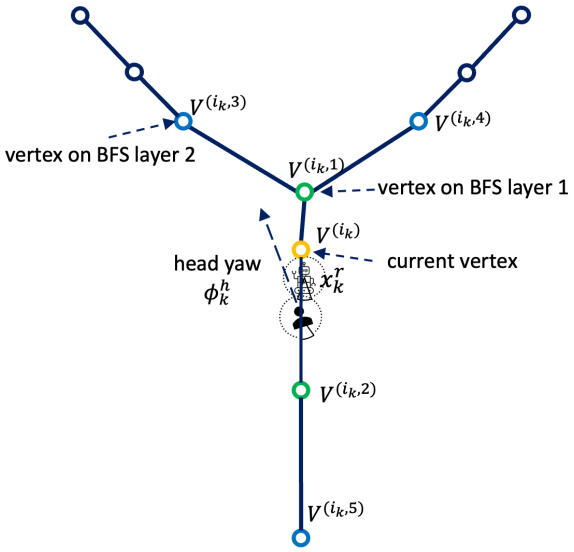
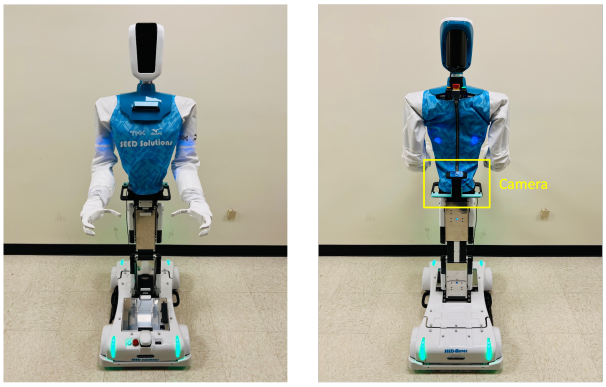


Figure 4.4. To infer human intent, the head direction of the human is compared with the directions between the current vertex and all the connected vertices obtained through the BFS algorithm.

4.4 Experimental Validation



(a) Robot (front) (b) Robot (back)

Figure 4.5. The experimental robot. A rear-facing camera is utilized for observing the human's head direction.

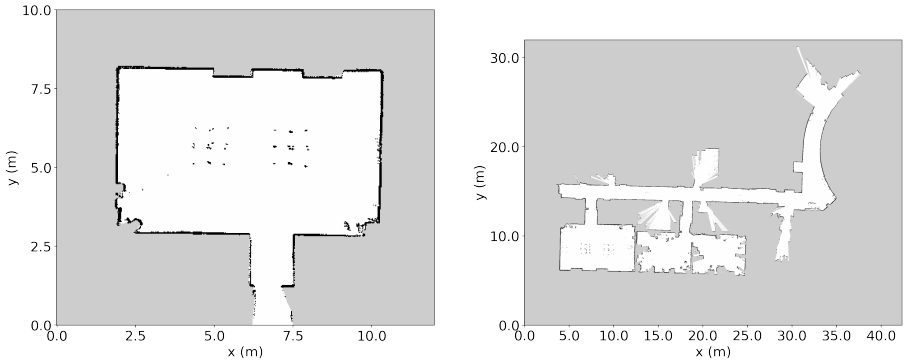
Figure 4.5 showcases the THK Seed robot that was utilized for the experimental validation. The robot is equipped with an omnidirectional wheeled platform, a 2D LiDAR in the front, and a rear-facing camera for observing the human's head direction. Table 4.1 presents the parameters employed in the proposed approach. The robot incorporates a navigation stack with the Gmapping SLAM technique, and the head direction is observed using Mediapipe [71]. The grid side length l and the level number of the BFS algorithm, n_{BFS}, were varied as part of a parametric study.

Figure 4.6 displays the maps of the two environments, and Table 4.2 presents the corresponding parameters. The first environment presents a small-scale office space with tables and chairs. The second environment encompasses an entire floor of a building, featuring corridors and rooms, thus representing a large-scale setting.

4.4.1 Graphical Representation of Modeling

This section focuses on the analysis of the graphical representation of the intent-pattern model. Figure 4.7 presents the demonstration trajectory in the first environment with the arrows depicting the robot poses and the human head yaws. The left figure illustrates the trajectory from 0 s to 70 s, while the right figure displays the remaining duration. Figure 4.8

(a) Office environment (b) Building environment

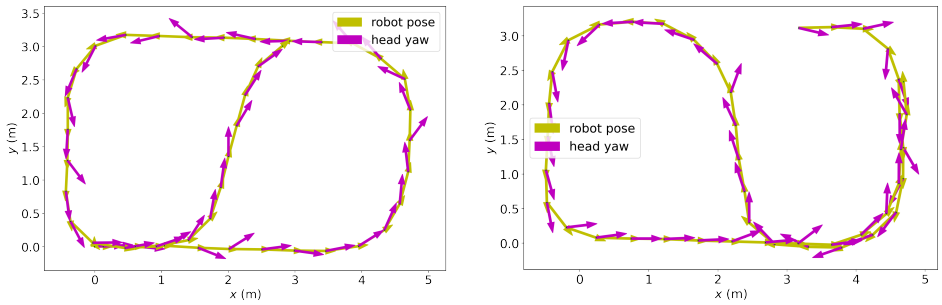


(c) Office map (d) Building map

Figure 4.6. Sample photos and maps of the two experimental environments

showcases the graphical representations of the intent-pattern model using different grid side lengths l. The left figure corresponds to $l = 0.5$ m, while the right figure represents $l = 1$ m. Notably, the left figure with a smaller grid side length exhibits higher vertex and edge density, providing more detailed information about the demonstration trajectory.

