UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

MULTIPLE EXTENDED OBJECT TRACKING WITH THE 3D-INSTANCE SEGMENTATION ALGORITHM

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO

NICOLÁS IGNACIO FIERRO FLORES

PROFESOR GUÍA:
MARTIN ADAMS

MIEMBROS DE LA COMISIÓN:
CLAUDIO PÉREZ FLORES
MIGUEL TORRES TORRITI

Este trabajo ha sido parcialmente financiado por ANID-Subdirección de Capital Humano/Magíster Nacional/2023 - 22230898.

SANTIAGO DE CHILE
2024

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
Y MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO
POR: NICOLÁS IGNACIO FIERRO FLORES
FECHA: 2024
PROF. GUÍA: MARTIN ADAMS

SEGUIMIENTO MULTIPLE DE OBJETOS EXTENDIDOS CON EL ALGORITMO DE SEGMENTACIÓN DE INSTANCIAS EN 3D

Los algoritmos clásicos de seguimiento de múltiples objetos (MOT) asumen la generación de mediciones únicas por objetivo, pero la evolución hacia el seguimiento de objetos extendidos (MEOT) asume que un objetivo pueda generar múltiples mediciones. El MEOT enfrenta desafíos en entornos congestionados, donde mediciones cercanas pueden interpretarse erróneamente.

Presentamos el algoritmo 3D-INSEG (Segmentación de Instancias en 3D) usando cámaras estéreo y redes neuronales (NNs) para segmentación y profundidad en 3D. La visión estéreo permite la obtención de profundidad, mediante la cual la segmentación 2D producida mediante NNs puede ser llevada a coordenadas en 3D, de esta forma cada píxel perteneciente al objetivo genera una medición 3D.

Validamos con datos LIDAR Velodyne, enfocándonos en el seguimiento humano. Aplicamos 3D-INSEG a secuencias estéreo, extrayendo información 3D para cada objeto detectado. Las mediciones se procesan con un filtro PMBM de objetivo extendido con implementación GGIW.

El MEOT se beneficia de los datos generados mediante el algoritmo 3D-INSEG, demostrado comparativamente con datos LIDAR Velodyne. Este trabajo mejora el seguimiento en entornos desafiantes con segmentación y estimación de profundidad en 3D.

RESUMEN EN INGLÉS DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA, MENCIÓN ELÉCTRICA
Y MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO
POR: NICOLÁS IGNACIO FIERRO FLORES
FECHA: 2024
PROF. GUÍA: MARTIN ADAMS

MULTIPLE EXTENDED OBJECT TRACKING WITH THE 3D-INSTANCE SEGMENTATION ALGORITHM

Classical multiple object tracking (MOT) assumes each target gives one measurement. Newer work considers extended object tracking (MEOT), where one target can generate multiple measurements. Good measurements are key for accurate tracking.

We propose 3D-INSEG (3D-INstance SEGmentation) using stereo cameras and neural networks for depth and 3D segmentation. Stereo vision helps with depth info, making 2D segmentation from CNNs better. We check this against traditional clustering with Velodyne LIDAR data.

We focus on tracking individual humans, estimating depth with RAFT-stereo, and using Mask-RCNN for 2D segmentation. We test MEOT with simulated and real laser data in open spaces, seeing limits in crowded or tight spots where close measurements can be wrong.

Then, we use 3D-INSEG for MEOT with stereo image sequences, getting 3D info for each target. We use an extended target PMBM filter with a GGIW setup to process measurements. MEOT does better with 3D-INSEG data, shown by comparing with Velodyne LiDAR-based MEOT in the same spots. This improves tracking accuracy in tough spots using segmentation and depth.

Et elle disait
On joue tous la même chanson
Mais on la joue pas d'la même façon
On écoute tous la même chanson
Mais on l'entend pas d'la même façon
-Jacques (Dans La Radio)

Acknowledgments

Quiero agradecer a mi madre por haberme apoyado siempre y haberme dado grandes momentos: largas conversaciones, consejos, idas al C.D.A., y una lista interminable de hermosos recuerdos que atesoro con mucho cariño.

A mi padre por su apoyo constante, cariño y consejos, así como a todos los Fierro Carmona. Agradezco a mis hermanos por su visita en el verano previo al término de mi tesis; fue un hermoso momento.

A mi familia por siempre estar ahí y haber sentido su cariño: a los de Santiago, Parral, Antofagasta y Tomé. En especial a mi Tata Lucho y Tío Pepe.

A mis amigos que han estado en este proceso universitario, que me ha llenado de alegrías y buenas experiencias: Matraqueo&W, Rancagua extendido, PxB, La Champa, los de Francia, el CC y Kongberzion. Una mención especial a Paul Lieutier, que desde que lo conocí siempre me apoyó y se convirtió en un gran amigo, gracias a él pude conocer a los amigos de La Champa y vivir dos años y medio muy valiosos en Francia, y ahora continuamos en Chile, compartiendo en el trabajo y en los entrenamientos de BJJ. Mención especial a Sebastián Brzovic, con quien comenzamos juntos el proyecto de paracaidismo y con quien nos motivamos mutuamente, Simón Vidal, que probablemente sea de las pocas personas que lean esto y Eduardo Agüero y Juan Grant que me pidieron ser mencionados.

Quiero agradecer de forma especial a mi polola Javiera Águila, quien me ha acompañado en todo este proceso: cuando envié los documentos para ser aceptado en el programa de magíster, cuando obtuve la beca Magíster Nacional de ANID, y ahora último, en el cierre del ciclo universitario, estando a mi lado en la redacción de varios capítulos de esta tesis.

Agradezco a mi profesor guía Martin Adams por su apoyo y guía en el desarrollo de este trabajo, así como a Leonardo Cament por su gran dedicación en los proyectos del Laboratorio de Visión Computacional y el apoyo que me dio en todas las etapas de la tesis. Sin su apoyo, esta tesis no vería la luz. Agradezco al resto de los estudiantes del laboratorio con quienes pasé muy buenos momentos, en particular a Ignacio Dassori, que me ayudó en todas las tomas de datos, en la instalación de sensores y registros, y que se encuentra en algunas de las fotos de esta tesis.

Table of Content

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Autonomous systems, such as self-driving vehicles [3], heavily depend on robust environmental perception to identify and monitor moving objects within their surveillance area. Scan-based sensors, such as RAdio Detection And Ranging (RADAR) [4] [5], and LIght Detection And Ranging (LIDAR) [6], are commonly employed for this task. While these sensors provide precise point measurements of detected objects, they lack information about the types of objects and are susceptible to noise, without detailing the origin of reflections. Consequently, the development of algorithms that can effectively mitigate sensor noise and extract labeled object tracks is imperative. This intricate problem is commonly known as Multiple Object Tracking (MOT).

The sensors utilized frequently offer high resolution, resulting in numerous detections for a single object. This situation presents the challenge of tracking multiple objects, where individual objects may generate multiple spatially dispersed detections without accompanying labels. This specific tracking challenge is referred to as Multiple Extended Object Tracking (MEOT). In MEOT, accurately determining the spatial boundaries of an object becomes crucial, because these extents aid in distinguishing between multiple objects. They facilitate robust tracking by ensuring that each object is uniquely identified and monitored over time.

Recently extended target tracking solutions based on Random Finite Sets (RFSs) have been proposed in which detection as well as state uncertainty is taken into account. [7] introduces a Generalised Labelled Multi-Bernoulli (GLMB) based extended object tracker in which the target estimate is composed of several Multi-Bernoulli (MB) components. In the GLMB tracker, the number of feasible associations, and consequently MB components, increases exponentially with the number of measurements and objects. In MEOT the data association task involves both delineating the origin-based clusters of measurements and determining the connections between these measurement cells and potential sources. The Methods for this task can be grouped into three categories: Gating [8], Clustering followed by assignment [9] and Sampling [9]. To deal with this problem previous work such as [10] has used clustering techniques based on distance such as DBSCAN [11] to generate the most probable associations.

The Poisson multi-Bernoulli mixture (PMBM) filter [12], has demonstrated its status as one of the leading methods in target tracking. This filter is grounded in the concept of RFS, where a potentially detected target is represented as a Bernoulli RFS, and the ensemble of potential targets is modeled as a Poisson Point Process (PPP).

The extended version of the PMBM filter for MEOT, incorporating the gamma Gaussian inverse-Wishart (GGIW) implementation, is detailed in [13]. This version utilizes a two-step clustering and assignment approach [14], to identify relevant global hypotheses during the update of each preceding global hypothesis. The process involves applying DB-SCAN [11] and subsequently employing Murty's algorithm [15] for each measurement partition and global hypothesis, facilitating the determination of optimal cluster-to-Bernoulli component assignments.

While this implementation demonstrates promising results in simulated scenarios, challenges arise when applied to real laser data. The clustering approach may become computationally expensive due to the sheer volume of data. Moreover, the abundance of objects and data in the scene can contribute to misdetections, thereby complicating the MEOT task.

The surge of Artificial Intelligence (AI) technologies in recent years has revolutionized the landscape of computer vision applications. This study harnesses the potential of these advancements by synergistically integrating depth estimation [16] and 2D segmentation techniques [17] employing state-of-the-art AI approaches to address the challenges posed by the assignment of measurements to a single target in multi-object tracking.

In this work, we introduce the 3D-INSEG (3D INstance SEGmentation) algorithm, aimed at mitigating the complexities of the data association problem. By detecting objects across various classes and identities in a 3D spatial context, the algorithm enables the effective grouping of multiple measurements and the establishment of connections with potential sources. The 3D-INSEG algorithm offers advantages over traditional techniques based on the clustering and gating of point cloud data. The primary goal of this article is to assess the effectiveness of the 3D-INSEG Algorithm via a comparative analysis with alternative methods. Additionally, we show the utility of the detections produced by the 3D-INSEG algorithm within an extended tracking algorithm [18], which estimates target extent, approximated as an ellipsoid, jointly with the target's kinematic state. The first part of this study focuses on Single Extended Object Tracking (SEOT). A SEOT algorithm is proposed based on the integrated depth estimation and 2D segmentation techniques are applied to enable robust 3D human tracking. The subsequent application of a SEOT algorithm, based on random matrices [18], will demonstrate the potential of AI-based methodologies in augmenting tracking precision. The success of this approach is gauged by assessing the accuracy of the 3D detections. This evaluation relies on comparisons with ground truth data acquired through LIDAR measurements and manual annotations. In the second part of this study, we implement the 3D INstance SEGmentation (3D-INSEG) algorithm [19] for MEOT to address the challenges posed by the data association problem. By detecting objects across various classes and identities within a 3D spatial context, the algorithm facilitates the effective grouping of multiple measurements and the establishment of connections with potential sources. The 3D-INSEG algorithm offers advantages over traditional techniques that rely on the clustering and gating of point cloud data. We demonstrate the benefits of utilizing masked and clustered data when integrated into the PMBM filter, showcasing improved performance compared to scenarios where Velodyne data alone is employed.

1.2 Hypothesis

Our hypothesis suggests that in environments with dense and extensive laser data, the task of MEOT becomes challenging due to the complexity arising from numerous object detections. To address this challenge, we propose integrating stereo vision with Neural Networks (NNs) for 3D depth estimation and segmentation. This integration aims to achieve 3D instance segmentation within both SEOT and MEOT contexts.

We expect that leveraging stereo vision combined with NNs-based segmentation will significantly enhance tracking accuracy and reduce computational complexity, especially in environments with dense object populations. By incorporating 3D instance segmentation into the tracking framework, we anticipate improved object identification and localization, which are essential for robust and reliable tracking performance. The structured data, modeled as a set of clusters, is anticipated to facilitate this enhancement.

1.3 General Objectives

The general objectives of this research are threefold. Firstly, the study aims to integrate stereo vision with CNN-based segmentation to achieve 3D instance segmentation. This integration will leverage the depth information from stereo vision and the object recognition capabilities of CNNs to enhance the segmentation process. Secondly, the research will evaluate the performance of the 3D-INSEG algorithm in scenarios involving extended object tracking. By testing this algorithm in various challenging environments, the study seeks to determine its effectiveness and accuracy in tracking objects over time. Thirdly, the research will focus on the implementation and validation of the proposed algorithm using real-world data. This will involve practical experiments and data collection to ensure that the algorithm performs well under real-world conditions. By addressing these objectives, the study aims to contribute to the development of more advanced and reliable object tracking systems, which are crucial for applications such as autonomous driving and surveillance.

1.4 Specific Objectives

The specific objectives of this thesis are as follows:

- Implement a stereo disparity estimation model for images obtained using the ZED 2 camera. This model will compute the disparity between stereo pairs of images, which will then be used for inverse projection to create a 3D representation of the scene.

- Implement instance segmentation model in 2D images. The instance segmentation model will then be applied to stereo images to extend the segmentation into 3D space.
- Develop the 3D-INSEG algorithm, which will use the stereo disparity estimation and 2D instance segmentation models to generate 3D segmentations from stereo image pairs. This algorithm will identify and segment objects in 3D, providing detailed measurements for each segmented object, which can then be associated with different sources.
- Integrate the 3D segmentations generated by the 3D-INSEG algorithm into MEOT filters. This integration aims to enhance the accuracy of tracking extended objects by utilizing the detailed 3D segmentations, and to simplify data association in MEOT by using clusters derived from the segmentation.
- Collect data in various scenarios using the ZED 2 stereo camera and Velodyne LiDAR. The data collection will include different environments with varying numbers of objects and different scene characteristics, ensuring a comprehensive dataset for testing and validation.
- Conduct a qualitative and quantitative comparison of MEOT performance using data from LIDAR and the 3D-INSEG algorithm. This comparison will evaluate the strengths and weaknesses of each approach in different scenarios, assessing their effectiveness in terms of tracking accuracy, data association complexity, and overall robustness.

1.5 Thesis Structure

The remainder of this thesis is structured as follows: Chapter 2 provides a comprehensive literature review. Chapter 3 details the 3D-INSEG algorithm. Chapter 4 focuses on the integration of 3D-INSEG with SEOT. Chapter 5 explores the integration of 3D-INSEG with MEOT. Chapter 6 presents the experimental results, analyses and discussion. Chapter 7 summarizes the findings and conclusions drawn from this research.

Chapter 2

Literature Review

2.1 Extended Object Tracking

Extended Object Tracking algorithms not only estimate the kinematic state of an object, but they also estimate its shape and orientation. In [7], SEOT algorithms and their MEOT extensions are surveyed. The initial theoretical foundations for a multiple extended object tracker was established based on the Probability Hypothesis Density (PHD) filter, incorporating the PPP model [20]. Subsequently, this filter was implemented utilizing the random matrix model, with an inverse Wishart distribution employed for estimating the shape matrix [21]. To enhance the approach, an estimate of the Poisson rate, governing the expected number of generated measurements for each target, was introduced using a gamma distribution [22]. The integration of these components led to the development of the GGIW model [23]. The PMBM filter [12] has been adapted for extended object tracking with a GGIW formulation [24]. For extended objects, the state of an object x_k is defined as

$$x_k = [\gamma_k^\top, m_k^\top]^\top, \tag{2.1}$$

where parameter vector γ_k describes the shape, orientation, and size of the object at discrete time k and m_k describes the kinematic state at time k.

2.1.1 Random Matrix Approaches

An elliptical object extent approach is presented in [25], based on a 2D shape matrix:

$$\mathbf{X}_k = \begin{bmatrix} \cos\alpha_k & -\sin\alpha_k \\ \sin\alpha_k & \cos\alpha_k \end{bmatrix} \begin{bmatrix} (l_k^1)^2 & 0 \\ 0 & (l_k^2)^2 \end{bmatrix} \begin{bmatrix} \cos\alpha_k & -\sin\alpha_k \\ \sin\alpha_k & \cos\alpha_k \end{bmatrix}^\top, \tag{2.2}$$

where α_k is the (counter-clockwise) orientation and l_k^1 and l_k^2 denote the semi-axis lengths of the ellipse at discrete time k. This matrix implicitly encodes the shape parameters of γ_k in (2.1). In this approach, it is assumed that measurements are scattered across the entire ellipse surface following a spatial Gaussian distribution. The distribution is modeled as

$$p(z_k|m_k, \mathbf{X}_k) = \mathcal{N}(z_k; \mathbf{H}m_k, \mathbf{X}_k), \tag{2.3}$$

where $\mathbf{H} = [\mathbf{I}_2, \mathbf{0}_2]$, the measurement source z_k is modeled as Gaussian distributed with mean $\mathbf{H}m_k$ and covariance matrix $\mathbf{X}_k$, With this, the likelihood to measure the set $\mathcal{Z}_k$ given both kinematic state and extension as well as the number of measurements, yields

$$p(\mathcal{Z}_k|n_k, x_k, X_k) = \prod \mathcal{N}(z_k; Hx_k, X_k). \tag{2.4}$$

Introducing the mean measurement and the measurement spread

$$\bar{z}_k = \frac{1}{n_k}\sum_{j=1}^{k} z_k^j, \bar{Z}_k = \sum_{j=1}^{n_k}(z_k^j - \bar{z}_k)(z_k^j - \bar{z}_k)^\top, \tag{2.5}$$

the equation 2.4 can be written as

$$p(\mathbf{Z}_k|n_k, x_k, X_k) \propto \mathcal{N}(\bar{z}_k; Hx_k, X_k/n_k) \times \mathcal{W}(\bar{Z}_k; n_k - 1, X_k), \tag{2.6}$$

where

$$\mathcal{W}(X; v, V) = \frac{|X|^{\frac{v-d-1}{2}}}{2^{\frac{vd}{2}}\Gamma_d(\frac{v}{2})|V|^{\frac{v}{2}}}\text{etr}\left(-\frac{1}{2}XV^{-1}\right), \tag{2.7}$$

with $m \geq d$ denotes the Wishart density [26] of a d-dimensional SPD random matrix X with expected SPD matrix vV; etr(.) is an abbreviation for exp(tr(.)) and Γ_d(.)is the multivariate gamma function. In [25], the concept of conjugate priors is applied to this equation to obtain update equations this is complemented with an evaluation of the Chapman-Kolmogorov theorem obtaining a recursive Bayesian estimation cycle. Only the results are discussed, for more details see [25], the resulting estimator is based on specific products form of prior and posterior densities. The posterior is factored as

$$p(x_k, X_k|\mathcal{Z}_k) = p(x_k|X_k, \mathcal{Z}_k)p(X_k|\mathcal{Z}_k), \tag{2.8}$$

within this product, the matrix-variate density is given by

$$p(X_k|\mathcal{Z}_k) = \mathcal{IW}(X_k; v_{k|k}, \tilde{X}_{k|k}), \tag{2.9}$$

where the inverse Wishart density [26] is given by

$$\mathcal{IW}_d(\mathbf{X}_k; v, \mathbf{V}) = \frac{|\mathbf{V}|^{\frac{v}{2}}}{2^{\frac{vd}{2}}\Gamma_d\left(\frac{v}{2}\right)} \cdot |\mathbf{X}_k|^{-\frac{v+d+1}{2}} \cdot \text{etr}\left(-\frac{1}{2}\left(\mathbf{V}^{-1}\mathbf{X}_k\right)\right), \tag{2.10}$$

where d denotes the dimension of the random matrix $\mathbf{X}_k$, v represents the degrees of freedom parameter, $\mathbf{V}$ signifies the scale matrix parameter of size $d \times d$, $\Gamma_d(\cdot)$ stands for the multivariate gamma function, and $|\cdot|$ indicates the determinant. The mean of the inverse Wishart density is given by

$$\frac{1}{v-d-1}V, \tag{2.11}$$

for $v > d + 1$. The vector-variate density reads

$$p(x_k|X_k, \mathcal{Z}_k) = \mathcal{N}(x_k; x_{k|k}, \tilde{P}_{k|k} \otimes X_k), \tag{2.12}$$

herein $\otimes$ denotes the Kronecker product [27]. The prediction equations given in [25] are

$$\mathbf{x}_{k|k-1} = \mathbf{F}\mathbf{x}_{k-1|k-1}, \tag{2.13}$$

$$\tilde{\mathbf{P}}_{k|k-1} = \tilde{\mathbf{F}}\tilde{\mathbf{P}}_{k-1|k-1}\tilde{\mathbf{F}}^T + \tilde{\mathbf{Q}}, \tag{2.14}$$

$$\tilde{\mathbf{X}}_{k|k-1} = \left(\frac{v_{k|k-1} - d - 1}{v_{k-1|k-1} - d - 1}\right)\tilde{\mathbf{X}}_{k-1|k-1}, \tag{2.15}$$

$$v_{k|k-1} = \exp\left(-\frac{T}{\tau}\right) v_{k-1|k-1}. \tag{2.16}$$

Based on the given forms of prior and posterior densities, the update equations are given by

$$\mathbf{S}_{k|k-1} = \tilde{\mathbf{S}}_{k|k-1}\mathbf{X}_k, \tag{2.17}$$

$$\tilde{\mathbf{S}}_{k|k-1} = \mathbf{H}\tilde{\mathbf{P}}_{k|k-1}\mathbf{H}^T + \frac{1}{n_k}, \tag{2.18}$$

$$\mathbf{K}_{k|k-1} = \tilde{\mathbf{K}}_{k|k-1} \otimes \mathbf{I}_d, \tag{2.19}$$

$$\tilde{\mathbf{K}}_{k|k-1} = \tilde{\mathbf{P}}_{k|k-1}\mathbf{H}^T\tilde{\mathbf{S}}^{-1}_{k|k-1}. \tag{2.20}$$

For the kinematic parameters:

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + (\tilde{\mathbf{K}}_{k|k-1} \otimes \mathbf{I}_d)((\bar{\mathbf{y}}_k) - \mathbf{H}\mathbf{x}_{k|k-1}), \tag{2.21}$$

$$\tilde{\mathbf{P}}_{k|k} = \tilde{\mathbf{P}}_{k|k-1} - \tilde{\mathbf{K}}_{k|k-1}\mathbf{S}_{k|k-1}\tilde{\mathbf{K}}^T_{k|k-1}. \tag{2.22}$$

For the extension parameters:

$$\mathbf{v}_{k|k} = \mathbf{v}_{k|k-1} + n_k, \tag{2.23}$$

$$\tilde{\mathbf{X}}_{k|k} = \tilde{\mathbf{X}}_{k|k-1} + \tilde{\mathbf{S}}^{-1}_{k|k-1}\mathbf{N}_{k|k-1} + \bar{\mathbf{Y}}_k, \tag{2.24}$$

with

$$N_{k|k-1} = (\hat{z}_k - Hx_{k|k-1})(\hat{z}_k - Hx_{k|k-1})^\top. \tag{2.25}$$

In [18], the spatial distribution for a single measurement z_k, is modeled as a Gaussian distribution with additive Gaussian sensor noise with covariance $\mathbf{R}$, being normally distributed with variance $\beta\mathbf{X}_k + \mathbf{R}$ and thus

$$p(\mathbf{Z}_k|n_k, x_k, X_k) = \prod_{j=1}^{n_k}\mathcal{N}(z_k^j; \mathbf{H}x_k, \beta\mathbf{X}_k + \mathbf{R}), \tag{2.26}$$

where β is a scaling factor. Typically, β is set to 1/4 to match a uniform spatial measurement source distribution. In this approach, the shape matrix $\mathbf{X}_k$ is estimated as an Inverse Wishart density given in eq. 2.10. In [18] the update is performed by ignoring the uncertainty coming with the predicted estimate, according to

$$x_{k|k} = x_{k|k-1} + K_{k|k-1}(\bar{z}_k - Hx_{k|k-1}), \tag{2.27}$$

$$P_{k|k} = P_{k|k-1} - K_{k|k-1}S_{k|k}K^\top_{k|k-1}, \tag{2.28}$$

with

$$S_{k|k-1} = HP_{k|k-1}H^\top + \frac{Z_{k|k-1}}{n_k}, \tag{2.29}$$

being an approximation of the true innovation covariance.

$$K_{k|k-1} = P_{k|k-1}H^\top S^{-1}_{k|k-1}, \tag{2.30}$$

denoting the corresponding gain and

$$\mathbf{Z}_{k|k-1} = \beta X_{k|k} + R, \tag{2.31}$$

indicating the predicted variance of a single measurement. The updated extension estimate is given by

$$X_{k|k} = \frac{1}{\alpha_{k|k}}(\alpha_{k|k-1} X_{k|k-1} + \hat{N}_{k|k-1} + \hat{Y}_{k|k-1}), \tag{2.32}$$

$$\alpha_{k|k} = \alpha_{k|k-1} + n_k, \tag{2.33}$$

with

$$\hat{N}_{k|k-1} = X^{1/2}_{k|k-1} S^{-1/2}_{k|k-1} N_{k|k-1} (S^{-1/2}_{k|k-1})^\top (X^{1/2}_{k|k-1})^\top, \tag{2.34}$$

$$\hat{Z}_{k|k-1} = X^{1/2}_{k|k-1} Z^{-1/2}_{k|k-1} \bar{Z}_k (Z^{-1/2}_{k|k-1})^\top (X^{1/2}_{k|k-1})^\top, \tag{2.35}$$

$$\alpha_{k|k} = v_{k|k} - d - 1. \tag{2.36}$$

The GGIW [18], [28] filter is a Bayesian tracking technique designed for scenarios involving Poisson random matrix models. The gamma distribution is the conjugate prior for the unknown Poisson rate. It employs a conjugate prior distribution that combines a gamma distribution with a Gaussian distribution and an inverse Wishart distribution. The filter's posterior density $f_{k|k}(\xi)$ is defined as:

$$f_{k|k}(\xi) = \mathcal{G}(\gamma_k; \alpha_{k|k}, \beta_{k|k}) \mathcal{N}(x_k; m_{k|k}, P_{k|k}) \times \mathrm{IW}_d(X_k; v_{k|k}, V_{k|k}), \tag{2.37}$$

where: $\mathcal{G}(\gamma_k; \alpha_{k|k}, \beta_{k|k})$ is a gamma distribution with shape $\alpha_{k|k}$ and rate $\beta_{k|k}$, $\mathcal{N}(x_k; m_{k|k}, P_{k|k})$ represents a Gaussian distribution with mean $m_{k|k}$ and covariance $P_{k|k}$, $\mathrm{IW}_d(X_k; v_{k|k}, V_{k|k})$ denotes an inverse Wishart distribution with degrees of freedom $v_{k|k}$ and scale matrix $V_{k|k}$.

2.2 Random Finite Set Modelling

This section provides an overview of the application of various types of RFSs in multiple extended object tracking. Specifically, three key types of RFSs are highlighted: the PPP, the Bernoulli process, and the multi-Bernoulli process. While only brief explanations are provided here, for further details, please refer to, e.g. [29, 30].

2.2.1 Poisson Point Process

A PPP is a type of RFS characterized by a Poisson-distributed cardinality, where each object within the set is independent and identically distributed (i.i.d.). The PPP is defined by a single parameter known as the intensity function, denoted as $D(x)$. This intensity function can be decomposed into two components: the Poisson rate μ and the spatial distribution $f(x)$, such that $D(x) = \mu f(x)$.

The Poisson rate μ represents the average number of objects per unit area, while the spatial distribution $f(x)$ describes the spatial arrangement of the objects within the region of interest. Higher intensity values imply a greater likelihood of finding objects in a specific area, while lower intensities indicate a lower probability of object presence.

In the context of multiple extended object tracking, PPPs find widespread application in modeling various scenarios, including:

1. False alarm detections: PPPs are commonly employed to model spurious detections or false alarms generated by the sensor.
2. Extended object detections: These are used to represent detections caused by actual extended objects within the scene.
3. Object birth: PPPs can also model the process by which new objects enter the sensor's field of view.

Furthermore, in the PMBM model, undetected objects are explicitly modeled as PPPs [12, 24, 31]. This modeling approach allows for a comprehensive representation of both detected and undetected objects within the tracking framework.

2.2.2 Bernoulli Process

A Bernoulli RFS X is characterized by the property that it is either empty with probability $1-r$ or, with probability r, contains a single element x with distribution $f(x)$. In simpler terms, the cardinality of the set follows a Bernoulli distribution with parameter r. The Bernoulli distribution is fully determined by the probability of existence r and the state density $f(x)$.

In the context of multiple object tracking, Bernoulli RFSs play a crucial role in modeling detected objects. They effectively capture the inherent uncertainties associated with object tracking tasks. The uncertainty arises from the ambiguity in determining whether an estimate corresponds to a genuine object. This ambiguity is quantified by the probability of existence r.

When an object does exist, its state x remains unknown, and the uncertainty regarding the object's state is encapsulated by the state density $f(x)$. This means that even when an object is detected, there can be uncertainty regarding its exact properties or characteristics, such as its position, velocity, or other attributes.

By utilizing Bernoulli RFSs, multiple object tracking algorithms can effectively represent and manage the uncertainties inherent in object detection and tracking processes.

2.2.3 Multi-Bernoulli Process

In the context of multiple object tracking, it is often assumed that objects are independent entities. A multi-Bernoulli (MB) RFS X represents this assumption by being the composite of independent Bernoulli RFSs X_i, denoted as $X = \cup_{i \in I} X_i$, where I is the index set for the

Bernoulli components of the MB. Each component X_i corresponds to a single object, and the parameters describing the ith Bernoulli RFS are r_i and f_i.

The MB distribution is fully characterized by the parameters $\{r_i, f_i\}_{i \in I}$, where r_i represents the probability of existence and f_i represents the state density for the ith object. Notably, $|I|$ defines the maximum number of objects that the MB RFS can effectively represent.

In the domain of multiple object tracking, the MB RFS serves as a suitable model for scenarios involving multiple detected objects. It efficiently captures the independence among objects and allows for the representation of uncertainty associated with the existence and state of each individual object.

2.2.4 Multi-Bernoulli Mixture

A Multi-Bernoulli mixture (MBM) is a RFS whose multi-object density is a normalized weighted sum of MB densities, forming a mixture of MB densities. The MBM density is fully characterized by the set of parameters $\{(W_j, \{r_{j,i}, f_{j,i}\}_{i \in I_j})\}_{j \in J}$, where J denotes the index set for the MBs in the MBM, also referred to as components of the MBM. Each component j has a weight W_j, representing the probability of the jth MB, and an index set I_j for the Bernoulli components within the jth MB. Additionally, the parameters of the ith Bernoulli within the jth MB are denoted as $r_{j,i}$ and $f_{j,i}$.

