Extended Target Tracking with 3D-INSEG and its Benefits in Dense Scenarios

1st Nicolás Fierro
Department of Electrical Engineering
Universidad de Chile
Santiago , Chile
nicolas.fierro@ug.uchile.cl

2nd Martin Adams
Department of Electrical Engineering
Universidad de Chile
Santiago , Chile
martin@ing.uchile.cl

3rd Leonardo Cament
Department of Electrical Engineering
Universidad de Chile
Santiago , Chile
lcament@ing.uchile.cl

***Abstract*—In Multiple Extended Object Tracking (MEOT), it is assumed that a solitary target can produce multiple measurements. The quality of these measurements is paramount for obtaining accurate estimates of tracks over time. To test state-of-the-art MEOT algorithms, both simulated and real laser data, recorded in open spaces, have been used. MEOT algorithms work well in these scenarios, but when applied in more cluttered or restricted spaces, they often fail to produce good results, because close proximity target measurements are considered as measurements with the same origin. To address these cases, this article applies the 3D INstance SEGmentation (3D-INSEG) algorithm to MEOT to process stereo image sequences, extracting 3D information corresponding to each detected target using cameras. The algorithm selects pixels from each detected target and calculates the disparity map from stereo pairs, projecting them into 3D space using this disparity map. Subsequently, these measurements undergo processing by an extended target Poisson multi-Bernoulli mixture (PMBM) filter with a gamma Gaussian inverse-Wishart (GGIW) implementation. The advantages of MEOT with the 3D-INSEG-generated data are demonstrated in this article via a comparison with MEOT based on Velodyne LiDAR data points recorded from the same scenario processed by the same MEOT algorithm.**

***Index Terms*—Stereo vision, tracking, Multiple Extended Object Tracking, Depth Estimation, Segmentation.**

I. Introduction

Autonomous systems, such as self-driving vehicles [1], rely on robust environmental perception to identify and monitor moving objects, using scan-based sensors such as radar [2] [3] and LiDAR [4]. These sensors provide precise point measurements but lack information about object types and are prone to noise, making the development of algorithms to extract labeled object tracks, known as Multiple Object Tracking (MOT), imperative. High-resolution sensors result in many detections for a single object, creating the challenge of tracking multiple objects (MEOT), where individual objects may generate multiple detections without labels. In MEOT, determining an object's spatial boundaries is crucial for distinguishing between multiple objects and ensuring robust tracking.

The PMBM filter [5], [6], has demonstrated its status as a robust method in target tracking. This filter is grounded in the concept of Random Finite Set (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 GGIW implementation, is detailed in [7]. This version utilizes a two-step clustering and assignment approach [8], to identify relevant global hypotheses during the update of each preceding global hypothesis. The process involves applying Density-Based Spatial Clustering of Applications with Noise (DB-SCAN) algorithm [9] and subsequently employing Murty's algorithm [10] for each measurement partition and global hypothesis, facilitating the determination of optimal cluster-to-Bernoulli component assignments.

While this implementation [7] 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.

In this article, we apply and implement the 3D-INSEG algorithm [11] 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 laser data alone is employed.

The remainder of the paper is structured as follows. In Section II, previous related work is presented. Section III demonstrates problems faced in MEOT when determining the data association and alternatives to solve it. In Section IV the application of 3D-INSEG in MEOT filtering is presented. Section V presents and discusses the results. Finally, Section VI presents the main conclusions of this work and future steps.

II. previous work

A. The 3D-INSEG Algorithm

The 3D-INSEG [11] algorithm generates 3D instance segments from pairs of stereo images. This process involves several key steps:

1) Undistortion: This corrects distortions in the stereo images, ensuring accurate representation of the scene. The distortion model in the context of the 3D-INSEG algorithm relies on the camera matrix K. This matrix encapsulates the intrinsic properties of the camera and is established through the calibration process, taking into account parameters such as the focal length and principal point. K is defined as:

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

where, f_x and f_y denote the focal lengths along the x and y axes within the camera coordinate system. The values c_x and c_y represent the coordinates of the principal points, signifying the optical center of the image. The third row is consistently set to $0, 0, 1$ to ensure homogeneity in the matrix.