The metrics of the motion patterns are presented in Figure 4.9, considering variations in the grid length l and the duration of the demonstration. These metrics include the number of vertices, the number of edges, the total length of the graph, and the average angle of connected edges. The average angle reflects the smoothness of the robot's movement when traversing the graph. A larger angle indicates less smooth movements. Figure 4.9(a) demonstrates that a larger grid length reduces the graph size but leads to a larger angle, implying less smooth movements. Based on this analysis, a grid length of approximately 0.8 m appears to be a reasonable choice.



(a) 0 s - 70 s (b) 70 s - 140 s

Figure 4.7. Demonstration data of the office environment. The arrows represent the robot poses and the human head yaws.

Table 4.1. Parameters of the implementation

Parameter	Value
Robot	THK Seed
Size of Robot Base [m]	0.7×0.5
Time step [s]	$\Delta t = 0.2$
Grid Side Length l [m]	0.2 - 2
Level Number of BFS n_{BFS}	1,2,3
Detecting Head Yaw	Mediapipe
Control	Navigation Stack
Global Planner	Navfn
Local Planner	Timed Elastic Band
SLAM	Gmapping

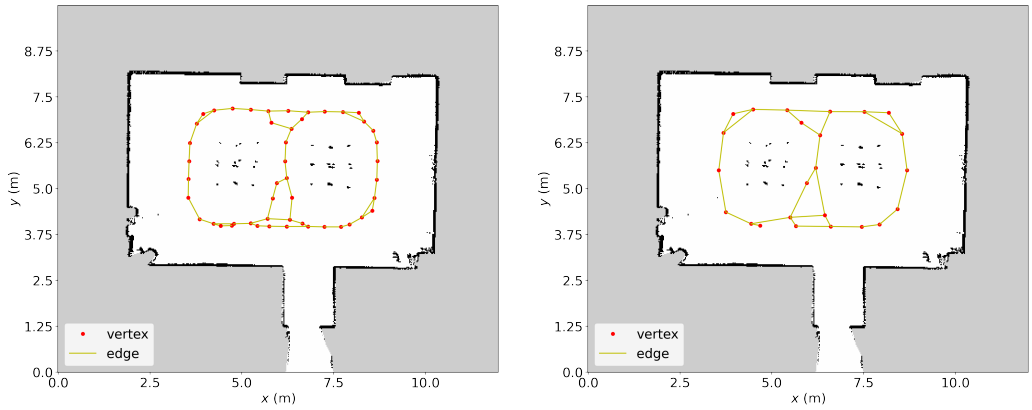
Table 4.2. Parameters for environments

Parameter	Value	Value
Environment	Office	Building
Size [m]	8×8	25×35
Duration of demonstration [s]	140	175
Characteristic	Small	Large

Figure 4.9(b) examines the metrics in relation to the demonstration duration, with l = 0.8 m. The number of vertices, the number of edges, and the total length of the graph increase as the duration of the demonstration extends, reaching a plateau after around 100 s. This indicates that the graph size grows with additional demonstration data but eventually reaches a limit due to the fixed number of grids in the map. The angle remains table after 80 s, suggesting that a certain amount of demonstration data is sufficient.

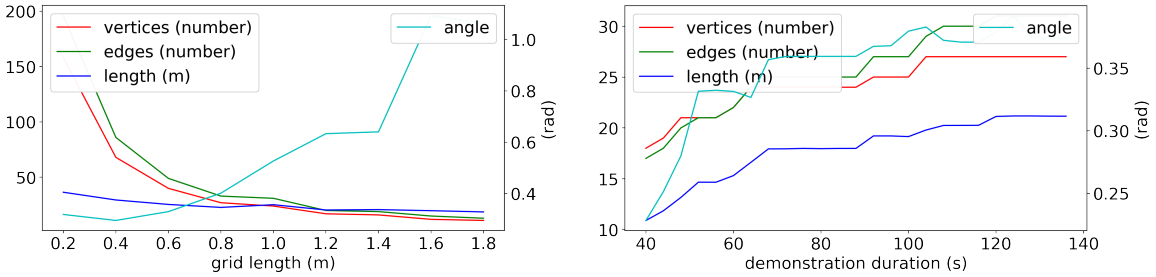
4.4.2 Intent Inference and Motion Prediction

This section aims to verify the performance of intent inference and motion prediction using the proposed model. Figure 4.10 presents a comparison of different approaches in predicting future robot positions. The proposed approach is compared with two conventional approaches: “orientation” and “head”. The conventional approaches require the human position, whereas the proposed approach only requires the head direction. For the sake of



(a) Grid length l is 0.5 m (b) Grid length l is 1 m

Figure 4.8. Graphical representation of intent-pattern model with different grid side lengths

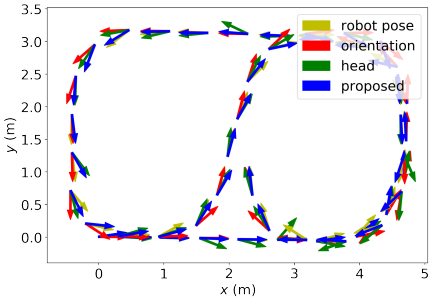


(a) With respect to grid length l (b) With respect to the duration of demonstration