In the context of multiple object tracking, each MB $j \in J$ corresponds to a unique global multi-object hypothesis. A global multi-object hypothesis encompasses a specific sequence of data associations for an entire sequence of measurement sets, involving one association for each time step from the initial time step to the current time step. The weight W_j corresponds to the probability of the associated sequence of associations.

2.2.5 Standard Extended Object Measurement Model

In the standard extended object measurement model, the set of measurements obtained at time step k is the union of object-generated measurements and false alarm measurements, given by $Z = \bigcup_i W_i \cup \mathcal{E}$, where $\mathcal{E}$ represents the set of false alarm measurements, and W_i represents the set of measurements from the ith object. The standard Multiple Target Tracking (MTT) assumptions assert that the sets are all independent, and the measurement origin is unknown. In other words, for the measurement set Z, it is not known which measurements are false alarms, nor is it known which measurements originated from which object.

The standard choice in multiple object tracking is to model the set of false alarm measurements $\mathcal{E}$ as a Poisson Point Process (PPP) with rate λ and spatial distribution $c(z)$, with the false alarm PPP intensity given by $\kappa(z) = \lambda c(z)$. A model for the detections W_i caused by an object with state x_i must account for both the number of measurements per object and the distribution of each measurement. A discussion of alternative single extended object measurement models is presented in [14]; the non-homogeneous Poisson Point Process (PPP) model [32] is the standard choice, largely due to its versatility and relative simplicity. In this model, individual measurements are spatially distributed around the object. Specifically, an extended object with state x_i is detected with a state-dependent probability of detection $p_D(x_i)$, and if detected, the object measurement set W_i is modeled as a PPP with a state-dependent Poisson rate $\gamma(x_i)$ and spatial distribution $\varphi(\cdot|x_i)$. Thus, conditioned on the state x_i, the PPP intensity is $\gamma(x_i)\varphi(\cdot|x_i)$.

For a non-empty set of measurements ($|W_i| > 0$), the conditional extended object measurement set likelihood is denoted as

$$\mathcal{L}(W_i|x_i) = p_D(x_i)e^{-\gamma(x_i)} \prod_{z \in W_i} \gamma(x_i)\varphi(z|x_i).$$

Note that this likelihood is the product of the probability of detection and the PPP density. The effective probability of detection for an extended object with state x_i is $p_D(x_i)(1 - e^{-\gamma(x_i)})$, where $1 - e^{-\gamma(x_i)}$ is the Poisson probability of generating at least one detection. Accordingly, the effective probability of missed detection, i.e., the probability that the object is not detected, is $q_D(x_i) = 1 - p_D(x_i) + p_D(x_i)e^{-\gamma(x_i)}$.

In summary, the standard extended object measurement model assumes that both the set of false alarm detections and the set of measurements from an object follow Poisson Point Process distributions.

2.2.6 The Poisson Multi-Bernoulli Mixture filter

The PMBM (Poisson Multi-Bernoulli Mixture) conjugate prior is a modeling approach for the multiple object tracking problem, initially developed for extended objects in [24,31], and for point objects in [12]. In this model, the set of objects X is partitioned into two disjoint subsets, $X = X_d \cup X_u$. The first subset, X_d, consists of objects detected by the sensor in at least one scan. The second subset, X_u, comprises undetected objects, i.e., objects that could be within the sensor's field of view but have not been detected.

In the PMBM model, the set of undetected objects is modeled as PPP (Poisson Point Process) distributed, while the set of detected objects is modeled as MBM (Multi-Bernoulli Mixture) distributed. Hence, this model is referred to as PMBM. The PMBM extended object filter [24], [31] estimates the PMBM multi-object density, which is defined entirely by the parameters $D_u, \{W_j, \{r_{j,i}, f_{j,i}\}_{i \in I_j}\}_{j \in J}$, where $D_u(\cdot)$ is the PPP intensity for the set of undetected objects X_u, and $\{(W_j, \{r_{j,i}, f_{j,i}\}_{i \in I_j})\}_{j \in J}$ are the MBM parameters for the set of detected objects X_d.

If the measurement model is of the standard form, and the predicted multi-object density is a PMBM density, then the posterior multi-object density is also a PMBM density [31, Th. 1]. For further details, please refer to [24], [31].

Typically, multiple object tracking algorithms only model the set of detected objects and do not explicitly model the set of undetected objects. However, modeling the set of undetected objects is useful, especially when dealing with occlusions, where an object may not be detected despite being within the sensor's field of view. In multiple extended object tracking, occlusions are often modeled via a non-homogeneous probability of detection $p_D(\cdot)$ [33], [21], [34], [35]. Integrating a non-homogeneous probability of detection into the PMBM filter leads to the undetected object intensity $D_u(\cdot)$ being higher in occluded parts of the sensor's field of view and lower in parts that are not occluded. This reflects the fact that an undetected object is more likely to be located in an occluded area than in a visible area [9].

2.3 The Challenge of Data Association in MEOT and its complexity

The case of multiple extended targets introduces a departure from the single target-single measurement assumption, allowing for measurements to be part of the same object. In the context outlined in [7], the number of possible associations, considering a set of previously detected objects $\mathcal{I}$ with $|\mathcal{I}|$ elements and a set of measurements $\mathcal{Z}$ with $|\mathcal{Z}|$ elements, is determined by the number of set partitions of $\mathcal{I} \cup \mathcal{Z}$. The number of cell-to-object assignments is given by:

$$N_{\mathcal{A}}(|\mathcal{Z}|, |\mathcal{I}|) = \sum_{C=1}^{|\mathcal{Z}|} \left[\left\{ {|\mathcal{Z}| \atop C} \right\} \sum_{T=0}^{\min(C,|\mathcal{I}|)} \binom{C}{T} \binom{|\mathcal{I}|}{T} T! \right]. \tag{2.38}$$

Here, $\left\{ {|\mathcal{Z}| \atop C} \right\}$ denotes the Stirling number of the second kind[1]. It represents the number of different possibilities to partition a set $\mathcal{Z}$ with $|\mathcal{Z}|$ elements into C cells. $\binom{C}{T}$ is the binomial coefficient, given by $\binom{C}{T} = \frac{C!}{(C-T)!T!}$, which represents the number of subsets with cardinality T that can be formed with C objects. Equation 2.38 demonstrates that for even small values of $|\mathcal{I}|$ and $|\mathcal{Z}|$, $N_{\mathcal{A}}(|\mathcal{Z}|, |\mathcal{I}|)$ becomes extremely large, rendering many MEOT algorithms intractable.

Consider the example in fig. 2.1 with $|\mathcal{Z}| = 15$ measurements and $|\mathcal{I}| = 2$ object predictions. In this case there would be 72'384'727'657 possible data associations: With the 3D Segmentation Algorithm it would be possible to drastically reduce the amount of possible

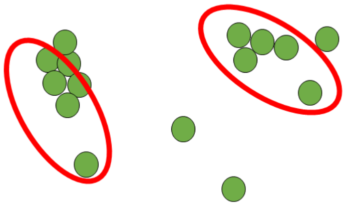


Figure 2.1: Example for data association with $|\mathcal{Z}| = 15$ measurements in green and $|\mathcal{I}| = 2$ predictions in red.

[1] $\left\{ {|\mathcal{Z}| \atop C} \right\} = \frac{1}{C!} \sum_{j=0}^{C} (-1)^{C-j} \binom{C}{j} j^{|\mathcal{Z}|}$

cell-to-object assignments by grouping points in the same detection with segmentation and thus model the same scenario with less parameters.

2.3.1 Extended PMBM filter

For a predicted PMBM, indexed by $\mathcal{J}$ the total number of possible associations is given by

$$N_A = \sum_{j \in \mathcal{J}} N_{A_j}(|\mathcal{Z}|, |\mathcal{I}_j|). \tag{2.39}$$

In [36] it is shown that, for the PMBM filter, for multiple extended target filtering considering the j-th predicted MB with Bernoulli components indexed by $\mathcal{I}_j$, and a set of measurements $\mathcal{Z}$, the complexity of the update operation is between exponential ($O(2^{|\mathcal{Z}|+|\mathcal{I}_j|})$) and factorial ($O((|\mathcal{Z}| + |\mathcal{I}_j|)!)$). A simple example demonstrates this high complexity. Let the PMBM filter be initialized at time $k = 0$ with $J_0 = \{j_1\}$, $W^0_{j_1} = 1$, and $\mathcal{I}^0_{j_1} = \emptyset$, i.e., an empty MBM. This corresponds to zero previously detected targets at initialization. Given a measurement set $\mathcal{Z}_1$ at time $k = 1$, the number of MB components in the updated PMBM density is given by the number of associations,

$$|J_1| = N_{Aj_1}(|\mathcal{Z}_1|, 0) = \sum_{C=1}^{|\mathcal{Z}_1|} \binom{|\mathcal{Z}_1|}{C} = B(|\mathcal{Z}_1|), \tag{2.40}$$

where B denotes the Bell number. The number of MBM components, given measurement sets up to and including time step k and an empty initial MBM, is given by the Bell number whose order n is the sum of the measurement set cardinality:

$$|J_k|^k = |J_{k+1}|^k = B\left(\sum_{t=1}^{k} |\mathcal{Z}_t|\right) = B\left(|\mathcal{Z}|^k\right). \tag{2.41}$$

The sequence of Bell numbers $B(n)$ is log-convex, and $B(n)$ grows very rapidly. For illustration the following table show the first values of the sequence:

Table 2.1: Values of the first ten Bell numbers.

B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
1	2	5	15	52	203	877	4140	21147	115975

In articles such as [13] and [36], gating, clustering and ranking of the association events are used to reduce the number of data associations. After the PMBM update, techniques including pruning, merging, and recycling are used to reduce the number of components. However when a significant quantity of data is processed, filtering becomes slow due to the complexity of these techniques. For example, for the Density-Based Spatial Clustering of Applications with Noise (DB-SCAN) clustering algorithm, [11] is used and its overall complexity is $\mathcal{O}(|\mathcal{Z}|^2)$ in the worst case. In this article, this is avoided by providing clustered data preprocessed by the 3D-INSEG algorithm [19].

2.4 Instance Segmentation

Instance segmentation is a computer vision task that aims to identify and differentiate individual object instances within an image. The primary goal of instance segmentation is to produce a pixel-wise segmentation map of the image, where each pixel is assigned to a specific object instance. AI approaches, as discussed in [37], have demonstrated good results. While these techniques allow us to locate objects in a 2D image space, in this work we perform instance segmentation in 3D using stereo vision. Therefore in this work, we address this challenge by using Mask R-CNN [1], a powerful architecture that excels in precisely delineating object boundaries at the pixel level.

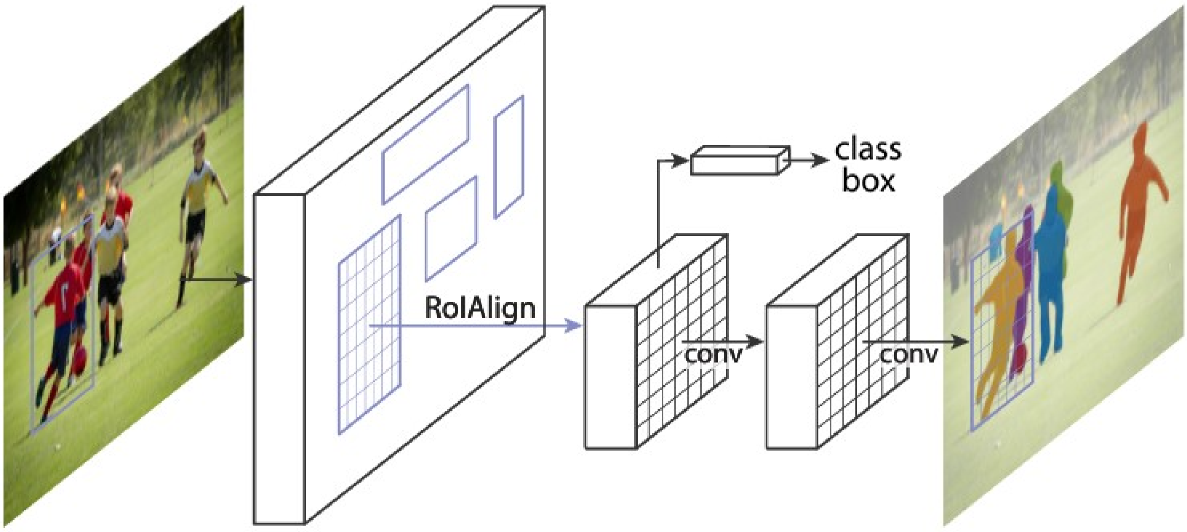


Figure 2.2: The MaskR-CNN framework for instance segmentation. Extracted from [1]

2.5 Stereo Matching

Stereo matching, or disparity estimation, is a fundamental technology in computer vision used to reconstruct 3D structures from 2D images of the real world. It finds extensive application in areas like autonomous driving, augmented reality, and robotics navigation. In stereo matching, given a pair of rectified stereo images, the objective is to calculate the disparity for every pixel in the reference image. Disparity refers to the horizontal displacement between corresponding pixels in the left and right images. In this work we use the Multilevel Recurrent Field Transforms for Stereo Matching (RAFT-Stereo) [2] algorithm that combines a feature encoder and a context encoder to extract features from input images. These features are used to create a correlation pyramid, generating a 3D correlation volume through efficient dot product computations. This technology enables us to derive disparity maps from pairs of images, which can then be processed to obtain depth information. While the RAFT-Stereo algorithm alone does not provide objects detections, it offers depth information, valuable in tracking applications.

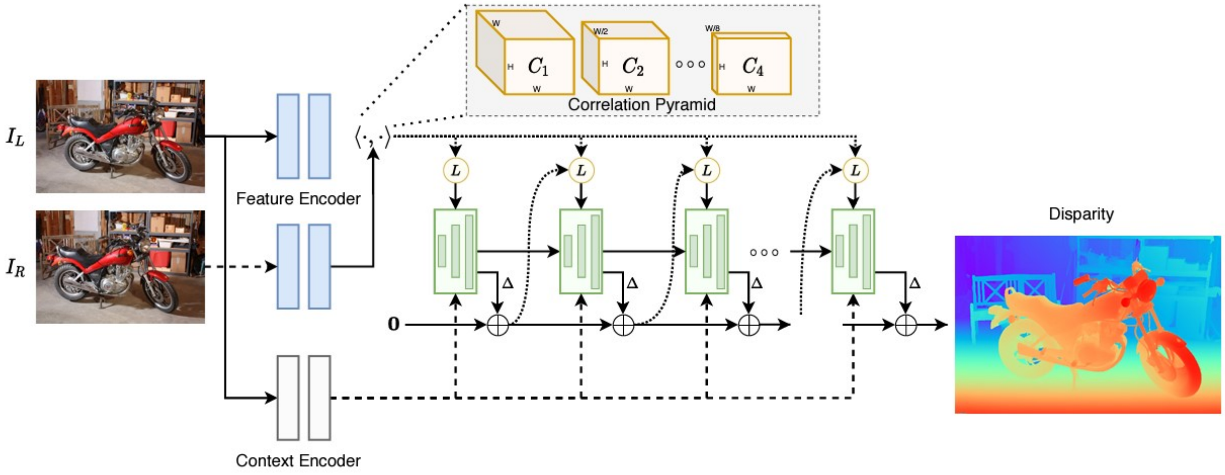


Figure 2.3: The RAFT-Stero framework for stereo matching extracted from [2]

Chapter 3

The 3D Instance Segmentation Algorithm

3.1 Introduction

The 3D-INSEG Algorithm is designed to address the challenges of object detection and tracking in environments with dense and extensive laser data. By integrating stereo vision and NNs for 3D depth estimation and segmentation, this algorithm aims to achieve robust 3D instance segmentation within both SEOT and MEOT contexts.

The algorithm consists of several key components, as shown in Fig. 3.1. First, it includes image undistortion to correct lens distortions and ensure accurate geometry and measurements. Next, depth estimation is performed using the RAFT-Stereo algorithm to obtain the disparity map from stereo image pairs. Following depth estimation, instance segmentation identifies and labels individual objects within the images, providing binary masks for each detected object, these masks are eroded to avoid depth discontinuities in detected objects' border.

Finally, the 3D Projection component transforms 2D inferences into 3D Cartesian space, aligning segmented instances with real-world coordinates based on depth information. The result is a set of clustered 3D detections (D), where each cluster (C) represents a distinct object or entity.

The use of clusterized data as measurements in tracking simplifies the data association process, improves tracking accuracy, and reduces computational complexity. This chapter will delve into each component of the 3D-INSEG Algorithm, highlighting its contributions and benefits in object detection and tracking tasks.

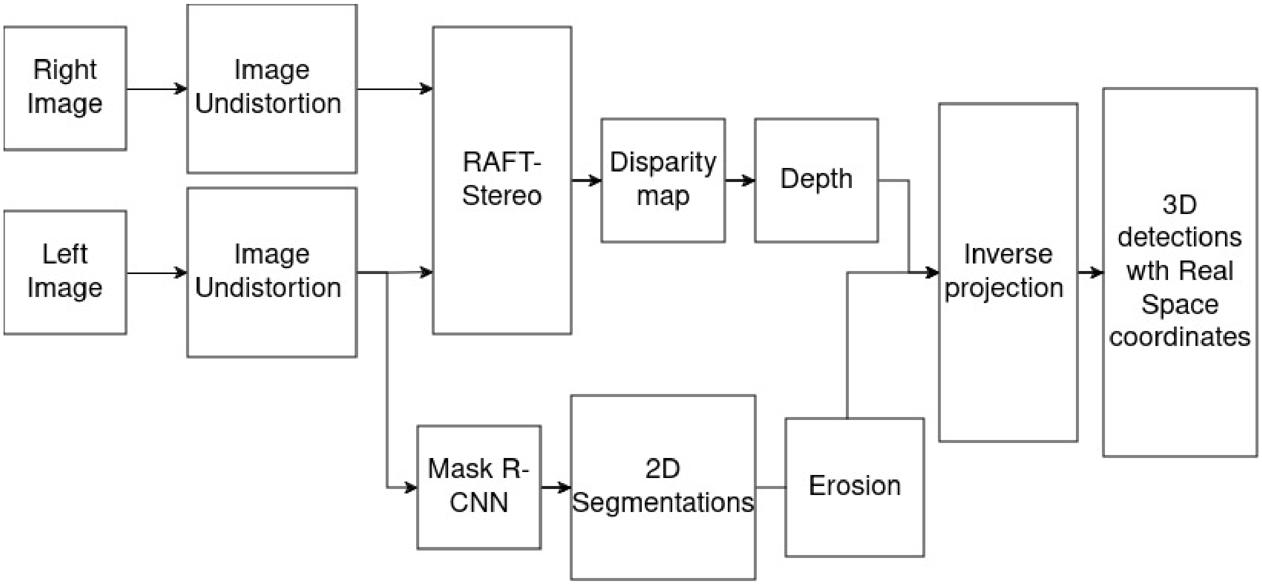


Figure 3.1: The 3D-INSEG algorithm diagram.

3.2 Image Undistortion

Undistorting images is essential for this task as we aim to project objects into 3D real space coordinates. It corrects lens distortions, ensuring accurate geometry and reliable measurements.

The distortion model uses the camera Matrix K. The camera matrix represents the intrinsic properties of the camera and is determined during calibration. It includes parameters such as the focal length, and principal point. The camera matrix is defined as in (3.1).

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.1}$$

where f_x and f_y are the focal lengths along the x and y axes with the camera coordinate system, respectively, and c_x and c_y are the coordinates of principal points, representing the optical center of the image. The third row is fixed as $0, 0, 1$ to maintain homogeneity in the matrix. The parameters are given in Appendix A.

The distortion correction process involves the following steps:

1. Capture an image using the camera. Denote the distorted image as $I_{\text{distorted}}$.
2. Use the camera matrix (K) to compute the undistorted and rectified image coordinates for each pixel in the original distorted image. This is carried out using the following equations:

$$X_u = \frac{X_d - c_x}{f_x}, \tag{3.2}$$

$$Y_u = \frac{Y_d - c_y}{f_y}, \tag{3.3}$$

where: (X_d, Y_d) are the distorted image coordinates. (X_u, Y_u) are the undistorted and rectified image coordinates.

3. Interpolate the pixel values from the original distorted image ($I_{\text{distorted}}$) to the undistorted and rectified image coordinates ((X_u, Y_u)). This step ensures that no information is lost during the correction. The result is a pair of rectified images ($I_{\text{rectified-left}}$ and $I_{\text{rectified-right}}$) that accurately represent the scene geometry.

The rectified images ($I_{\text{rectified-left}}$ and $I_{\text{rectified-right}}$) are ready for use in Depth Estimation and Instance Segmentation.

3.3 Depth Estimation Using the RAFT-Stereo algorithm

We employ the RAFT-Stereo [2] architecture to obtain the disparity map of two images. The depth z from the disparity map is

$$z = \frac{Bf}{d}, \tag{3.4}$$

where B is the baseline distance between the two cameras, f is the focal length of the cameras and d is the disparity value for a corresponding point in the disparity map.

3.4 Instance Segmentation

Instance segmentation identifies and labels individual objects or instances within the images, distinguishing them from the background. A mask is obtained for each detected object from the left image. Modelled as a binary matrix M that indicates if a pixel is part of an object:

$$M(i,j) = \begin{cases} 1, & \text{if pixel } (i,j) \text{ is part of the object} \\ 0, & \text{otherwise.} \end{cases} \tag{3.5}$$

Here, the binary matrix is obtained during instance segmentation, where each entry indicates whether the corresponding pixel belongs to the detected object. To obtain a set of object points from M:

$$\mathcal{M}_{\text{object}} = \{(i,j) \mid M(i,j) = 1\}. \tag{3.6}$$

$\mathcal{M}_{\text{object}}$, consists of pixel coordinates that belong to the detected object.

$$\mathcal{M}_{\text{object}} = \{p_1, p_2, \ldots, p_n\}. \tag{3.7}$$

3.5 Erosion

The erosion of a binary image A by a structuring element B, denoted as $A \ominus B$, is mathematically defined as:

$$A \ominus B = \{z \mid (B)_z \subseteq A\}, \tag{3.8}$$

where$(B)_z$ represents the translation of the structuring element B by the vector z. $B \subseteq A$ indicates that all points of B lie within A. $A \ominus B$ denotes the set of all points z such that $(B)_z$ is entirely contained within A. In practical terms, this operation involves sliding the structuring element B over the binary image A, and for each position, checking if B fits entirely within A. This process helps in smoothing object boundaries and refining object masks for efficient processing and analysis of images. By applying erosion before depth estimation, we create smoother transitions in depth values at object borders, reducing discontinuities between objects and the background.

3.6 3D Projection: Transforming 2D Inferences to 3D Cartesian Space

Depth estimation determines the depth information for each pixel, providing a 3D representation of the scene. Inverse projection maps the segmented instances and their associated depth information back into the 3D space, aligning them with the real-world coordinates. By having the depth of each pixel and a set of masks, each mask is projected into 3D space. Given the pixel coordinates (u, v) of a 2D inference and its corresponding depth value (z), the transformation from camera to world coordinates is:

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \text{depth} \times \text{inv}(K) \times \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{3.9}$$

where:

- $[x \quad y \quad z]^\top$ represents the 3D Cartesian coordinates of the projected point,
- $[x_c \quad y_c \quad z_c]^\top$ represents the 3D homogeneous coordinates in camera coordinates,
- $\text{inv}(K)$ denotes the inverse of the camera matrix K.

For each pair of images in a sequence of stereo images, a set of detections, denoted as $\mathcal{D} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$, is generated. Here, each set $\mathcal{C}$ represents a cluster of 3D points corresponding to the detected object. Algorithm 1 summarizes the 3D-INSEG algorithm where b is the baseline and f is the focal length of the stereo cameras.

In the following chapters the set of detections will be used as measurements in tracking, simplifying the data association process by having clusterized data. The purpose of generating these detections is to simplify the data association process in subsequent tracking

Algorithm 1 3D-INSEG

```
Input: $I_L$ (Left Image), $I_R$ (Right Image)
Output: $\mathcal{D} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$
  1. 2d_masks ← MaskRCNN($I_L$)
  2. disp_map ← RAFT_Stereo($I_L, I_R$)
  3. For each 2d_mask in 2d_masks
       • Initialize cluster: $\mathcal{C} = \{\}$
       • For each pixel = $(u, v)$ in 2d_mask:
           – $d$ = disp_map($u, v$)
           – depth = $\frac{bf}{d}$
           – $(x, y, z)$ = 3dProjection(K,depth,u,v)
           – $\mathcal{C} := \mathcal{C} \cup \{(x, y, z)\}$
       • $\mathcal{D} := \mathcal{D} \cup \mathcal{C}$
Return $\mathcal{D}$
```

tasks. By clustering the detected points into sets $\mathcal{C}$, we organize the data such that each cluster represents a distinct object or entity within the environment. This clustering helps in reducing the complexity of data association, as measurements ($\mathcal{D}$) are already grouped based on their spatial proximity and similarity in characteristics. The use of clusterized data as measurements in tracking facilitates the following benefits:

- Simplification of Data Association: Clustering aggregates 3D points into coherent object representations, making it easier to associate these clusters with corresponding objects across frames.
- Improved Tracking Accuracy: By working with pre-clustered detections, the tracking algorithm can focus on matching entire object clusters rather than individual points, leading to more robust and accurate tracking results.
- Reduction of Computational Complexity: Grouping detections into clusters reduces the number of individual data points that need to be considered during data association, thereby improving computational efficiency.

Chapter 4

3D Instance Segmentation for Single Extended Object Tracking

In this chapter the SEOT algorithm presented in [18] is implemented to demonstrate the usefulness of the 3D-INSEG algorithm in SEOT. For extended objects, the state of an object x_k can be defined as

$$x_k = [\gamma_k^\top, m_k^\top]^\top, \tag{4.1}$$

where parameter vector γ_k describes the shape, orientation, and size of the object at discrete time k and m_k describes the kinematic state at time k.

An elliptical object extent approach is presented in [18], based on a shape matrix:

$$\mathbf{X}_k = \begin{bmatrix} \cos\alpha_k & -\sin\alpha_k \\ \sin\alpha_k & \cos\alpha_k \end{bmatrix} \begin{bmatrix} (l_k^1)^2 & 0 \\ 0 & (l_k^2)^2 \end{bmatrix} \begin{bmatrix} \cos\alpha_k & -\sin\alpha_k \\ \sin\alpha_k & \cos\alpha_k \end{bmatrix}^\top, \tag{4.2}$$

where α_k is the (counter-clockwise) orientation and l_k^1 and l_k^2 denote the semi-axis lengths of the ellipse at discrete time k. This matrix implicitly encodes the shape parameters of γ_k in (4.1). In this approach, it is assumed that measurements are scattered across the entire surface of the ellipse following a spatial Gaussian distribution. The distribution is modeled as $p(y_k|m_k, \mathbf{X}_k) = N(y_k; \mathbf{H}m_k, \beta\mathbf{X}_k)$, where $\mathbf{H} = [\mathbf{I}_2, \mathbf{0}_2]$, the measurement source y_k is modeled as a Gaussian distributed with mean $\mathbf{H}m_k$ and covariance matrix $\beta\mathbf{X}_k$, where β is a scaling factor. Typically, β is set to 1/4 to match a uniform spatial measurement source distribution.

Additionally the spatial distribution for a single measurement z_k, $\phi(z|m, X)$, is modeled as a Gaussian distribution with additive Gaussian sensor noise with covariance $\mathbf{R}$, i.e. $\phi(z_k|m_k, \mathbf{X}_k) = N(z_k; \mathbf{H}m_k, \beta\mathbf{X}_k + \mathbf{R})$.

In this approach, the shape matrix $\mathbf{X}_k$ is modelled as an Inverse Wishart density [26] $IW_d(\mathbf{X}_k; v, \mathbf{V})$, where v is the degree of freedom and $\mathbf{V}$ is a symmetric positive definite matrix in $\mathbb{R}^{2\times 2}$. The update equations are shown in Algorithm 2, where m_- is the mean of the kinematic state and $\mathbf{P}_-$ its covariance matrix, $\mathbf{V}_-$ is the extent estimate and v_- its degree of freedom, m_+ is the is the posterior mean of the kinematic state and $\mathbf{P}_+$ its covariance matrix, $\mathbf{V}_+$ is the posterior extent estimate and v_+ its degree of freedom, τ denotes some time constant related to the agility with which the object may change its extension over time and T the prediction time interval.