2) 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:

$$M(i,j) = \begin{cases} 1, & \text{if pixel } (i,j) \text{ is part of the object} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

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:

$$\mathcal{M}_{\text{object}} = \{(i,j) \,|\, M(i,j) = 1\}. \tag{3}$$

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

$$\mathcal{M}_{\text{object}} = \{p_1, p_2, \ldots, p_n\}. \tag{4}$$

3) Depth estimation determines the depth information for each pixel, providing a 3D representation of the scene.
4) 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:

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \text{depth} \times \text{inv}(K) \times \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{5}$$

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.

5) 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.

B. MEOT

MEOT algorithms not only estimate the kinematic state of an object, but they also estimate its shape and orientation. In [12], Single Extended Object Tracking (SEOT) algorithms and their MEOT extensions are surveyed. The initial theoretical foundation for a multiple extended object tracker was established based on the Probability Hypothesis Density (PHD) filter, incorporating the aforementioned PPP model [13]. Subsequently, this filter was implemented utilizing the random matrix model, with an inverse Wishart distribution employed for estimating the shape matrix [14]. 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 [15]. The integration of these components led to the development of the GGIW model [16]. The PMBM filter [5] has been adapted for extended object tracking with a GGIW formulation [17].

III. 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 [12], the number of possible associations, considering a set of previously detected objects $\mathcal{I}$ with $|\mathcal{I}|$ elements and a set of measurements $\mathcal{Z}$[1] with $|\mathcal{Z}|$ elements, is determined by the number of set partitions of $\mathcal{I} \cup \mathcal{Z}$.A cell refers to a subset of measurements that are grouped together. Each cell contains measurements that are likely to belong to the same object. The number of cell-to-object assignments is given by:

$$N_{\mathcal{A}}(|\mathcal{Z}|, |\mathcal{I}|) = \sum_{C=1}^{|\mathcal{Z}|} \left[\left\{ {|\mathcal{Z}| \atop C} \right\} \sum_{T=0}^{\min(C,|\mathcal{I}|)} \binom{C}{T} \binom{|\mathcal{I}|}{T} T! \right]. \tag{6}$$

Here, $\left\{ {|\mathcal{Z}| \atop C} \right\}$ denotes the Stirling number of the second kind, defined as $\left\{ {|\mathcal{Z}| \atop C} \right\} = \frac{1}{C!} \sum_{j=0}^{C} (-1)^{C-j} \binom{C}{j} j^{|\mathcal{Z}|}$. 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

[1] $\mathcal{Z}$ differs from $\mathcal{D}$ because for $\mathcal{Z}$ there is no notion of clusters, and measurements can be from different objects. Conversely, for $\mathcal{D}$, each element represents a cluster of points, where all the points within a cluster correspond to measurements of one specific object.

subsets with cardinality T that can be formed with C objects. Equation (6) 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.

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

$$N_A = \sum_{j \in \mathcal{J}} N_{A_j}(|\mathcal{Z}|, |\mathcal{I}_j|). \tag{7}$$

In [18] 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 $J_0 = \{j_1\}$, $W_{j_1}^0 = 1$, and $\mathcal{I}_{j_1}^0 = \emptyset$, 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,

$$|J_1| = N_{Aj_1}(|\mathcal{Z}_1|, 0) = \sum_{C=1}^{|\mathcal{Z}_1|} \binom{|\mathcal{Z}_1|}{C} = B(|\mathcal{Z}_1|), \tag{8}$$

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:

$$|J_k|^k = |J_{k+1}|^k = B\left(\sum_{t=1}^{k} |\mathcal{Z}_t|\right) = B\left(|\mathcal{Z}|^k\right). \tag{9}$$

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:

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

In articles such as [7] and [18], 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 DB-SCAN clustering algorithm, [9] 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 [11].