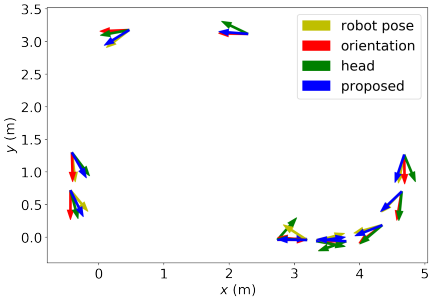
Figure 4.9. Metrics of the graphical representation with respect to different parameters: number of vertices, number of edges, and edge length are plotted on the left vertical axis, while the average angle of connected edges is shown on the right vertical axis.

comparison, this paper approximates the human position to be the robot position since the human is close proximity to the robot during robotic escorting. The "orientation" approach predicts future positions based on the robot's current position and orientation using the differential drive model of Eq. (4.4). The "head" approach predicts future positions based on the robot's current position and the head yaw of the human using Eq. (4.5).

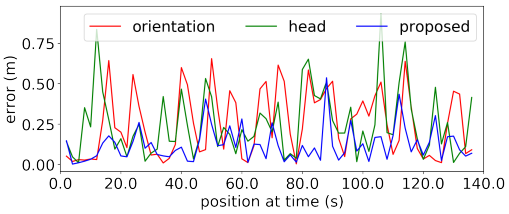
Figure 4.10(a) provides a trajectory view of the predictions. The arrows representing "robot pose" depict the robot trajectory from the demonstrations, while the other arrows



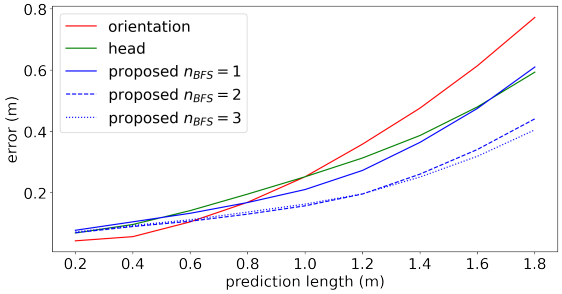
(a) Trajectory view



(b) Trajectory view for distracted behavior



(c) Error of prediction compared to the demonstration data



(d) Prediction error with respect to the prediction length and the number of BFS layers n_{BFS}

Figure 4.10. Comparison of prediction performance among different approaches for future robot positions

show the predictions of the different approaches. Figure 4.10(c) plots the average error of the prediction endpoint, which is 1 m ahead of the human. The average errors of the three approaches ("orientation", "head", "proposed") are 0.26 m, 0.25 m, and 0.11 m, respectively. It is evident that the arrows of the proposed approach align closely with the demonstration trajectory and exhibit smaller deviations.

A peak in the performance of the "head" approach occurs around 12 s. This is because sometimes the human looks in directions other than the moving direction to check their surroundings. This behavior can be considered **distracted behavior**. Figure 4.10(b) presents the trajectory view of the prediction for sample positions where the absolute angle between the head yaw and the moving direction is larger than 0.4 rad. For distracted behavior, the average error of the three approaches ("orientation", "head", "proposed") are 0.33 m, 0.52 m,

and 0.17 m, respectively. The proposed approach still yields the smallest error, highlighting its advantage in addressing distracted behavior.

Figure 4.10(d) demonstrates the prediction error with respect to the prediction length and the number of BFS layers n_{BFS}. The "orientation" approach performs best in the range of 0.2 m to 0.6 m. Beyond 0.8 m, the proposed approach exhibits smaller errors, showcasing its ability to predict future position over longer periods. The sparsely dotted blue line (n_{BFS} = 2) and densely dotted blue line (n_{BFS} = 3) outperforms the solid blue line (n_{BFS} = 1) after 0.8 m. This confirms the effectiveness of using BFS to find more connected vertices, leading to improved predictions of future positions.

4.4.3 User Study

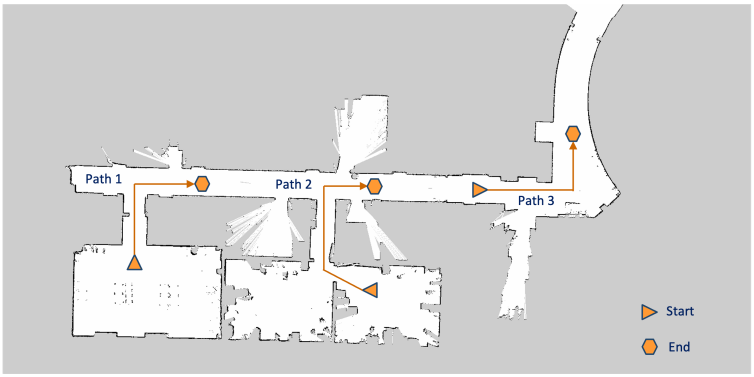
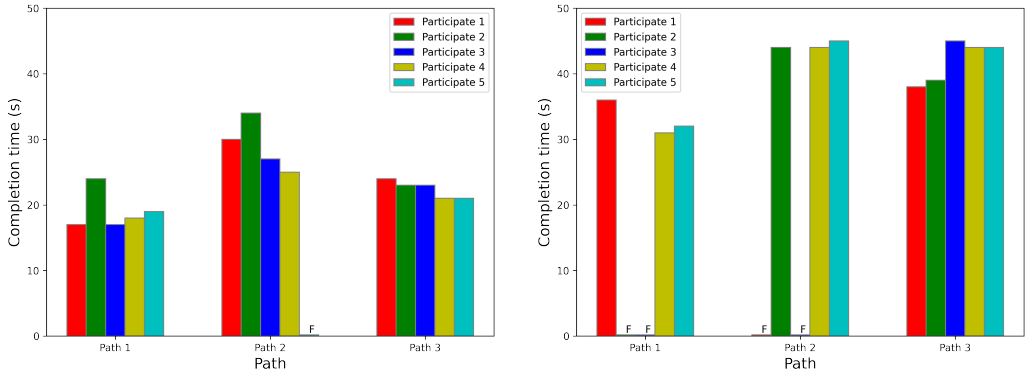