Algorithm 2 Update

Input: Prior density specified by m_-, $\mathbf{P}_-$ and v_-, $\mathbf{V}_-$, set of detections $\mathcal{W}$, $n = |\mathcal{W}|$, measurement noise covariance $\mathbf{R}$.
Output: Posterior density parameterized by m_+, $\mathbf{P}_+$, v_+, $\mathbf{V}_+$.

```
    $\bar{z} = \frac{1}{n}\sum_{z_i \in W} z_i$
    $\varepsilon = \bar{z} - \mathbf{H}m_-$
    $\mathbf{Z} = \sum_{z^i \in W}(z^i - \bar{z})(z^i - \bar{z})^\top$
    $\hat{\mathbf{X}} = \mathbf{V}_-(v_- - 3)^{-1}$
    $\mathbf{Y} = z\hat{\mathbf{X}} + \mathbf{R}$
    $\mathbf{S} = \mathbf{H}\mathbf{P}_-\mathbf{H}^\top + \frac{\mathbf{Y}}{n}$
    $\mathbf{K} = \mathbf{P}_-\mathbf{H}^\top\mathbf{S}^{-2}$
    $\hat{\mathbf{N}} = \hat{\mathbf{X}}^{1/2}\mathbf{S}^{-1/2}\varepsilon\varepsilon^\top\mathbf{S}^{-\top/2}\hat{\mathbf{X}}^{\top/2}$
    $\hat{\mathbf{Z}} = \hat{\mathbf{X}}^{1/2}\mathbf{Y}^{-1/2}\mathbf{Z}\mathbf{Y}^{-\top/2}\hat{\mathbf{X}}^{\top/2}$
    $\alpha_- = v_- - 3$
    $\alpha_+ = 2 + \exp(-T/\tau)(\alpha_- - 2) + n$
    $m_+ = m_- + \mathbf{K}\varepsilon$
    $\mathbf{P}_+ = \mathbf{P}_- - \mathbf{K}\mathbf{S}\mathbf{K}^\top$
    $v_+ = \alpha_+ + 3$
    $\mathbf{V}_+ = \frac{1}{\alpha_+}(\alpha_-\mathbf{V}_- + \hat{\mathbf{N}} + \hat{\mathbf{Z}})$
```

4.1 Implementation

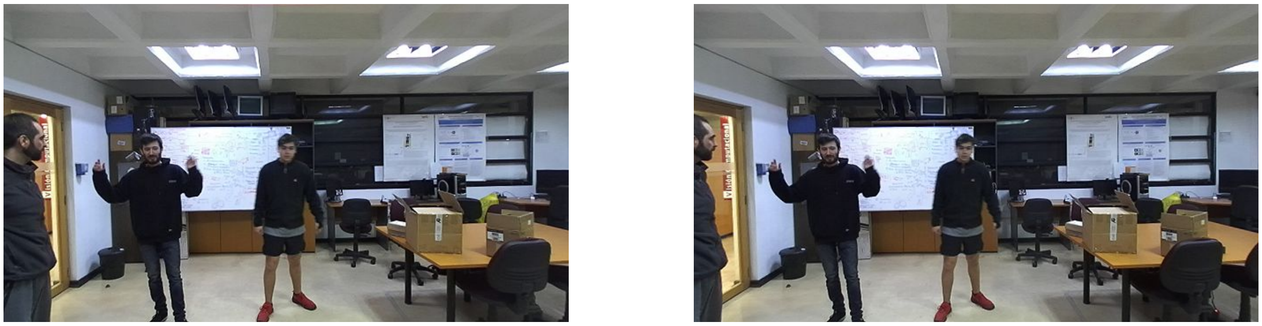
Algorithm 1 summarizes the 3D segmentation algorithm.

The algorithmic form of the single object tracking update is given in Algorithm 2 and the algorithm for the random matrix approach using the 3D-INSEG algorithm for SEOT is explained in Algorithm 3.

```
Algorithm 3 SEOT
Input: Set of labeled 3D Point Measurements $\mathcal{D}$, Initial State Estimate (StateEstimate)
Output: Updated State Estimate (UpdatedStateEstimate)
  1. Extract the person measurements $W = \mathcal{D}(person)$
  2. UpdatedStateEstimate = StateEstimate.
  3. For each time step:
      (a) Predict(StateEstimate)
      (b) Update(PredictedStateEstimate,$W$)
      (c) Store UpdatedStateEstimate
Return UpdatedStateEstimate.
```

4.1.1 Demonstration of the 3D-INSEG algorithm

Two input images are shown in Figs. 4.1a and 4.1b, serving as the foundation for deriving the disparity map shown in Fig. 4.2 (warm colors represent objects closer to the camera, while cold colors represent objects farther away). Subsequent segmentation in Fig. 4.3 results in the detection of three persons and two chairs, with each pixel belonging to the same object assigned the same color. These combined outcomes are then projected into 3D coordinates, resulting in 3D segmentation. This 3D segmentation is illustrated in Fig. 4.4 alongside the LIDAR Pointcloud, where blue dots represent LIDAR points, and colored regions represent 3D segmentation instances of different objects in the scene (in this case, two chairs in purple and yellow, and three persons in green, blue, and red).



(a) Demonstration of the 3D-INSEG algorithm : Left Image.

(b) Demonstration of the 3D-INSEG algorithm: Right Image.

Figure 4.1: Demonstration of the 3D-INSEG algorithm: Stereo pair of images.

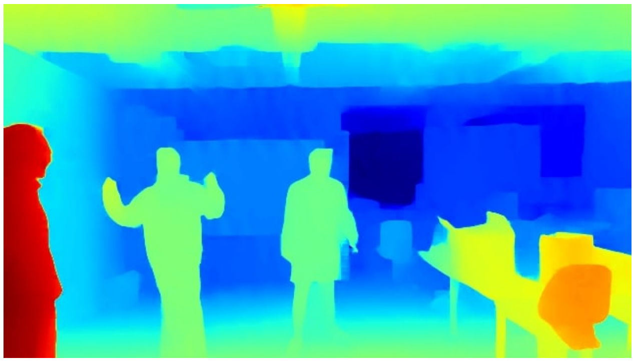


Figure 4.2: Demonstration of the 3D-INSEG algorithm Disparity Map.

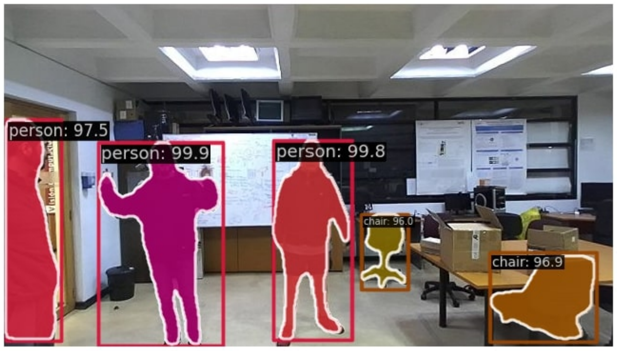


Figure 4.3: Demonstration of the 3D-INSEG algorithm: Segmentation.

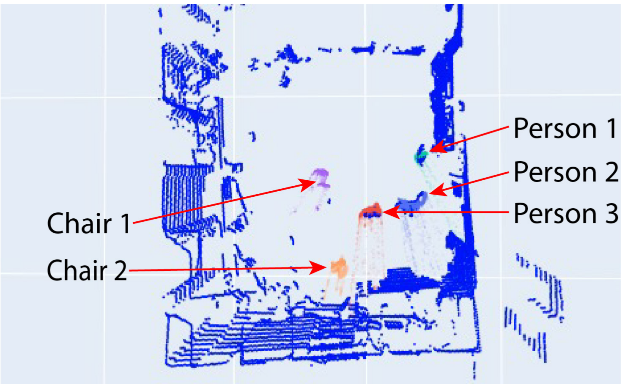


Figure 4.4: Demonstration of the 3D-INSEG algorithm: 3D Segmentation with LiDAR Pointcloud data.

4.2 Benchmark Clustering Approach

To assess the performance of the visual 3D-INSEG algorithm for human detection, we compare it with a leading clustering algorithm based on 3D LIDAR data [38]. This approach evaluates candidate points at similar bearing angles across different elevations to gauge their consistency with a human subject based on range values. Subsequently, the DBSCAN algorithm generates a cluster of points presumed to belong to a human subject.

The algorithm 4 begins by partitioning the input 3D point cloud into clusters based on height, where each cluster represents points within a specific height range. These height-based clusters are then consolidated into general clusters ($\mathcal{C}$), which are considered potential instances of human subjects within the environment. This clustering-based detection method provides a benchmark against which the performance of the 3D-INSEG algorithm can be evaluated, particularly in scenarios where traditional LIDAR-based clustering approaches are employed for human detection.

The comparison between the visual 3D-INSEG algorithm and this benchmark clustering approach offers insights into the effectiveness and advantages of leveraging stereo vision and neural networks for human detection in complex environments. Through this comparative analysis, we aim to identify the strengths and limitations of each approach, contributing to the advancement of state-of-the-art techniques in 3D instance segmentation and object detection.

Algorithm 4 3D human detector using clustering

```
Input: 3D Point Cloud
Output: Clusters $\mathcal{C} = \mathcal{C}_1, \ldots, \mathcal{C}_n$
  1. Initialize an empty list $height\_Clusters$
  2. For each height in Point Cloud
         height_Clusters.Add(Cluster(points,threshold))
  3. $\mathcal{C}$ =General_cluster(height_Cluster)
Return Clusters $\mathcal{C}$.
```

Chapter 5

3D Instance Segmentation for Multiple Extended Object Tracking

5.1 Introduction

The PMBM filter for MEOT is detailed in [13]. While the algorithm demonstrates satisfactory performance in simulations and when applied to 2D laser data characterized by relatively low clutter, its effectiveness diminishes in real-world scenarios with a high density of objects. In such situations, the filter's performance deteriorates, leading to inaccurate estimates due to the proximity of objects and the abundance of laser points within the surveillance area. This is where the 3D-INSEG proves valuable thanks to its capability to detect different objects and provide clusters. In this chapter the PMBM extended filter implementation with the GGIW defined in 2.1.1 for MEOT is implemented to demonstrate the usefulness of the 3D-INSEG algorithm in MEOT.

5.2 GGIW implementation for a single target

The state representation for a single extended target at time step k, denoted by ξ_k, comprises three components: a scalar γ_k, a vector x_k, and a matrix $\mathbf{X}_k$. The vector $x_k \in \mathbb{R}^{n_x}$ represents the kinematic state, encapsulating parameters related to the target's position and motion, such as velocity, acceleration, and turn-rate. The random matrix $\mathbf{X}_k \in \mathcal{S}_d^{++}$ characterizes the extent state, delineating the size and shape of the target, where d denotes the dimension of the extent (typically $d = 2$ or 3) and $\mathcal{S}_d^{++}$ denotes the set of symmetric positive definite matrices. The representation of an extended target's state at time step k, denoted by ξ_k, involves several random variables. The scalar variable $\gamma_k > 0$ serves as the Poisson rate in the measurement model. The likelihood of a single measurement z given the target state ξ_k is described by the Gaussian distribution:

$$\phi(z_k|\xi_k) = \mathcal{N}(z_k; \mathbf{H}_k x_k, \mathbf{X}_k), \tag{5.1}$$

where $\mathbf{H}_k$ represents the known measurement model. The single-target conjugate prior for the Poisson random matrix model is the gamma-Gaussian-inverse Wishart (GGIW) distribution:

$$f_{k|k}(\xi_k) = \mathrm{GGIW}(\xi_k; \zeta_{k|k}), \tag{5.2}$$

where $\zeta_{k|k} = (\alpha_{k|k}, \beta_{k|k}, m_{k|k}, \mathbf{P}_{k|k}, v_{k|k}, \mathbf{V}_{k|k})$ represents the set of GGIW density parameters and α represents the shape parameter of the gamma distribution, β the rate parameter of the gamma distribution, m the mean and $\mathbf{P}$ the covariance for the Gaussian distribution, v the number of degrees of freedom for the inverse Wishart distribution and $\mathbf{V}$ the shape matrix for the inverse Wishart distribution. The updated parameters $\zeta_{k|k}$ and the corresponding predicted likelihood for a GGIW distribution with prior parameters $\zeta_{k-1|k-1}$ that are updated with a set of measurements $\mathcal{Z}$ under the linear Gaussian model, are detailed in Algorithm 5. These parameters and their updates play a critical role in the measurement update within the random matrix extended target model see, e.g., [25], [18]. The motion models for the kinematic state, extent, and measurement rate are characterized as follows:

1. Kinematic State: The evolution of the kinematic state x_k from time step k to $k+1$ follows the model:

$$x_{k+1} = f(x_k) + w_k, \tag{5.3}$$

where w_k represents Gaussian process noise with zero mean and covariance $\mathbf{Q}$, and f is the state transition function.

2. Extent: The evolution of the extent state matrix $\mathbf{X}_k$ from time step k to $k+1$ is governed by the transformation:

$$\mathbf{X}_{k+1} = \mathbf{M}(x_k)\mathbf{X}_k\mathbf{M}(x_k)^\top, \tag{5.4}$$

where $\mathbf{M}(x_k)$ denotes a transformation matrix.

3. The measurement rate γ_{k+1} represents the expected number of measurements per target. It is assumed to remain constant and is equal to γ_k.

The predicted parameters $\zeta_{k+1|k}$ for a GGIW distribution with prior parameters $\zeta_{k|k}$, under these models, are detailed in Algorithm 6. For more extensive discussions regarding prediction within the random matrix extended target model see [25], [18].

5.3 GGIW-PMBM filter for MEOT

The GGIW-PMBM filter [24] operates through a recursive process, encompassing an update and a prediction phase. The update step involves integrating the GGIW-PMBM density parameters, comprising three main procedures: PPP update, MBM update, and the creation of new MB components from the PPP. The PPP update manages missed detections and incorporates new measurements associated with undetected targets, while the MBM update processes detected targets using extended target likelihood and data association probabilities. New MB components are created by converting PPP components associated with measurements into Bernoulli components. In the prediction phase, the filter anticipates future target behavior based on the current state and past observations. The PPP prediction

Algorithm 5 GGIW Update

Input: GGIW parameter ζ_+, set of measurements $\mathcal{Z}$, measurement model $\mathbf{H}$.
Output: Updated GGIW parameter ζ and predicted likelihood l:

$$\zeta = \begin{cases} \alpha = \alpha_+ + |\mathcal{Z}|, \\ \beta = \beta_+ + 1, \\ m = m_+ + \mathbf{K}\varepsilon, \\ \mathbf{P} = \mathbf{P}_+ - \mathbf{K}\mathbf{H}\mathbf{P}^+, \\ v = v_+ + |\mathcal{Z}|, \\ \mathbf{V} = \mathbf{V}_+ + \mathbf{N} + Z \end{cases}$$

where
$\bar{z} = \frac{1}{|\mathcal{Z}|}\sum_{z_i \in \mathcal{Z}} z_i$,
$Z = \sum_{z_i \in \mathcal{Z}}(z_i - \bar{z})(z_i - \bar{z})^\top$,
$\hat{\mathbf{X}} = \mathbf{V}_+(v_+ - 2d - 2)^{-1}$,
$\varepsilon = \bar{z} - \mathbf{H}m_+$,
$\mathbf{S} = \mathbf{H}\mathbf{P}^+\mathbf{H}^\top + \hat{\mathbf{X}}$,
$\mathbf{K} = \mathbf{P}^+\mathbf{H}^\top(\mathbf{S})^{-1}$,
$\mathbf{N} = \hat{\mathbf{X}}^{1/2}\mathbf{S}^{-1/2}\varepsilon\varepsilon^\top\mathbf{S}^{-T/2}\hat{\mathbf{X}}^{T/2}$.
Predicted likelihood, where $\Gamma(\cdot)$ is the Gamma function, and $\Gamma_d(\cdot)$ is the multivariate Gamma function,

$$l = (\pi^{|\mathcal{Z}|}|\mathcal{Z}|)^{-\frac{d}{2}} \frac{|\mathbf{V}_+|^{\frac{v_+ - d - 1}{2}}\Gamma_d(\frac{v-d-1}{2})|\hat{\mathbf{X}}|^{\frac{1}{2}}\Gamma(\alpha)(\beta_+)^{\alpha_+}}{|\mathbf{V}|^{\frac{v-d-1}{2}}\Gamma_d(\frac{v_+ - d - 1}{2})|\mathbf{S}|^{\frac{1}{2}}\Gamma(\alpha_+)(\beta)^{\alpha}}$$

Algorithm 6 GGIW Prediction

Input: $\zeta_{k|k}$
Output: Predicted GGIW parameters $\zeta_{k+1|k}$

$$\zeta_{k+1|k} = \begin{cases} \alpha_{k+1|k} = \alpha_{k|k}\eta_k, \\ \beta_{k+1|k} = \beta_{k|k}\eta_k, \\ m_{k+1|k} = f(m_{k|k}), \\ \mathbf{P}_{k+1|k} = \mathbf{F}_{k|k}\mathbf{P}_{k|k}\mathbf{F}_{k|k}^\top + \mathbf{Q}, \\ v_{k+1|k} = 2d + 2 + e^{-Ts/\tau}\frac{v_{k|k} - 2d - 2}{v_{k|k} - 2d - 2}, \\ \mathbf{V}_{k+1|k} = \left(\frac{v_{k+1|k} - 2d - 2}{v_{k|k} - 2d - 2}\right)^{-1} \times \mathbf{M}_{k|k}\mathbf{V}_{k|k}\mathbf{M}_{k|k}^\top \end{cases}$$

where $\mathbf{F}_{k|k} = \nabla_x f(x)|_{x = m_{k|k}}$

processes target birth and undetected target propagation, while the MBM predictor predicts detected target trajectories. Key variables include the PPP intensity $D^u(x)$, MBM parameters $\{\mathcal{W}_j, \{r_{j,i}, f_{j,i}\}_{i\in I_j}\}_{j\in J}$, and data association probabilities $\mathcal{W}_j(A)$. Algorithm 7 describes the GGIW-PMBM prediction, for the update and the other components of the GGIW-PMBM filter. See [36] for more details. Demonstrations and the code of the implementation are available at `https://github.com/nfierroflo/3D-INSEG-for-MEOT`.

Algorithm 7 GGIW PMBM prediction

Input: D^u, $\{\mathcal{W}_j, \{r_{j,i}, f_{j,i}\}_{i\in I_j}\}_{j\in J}$.
Output: D^{u+}, $\{(\mathcal{W}_j^+, \{(r_{j,i}^+, f_{j,i}^+)\}_{i\in I_j})\}_{j\in J}$

$$D^{u+}(x) = \sum_{n=1}^{\mathbf{N}_b} w_{b,n} \mathrm{GGIW}(x;\zeta_{b,n})$$

$$+ \sum_{n=1}^{\mathbf{N}_u} w_{u,n} p_S(\hat{x}_{u,n}) \mathrm{GGIW}(x;\zeta_{u,n})$$

$$r_{j,i}^+ = p_S(\hat{x}_{j,i}) r_{j,i} \quad f_{j,i}^+(x) = \mathrm{GGIW}(x;\zeta_{j,i}^+)$$

and $\mathcal{W}_j^+ = \mathcal{W}_j$, where $\zeta_{u,n}^+$ and $\zeta_{j,i}^+$ are computed as in algorithm 6.

Chapter 6

Results

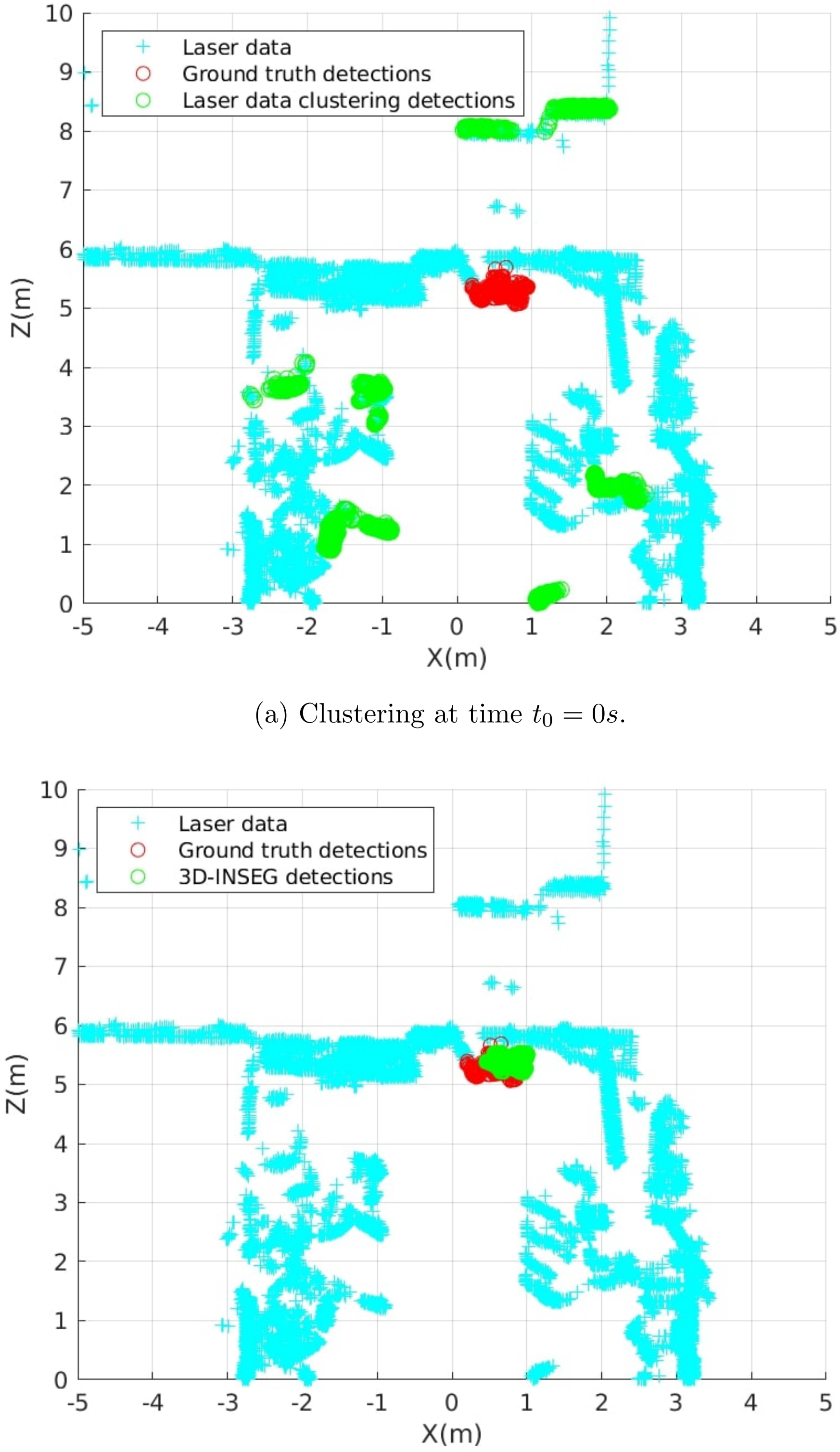
6.1 SEOT scenarios.

6.1.1 Human Detection

Figs. 6.1a-6.3b show human detection using both the clustering approach explained in section 4.2 in subfigures (a) and 3D-INSEG explained in section 4.1 in subfigures (b) for a sequence of a human walking from the door to the front of the camera in an indoor dense scenario at different time stamps ($t_0 = 0s$, $t_1 = 0.6s$, $t_2 = 1.31s$), the experiment is shown in 2D (plan view) for clarity. The laser data is represented in blue, the clustering approach and 3D-INSEG are represented in green. Table 6.1 summarizes the dectections over time for both approaches.

Table 6.1: Comparison of human detection over time using the Clustering approach and 3D-INSEG data.

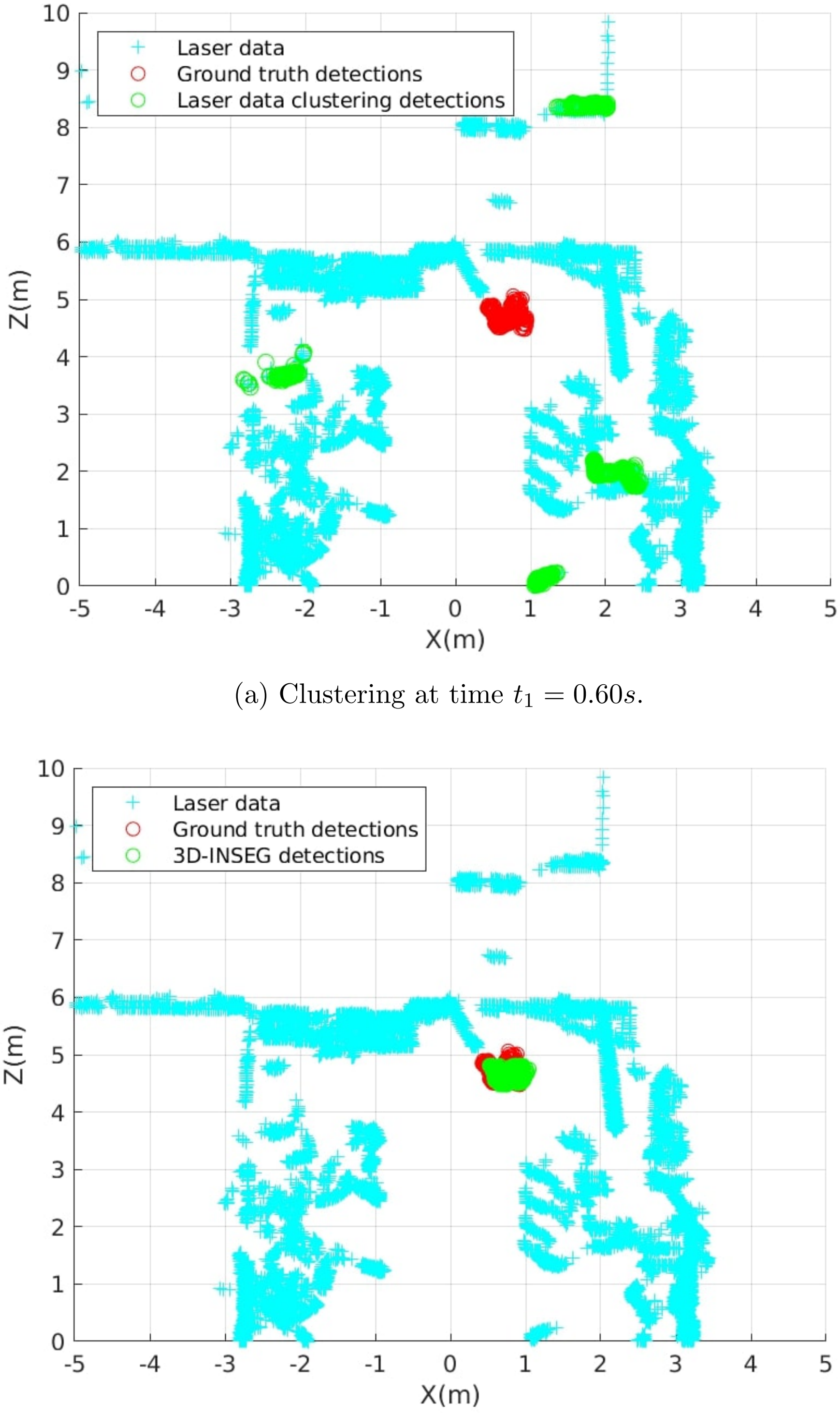
Method	Time stamp	No. of detections	Human detected
Clustering	0 s	17	No
3D-INSEG	0 s	1	Yes
Clustering	0.60s	13	No
3D-INSEG	0.60s	1	Yes
Clustering	1.31s	15	Yes
3D-INSEG	1.31s	1	Yes
Clustering	2.62s	17	Yes
3D-INSEG	2.62s	1	Yes
Clustering	3.33s	17	Yes
3D-INSEG	3.33s	1	Yes
Clustering	3.93s	19	Yes
3D-INSEG	3.93s	1	Yes
Clustering	4.64s	16	Yes
3D-INSEG	4.64s	1	Yes



(a) Clustering at time $t_0 = 0s$.

(b) 3D-INSEG at time $t_0 = 0s$.

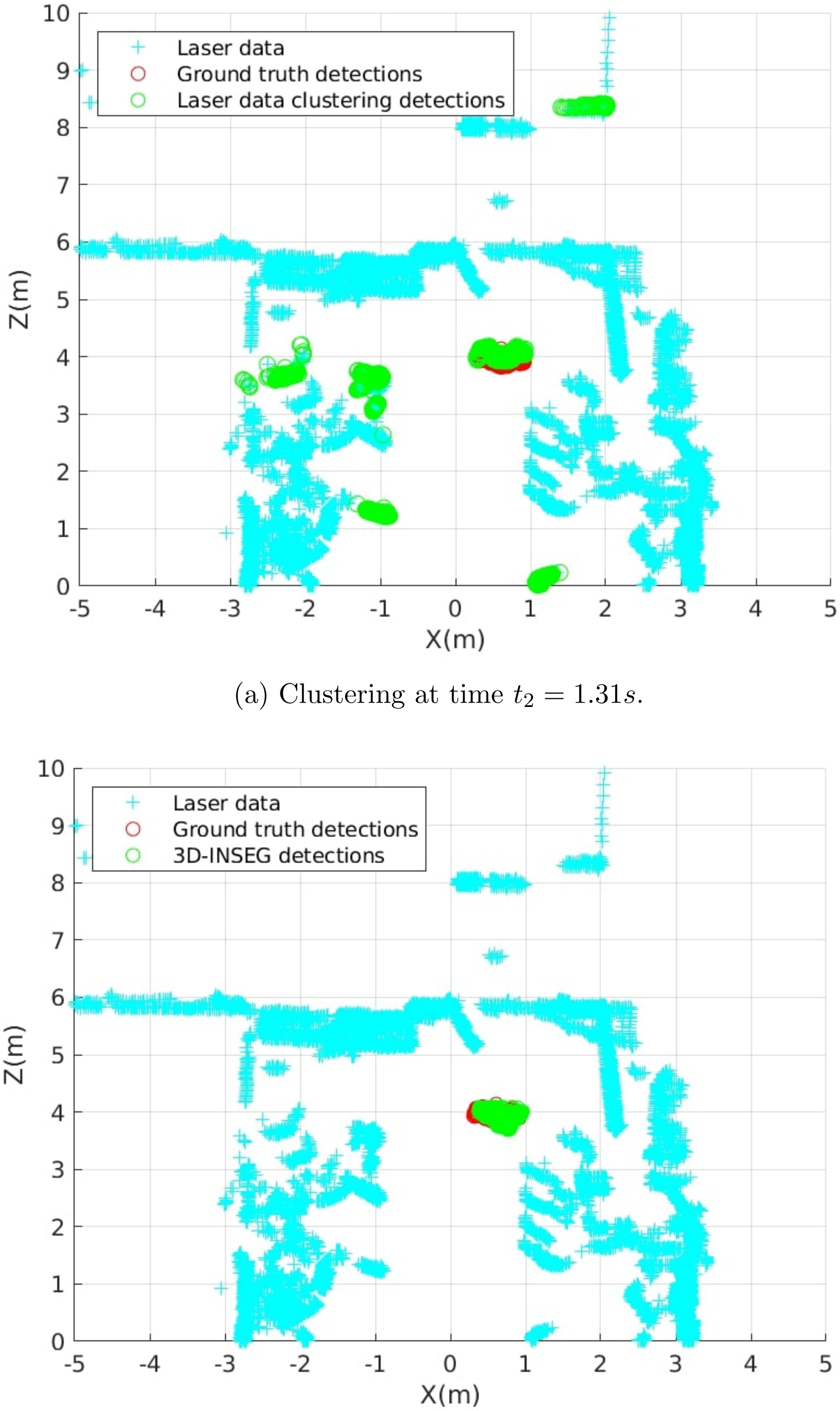
Figure 6.1: Human detection comparison at time $t_0 = 0s$.