Figure 4.11. Three test paths for the user study

This section evaluates the performance of the proposed escorting approach in real environments through a user study. Five participants tested the escorting performance and compared it with a conventional approach as described in Section 4.2.2. They were instructed to move from one position to another within the environment. Three specific paths were selected from the building map, as shown in figure 4.11. Prior to the user study, a demonstration was performed to construct the intent-pattern model, and the graphical representation was shown to the participants. Then, the instructions for each approach were explained to the participants.

For the proposed approach, the instructions included: the robot will follow the head direction and use motion patterns to determine future directions based on the head direction. For conventional escorting approach, the instructions included: the robot will rely solely on the head direction for determining directions, and increased head rotation result in increased robot rotation. The completion time and success rate were recorded for each participant, as these metrics play a crucial role in determining the overall user experience. If the robot collided or deviated from the designated route more than twice, the test was considered a failure. After completing the tests, the participants were asked a comparison question: which approach do you find easier to operate, and which approach would you prefer to use in a supermarket setting?

Figure 4.12 presents the completion time results of the user study, with failures indicated by the symbol "F". The failure rates for the proposed and conventional approaches were 1/15 and 4/15, respectively. The average completion time for the proposed approach was 27.1 s, while for the conventional approach it was 40.2 s. The proposed approach demonstrated shorted completion time and a higher success rate, indicating its advantages over the conventional approach. In response to the comparison question, all five participants chose the proposed approach and rated it as easier to operate. One participant noted that the conventional approach allowed for more precise control of the robot, but it took time to become familiar with the controls and requires greater attention. For complex environments like a supermarket, all participants expressed a preference for using the proposed approach.



(a) Proposed (b) Conventional

Figure 4.12. Completion time of the user study. The symbol "F" indicates failure.

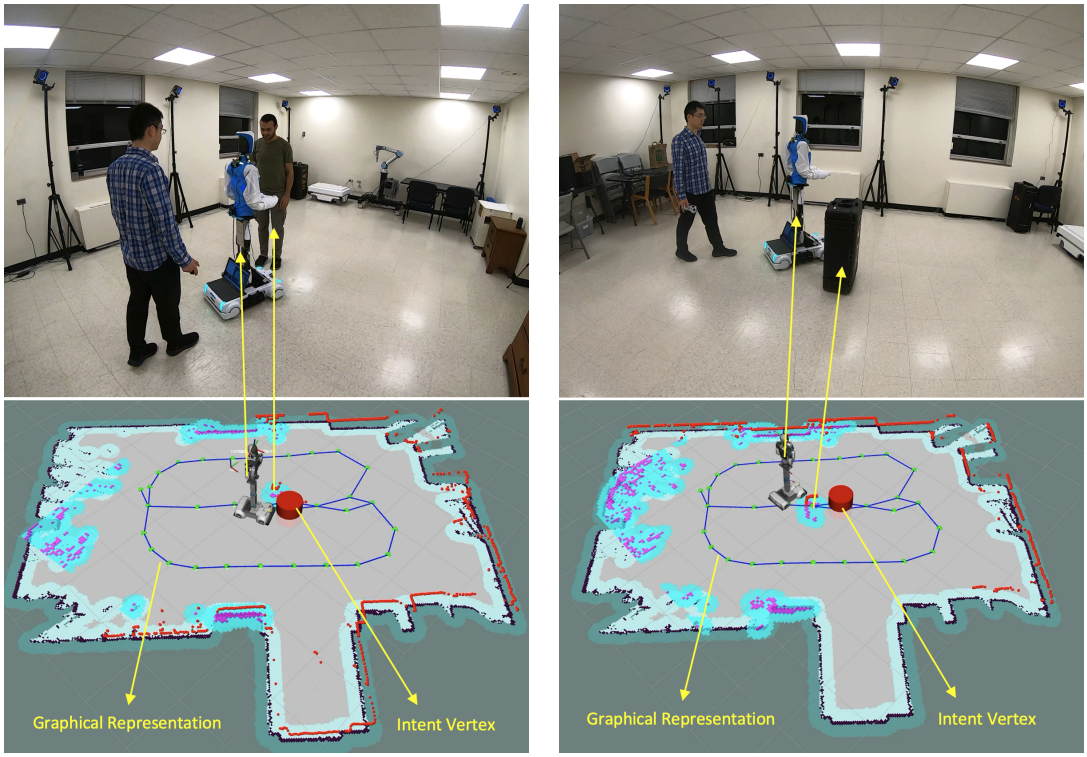
4.4.4 Collision Avoidance

This section showcases the effectiveness of collision avoidance for ensuring safety during the escorting process. Two specific cases were considered: when another person unexpectedly appears in the path of the robot and when an object not present in the map suddenly appears in the environment. According to the design, the robot is expected to navigate around these obstacles if there is sufficient space to do so. Figure 4.13 presents the results of the safety test, with the upper row showing corresponding photos and the lower row displaying the navigation view. The photos capture a moment in which the robot is in motion, while the navigation views provide insights into the status of the navigation stack. The graphical representation of the intent-pattern model is represented by the vertices and connected edges, with the goal vertex of human intent depicted as a red column. The robot itself is visualized as a 3D model.

During the test, the robot behaved as intended, demonstrating effective obstacle avoidance capabilities. In the first case, the robot successfully maneuvered around the travel case obstacle. In the second case, the robot initially halted in front of the person and then attempted to navigate around them. These tests effectively validate the safety aspects of the proposed approach in real-world applications.

4.5 Summary

This chapter introduces a novel approach to robotic escorting that enables autonomous movement of the robot in front of a human while aligning with the human's aim without direct communication. The proposed approach leverages demonstration data to model human intents and motion patterns as a graph. By inferring the human's intent based on the current head direction and the modeled graph, the robot effectively identifies the most probable vertex as an indication of the future human position. Subsequently, the robot autonomously moves towards the corresponding vertex, eliminating the need for the human to continuously provide accurate head direction at every time step. This capability significantly improves the adaptability of the system to accommodate distracted human behavior and reduces the cognitive burden on the human operator. Parametric studies and experimental validations



(a) Escorting with other individuals (b) Escorting with objects

Figure 4.13. The safety test of collision avoidance with dynamic obstacles. The robot successfully avoids collisions with other individuals obstructing its path during the escorting task and with objects not present in the map.

demonstrate the enhanced position prediction capabilities and the ability of the proposed approach to handle distractions effectively. Additionally, a user study was conducted, with participants consistently rating the proposed approach as easier to operate and requiring less attention compared to conventional methods. The proposed approach paves the way for more efficient and natural human-robot interaction in escorting tasks.

5. Conclusions

The proposed research has provided a comprehensive and systematic study of considering human intention in robotic applications, covering modeling, inferring, and incorporating human intention. The proposed approaches have been extensively validated through simulations and real experiments, contributing to a deeper understanding of building more customized human-robot coordination based on human intention. The findings open up new avenues for future research and hold promise for developing more advanced and personalized human-robot coordination systems.

In the area of intention estimation, this work has made significant progress by introducing the proposed framework. However, there is still much future work to be done. One potential direction is extending the framework to handle high-dimensional data where the relationships among variables and intentions are too complex to be captured by grid-based probabilistic representation. Deep learning approaches hold promise in addressing this challenge. Additionally, since intentions are long-term concepts that do not change frequently, further research is needed to process the raw step-based results and incorporate the probabilities of all intentions. Given the success of data-driven approaches in robotics, the related approaches of machine learning will play more important role in the future [72, 73, 74].

The proposed work has also made important strides in state estimation of a human-maneuvered target using human intention. However, there are ongoing efforts to expand this approach to address partially observable problems and leverage model predictive control. While observations are necessary for constructing the intention-pattern model, the state may not be fully observable. The proposed approach, which is effective in prediction-driven estimation, offers great potential for autonomous robots, making model predictive control one of the most promising extensions.

Regarding robotic escorting, this work represents an important step forward in the field of robotic escorting. However, the current model construction is limited to the demonstrated area, which restricts the robot's ability to navigate to positions outside of this predefined region. Although this limitation ensures safety and alleviates concerns for both users and building owners, there is room for further development. Future work will focus on extending

the framework to encompass unknown areas by integrating the proposed approach with conventional techniques. By combining the strengths of both approaches, the robot will be able to navigate through unexplored regions while still benefiting from the intent-pattern model. The proposed framework is designed to facilitate collaboration between a human operator and the robot. However, for individuals in the environment who are not actively participating and whose behavior is not accessible, as well as their level of distraction being more uncertain, the effectiveness of the proposed framework is limited. Further research is required to infer the intentions of non-cooperative individuals in the presence of the robot [75, 76, 77, 78].

Human actions show causal relations in many tasks such as using tools to assemble a table. In light of the research conducted on inferring human intentions with causal relations based on tool usage for human-robot cooperation, one area of future work involves refining and generalizing the representation of causal relations as motion patterns. To enhance the applicability and adaptability of the framework, it is important to explore methods for automatically discovering and representing motion patterns across a broader range of tasks. This would involve investigating techniques such as machine learning and data mining to uncover common patterns to represent the causal relations.

Another promising direction for future research is the integration of contextual information to enhance the accuracy and robustness of inferring human intentions. Contextual factors such as environmental conditions, social cues, and user preferences can significantly influence human behavior [79, 80]. By incorporating contextual information into the inference process, the framework can adapt to dynamic and diverse situations, leading to more precise predictions of human intentions and more effective assistance from robots.

REFERENCES

[1] Amir Rasouli and John K Tsotsos. "Autonomous vehicles that interact with pedestrians: A survey of theory and practice". In: *IEEE transactions on intelligent transportation systems* 21.3 (2019), pp. 900–918.

[2] Kyle Brown, Katherine Driggs-Campbell, and Mykel J Kochenderfer. "Modeling and prediction of human driver behavior: A survey". In: *arXiv e-prints* (2020), arXiv–2006.

[3] A. Morris et al. "A Robotic Walker That Provides Guidance". In: *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*. 2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422). Vol. 1. Sept. 2003, 25–30 vol.1. DOI: 10.1109/ROBOT.2003.1241568.