(a) Clustering at time $t_1 = 0.60s$.

(b) 3D-INSEG at time $t_1 = 0.60s$.

Figure 6.2: Human detection comparison at time $t_0 = 0.60s$.



(a) Clustering at time $t_2 = 1.31s$.

(b) 3D-INSEG at time $t_2 = 1.31s$.

Figure 6.3: Human detection comparison at time $t_0 = 1.31s$.

Detection accuracy varies significantly between the two methods. At $t_0 = 0s$ and $t_1 = 0.6s$, the clustering approach fails to detect the human, indicating initial difficulties in distinguishing the target from background noise. In contrast, the 3D-INSEG method consistently detects the human at all timestamps. This stable performance across different time points demonstrates the robustness of the 3D-INSEG algorithm in accurately identifying the human subject from the beginning.

Comparatively, the detection rate of 3D-INSEG is consistently at 100% for all timestamps, while the clustering approach shows a lower and more variable detection rate. The clustering approach also exhibits a higher number of false positives, with detections that do not correspond to the human (e.g., 17 detections at $t_0 = 0s$ with no human detected), indicating a higher false positive rate compared to 3D-INSEG.

Qualitative analysis of the visual representations in Figs. 6.1a-6.3b reveals that 3D-INSEG provides clearer and more accurate delineation of the human subject compared to the clustering approach. The measurements generated by 3D-INSEG are more tightly fitted to the actual human.

Quantitatively, Table 6.1 offers a clear numerical comparison, highlighting the superiority of 3D-INSEG in both detection accuracy and consistency. For instance, at $t_0 = 0s$, 3D-INSEG correctly detects the human while clustering fails to do so, with similar patterns observed at subsequent timestamps. This analysis underscores the effectiveness of the 3D-INSEG algorithm over the clustering approach in various aspects of human detection and tracking, particularly in scenarios with potential occlusions and background noise.

6.1.2 Simulations on SEOT

In [18] the physical extension of an extended object is represented by a symmetric positive definite (SPD) random matrix X_k thus considering some ellipsoidal shape. Our implementation of [18] is simulated to show its potential in SEOT scenarios. Figure 6.4 shows a curved trajectory for an ellipse where the red ellipses represents the ground-truth, the blue ellipses represent the estimates using the algorithm, and green points are used for the measurements. Figure 6.5 shows the evolution of the Root Mean Square (RMS) Target Extension Error. Calculated as:

$$RMSE_X = \sqrt{\frac{1}{M}\sum_{\mu=1}^{M}[||X_{k|k} - X_k||_F^2]_\mu} \tag{6.1}$$

where M is the number of runs, $X_{k|k}$ the prediction of the extension matrix, X_k true extension matrix, and $||.||_F$ denotes the Frobenius norm. in this case $M = 100$.

Figure 6.6 shows the evolution of the RMS Target Location Error, calculated as:

$$RMSE_x = \sqrt{\frac{1}{M}\sum_{\mu=1}^{M}[||x_{k|k} - x_k||_2^2]_\mu} \tag{6.2}$$

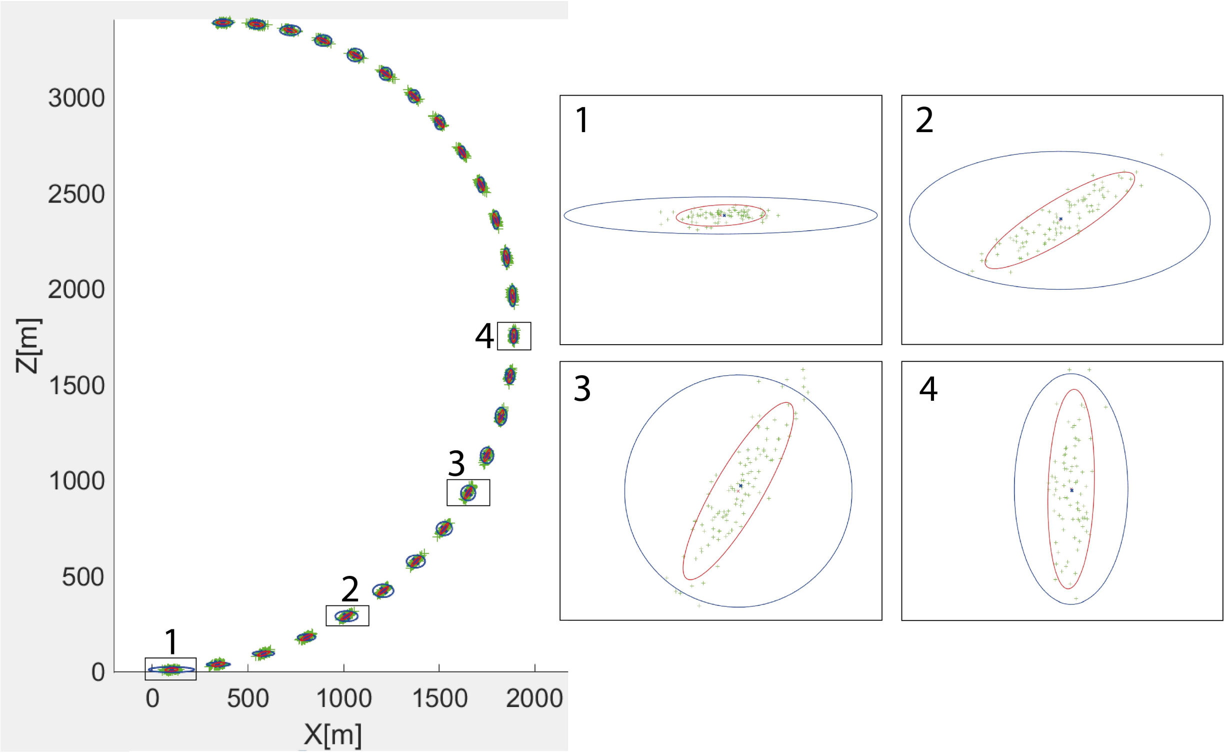


Figure 6.4: Simulation of the SEOT algorithm.

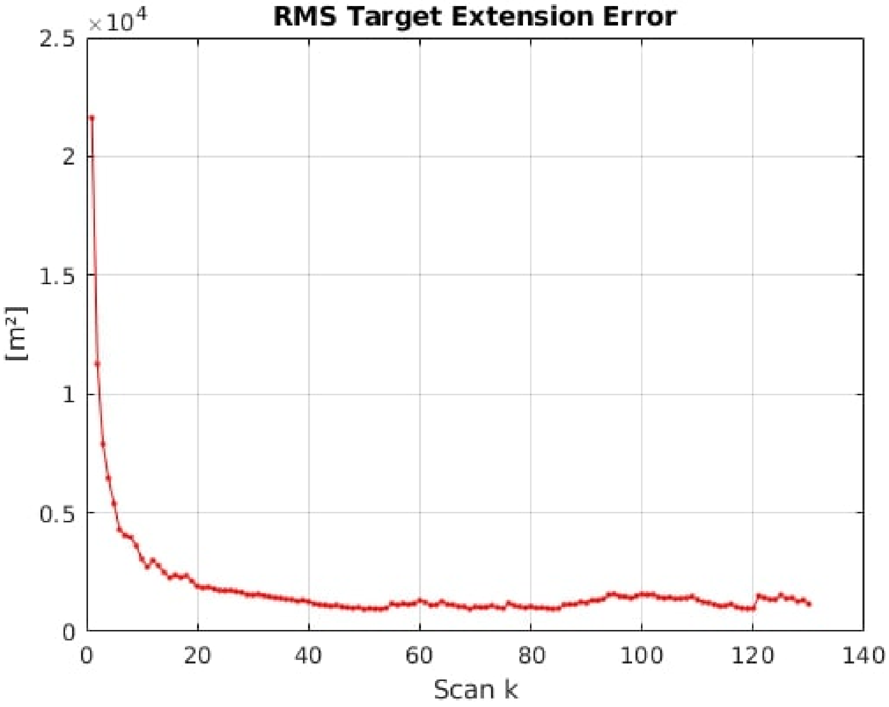


Figure 6.5: RMS target extension error versus time k.

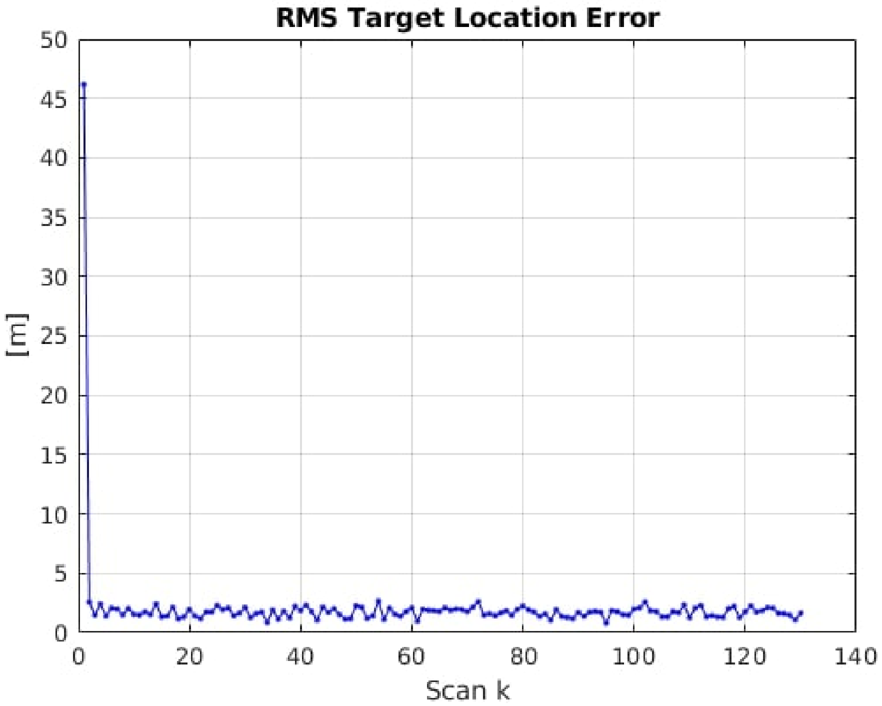


Figure 6.6: RMS target location error versus time k.

6.1.3 SEOT using 3D-INSEG data

The complete sequence of a human walking from the door towards the front of the camera explained in section 6.1.1 is used to show the usefulness of our 3D Segmentation Algorithm withim the SEOT algorithm from [18]. Figure 6.7 shows the measurements in green, the ground-truth shape and its center in blue together with the estimation and its center in red. For better comprehension the timestamp period of sampling is chosen as $t_{sampling} = 0.6s$. Figure 6.8 shows the evolution of the Target Extension Error calculated according to (6.3).

$$Error_X = ||X_{k|k} - X_k||_F \tag{6.3}$$

Figure 6.9 shows the evolution of the Target Location Error Calculated according to (6.4).

$$Error_x = ||x_{k|k} - x_k|| \tag{6.4}$$

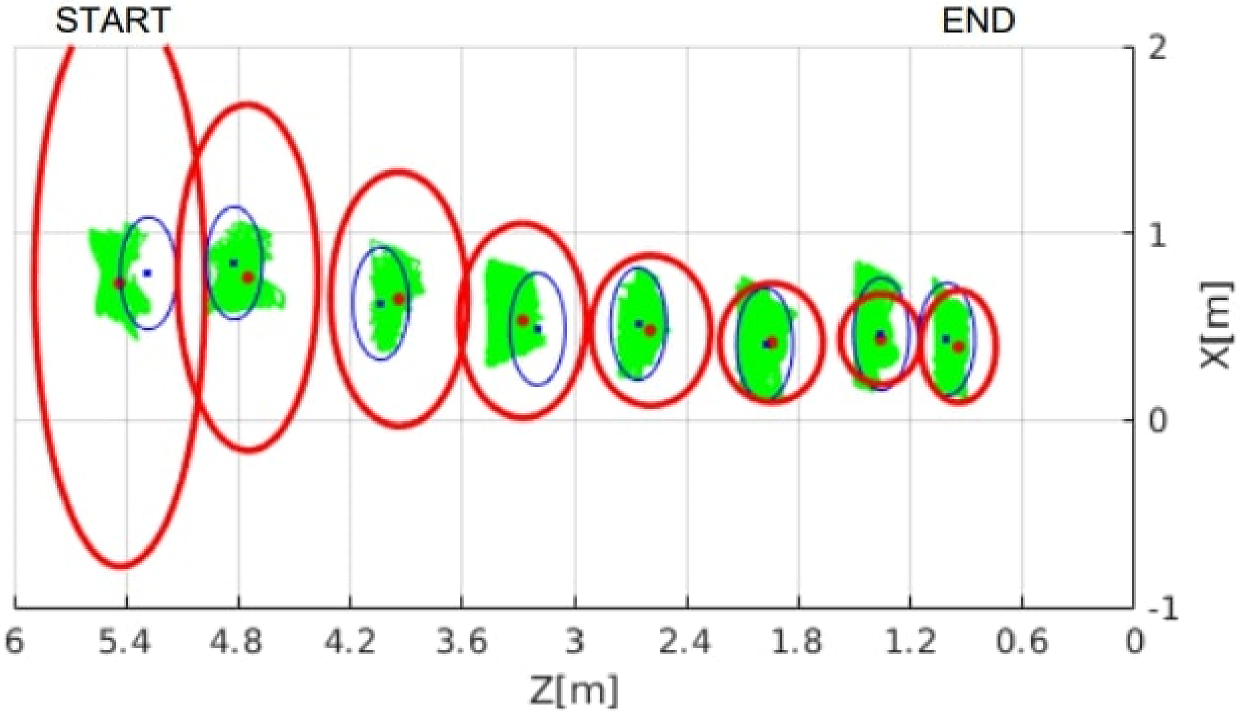


Figure 6.7: Human Tracking with the 3D-INSEG and SEOT algorithms.

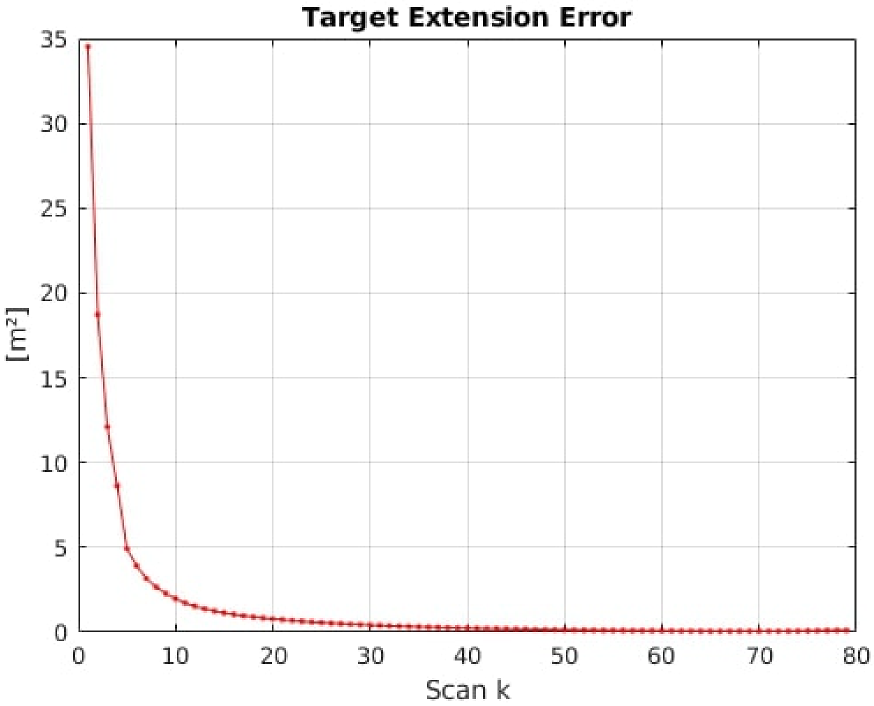


Figure 6.8: RMS target extension error versus time k in human tracking scenario.

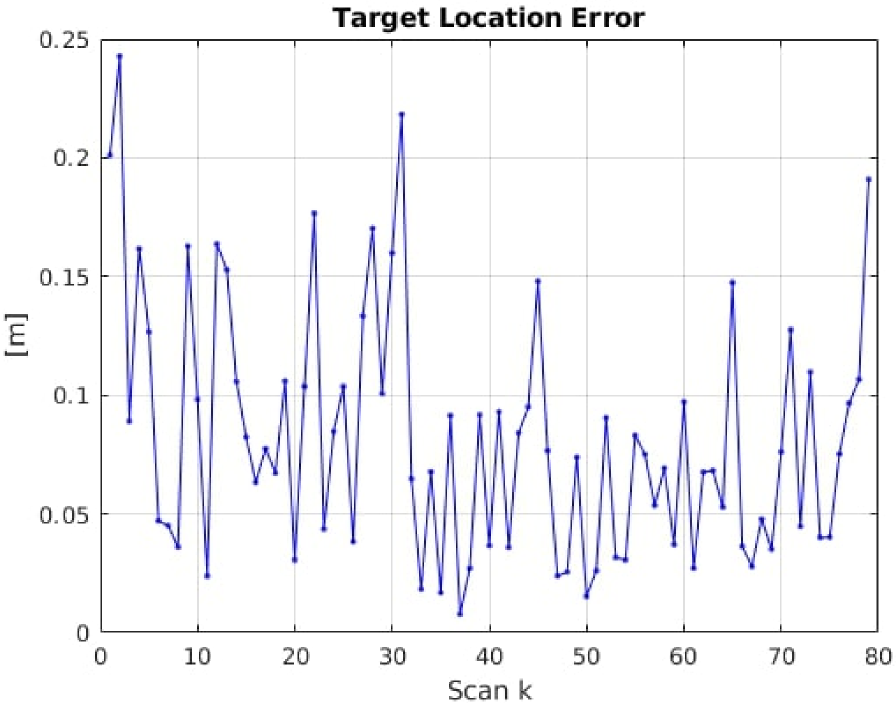


Figure 6.9: RMS target location error versus time k in human tracking scenario.

6.1.4 Discussion

The experiment detailed in Section 6.1.1 demonstrates the superior performance of human detection with the 3D-INSEG algorithm over the clustering algorithm at each timestamp. The 3D-INSEG algorithm accurately identifies the intended target and significantly reduces the number of false positives. A summary of this comparison is presented in Figs. 6.1a-6.3b, for three timestamps and in table 6.1 more timestamps are included.

The precision of the 3D-INSEG algorithm in detecting a specific object in 3D space is evident. It consistently identifies the human target across timestamps without any false positives. This contrasts with the alternative algorithm, which initially fails to detect the human and generates numerous false positive detections. This can be seen in Figs. 6.1a, 6.2a and 6.3a where multiple false detections (shown as green circles are present) arise from the use of the clustering algorithm. This is in contrast with the proposed 3D-INSEG algorithm (figs.6.1b, 6.2b and 6.3b) which detect the correct target without false positive detections. As highlighted in Section 2.3, reducing missdetections and clutter is advantageous for simplifying the data association process.

The simulations on SEOT implemented based on [18] showcase the efficacy of our approach in estimating the state of an ellipsoidal object following a curved trajectory. The root-mean-square error (RMSE) values discussed in Section 6.1.2 illustrate how the algorithm converges in terms of extension and location, it is shown in Fig. 6.5 for the extension and in Fig. 6.6 for location. However, a limitation of the algorithm is its relatively slow response to changes in orientation. Additionally, due to the representation of the extension using a matrix, the angle and axis length are not fully decoupled. For instance, during rotations, not only does the angle change as expected, but the axis length is also affected. This is observed in Fig. 6.4.

Finally, the experiment detailed in Section 6.1.3 highlights the practicality of using the 3D-INSEG algorithm with the SEOT algorithm. The human target is accurately tracked, the error value for the estimated extension gradually decreases over time (Fig. 6.8), and the error value for the estimated location is smaller than 0.25 [m] (Fig. 6.9).

6.2 MEOT scenarios

6.2.1 GOSPA

To evaluate the performance of extended object estimates with ellipsoidal extents, [39] showed that the Gaussian Wasserstein Distance (GWD) is the optimal choice. The GWD is defined as:

$$d_{GW}(x,\hat{x}) = \|H\xi - H\hat{\xi}\|_2^2 + \mathrm{Tr}\left(X + \hat{X} - 2\sqrt{X}\sqrt{\hat{X}}\right)^{1/2}, \tag{6.5}$$

where H corresponds to H_k in Eq. 5.1. It selects the position from the state vector and serves as the single target metric integrated into the Generalized Optimal Sub-Pattern Assignment (GOSPA) multi-object metric [40]. The GOSPA metric, formulated as:

$$\begin{aligned} d^p_{(c,\alpha)}(X,\hat{X}) = & \min_{\theta\in\Theta(|X|,|\hat{X}|)} \sum_{(i,j)\in\theta} d^{GW}_{(c)}(x_i,\hat{x}_j)^p \\ & + \frac{c^p}{\alpha}\left(|X| - |\theta| + |\hat{X}| - |\theta|\right)^{1/p}, \end{aligned} \tag{6.6}$$

incorporates $d^{GW}_{(c)}(x_i,\hat{x}_j) = \min(c, d_{GW}(x_i,\hat{x}_j))$, where $\Theta(|X|,|\hat{X}|)$ is the set of all possible 2D assignments, c denotes the base distance cut-off and p determines the severity of penalizing outliers in the localization component. In our experiments, $c = 1$, $p = 2$ and $\alpha = 2$. The GOSPA metric was introduced in [40] as an extension of the OSPA metric [41], and allows for the decomposition of multi-object error into three components: 1) state estimation error, 2) missed targets, and 3) false targets.

6.2.2 MEOT Simulation Results

The GGIW-PMBM filter is used under simulated scenarios. In this first simulation three tracks are estimated. These tracks are from 3 objects which have linear and angular velocity components.

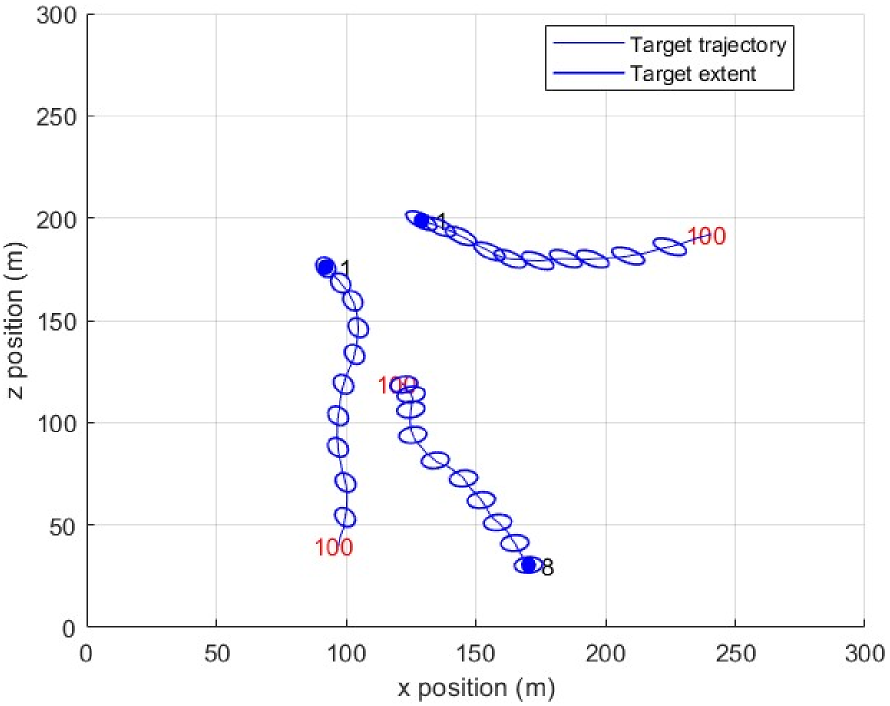


Figure 6.10: GGIW-PMBM simulation 1: Ground truth.

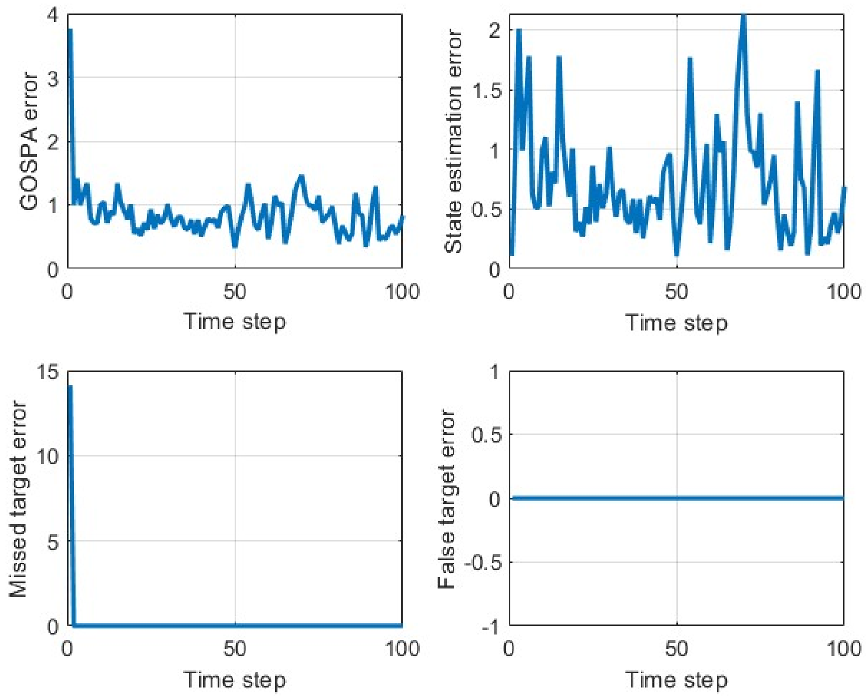


Figure 6.11: GGIW-PMBM simulation 1: GOSPA.

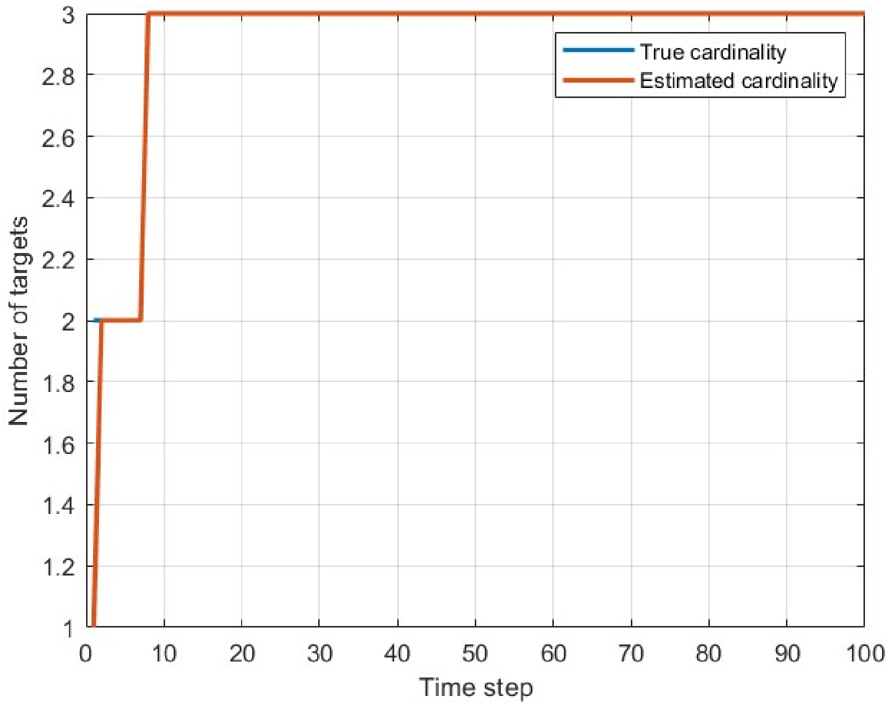


Figure 6.12: GGIW-PMBM simulation 1 Cardinality estimation.

In simulation 2, 5 tracks are estimated.

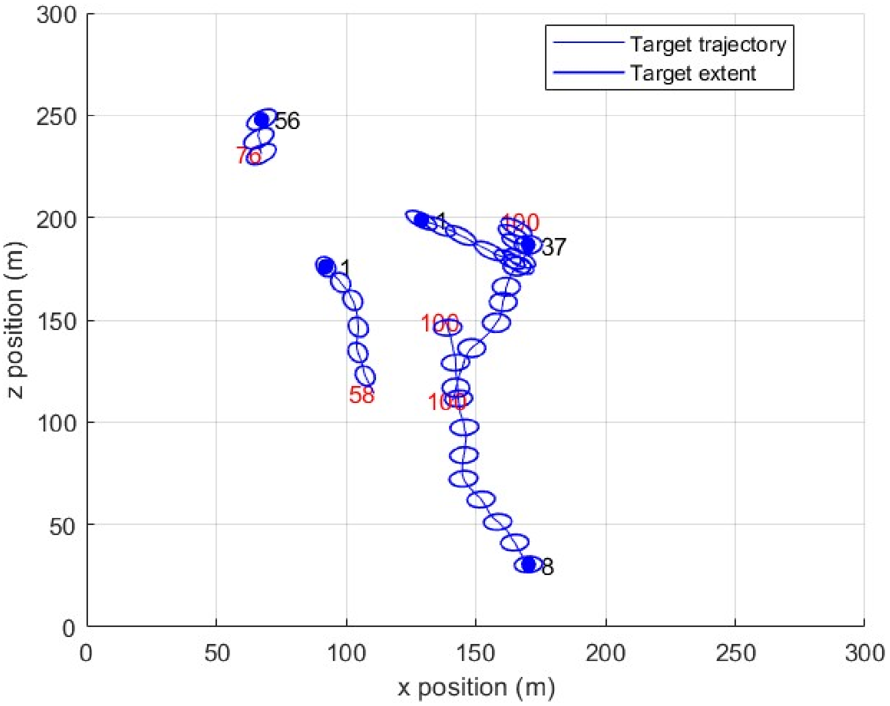


Figure 6.13: GGIW-PMBM simulation 2: Ground truth.