[4] K. Kosuge and Y. Hirata. "Human-Robot Interaction". In: *2004 IEEE International Conference on Robotics and Biomimetics*. 2004 IEEE International Conference on Robotics and Biomimetics. Aug. 2004, pp. 8–11. DOI: 10.1109/ROBIO.2004.1521743.

[5] Seung Yeol Lee et al. "Human-Robot Cooperation Control for Installing Heavy Construction Materials". In: *Autonomous Robots* 22.3 (Sept. 13, 2006), p. 305. ISSN: 1573-7527. DOI: 10.1007/s10514-006-9722-z. URL: https://doi.org/10.1007/s10514-006-9722-z (visited on 07/19/2022).

[6] Daniel M. Ho, Jwu-Sheng Hu, and Jyun-Ji Wang. "Behavior Control of the Mobile Robot for Accompanying in Front of a Human". In: *2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. 2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). July 2012, pp. 377–382. DOI: 10.1109/AIM.2012.6265891.

[7] Eui-Jung Jung, Byung-Ju Yi, and Shin'ichi Yuta. "Control Algorithms for a Mobile Robot Tracking a Human in Front". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Oct. 2012, pp. 2411–2416. DOI: 10.1109/IROS.2012.6386200.

[8] Przemyslaw A. Lasota, Terrence Fong, and Julie A. Shah. "A Survey of Methods for Safe Human-Robot Interaction". In: *Foundations and Trends in Robotics* 5.3 (2017), pp. 261–349. ISSN: 1935-8253, 1935-8261. DOI: 10.1561/2300000052. URL: http://www.nowpublishers.com/article/Details/ROB-052 (visited on 09/13/2019).

[9] Jos Elfring, René Molengraft, and Maarten Steinbuch. "Learning Intentions for Improved Human Motion Prediction". In: *Robotics and Autonomous Systems* 62.4 (Apr. 1, 2014), pp. 591–602. ISSN: 0921-8890. DOI: 10.1016/j.robot.2014.01.003. URL: https://www.sciencedirect.com/science/article/pii/S0921889014000062 (visited on 05/17/2021).

[10] Yanan Li and Shuzhi Sam Ge. "Human–robot collaboration based on motion intention estimation". In: *IEEE/ASME Transactions on Mechatronics* 19.3 (2013), pp. 1007–1014.

[11] Haoyu Bai et al. "Intention-aware online POMDP planning for autonomous driving in a crowd". In: *2015 ieee international conference on robotics and automation (icra)*. IEEE. 2015, pp. 454–460.

[12] Andrey Rudenko et al. "Human motion trajectory prediction: A survey". In: *The International Journal of Robotics Research* 39.8 (2020), pp. 895–935.

[13] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT Press, 2005.

[14] X. Rong Li and V.P. Jilkov. "Survey of Maneuvering Target Tracking. Part I. Dynamic Models". In: *IEEE Transactions on Aerospace and Electronic Systems* 39.4 (Oct. 2003), pp. 1333–1364. ISSN: 1557-9603. DOI: 10.1109/TAES.2003.1261132.

[15] Tetsuya Kawase et al. "Two-stage kalman estimator using advanced circular prediction for maneuvering target tracking". In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. Vol. 4. IEEE. 1998, pp. 2453–2456.

[16] Dean Conte and Tomonari Furukawa. "Autonomous Robotic Escort Incorporating Motion Prediction and Human Intention". In: *Proceedings of the 2021 IEEE International Conference on Robotic and Automation*. IEEE. 2021, 7 pages.

[17] Yu Liu and X Rong Li. "Intent based trajectory prediction by multiple model prediction and smoothing". In: *AIAA Guidance, Navigation, and Control Conference*. 2015, p. 1324.

[18] Yongming Qin, Makoto Kumon, and Tomonari Furukawa. "Estimation of a Human-Maneuvered Target Incorporating Human Intention". In: *Sensors* 21.16 (2021), p. 5316.

[19] Maren Bennewitz et al. “Learning motion patterns of people for compliant robot motion”. In: *The International Journal of Robotics Research* 24.1 (2005), pp. 31–48.

[20] Harish chaandar Ravichandar and Ashwin Dani. “Human intention inference through interacting multiple model filtering”. In: *2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE. 2015, pp. 220–225.

[21] Jos Elfring, René Van De Molengraft, and Maarten Steinbuch. “Learning intentions for improved human motion prediction”. In: *Robotics and Autonomous Systems* 62.4 (2014), pp. 591–602.

[22] Oriane Dermy et al. “Prediction of intention during interaction with icub with probabilistic movement primitives”. In: *Frontiers in Robotics and AI* 4 (2017), p. 45.

[23] Przemyslaw A Lasota and Julie A Shah. “A multiple-predictor approach to human motion prediction”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 2300–2307.

[24] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[25] X Rong Li and Vesselin P Jilkov. “Survey of maneuvering target tracking. Part V. Multiple-model methods”. In: *IEEE Transactions on Aerospace and Electronic Systems* 41.4 (2005), pp. 1255–1321.

[26] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O’Reilly Media, 2019.

[27] Yaakov Bar-Shalom, X Rong Li, and Thiagalingam Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[28] Matouš Vrba, Daniel Heřt, and Martin Saska. “Onboard marker-less detection and localization of non-cooperating drones for their safe interception by an autonomous aerial system”. In: *IEEE Robotics and Automation Letters* 4.4 (2019), pp. 3402–3409.