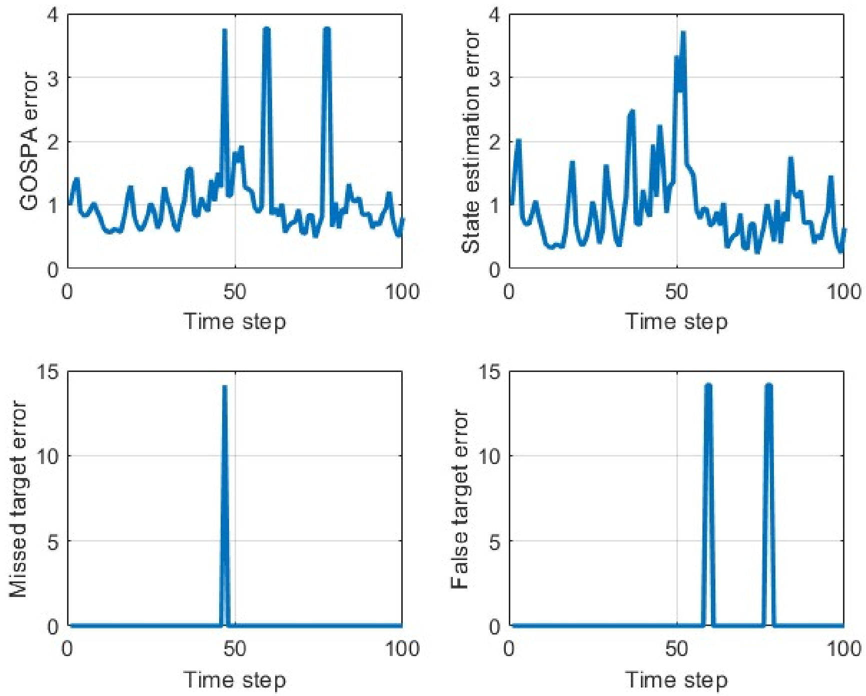


Figure 6.14: GGIW-PMBM simulation 2: GOSPA.

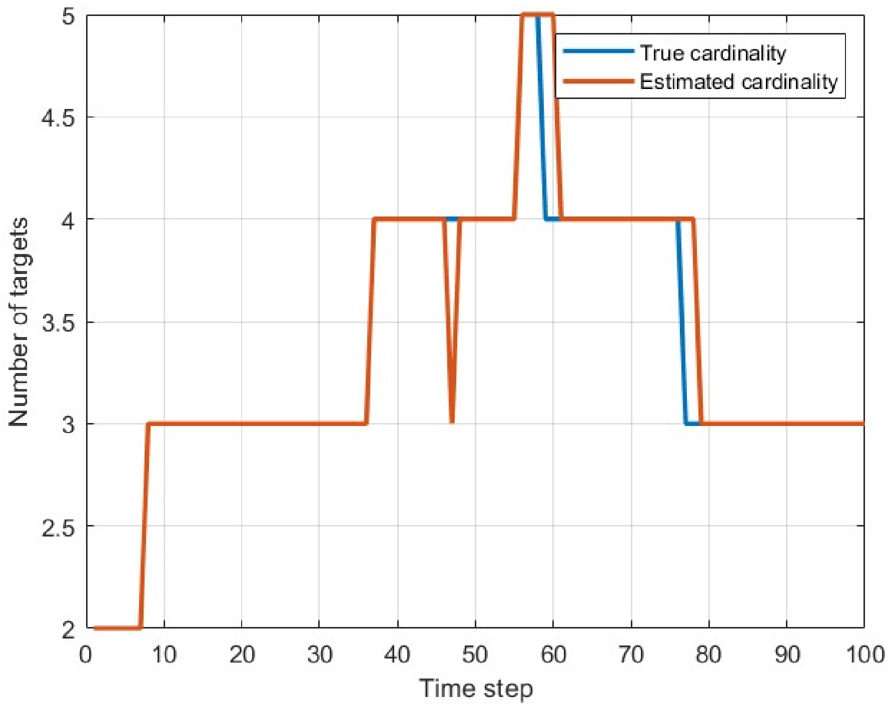


Figure 6.15: GGIW-PMBM simulation 2: Cardinality estimation.

In simulation 3 , 8 tracks are estimated.

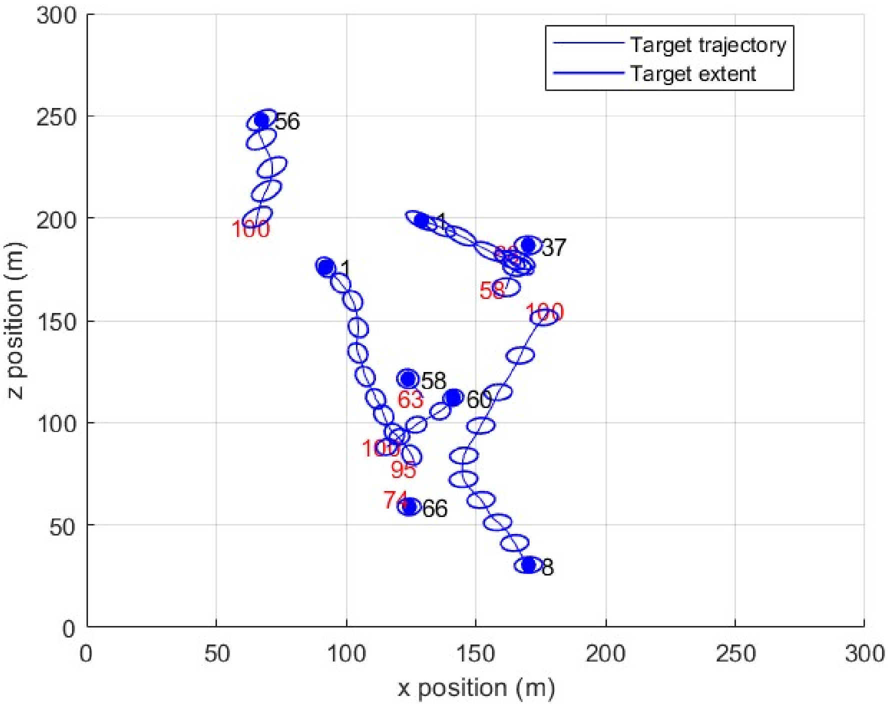


Figure 6.16: GGIW-PMBM simulation 3: Ground truth.

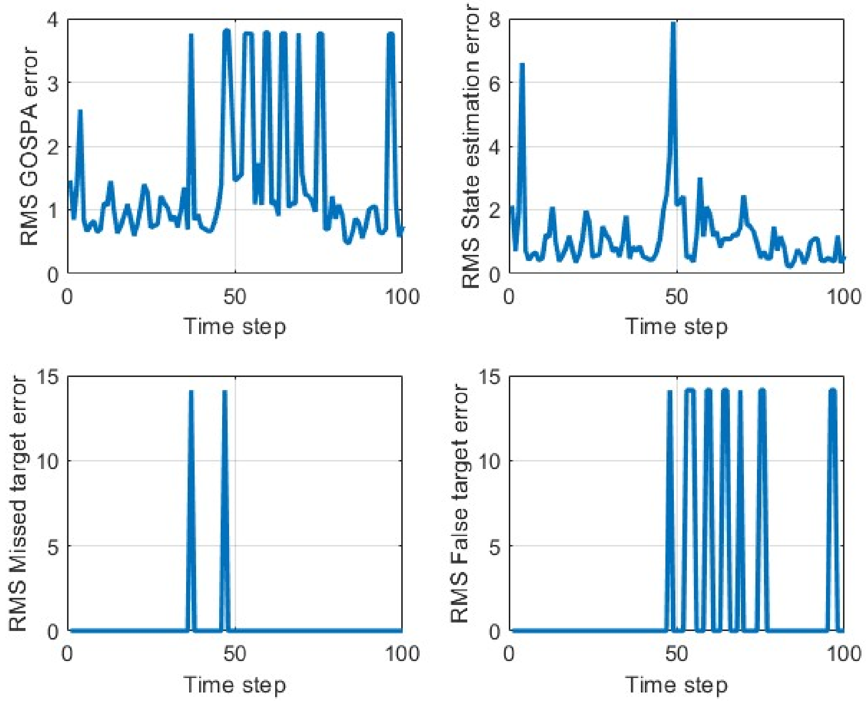


Figure 6.17: GGIW-PMBM simulation 3: GOSPA.

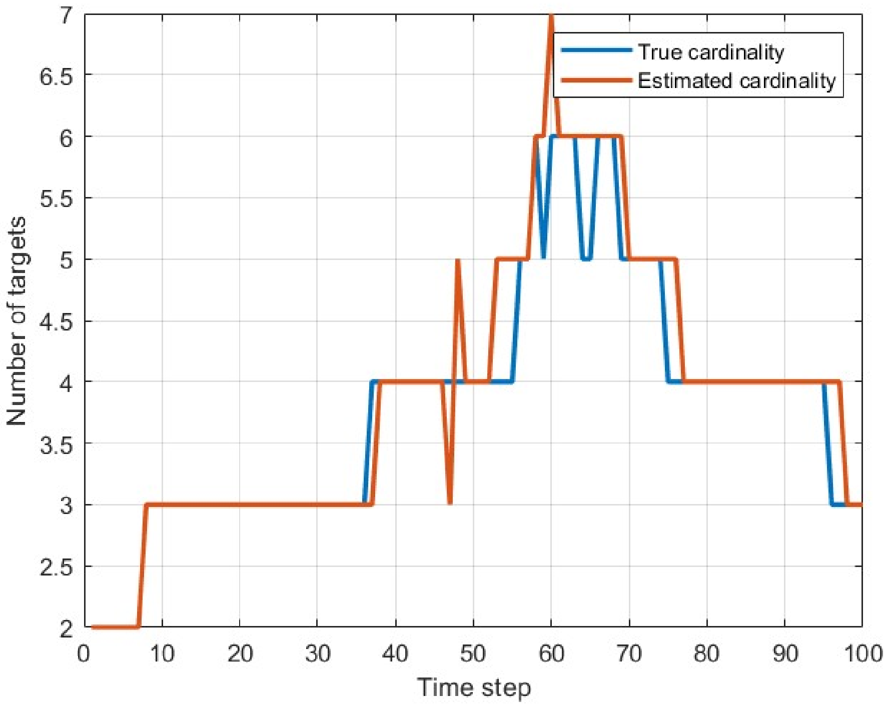


Figure 6.18: GGIW-PMBM simulation 3: Cardinality estimation.

6.2.3 MEOT Experimental Results

Four experiments in different scenarios with different target densities are used to test and compare the estimates from the velodyne and 3D-INSEG data. Table 6.2 summarizes the results. To illustrate this, data produced by the 3D-INSEG algorithm is shown in Appendix B.

Table 6.2: MEOT results summary.

	Velodyne		3D-inseg	
Experiment	Mean GOSPA	runtime	Mean GOSPA	runtime
Experiment 1	1.22	81.65	**0.2**	8.41
Experiment 2	0.95	4.87	**0.61**	3.1
Experiment 3	1.17	10.56	**0.79**	2.12
Experiment 4	1.13	25.1	0.98	**6.95**

Estimation is performed using both raw laser data and data that has been preprocessed with the 3D-INSEG algorithm. We compare the results using the standard extended target PMBM filter with its GGIW implementation. The 3D-INSEG clusters and the laser data are projected into 2D, so that the GGIW-PMBM tracking filter runs in 2D.

To process the laser data, we employed a two-step clustering and assignment approach aimed at updating each previous global hypothesis with relevant information [14]. Initially, we applied DB-SCAN [11] using five different distance values, evenly distributed between 0.1 and

0.5(m), to generate multiple measurement partitions. Subsequently, for each measurement partition and global hypothesis a, we utilized Murty's algorithm [15] to identify the $\lceil w^a_{k|k} \rceil$[1] best cluster-to-Bernoulli assignments, where $w^a_{k|k}$ denotes the weight of global hypothesis a. Additional parameters included a maximum of 20 global hypotheses ($N_h = 20$), thresholds for MBM pruning (10^{-2}), PPP weight pruning ($\Gamma_p = 10^{-3}$), and Bernoulli density pruning ($\Gamma_b = 10^{-3}$).

In the GGIW implementation, the target state $x = (\gamma, \xi, X)$ was defined, where γ represented the expected number of measurements per target, $\xi = [p_x, v_x, p_y, v_y]^T$ encapsulated the target's current position and velocity, and X was a 2×2 positive definite matrix describing the target's ellipsoidal shape. The kinematic state motion model is assumed to evolve under a constant velocity. Therefore the extent transformation function $\mathbf{M}$ is an identity matrix $\mathbf{M}(x_k) = \mathbf{I}_2$. The survival probability was set to $p_S = 0.99$. For the birth process, we adopted a Poisson Point Process (PPP) with a GGIW intensity featuring a weight of $w^k_b = 0.1$ for all time steps. Its GGIW density comprised a gamma distribution with a mean of 5 and a shape of 100, a Gaussian distribution with a mean vector $\bar{x}^k_b = [0\,\mathrm{m}, 0\,\mathrm{m/s}, 0\,\mathrm{m}, 0\,\mathrm{m/s}]^T$, and a covariance matrix $P^k_b = \mathrm{diag}([50^2\,(\mathrm{m}^2), 1\,(\mathrm{m}^2/\mathrm{s}^2), 50^2\,(\mathrm{m}^2), 1\,(\mathrm{m}^2/\mathrm{s}^2)])$, along with an inverse-Wishart distribution with a mean of $\mathrm{diag}([2, 2])\,(\mathrm{m}^2)$ and 100 degrees of freedom.

The ground truth for each experiment was manually marked, resulting in some variability. Ideally, the actual curves should be smoother.

6.2.4 Experiment 1

In the first experiment we explore a densely populated scenario involving extended targets. The experiment showcases the goal of estimating the state of two humans in an indoor environment for 11 seconds (160 frames) as they navigate their surroundings and eventually intersect paths. Fig. 6.19 shows different frames for a better understanding of the scenario.

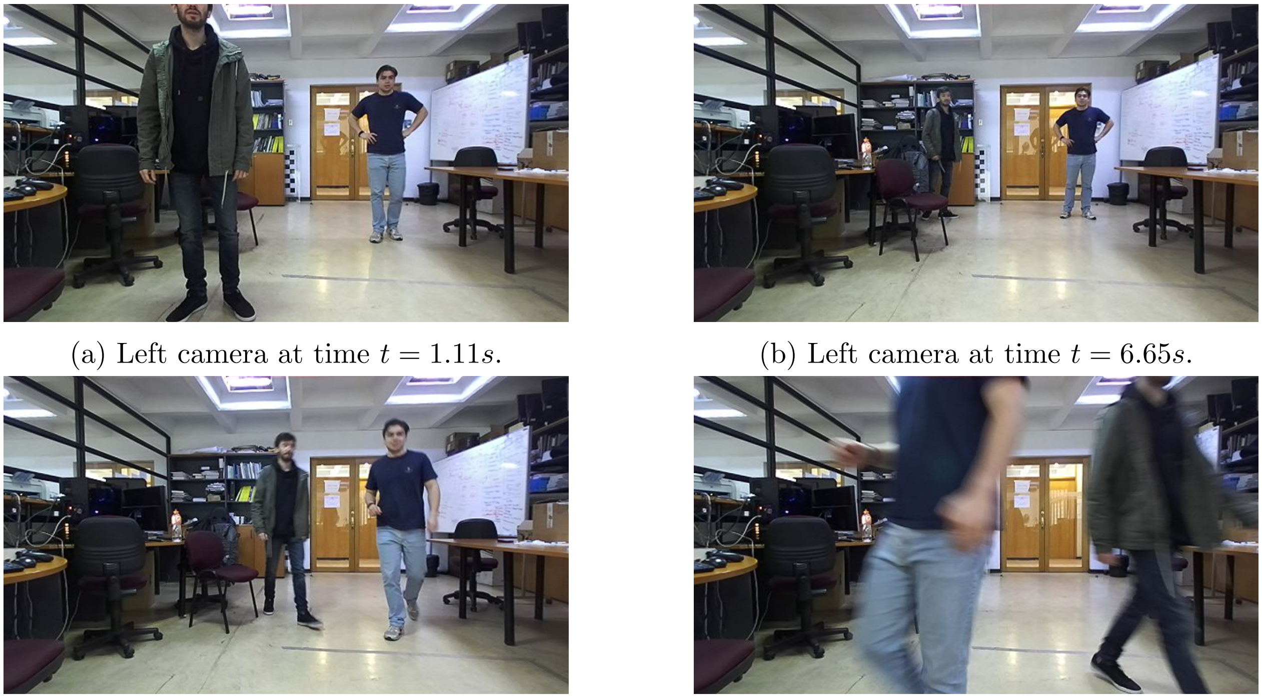
The ground truth trajectories were manually marked and are visualized in Fig. 6.20, where the green and red tracks correspond to separate target trajectories. For clarity, only the shapes (ellipses) corresponding to certain times t in seconds are shown.

Figures 6.21a, 6.22a and 6.23a show estimates at different times using laser data. Figures 6.21b, 6.22b and 6.23b show estimates at different times based on 3D-INSEG generated data. The laser data is represented in blue, the 3D-INSEG generated data in yellow, the estimated shapes and centers in red, and the ground truth shapes and centers in green.

Fig. 6.24 illustrates the GOSPA metric and its components over time, while Fig. 6.25 displays the estimated number of targets. Blue is used for MEOT with laser data as measurements, and yellow is used when it is performed using 3D-INSEG generated data.

The computational times in seconds to run the extended PMBM filter on a 12th Gen Intel(R) Core(TM) i7-12650H 2.30 GHz are 161.80 (laser data), and 0.75 (3D-INSEG data).

[1] where $\lceil\,\rceil$ corresponds to rounding up to the nearest integer.



(a) Left camera at time $t = 1.11s$.

(b) Left camera at time $t = 6.65s$.

(c) Left camera at time $t = 8.8s$.

(d) Left camera at time $t = 10.58s$.

Figure 6.19: Experiment 1: Reference images from the sequence of the left camera images.

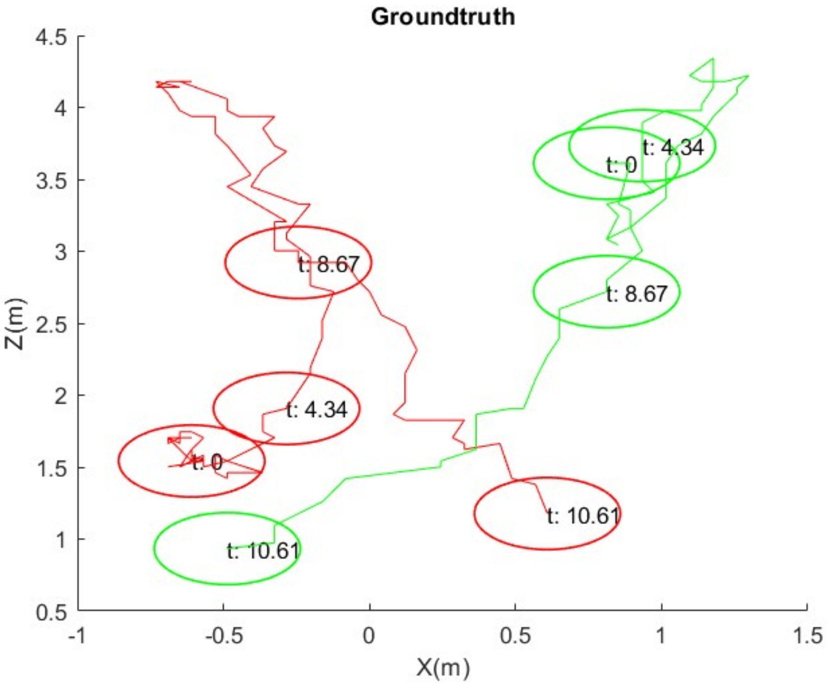
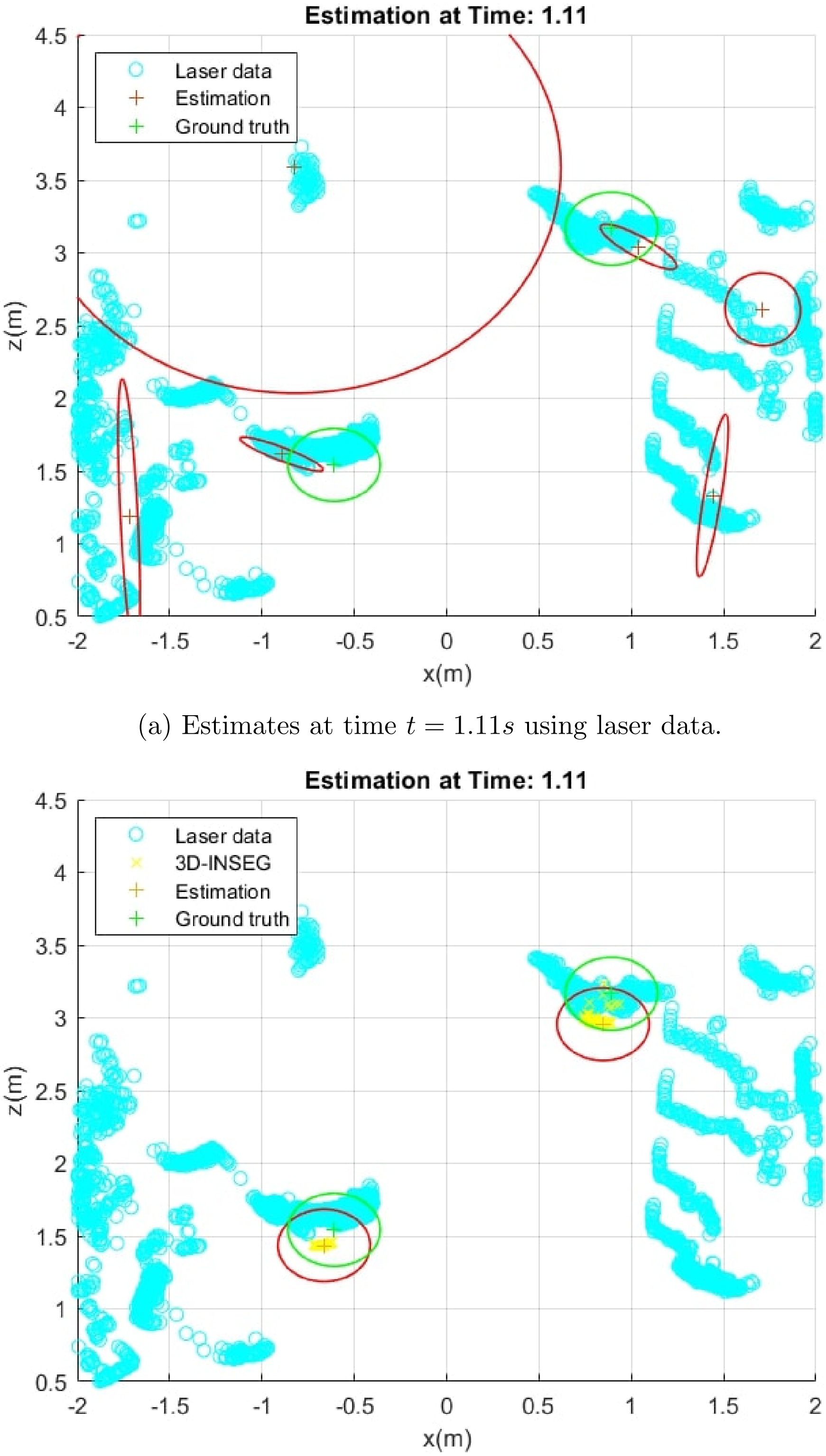


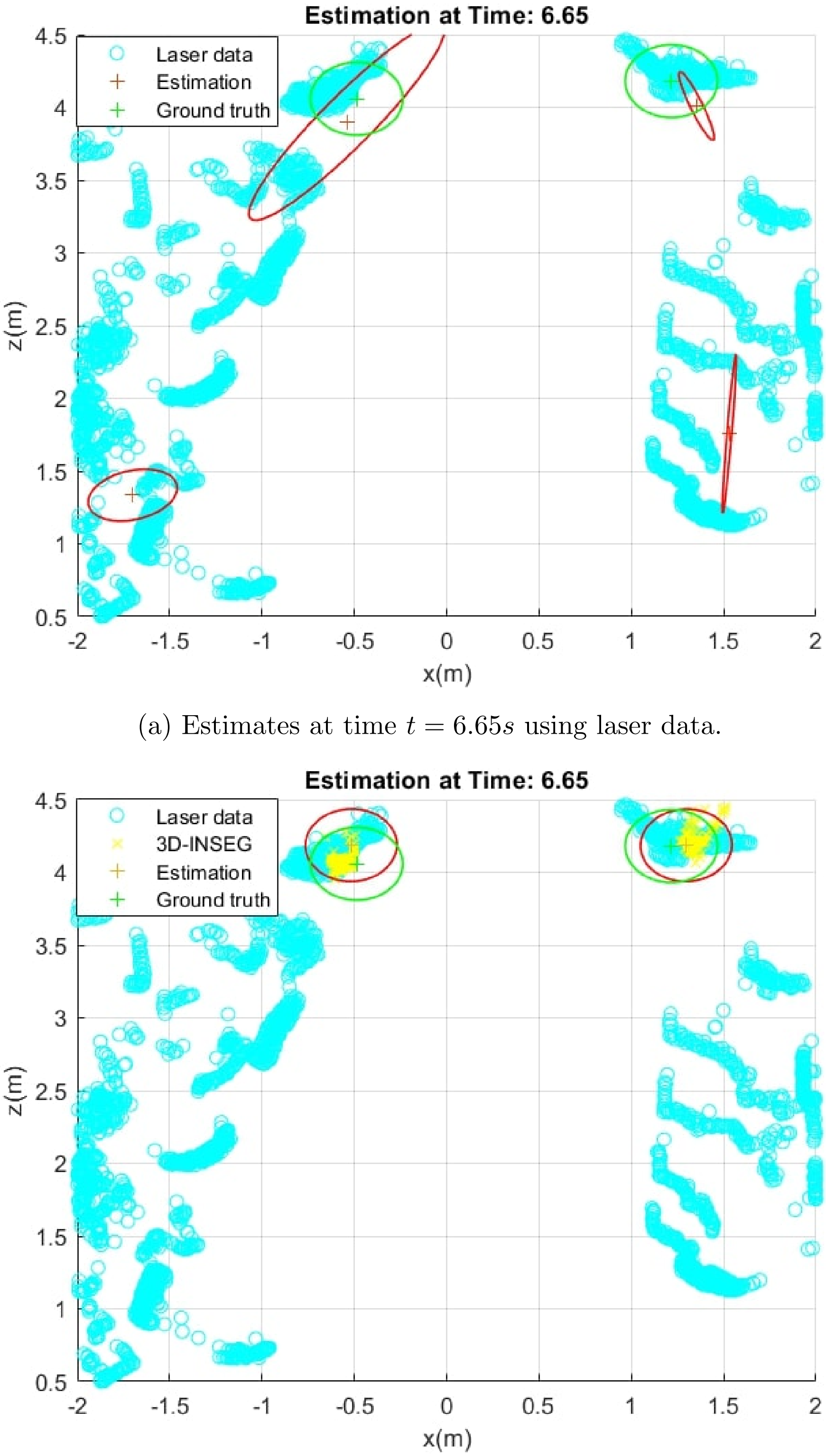
Figure 6.20: Experiment 1: Ground truth.



(a) Estimates at time $t = 1.11s$ using laser data.

(b) Estimates at time $t = 1.11s$ using 3D-INSEG data.

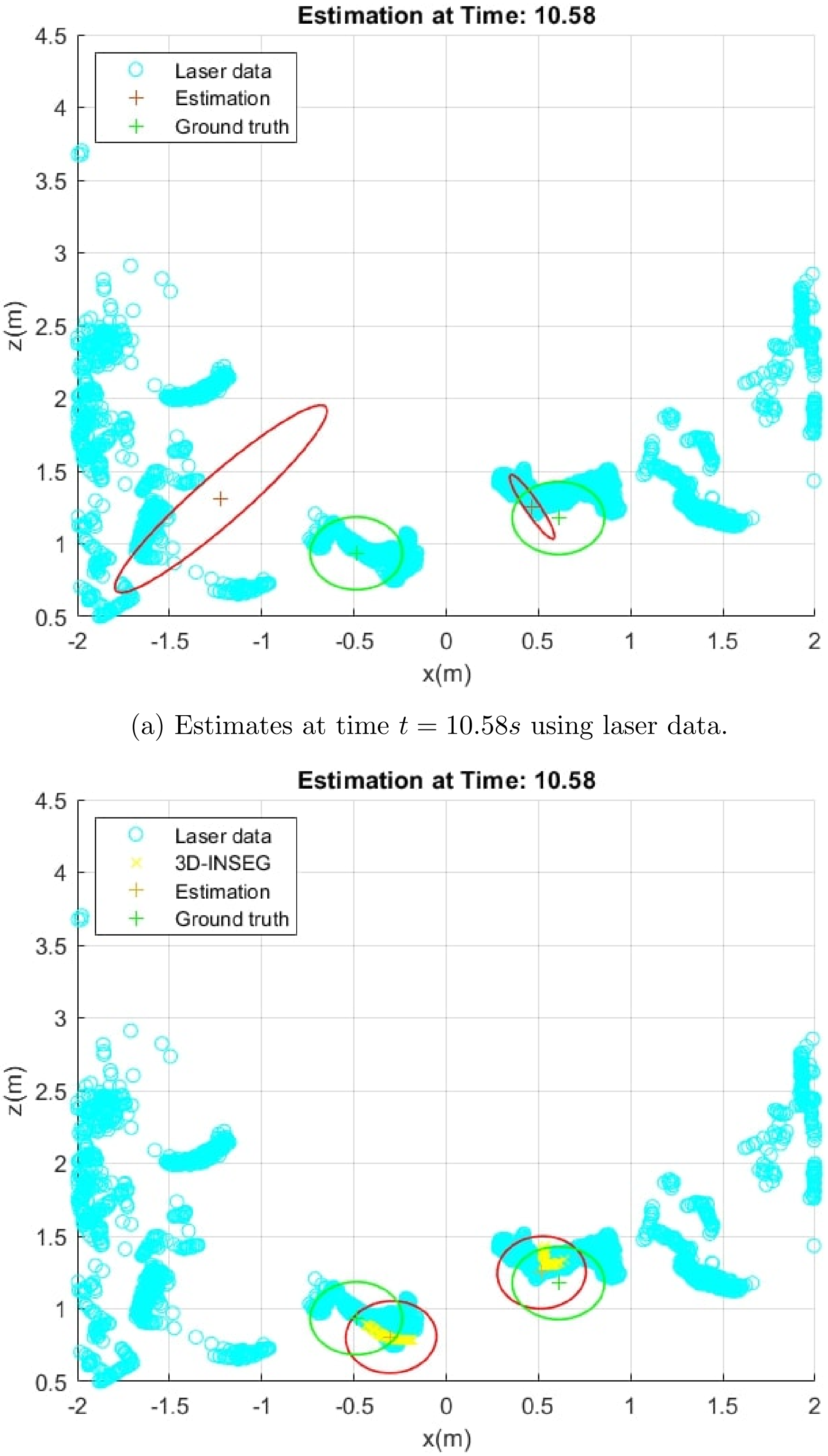
Figure 6.21: Experiment 1: Comparison of estimates at time $t = 1.11s$.