[29] Nuno Pessanha Santos, Victor Lobo, and Alexandre Bernardino. “Two-stage 3D model-based UAV pose estimation: A comparison of methods for optimization”. In: *Journal of Field Robotics* 37.4 (2020), pp. 580–605.

[30] Ren Jin et al. “Drone detection and pose estimation using relational graph networks”. In: *Sensors* 19.6 (2019), p. 1479.

[31] X Rong Li and Vesselin P Jilkov. “Survey of maneuvering target tracking. Part I. Dynamic models”. In: *IEEE Transactions on aerospace and electronic systems* 39.4 (2003), pp. 1333–1364.

[32] Peter J Costa. “Adaptive model architecture and extended Kalman-Bucy filters”. In: *IEEE Transactions on Aerospace and Electronic Systems* 30.2 (1994), pp. 525–533.

[33] William F Leven and Aaron D Lanterman. “Unscented Kalman filters for multiple target tracking with symmetric measurement equations”. In: *IEEE Transactions on Automatic Control* 54.2 (2009), pp. 370–375.

[34] Changyun Liu, Penglang Shui, and Song Li. “Unscented extended Kalman filter for target tracking”. In: *Journal of Systems Engineering and Electronics* 22.2 (2011), pp. 188–192.

[35] M Sanjeev Arulampalam et al. “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking”. In: *IEEE Transactions on signal processing* 50.2 (2002), pp. 174–188.

[36] Miguel Martínez-García et al. “Communication and interaction with semiautonomous ground vehicles by force control steering”. In: *IEEE transactions on cybernetics* (2020).

[37] Miguel Martínez-García, Yu Zhang, and Timothy Gordon. “Memory pattern identification for feedback tracking control in human–machine systems”. In: *Human factors* 63.2 (2021), pp. 210–226.

[38] J Josiah Steckenrider and Tomonari Furukawa. “A Probabilistic Model-adaptive Approach for Tracking of Motion with Heightened Uncertainty”. In: *International Journal of Control, Automation and Systems* 18 (2020), pp. 2687–2698.

[39] Tobias Gindele, Sebastian Brechtel, and Rüdiger Dillmann. “A probabilistic model for estimating driver behaviors and vehicle trajectories in traffic environments”. In: *13th International IEEE Conference on Intelligent Transportation Systems*. IEEE. 2010, pp. 1625–1631.

[40] Raman Mehra. “On the identification of variances and adaptive Kalman filtering”. In: *IEEE Transactions on automatic control* 15.2 (1970), pp. 175–184.

[41] Ali Almagbile, Jinling Wang, and Weidong Ding. “Evaluating the performances of adaptive Kalman filter methods in GPS/INS integration”. In: *Journal of Global Positioning Systems* 9.1 (2010), pp. 33–40.

[42] Philip L Bogler. “Tracking a maneuvering target using input estimation”. In: *IEEE transactions on Aerospace and Electronic Systems* 3 (1987), pp. 298–310.

[43] Ankush Chakrabarty et al. “State and unknown input observers for nonlinear systems with bounded exogenous inputs”. In: *IEEE Transactions on Automatic Control* 62.11 (2017), pp. 5497–5510.

[44] Henk AP Blom and Yaakov Bar-Shalom. “The interacting multiple model algorithm for systems with Markovian switching coefficients”. In: *IEEE transactions on Automatic Control* 33.8 (1988), pp. 780–783.

[45] Alper Akca and M Önder Efe. “Multiple model Kalman and Particle filters and applications: a survey”. In: *IFAC-PapersOnLine* 52.3 (2019), pp. 73–78.

[46] Xin Wang et al. “Combination of interacting multiple models with the particle filter for three-dimensional target tracking in underwater wireless sensor networks”. In: *Mathematical Problems in Engineering* 2012 (2012).

[47] Yanjun Ma, Shunyi Zhao, and Biao Huang. “Multiple-model state estimation based on variational Bayesian inference”. In: *IEEE Transactions on Automatic Control* 64.4 (2018), pp. 1679–1685.

[48] Miao Yu, Hyondong Oh, and Wen-Hua Chen. “An improved multiple model particle filtering approach for manoeuvring target tracking using airborne GMTI with geographic information”. In: *Aerospace Science and Technology* 52 (2016), pp. 62–69.

[49] Xiao-Rong Li and Yaakov Bar-Shalom. “Multiple-model estimation with variable structure”. In: *IEEE Transactions on Automatic control* 41.4 (1996), pp. 478–493.

[50] Linfeng Xu, X Rong Li, and Zhansheng Duan. “Hybrid grid multiple-model estimation with application to maneuvering target tracking”. In: *IEEE Transactions on Aerospace and Electronic Systems* 52.1 (2016), pp. 122–136.

[51] Xiao-Rong Li, Yaakov Bar-Shalom, and William Dale Blair. “Engineer’s guide to variable-structure multiple-model estimation for tracking”. In: *Multitarget-multisensor tracking: Applications and advances.* 3 (2000), pp. 499–567.

[52] Jemin Hwangbo et al. “Learning agile and dynamic motor skills for legged robots”. In: *Science Robotics* 4.26 (2019), eaau5872.

[53] Luis Yoichi Morales Saiki et al. “How Do People Walk Side-by-Side?: Using a Computational Model of Human Behavior for a Social Robot”. In: (2012), p. 8.