(a) Estimates at time $t = 6.65s$ using laser data.

(b) Estimates at time $t = 6.65s$ using 3D-INSEG data.

Figure 6.22: Experiment 1: Comparison of estimates at time $t = 6.65s$.



(a) Estimates at time $t = 10.58s$ using laser data.

(b) Estimates at time $t = 10.58s$ using 3D-INSEG data.

Figure 6.23: Experiment 1: Comparison of estimates at time $t = 10.58s$.

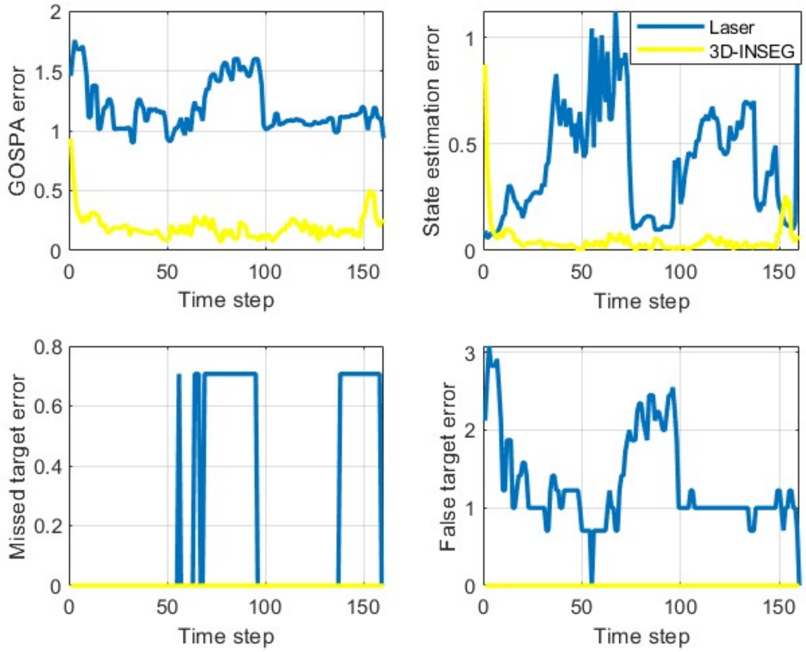


Figure 6.24: Experiment 1: GOSPA errors and their decomposition against time for the extended target scenario using the extended PMBM with laser data and the 3D-INSEG data.

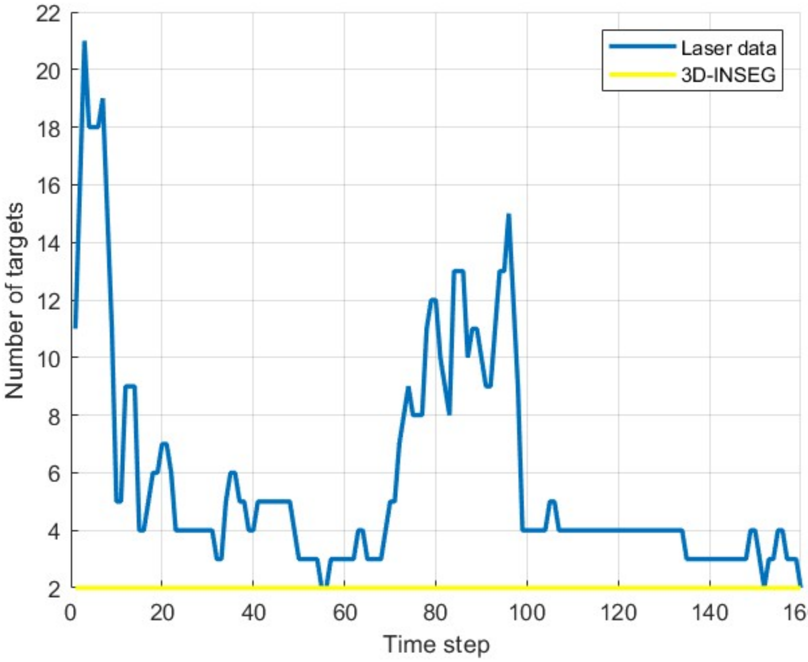
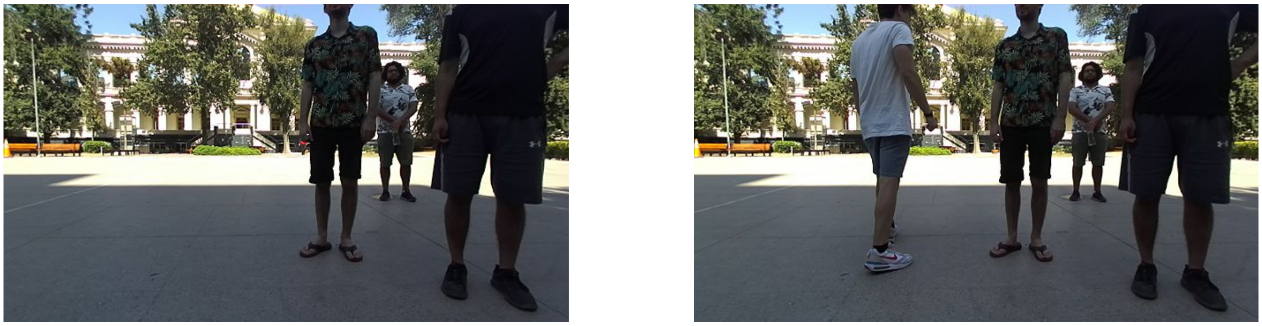


Figure 6.25: Experiment 1: Target cardinality estimated by the extended PMBM filter using the laser data and the 3D-INSEG data.

6.2.5 Experiment 2

For the second experiment we explore an outdoor scenario with low density in the surveillance area involving extended targets. The experiment showcases the goal of estimating the state of four humans in an outdoor environment for 7.39 (112 frames) seconds with three persons remaining in the same position and a fourth one approaching to them. Fig. 6.26 shows different frames for a better understanding of the scenario.



(a) Left camera at time $t = 0.7s$.

(b) Left camera at time $t = 6.58s$.

Figure 6.26: Experiment 2: Reference images from the sequence of the left camera images.

The ground truth trajectories were manually marked and are visualized in Fig. 6.27, where the blue tracks correspond to separate target trajectories. For clarity, only the shapes (ellipses in blue) for the birth are shown.

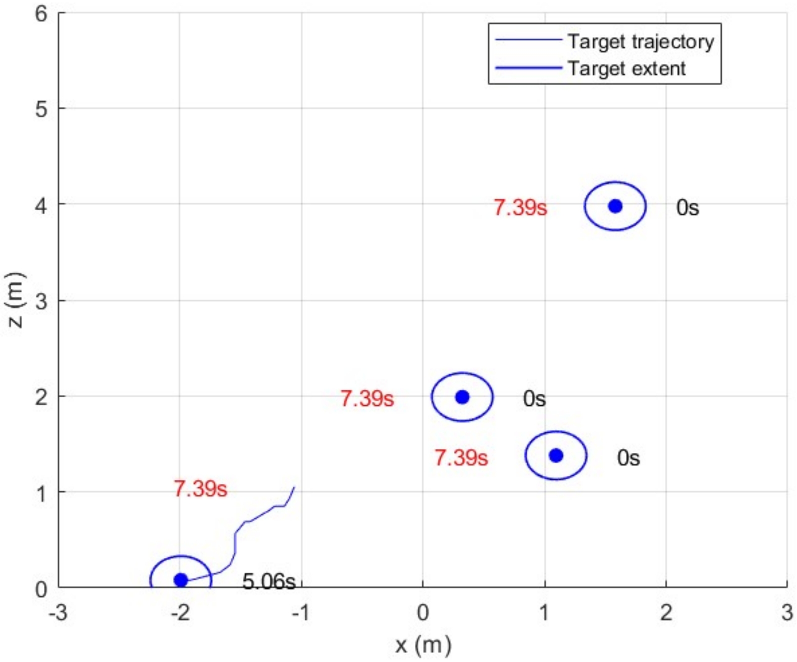


Figure 6.27: Experiment 2: Ground truth.

Figures 6.28a, 6.29a and 6.30a show estimates at different times using laser data. Figures 6.28b, 6.29b and 6.30b show estimates at different times based on 3D-INSEG generated data.

The laser data is represented in blue, the 3D-INSEG generated data in yellow, the estimated shapes and centers in red.

Fig. 6.31 illustrates the GOSPA metric and its components over time, while Fig. 6.32 displays the estimated number of targets. Blue is used for MEOT with laser data as measurements, and yellow is used when it is performed using 3D-INSEG generated data.

6.2.6 Experiment 3

For experiment 3 we explore an outdoor scenario with higher density compared to experiment 2 but still lower than the experiment 1 in the surveillance area involving extended targets. The experiment showcases the goal of estimating the state of three humans in an outdoor environment for 19.92 (300 frames) seconds with two persons remaining in the same position and a third one approaching to them. Fig. 6.33 shows different frames for a better understanding of the scenario.

The ground truth trajectories were manually marked and are visualized in Fig. 6.34, where the blue tracks correspond to separate target trajectories. For clarity, only the shapes (ellipses in blue) for the birth are shown.

Figures 6.35a, 6.36a and 6.37a show estimates at different times using laser data. Figures 6.35b, 6.36b and 6.37b show estimates at different times based on 3D-INSEG generated data. The laser data is represented in blue, the 3D-INSEG generated data in yellow, the estimated shapes and centers in red.

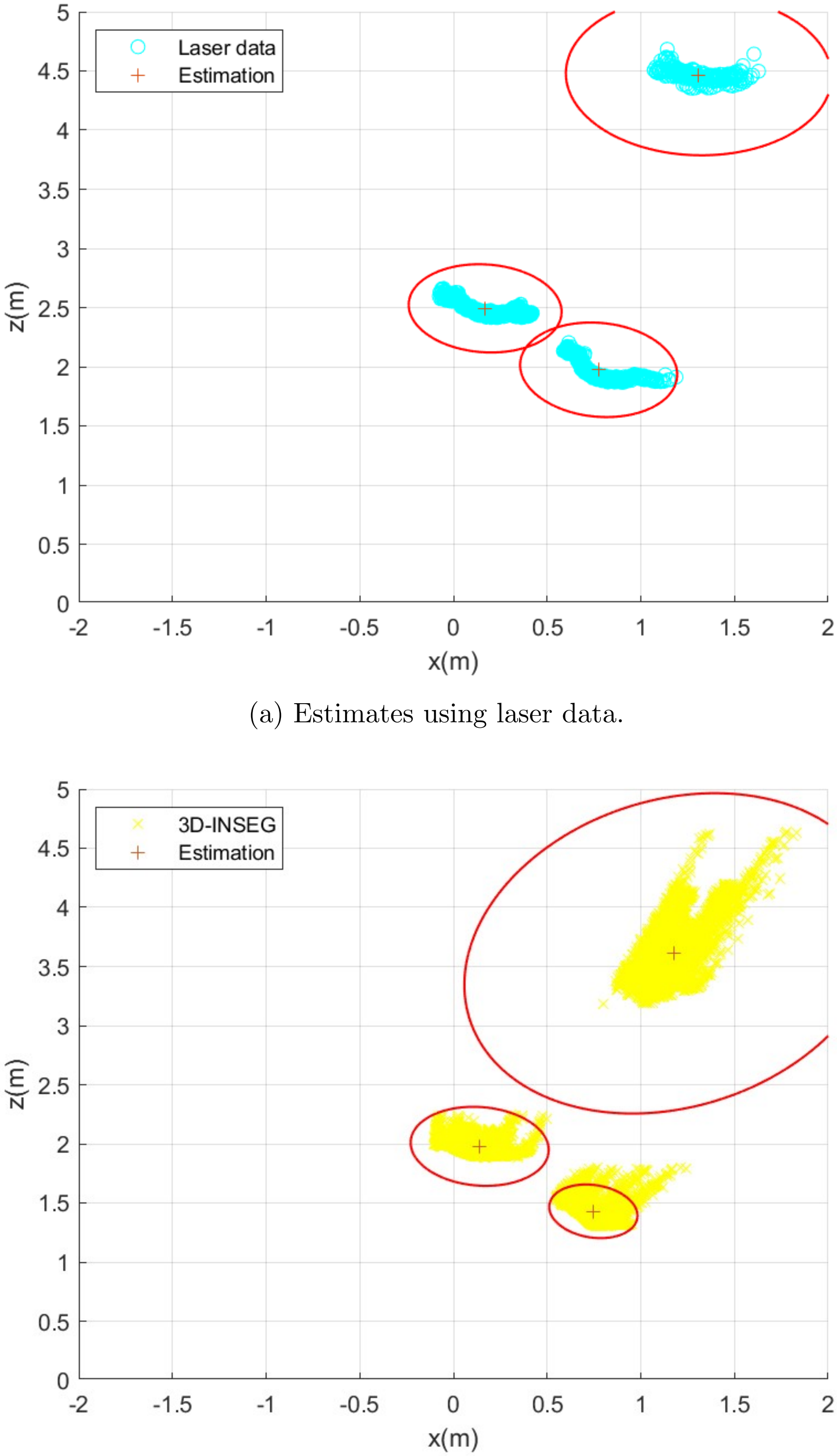
Fig. 6.38 illustrates the GOSPA metric and its components over time, while Fig. 6.39 displays the estimated number of targets. Blue is used for MEOT with laser data as measurements, and yellow is used when it is performed using 3D-INSEG generated data.

6.2.7 Experiment 4

For experiment 4 we explore an outdoor scenario with low density in the surveillance area involving extended targets. The experiment showcases the goal of estimating the state of five humans in an outdoor environment for 21.94 (330 frames) seconds with four persons remaining in the same position and a fifth one approaching to them. Fig. 6.40 shows different frames for a better understanding of the scenario.

The ground truth trajectories were manually marked and are visualized in Fig. 6.41, where the blue tracks correspond to separate target trajectories. For clarity, only the shapes (ellipses in blue) for the birth are shown.

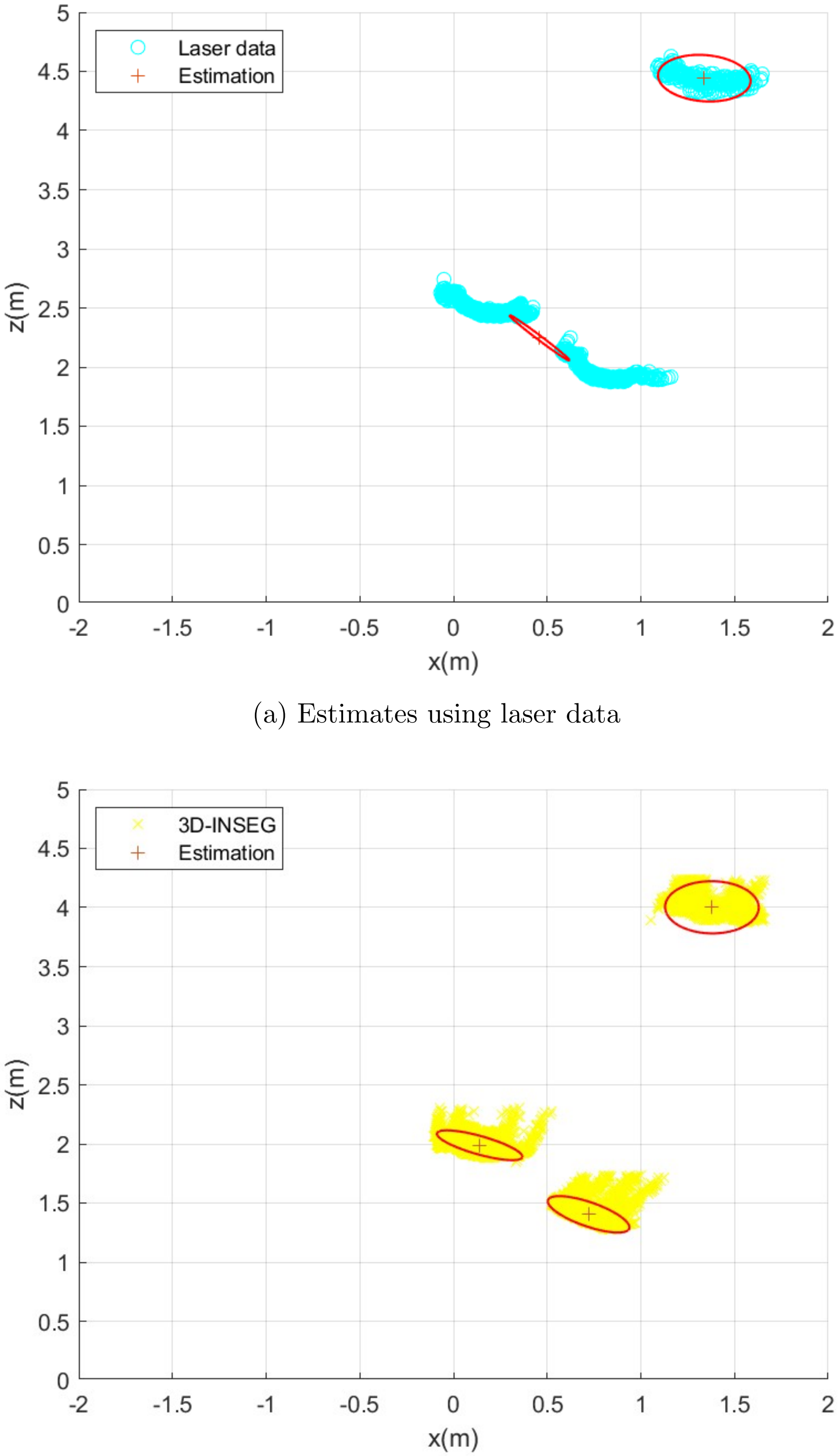
Figures 6.42a, 6.43a and 6.44a show estimates at different times using laser data. Figures 6.42b, 6.43b and 6.44b show estimates at different times based on 3D-INSEG generated data. The laser data is represented in blue, the 3D-INSEG generated data in yellow, the estimated shapes and centers in red. Fig. 6.45 illustrates the GOSPA metric and its components over time, while Fig. 6.46 displays the estimated number of targets. Blue is used for MEOT



(a) Estimates using laser data.

(b) Estimates using 3D-INSEG data.

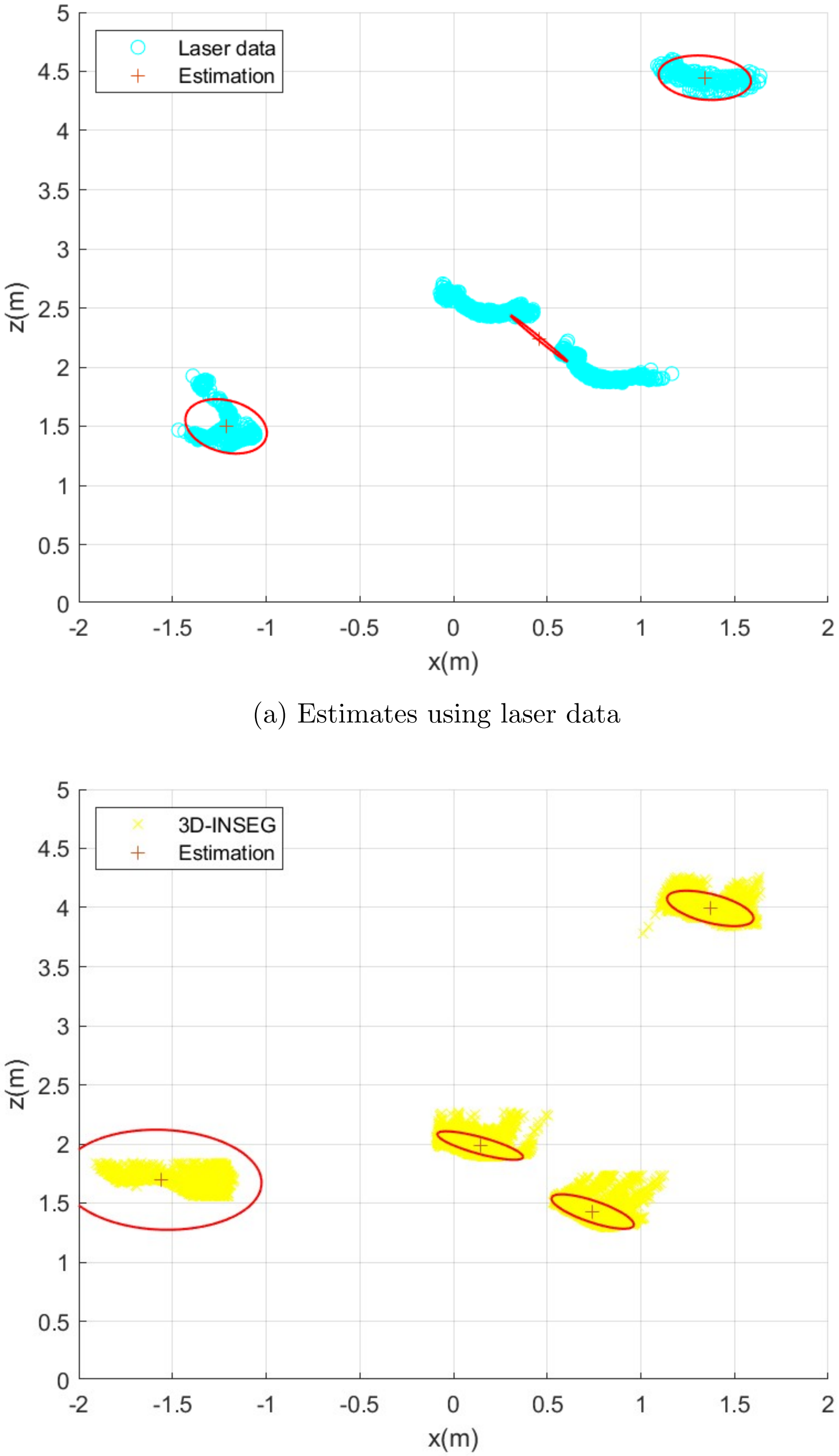
Figure 6.28: Experiment 2: Comparison of estimates at time t=0s.



(a) Estimates using laser data

(b) Estimates using 3D-INSEG data.

Figure 6.29: Experiment 2: Comparison of estimates at time t=3.26s.



(a) Estimates using laser data

(b) Estimates using 3D-INSEG data.

Figure 6.30: Experiment 2: Comparison of estimates at time t=6.58s.

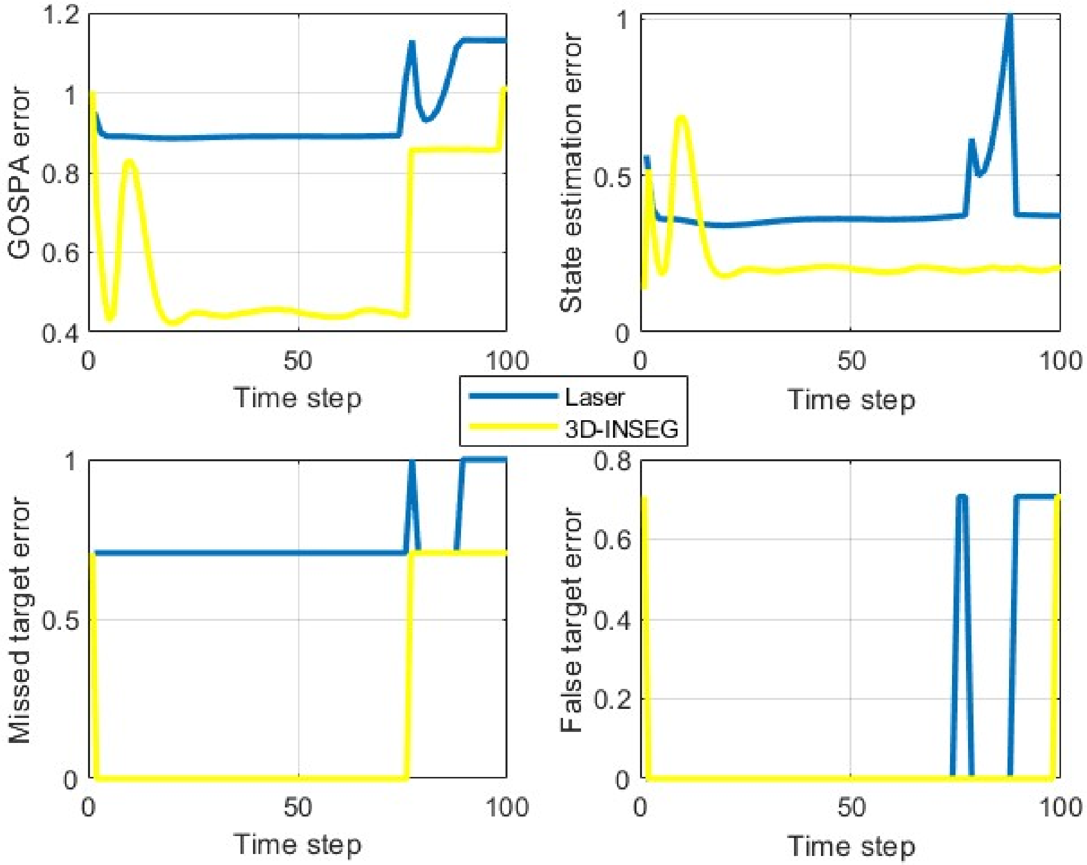


Figure 6.31: Experiment 2: GOSPA errors and their decomposition against time for the extended target scenario using the extended PMBM with laser data and the 3D-INSEG data.

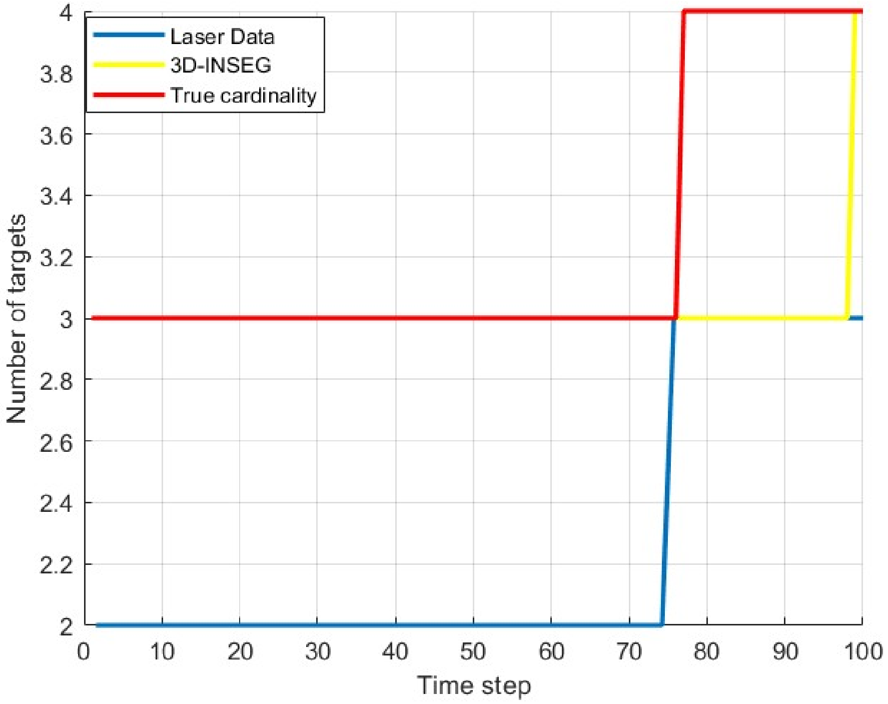
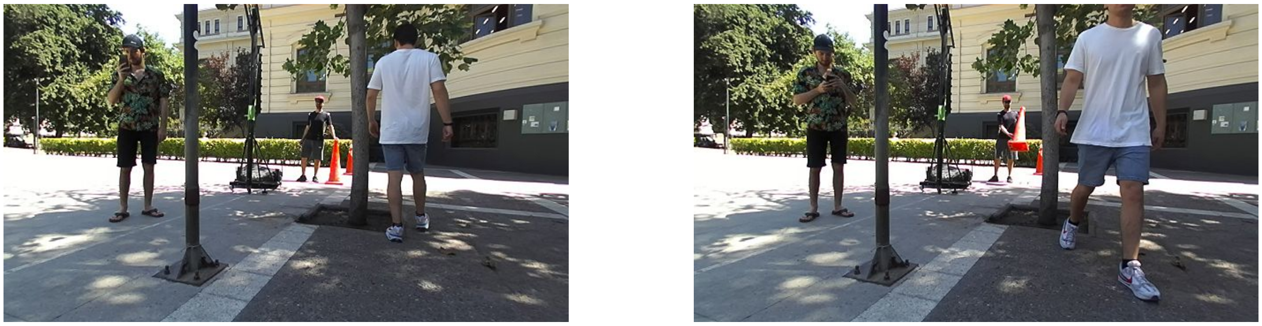


Figure 6.32: Experiment 2: Target cardinality estimated by the extended PMBM filter using the laser data and the 3D-INSEG data.



(a) Left camera at time $t = 8.59s$.

(b) Left camera at time $t = 17.31s$.

Figure 6.33: Experiment 3: Reference images from the sequence of the left camera images.

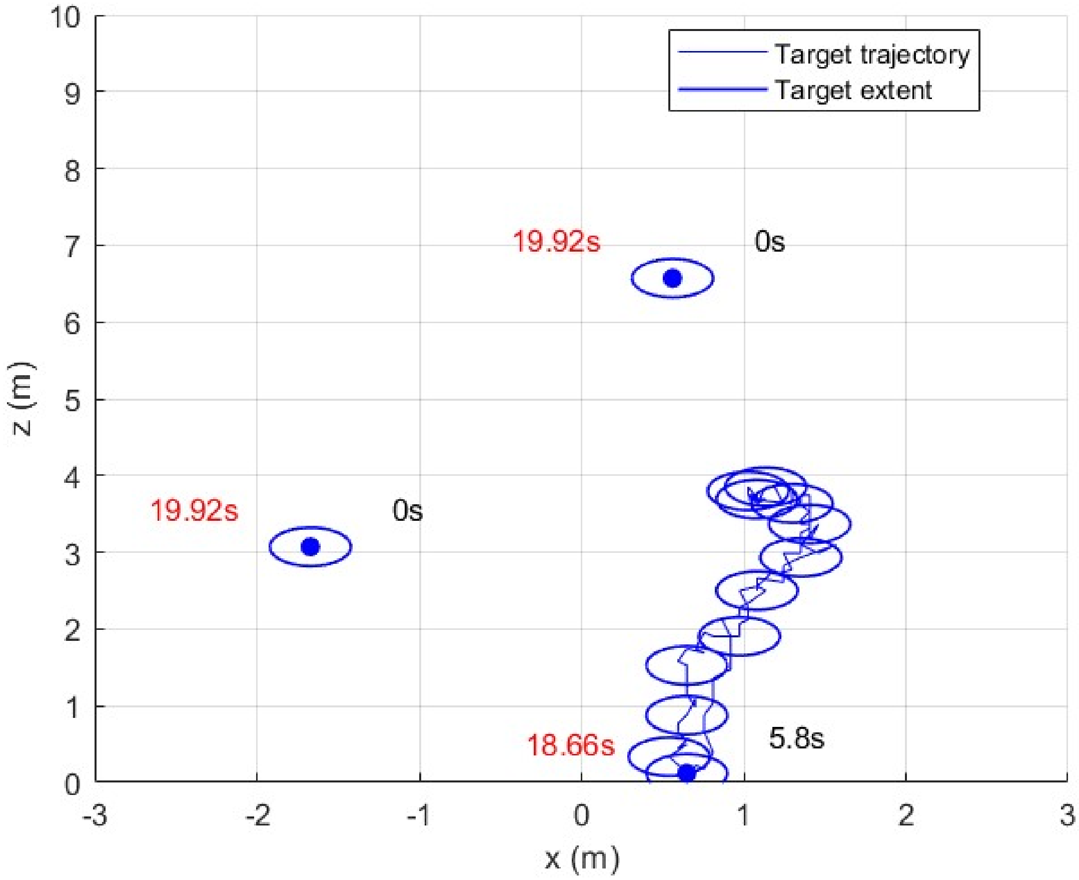
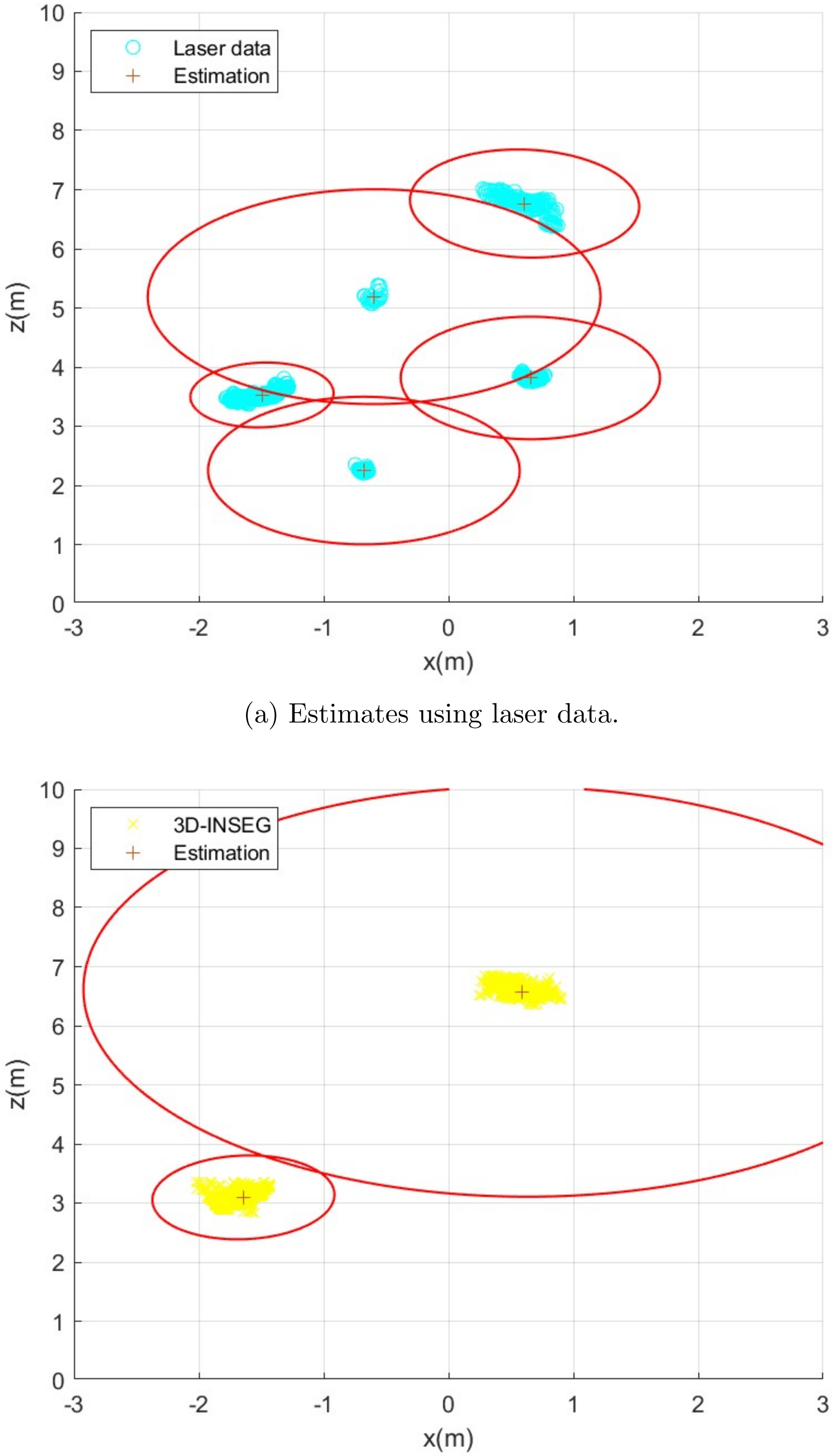


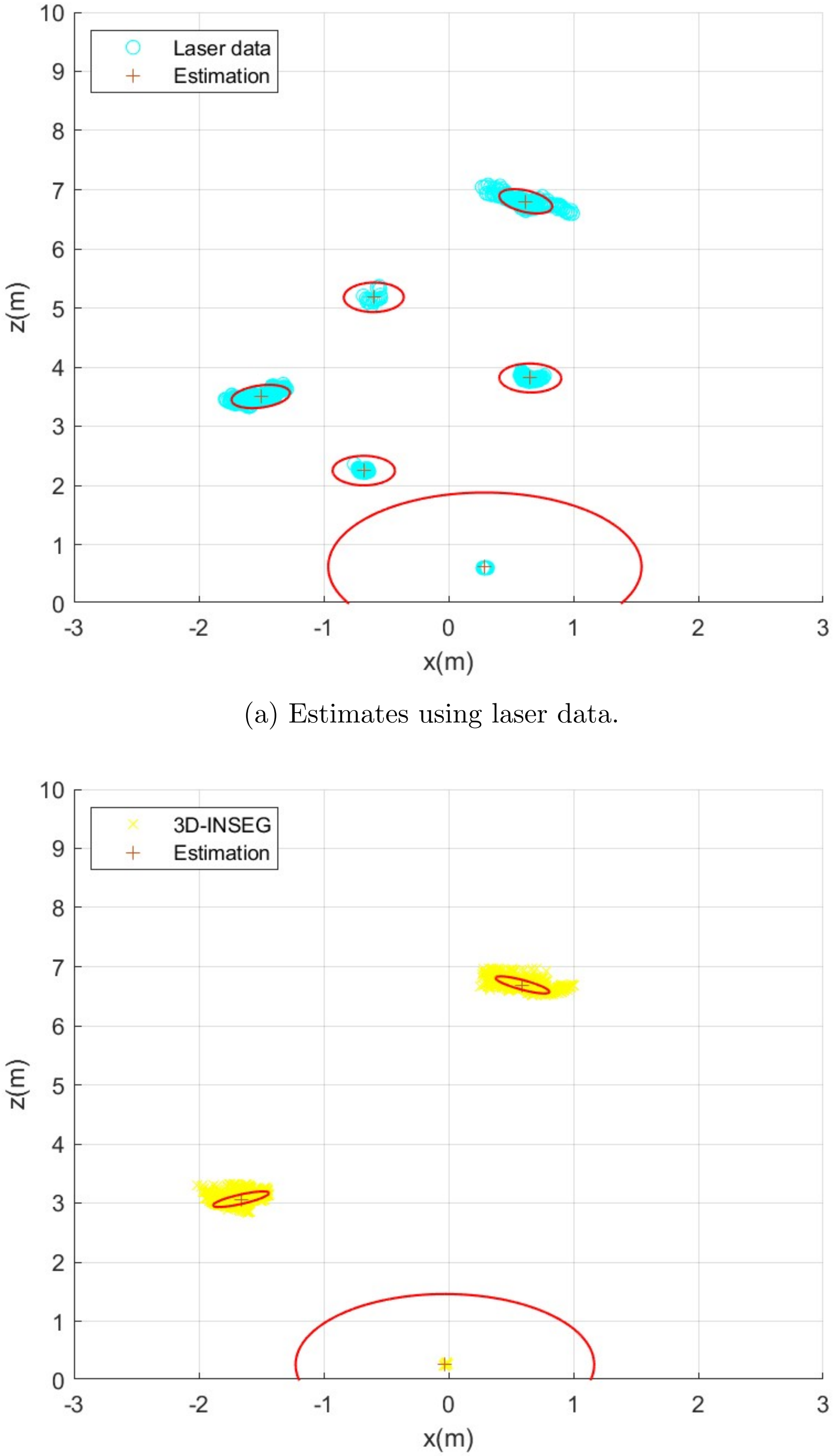
Figure 6.34: Experiment 3: Ground truth.



(a) Estimates using laser data.

(b) Estimates using 3D-INSEG data.

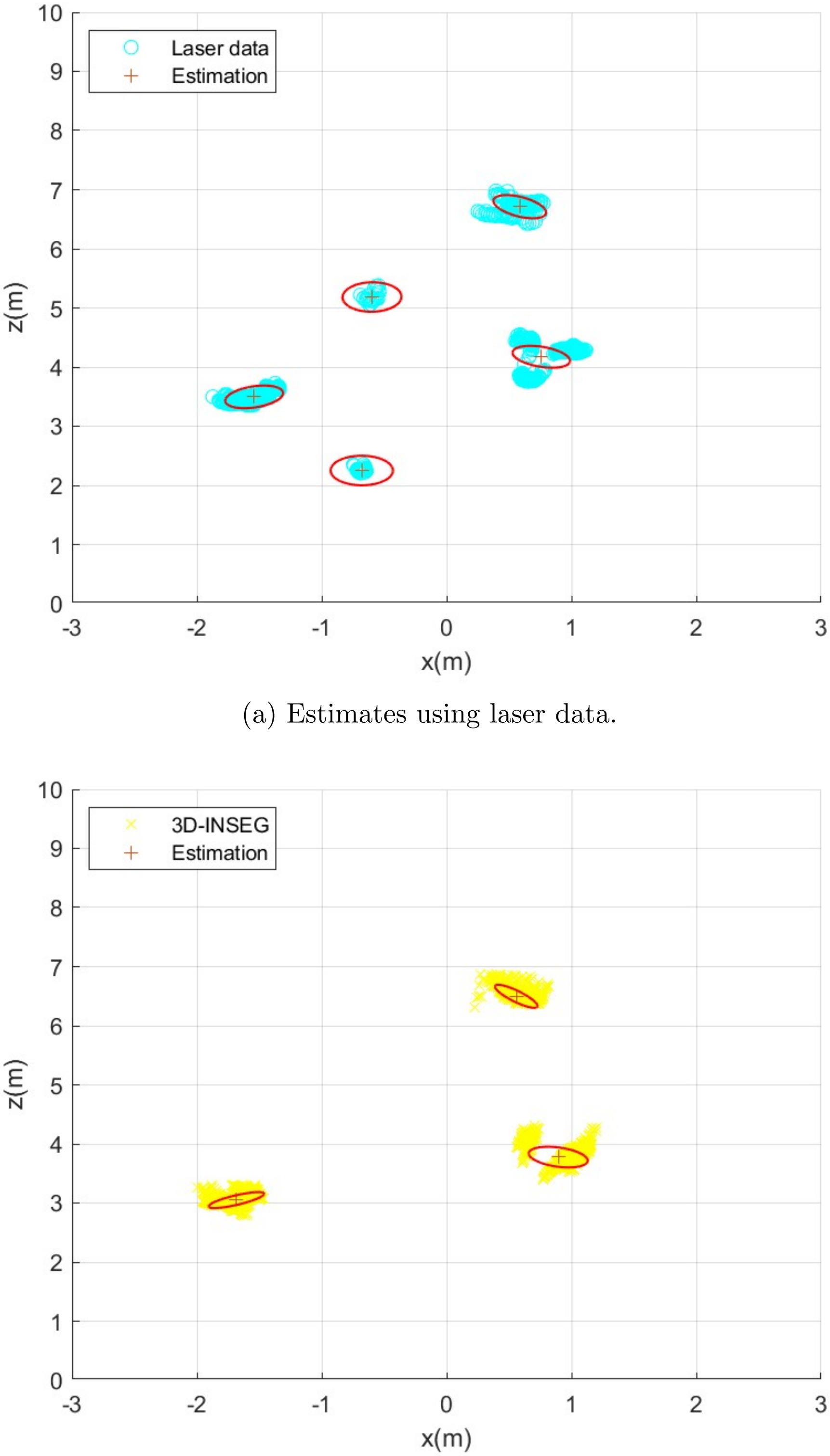
Figure 6.35: Experiment 3: Comparison of estimates at time t=0s.



(a) Estimates using laser data.

(b) Estimates using 3D-INSEG data.

Figure 6.36: Experiment 3: Comparison of estimates at time t=5.92s.



(a) Estimates using laser data.

(b) Estimates using 3D-INSEG data.

Figure 6.37: Experiment 3: Comparison of estimates at time t=11.72s.

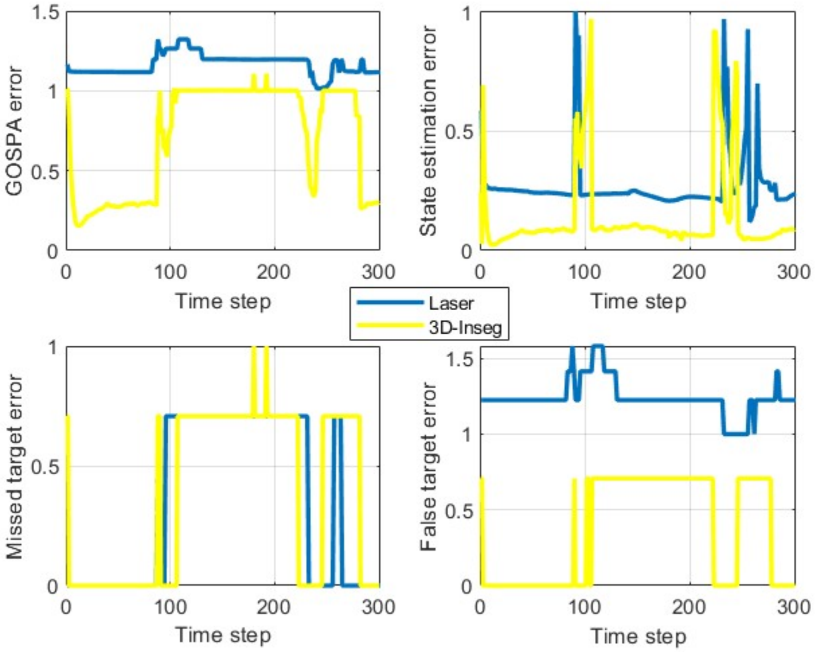


Figure 6.38: Experiment 3: GOSPA errors and their decomposition against time for the extended target scenario using the extended PMBM with laser data and the 3D-INSEG data.

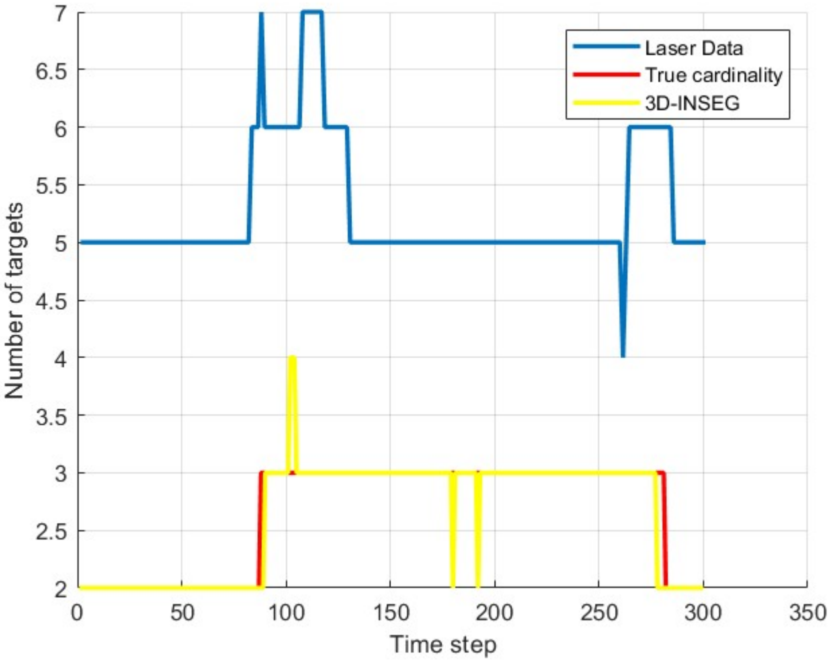
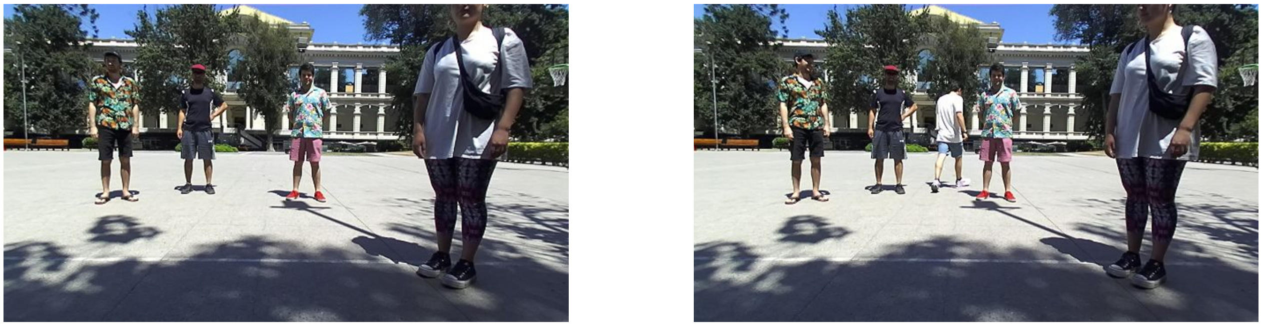


Figure 6.39: Experiment 3: Target cardinality estimated by the extended PMBM filter using the laser data and the 3D-INSEG data.



(a) Left camera at time $t = 0s$.

(b) Left camera at time $t = 7.26s$.

Figure 6.40: Experiment 4: Reference images from the sequence of the left camera images.

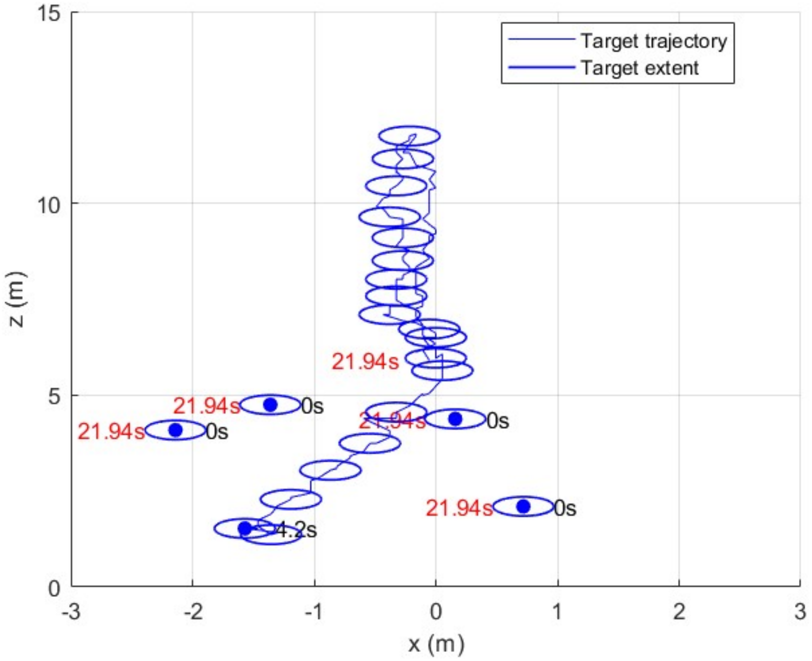
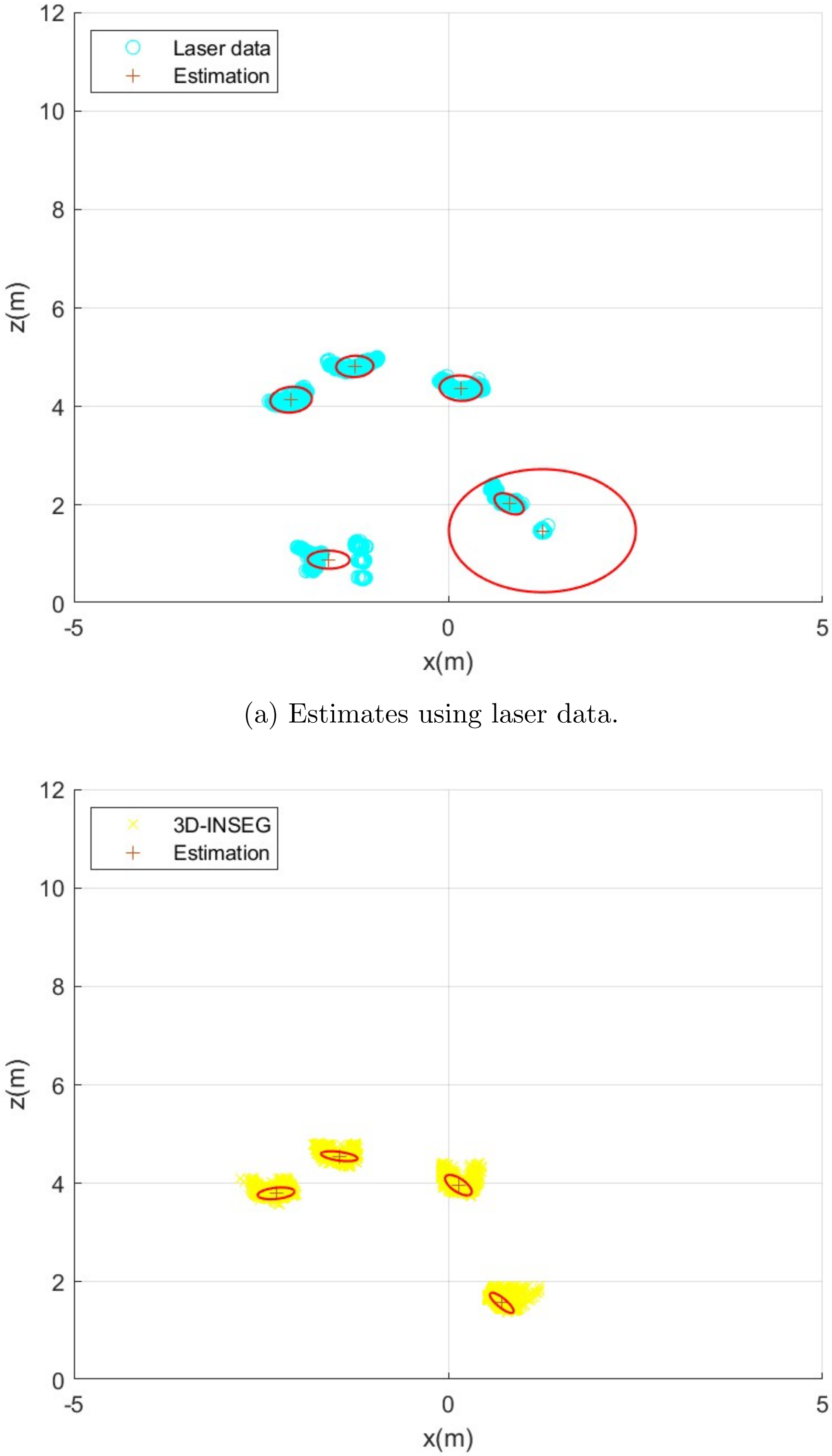


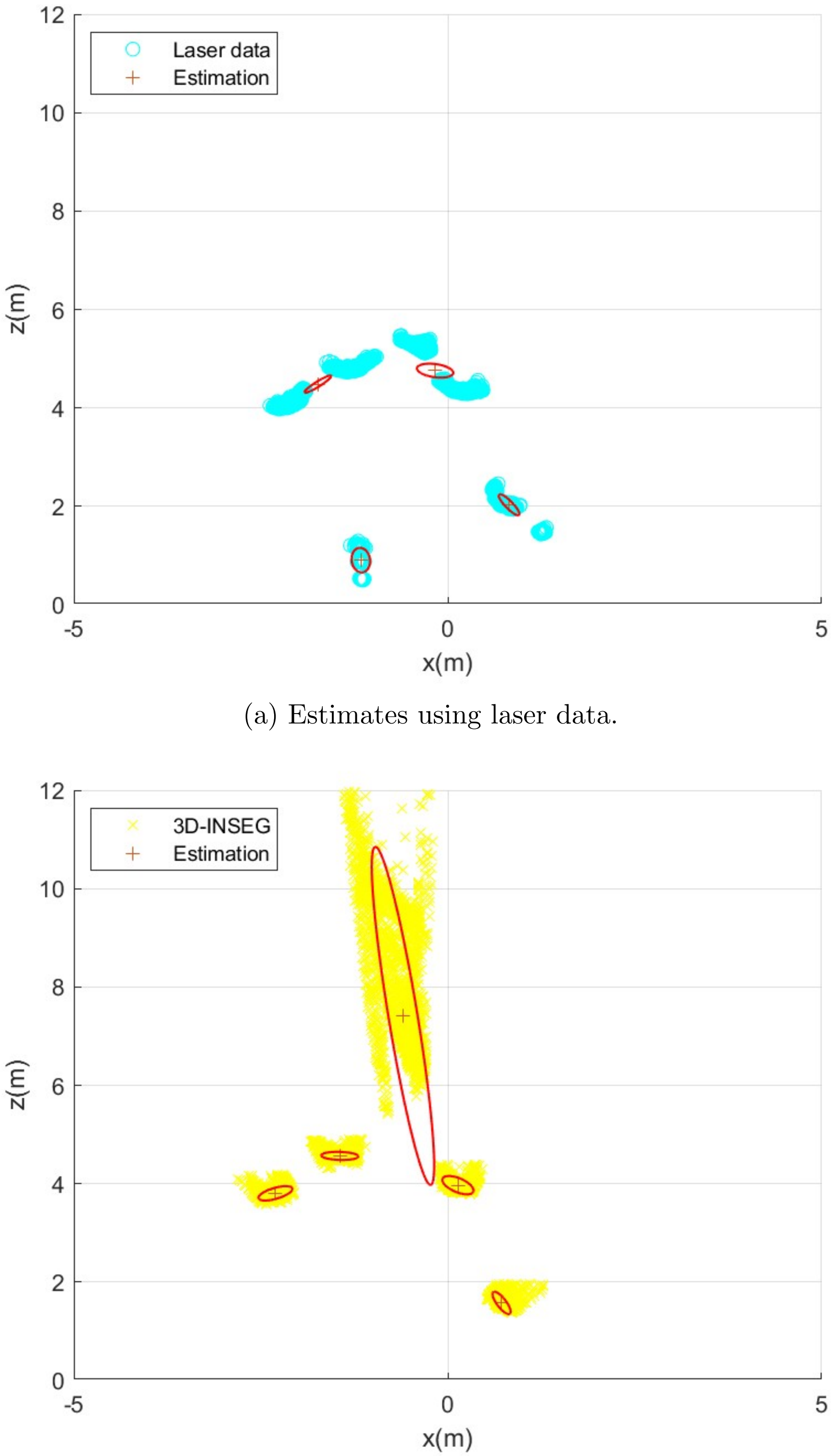
Figure 6.41: Experiment 4: Ground truth.



(a) Estimates using laser data.

(b) Estimates using 3D-INSEG data.