[54] Md Jahidul Islam, Jungseok Hong, and Junaed Sattar. “Person-following by autonomous robots: A categorical overview”. In: *The International Journal of Robotics Research* 38.14 (2019), pp. 1581–1618.

[55] Eui-Jung Jung, Byung-Ju Yi, et al. “Control algorithms for a mobile robot tracking a human in front”. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2012, pp. 2411–2416.

[56] Dean Conte and Tomonari Furukawa. “Autonomous Robotic Escort Incorporating Motion Prediction and Human Intention”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 3480–3486.

[57] Jwu-Sheng Hu, Jyun-Ji Wang, and Daniel Minare Ho. “Design of Sensing System and Anticipative Behavior for Human Following of Mobile Robots”. In: *IEEE Transactions on Industrial Electronics* 61.4 (Apr. 2014), pp. 1916–1927. ISSN: 1557-9948. DOI: 10.1109/TIE.2013.2262758.

[58] Andrey Rudenko et al. “Human Motion Trajectory Prediction: A Survey”. In: *The International Journal of Robotics Research* 39.8 (July 1, 2020), pp. 895–935. ISSN: 0278-3649. DOI: 10.1177/0278364920917446. URL: https://doi.org/10.1177/0278364920917446 (visited on 10/02/2020).

[59] D. M. Ho, J. Hu, and J. Wang. “Behavior control of the mobile robot for accompanying in front of a human”. In: *2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. 2012, pp. 377–382.

[60] E. Jung, B. Yi, and S. Yuta. “Control algorithms for a mobile robot tracking a human in front”. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012, pp. 2411–2416.

[61] Carlos A. Cifuentes et al. “Human-Robot Interaction Based on Wearable IMU Sensor and Laser Range Finder”. In: *Robot. Auton. Syst.* 62.10 (Oct. 2014), pp. 14251439. ISSN: 0921-8890. DOI: 10.1016/j.robot.2014.06.001. URL: https://doi.org/10.1016/j.robot.2014.06.001.

[62] Xiaobo Ma, Abolfazl Karimpour, and Yao-Jan Wu. "Statistical evaluation of data requirement for ramp metering performance assessment". In: *Transportation Research Part A: Policy and Practice* 141 (2020), pp. 248–261.

[63] Jiacheng Yuan, Nicolai Häni, and Volkan Isler. "Multi-step recurrent Q-learning for robotic velcro peeling". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 6657–6663.

[64] Xugui Zhou et al. "Data-driven design of context-aware monitors for hazard prediction in artificial pancreas systems". In: *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE. 2021, pp. 484–496.

[65] Xugui Zhou et al. "Hybrid Knowledge and Data Driven Synthesis of Runtime Monitors for Cyber-Physical Systems". In: *IEEE Transactions on Dependable and Secure Computing* (2023).

[66] Jin Xu and Ayanna Howard. "How much do you trust your self-driving car? exploring human-robot trust in high-risk scenarios". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2020, pp. 4273–4280.

[67] Dean Conte and Tomonari Furukawa. "Autonomous Bayesian Escorting of a Human Integrating Intention and Obstacle Avoidance". In: vol. 39. 6. Wiley, 2022, pp. 679–693.

[68] Laurel D. Riek. "Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines". In: *Journal of Human-Robot Interaction* 1.1 (July 28, 2012), pp. 119–136. DOI: 10.5898/JHRI.1.1.Riek. URL: https://doi.org/10.5898/JHRI.1.1.Riek (visited on 10/11/2022).

[69] Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

[70] Eitan Marder-Eppstein et al. "The office marathon: Robust navigation in an indoor office environment". In: *2010 IEEE international conference on robotics and automation*. IEEE. 2010, pp. 300–307.

[71] Camillo Lugaresi et al. "Mediapipe: A framework for building perception pipelines". In: *arXiv preprint arXiv:1906.08172* (2019).

[72] Xiaoling Luo et al. "A multisource data approach for estimating vehicle queue length at metered on-ramps". In: *Journal of Transportation Engineering, Part A: Systems* 148.2 (2022), p. 04021117.

[73] Xinrui Wang and Yan Jin. "Work Process Transfer Reinforcement Learning: Feature Extraction and Finetuning in Ship Collision Avoidance". In: *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Vol. 86212. American Society of Mechanical Engineers. 2022, V002T02A069.

[74] Shijie Gao and Nicola Bezzo. "A conformal mapping-based framework for robot-to-robot and sim-to-real transfer learning". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 1289–1295.

[75] Rahul Peddi et al. "A data-driven framework for proactive intention-aware motion planning of a robot in a human environment". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 5738–5744.

[76] Xugui Zhou et al. "Strategic safety-critical attacks against an advanced driver assistance system". In: *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE. 2022, pp. 79–87.

[77] Jixuan Zhi, Lap-Fai Yu, and Jyh-Ming Lien. "Designing human-robot coexistence space". In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 7161–7168.

[78] Jixuan Zhi and Jyh-Ming Lien. "Improving Human-Robot Collaboration via Computational Design". In: *arXiv preprint arXiv:2303.11425* (2023).

[79] Yunzhong He et al. "Que2Engage: Embedding-based Retrieval for Relevant and Engaging Products at Facebook Marketplace". In: *Companion Proceedings of the ACM Web Conference 2023*. 2023, pp. 386–390.

[80] Yuxin Tian, Shawn Newsam, and Kofi Boakye. "Fashion Image Retrieval With Text Feedback by Additive Attention Compositional Learning". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 1011–1021.