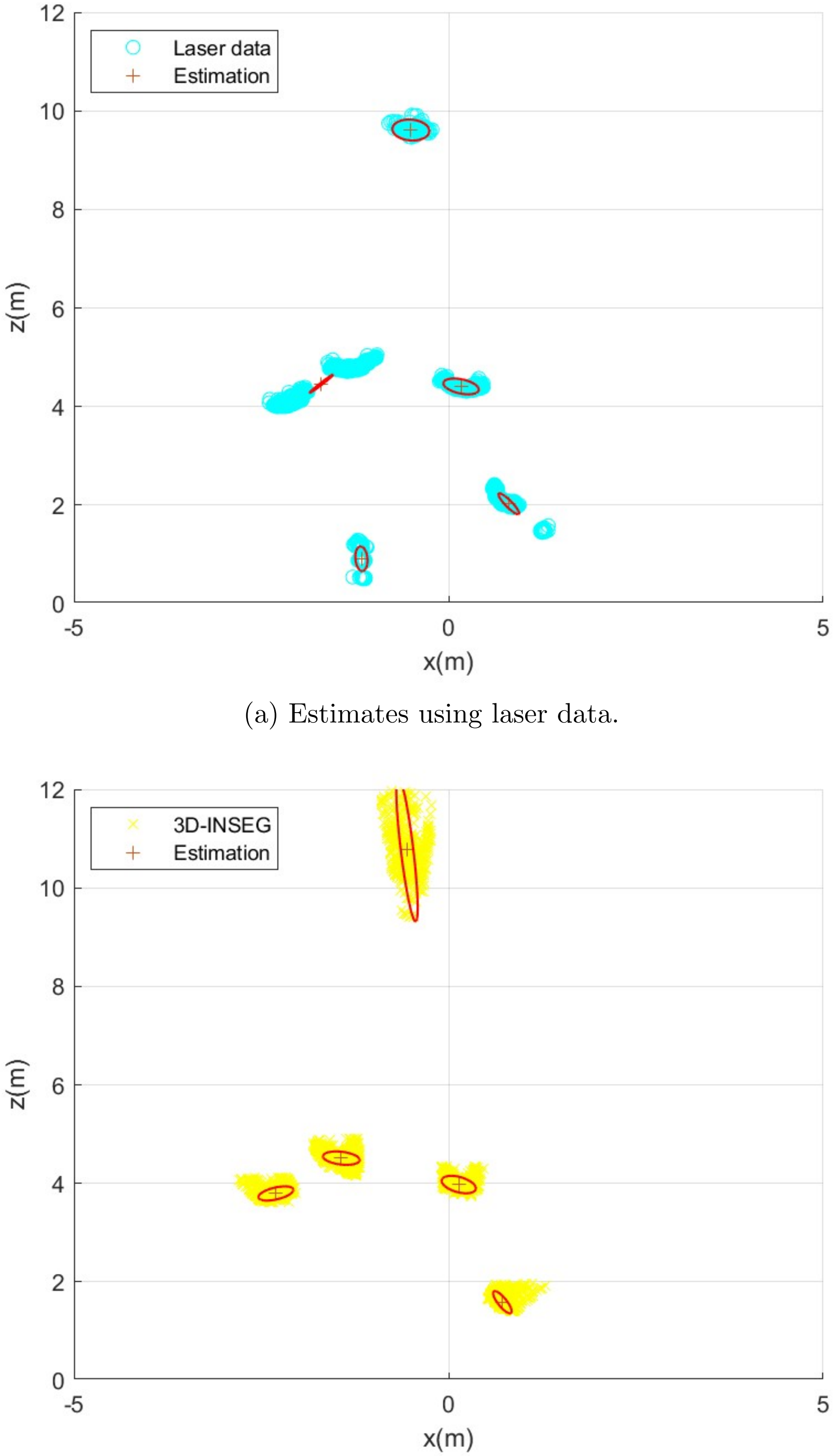
Figure 6.42: Experiment 4: Comparison of estimates at time t=0s.



(a) Estimates using laser data.

(b) Estimates using 3D-INSEG data.

Figure 6.43: Experiment 4: Comparison of estimates at time t=7.26s.



(a) Estimates using laser data.

(b) Estimates using 3D-INSEG data.

Figure 6.44: Experiment 4: Comparison of estimates at time t=14.53s.

with laser data as measurements, and yellow is used when it is performed using 3D-INSEG generated data.

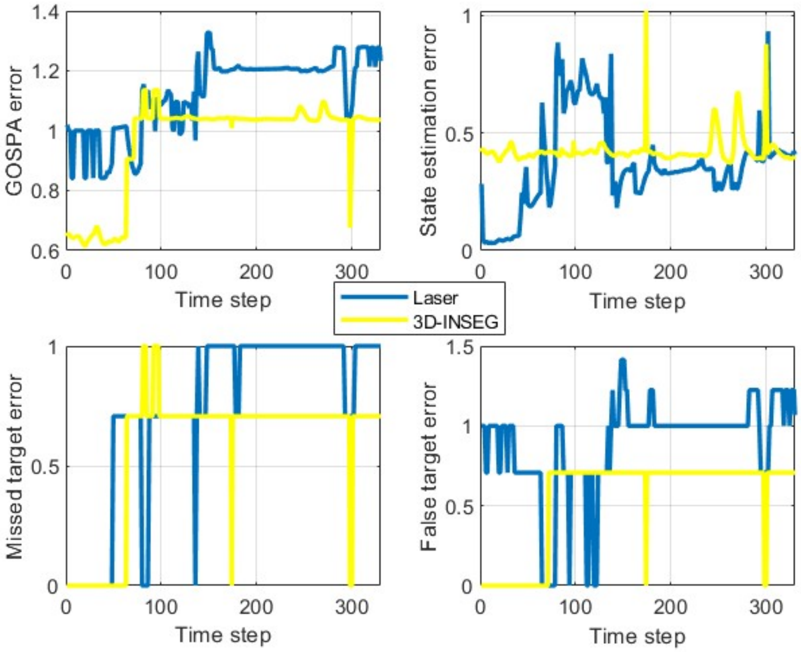


Figure 6.45: Experiment 4: GOSPA errors and their decomposition against time for the extended target scenario using the the extended PMBM with laser data and the 3D-INSEG data.

6.2.8 Discussion

From the MEOT simulation results presented in section 6.2.2, it is evident that the GGIW-PMBM filter effectively estimates tracks while maintaining good performance in terms of the GOSPA error. In Simulation 1, which involves three targets with angular and linear velocity, the estimates from the filter exhibit a low GOSPA error, albeit slightly higher at the beginning due to birth parameters. Moreover, the cardinality estimate from the filter remains accurate after the initial time steps.

In Simulation 2, where tracks are augmented, the MEOT task becomes more challenging, resulting in higher GOSPA error and less accurate estimated cardinality. The introduction of augmented tracks complicates the tracking process, reflected in the observed metrics.

Simulation 3 presents even more frequent peaks in errors, particularly in estimated cardinality, which aligns with expectations given the higher number of tracks being handled.

These simulation findings provide valuable insights into the performance of the GGIW-PMBM filter under varying conditions, highlighting its strengths and areas for improvement in handling complex multi-object tracking scenarios.

From the experimental results presented in Section 6.2.3, it is evident that the multi-

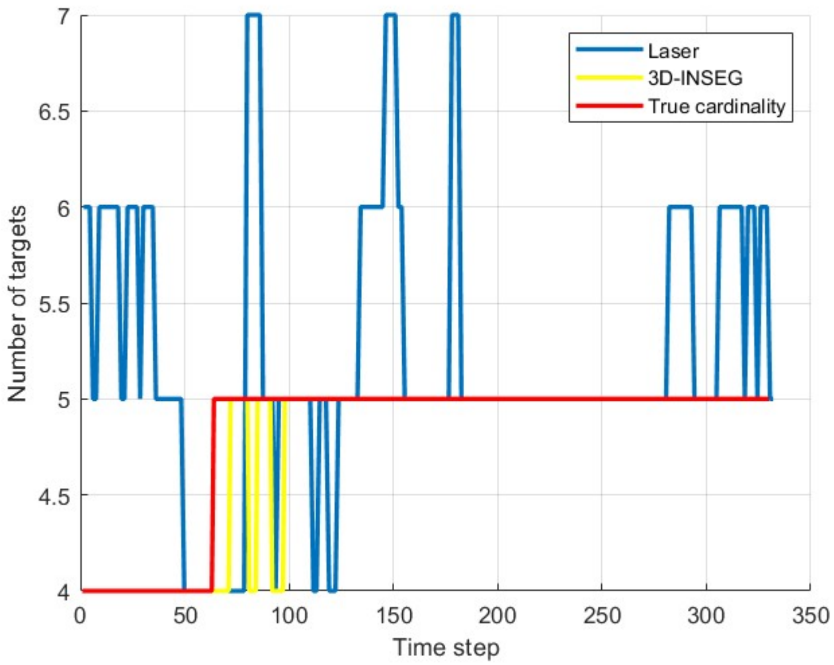


Figure 6.46: Experiment 4: Target cardinality estimated by the extended PMBM filter using the laser data and the 3D-INSEG data.

object extended object tracking (MEOT) task becomes significantly challenging in denser scenarios with closer laser points, such as indoor environments where a substantial amount of data is generated. In Experiment 1, despite reducing the surveillance area, the filter fails to track two persons effectively. Conversely, leveraging the 3D-INSEG algorithm enables accurate detection of persons and uses the detected points as measurements, leading to more correct estimates. This improvement is reflected in the metrics, with the GOSPA error being lower when utilizing 3D-INSEG data.

When employing the 3D-INSEG algorithm-generated data, the GGIW-PMBM filter exhibits fewer false positives and produces estimates with lower errors compared to estimates derived from laser data using the same filter. Additionally, across all experiments, the filter operates more efficiently due to the avoidance of DB-SCAN clustering.

These experimental findings underscore the effectiveness of integrating the 3D-INSEG algorithm into the MEOT framework, particularly in dense environments where traditional laser-based approaches encounter limitations. The ability of 3D-INSEG to provide accurate object detection and segmentation significantly enhances tracking performance and efficiency.

Chapter 7

Conclusion

This study has demonstrated the effectiveness of employing the 3D-INSEG algorithm in densely populated environments, emphasizing the critical role of object detection and segmentation in robust tracking. The integration of stereo vision with neural networks (NNs) for 3D depth estimation and segmentation, as showcased by the 3D-INSEG algorithm, has led to significant advancements in both single extended object tracking (SEOT) and multi-object extended object tracking (MEOT).

The experimental findings presented in Section 6.2.3 highlight the benefits of leveraging 3D-INSEG in high-density scenarios where traditional laser-based methods encounter limitations due to data volume. The 3D-INSEG algorithm not only enables accurate object detection and segmentation but also enhances the performance of the extended GGIW-PMBM filter, resulting in fewer false positives and reduced estimation errors compared to laser-based techniques. This integrated approach presents a promising solution for robust and dependable object tracking across diverse conditions.

Moreover, the successful implementation of the SEOT algorithm, complemented by the 3D-INSEG algorithm, underscores the potential of these methodologies to provide precise and reliable estimates in simulation and real-world tracking scenarios. The detailed object identity information offered by 3D-INSEG improves extended object tracking, serving as a solid foundation for state estimation in complex environments.

In conclusion, our hypothesis regarding the integration of stereo vision and NN-based segmentation has been validated through experimental validation. This approach not only enhances tracking accuracy and reduces computational complexity but also opens new possibilities for advanced tracking applications in environments with complex object shapes and high clutter conditions.

7.1 Future Work

An interesting avenue for future research involves exploring alternative algorithms for extended object tracking that offer distinct approaches to modeling object extensions beyond traditional methods, as discussed in [42], [43], and [44]. Conducting comparative studies to evaluate these methodologies using the proposed 3D-INSEG algorithm data would provide valuable insights into the strengths and limitations of different target shape and motion modeling concepts. This analysis could inform the development of more robust and versatile tracking frameworks capable of addressing diverse object tracking challenges.

Furthermore, there is potential for investigating a hybrid approach that integrates lidar measurements with 3D-INSEG detections, considering the density of targets in the environment. By combining the long-range capabilities of lidar with the precise object detection capabilities of the 3D-INSEG algorithm, researchers can aim to achieve superior tracking performance across varying environmental conditions. This hybridization strategy offers an opportunity to optimize object detection and segmentation, enhancing the overall effectiveness and reliability of extended object tracking systems.

Additionally, exploring advancements in machine learning techniques for object recognition and motion prediction could further enhance the capabilities of integrated tracking systems. By leveraging cutting-edge methodologies in neural networks and deep learning, researchers can push the boundaries of object tracking accuracy and efficiency, ultimately advancing the state-of-the-art in extended object tracking.

In summary, future research endeavors should focus on exploring alternative tracking algorithms, integrating sensor data fusion strategies, and leveraging advanced machine learning techniques to enhance the performance, robustness, and adaptability of extended object tracking systems in complex and dynamic environments.

Bibliography

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[2] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*, pp. 218–227, 2021.

[3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE access*, vol. 8, pp. 58443–58469, 2020.

[4] F. Engels, P. Heidenreich, M. Wintermantel, L. Stäcker, M. Al Kadi, and A. M. Zoubir, "Automotive radar signal processing: Research directions and practical challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 865–878, 2021.

[5] G. Hakobyan and B. Yang, "High-performance automotive radar: A review of signal processing algorithms and modulation schemes," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 32–44, 2019.

[6] Y. Li and J. Ibanez-Guzman, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, 2020.

[7] K. Granström and M. Baum, "A tutorial on multiple extended object tracking," *Authorea Preprints*, 2023.

[8] Y. Bar-Shalom and W. D. Blair, "Multitarget-multisensor tracking: applications and advances," *(No Title)*, 1992.

[9] K. Granström, M. Fatemi, and L. Svensson, "Poisson multi-bernoulli conjugate prior for multiple extended object estimation," *ArXiv e-prints*, 2016.

[10] M. Beard, S. Reuter, K. Granström, B.-T. Vo, B.-N. Vo, and A. Scheel, "Multiple extended target tracking with labeled random finite sets," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1638–1653, 2015.

[11] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, pp. 226–231, 1996.

[12] J. L. Williams, "Marginal multi-Bernoulli filters: RFS derivation of MHT, JIPDA, and association-based MeMBer," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 3, pp. 1664–1687, 2015.

[13] Á. F. García-Fernández, Y. Xia, and L. Svensson, "Poisson multi-bernoulli mixture filter with general target-generated measurements and arbitrary clutter," *IEEE Transactions on Signal Processing*, 2023.

[14] K. Granström, M. Baum, and S. Reuter, "Extended object tracking: Introduction, overview, and applications," *Journal of Advances in Information Fusion*, vol. 12, 12 2017.

[15] K. G. Murty, "An algorithm for ranking all the assignment in order of increasing cost," *Operations Research*, vol. 16, 1968.

[16] M. Poggi, F. Tosi, K. Batsos, P. Mordohai, and S. Mattoccia, "On the synergies between machine learning and binocular stereo for depth estimation from images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5314–5334, 2022.

[17] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[18] M. Feldmann, D. Fränken, and W. Koch, "Tracking of extended objects and group targets using random matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1409–1420, 2011.

[19] N. Fierro, M. Adams, and L. Cament, "3D-INSEG: A 3D Instance Segmentation Algorithm for Extended Object Tracking," in *2023 12th International Conference on Control, Automation and Information Sciences (ICCAIS)*, (Vietnam, Hanoi.), pp. 704–711, 11 2023.

[20] R. Mahler, "Phd filters for nonstandard targets, i: Extended targets," in *2009 12th International Conference on Information Fusion*, pp. 915–921, IEEE, 2009.

[21] K. Granstrom and U. Orguner, "A phd filter for tracking multiple extended targets using random matrices," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5657–5671, 2012.

[22] K. Granström and U. Orguner, "Estimation and maintenance of measurement rates for multiple extended target tracking," in *2012 15th International Conference on Information Fusion*, pp. 2170–2176, IEEE, 2012.

[23] C. Lundquist, K. Granström, and U. Orguner, "An extended target cphd filter and a gamma gaussian inverse wishart implementation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 3, pp. 472–483, 2013.

[24] K. Granström, M. Fatemi, and L. Svensson, "Gamma gaussian inverse-wishart poisson multi-bernoulli filter for extended target tracking," in *2016 19th International Conference on Information Fusion (FUSION)*, pp. 893–900, IEEE, 2016.

[25] J. W. Koch, “Bayesian approach to extended object and cluster tracking using random matrices,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 3, pp. 1042–1059, 2008.

[26] A. Gupta and D. Nagar, “Monographs and surveys in pure and applied mathematics,” in *Matrix Variate Distributions*, Chapman and Hall, 2000.

[27] D. A. Harville, “Matrix algebra from a statistician’s perspective,” 1998.

[28] K. Granström and U. Orguner, “Estimation and maintenance of measurement rates for multiple extended target tracking,” in *2012 15th International Conference on Information Fusion*, pp. 2170–2176, 2012.

[29] R. Mahler, *Statistical multisource-multitarget information fusion.* Artech, 2007.

[30] R. P. Mahler, *Advances in statistical multisource-multitarget information fusion.* Artech House, 2014.

[31] K. Granström, M. Fatemi, and L. Svensson, “Poisson multi-bernoulli mixture conjugate prior for multiple extended target filtering,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 1, pp. 208–225, 2019.

[32] K. Gilholm and D. Salmond, “Spatial distribution model for tracking extended objects,” *IEE Proceedings-Radar, Sonar and Navigation*, vol. 152, no. 5, pp. 364–371, 2005.

[33] K. Granstrom, C. Lundquist, and O. Orguner, “Extended target tracking using a gaussian-mixture phd filter,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 3268–3286, 2012.

[34] K. Wyffels and M. Campbell, “Negative information for occlusion reasoning in dynamic extended multiobject tracking,” *IEEE Transactions on Robotics*, vol. 31, no. 2, pp. 425–442, 2015.

[35] S. Reuter and K. Dietmayer, “Pedestrian tracking using random finite sets,” in *14th International Conference on Information Fusion*, pp. 1–8, IEEE, 2011.

[36] K. Granström, M. Fatemi, and L. Svensson, “Poisson multi-bernoulli mixture conjugate prior for multiple extended target filtering,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 1, pp. 208–225, 2020.

[37] A. M. Hafiz and G. M. Bhat, “A survey on instance segmentation: state of the art,” *International journal of multimedia information retrieval*, vol. 9, no. 3, pp. 171–189, 2020.

[38] L. Cament, M. Adams, J. Correa, and C. Perez, “The δ-generalized multi-bernoulli poisson filter in a multi-sensor application,” in *2017 International Conference on Control, Automation and Information Sciences (ICCAIS)*, pp. 32–37, IEEE, 2017.

[39] S. Yang, M. Baum, and K. Granström, “Metrics for performance evaluation of elliptic extended object tracking methods,” in *2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 523–528, 2016.

[40] A. S. Rahmathullah, Á. F. García-Fernández, and L. Svensson, "Generalized optimal sub-pattern assignment metric," in *2017 20th International Conference on Information Fusion (Fusion)*, pp. 1–8, IEEE, 2017.

[41] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

[42] J. S. Fowdur, M. Baum, and F. Heymann, "Tracking targets with known spatial extent using experimental marine radar data," in *2019 22th International Conference on Information Fusion (FUSION)*, pp. 1–8, IEEE, 2019.

[43] N. Wahlström and E. Özkan, "Extended target tracking using gaussian processes," *IEEE Transactions on Signal Processing*, vol. 63, no. 16, pp. 4165–4178, 2015.

[44] M. Baum and U. D. Hanebeck, "Extended object tracking with random hypersurface models," *IEEE Transactions on Aerospace and Electronic systems*, vol. 50, no. 1, pp. 149–159, 2014.

[45] S. Reuter, B.-T. Vo, B.-N. Vo, and K. Dietmayer, "The labeled multi-bernoulli filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3246–3260, 2014.

Annex A

Camera Intrinsic Parameters and Camera Matrix

The camera intrinsic parameters include the focal lengths (f_x, f_y) and the principal points (c_x, c_y), which are used to define the camera matrix (K). For the zed camera at 640x360 resolution these values in pixels are:

$$f_x = 342, \quad f_y = 342, \quad c_x = 308, \quad c_y = 183$$

The camera matrix K is defined as:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

This matrix represents the intrinsic properties of the camera, including the focal lengths and principal points, which are essential for tasks such as image rectification, stereo vision, and 3D reconstruction. The baseline distance is 120mm.

Annex B

3D-INSEG detections animation

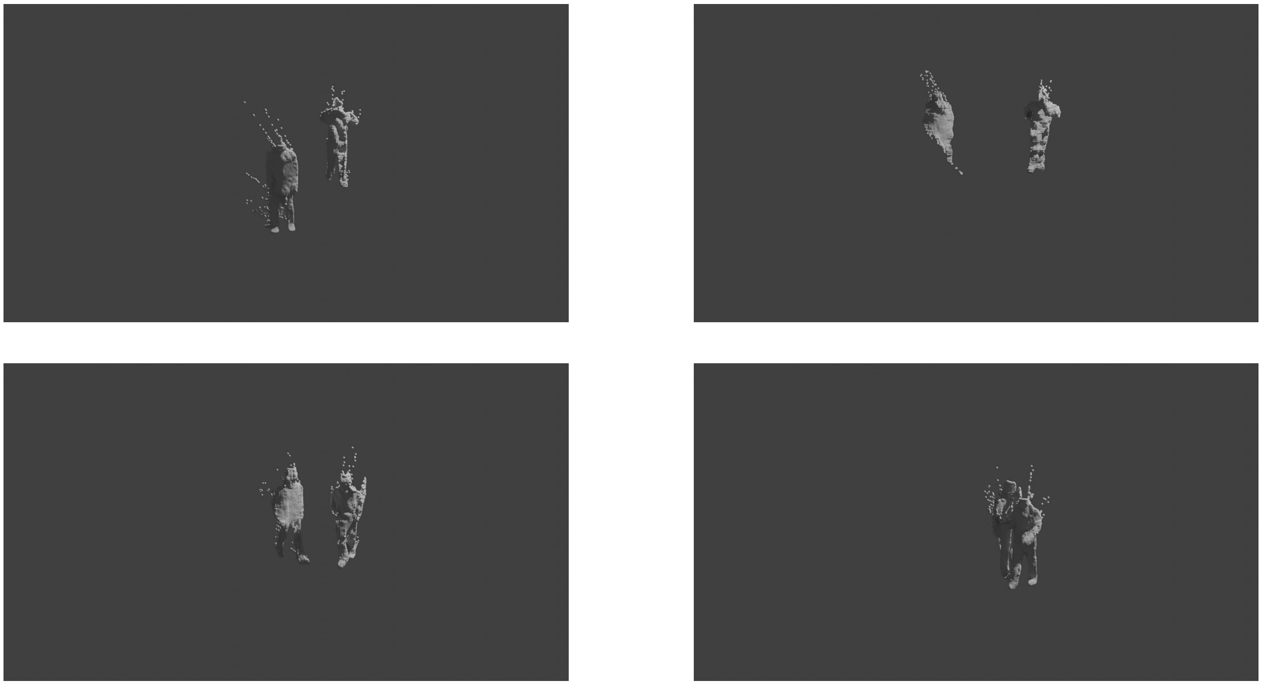


Figure B.1: Experiment 1: 3D-INSEG detections animation.

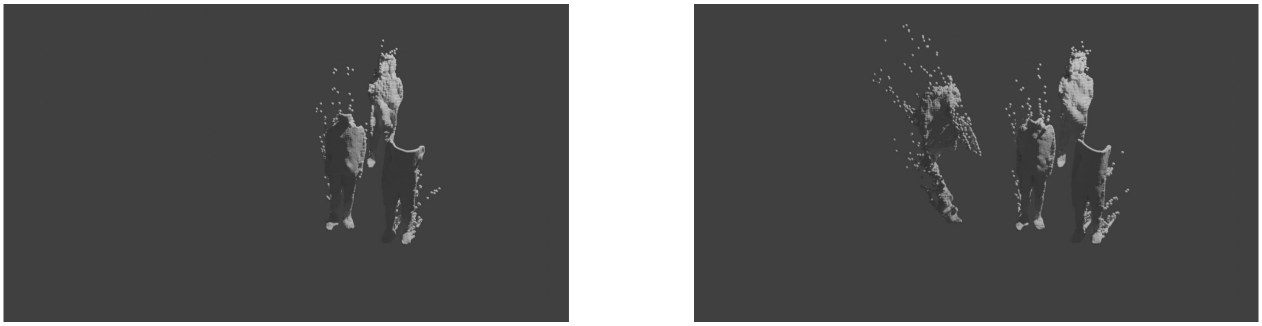


Figure B.2: Experiment 2: 3D-INSEG detections animation.

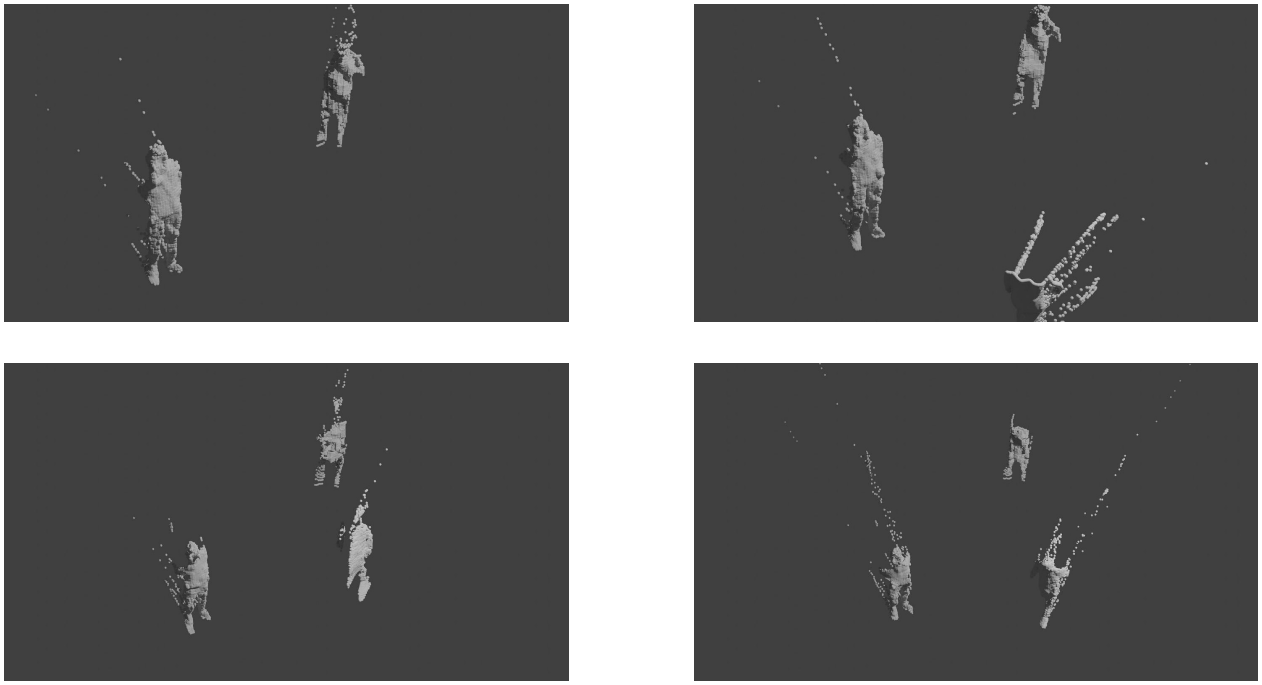


Figure B.3: Experiment 3: 3D-INSEG detections animation.

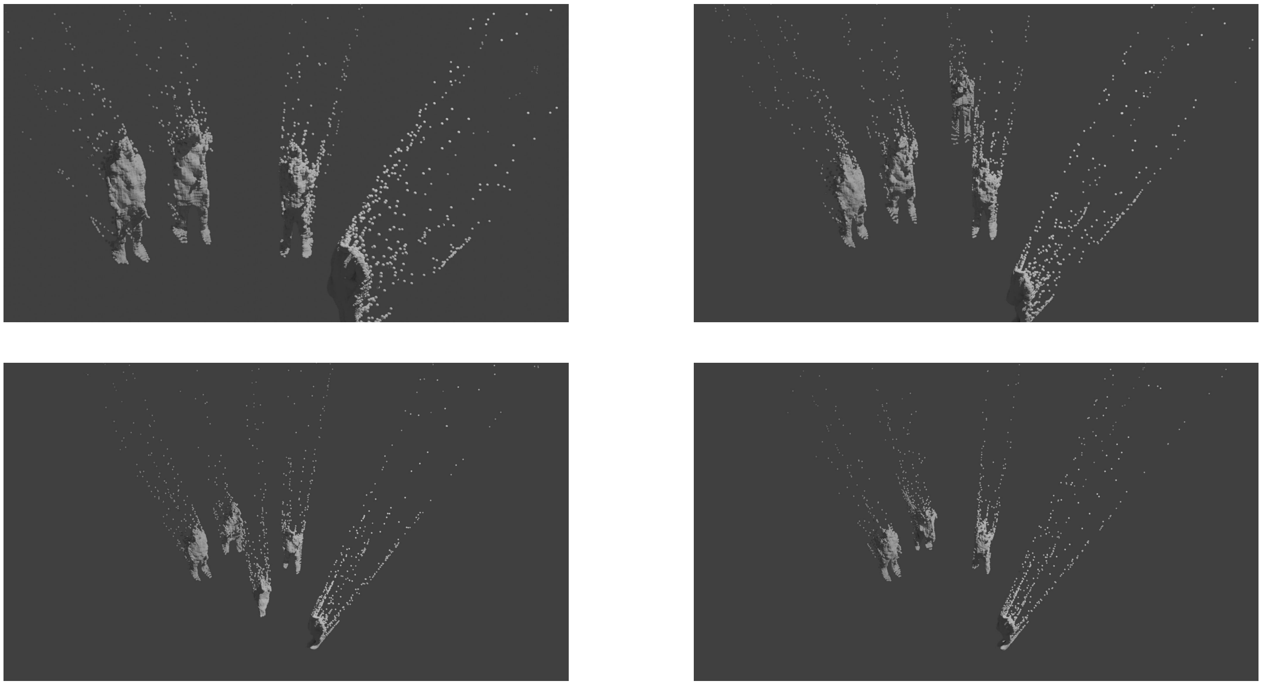


Figure B.4: Experiment 4: 3D-INSEG detections animation.