IV. The 3D-INSEG Algorithm for MEOT

The PMBM filter for MEOT is detailed in [7]. 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 article the PMBM extended object tracking filter implementation with the GGIW for MEOT is implemented to demonstrate the usefulness of the 3D-INSEG algorithm in MEOT.

Algorithm 1 summarizes the 3D-INSEG algorithm where b is the baseline and f is the focal length of the stereo cameras.

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 $\leftarrow$ **MaskRCNN**(I_L)
2) disp_map $\leftarrow$ **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 $\mathcal{D}$

A. 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 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 $X_k \in \mathcal{S}_d^{++}$ characterizes the extent state, delineating the size and shape of the target, where d denotes the dimension of the extent (typically $d = 2$ or $d = 3$) 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. Firstly, 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:

$$\phi(z_k|\xi_k) = \mathcal{N}(z_k; H_k x_k, X_k), \tag{10}$$

where 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:

$$f_{k|k}(\xi_k) = \text{GGIW}(\xi_k; \zeta_{k|k}), \tag{11}$$

where $\zeta_{k|k} = (\alpha_{k|k}, \beta_{k|k}, m_{k|k}, P_{k|k}, v_{k|k}, V_{k|k})$ 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 P the covariance for the Gaussian distribution, v the number of degrees of freedom

for the inverse Wishart distribution and 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 $\zeta_{k|k-1}$ that are updated with a set of measurements $\mathcal{Z}$ under the linear Gaussian model, are detailed in algorithm 2. These parameters and their updates play a critical role in the measurement update within the random matrix extended target model see, e.g., [19], [20]. 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:

$$x_{k+1} = f(x_k) + w_k, \tag{12}$$

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

2. Extent: The evolution of the extent state matrix X_k from time step k to $k+1$ is governed by the transformation:

$$X_{k+1} = M(x_k) X_k M(x_k)^T, \tag{13}$$

where $M(x_k)$ denotes a transformation matrix.

3. The measurement rate γ_{k+1} represents the expected number of measurements per target. It remains 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 motion models, are detailed in algorithm 3. For more extensive discussions regarding prediction within the random matrix extended target model, see [19], [20].

B. GGIW-PMBM filter for MEOT

The GGIW-PMBM filter [17] 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 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 processes target birth and undetected target propagation, while the MBM prediction forecasts detected target trajectories. Key variables include the PPP intensity $D^u(x)$, MBM parameters $\{\mathcal{W}_j, \{r_{j,i}, g_{j,i}\}_{i\in I_j}\}_{j\in J}$, where r and g are the components of the Bernoulli RFS, J is an index set for the MBs in the MBM (also called components of the MBM), I_j is an index set for the Bernoullis in the j th MB, and $\mathcal{W}_j(A)$ the data association probabilities. Algorithm 4 describes the GGIW-PMBM prediction, for the update and the other components of the GGIW-PMBM filter. See [18] for more details. Demonstrations and the code of the implementation are available at https://github.com/nfierroflo/3D-INSEG-for-MEOT.

Algorithm 2 GGIW Update

Input: GGIW parameter ζ_+, set of measurements $\mathcal{Z}$, measurement model H.

Output: Updated GGIW parameter ζ and predicted likelihood l:

$$\zeta = \begin{cases} \alpha = \alpha_+ + |\mathcal{Z}|, \\ \beta = \beta_+ + 1, \\ m = m_+ + K\epsilon, \\ P = P_+ - KHP^+, \\ v = v_+ + |\mathcal{Z}|, \\ V = V_+ + N + Z \end{cases}$$

where

$\bar{z} = \frac{1}{|\mathcal{Z}|}\sum_{z_i\in\mathcal{Z}} z_i$,

$Z = \sum_{z_i\in\mathcal{Z}}(z_i - \bar{z})(z_i - \bar{z})^T$,

$\hat{X} = V_+(v_+ - 2d - 2)^{-1}$,

$\epsilon = \bar{z} - Hm_+$,

$S = HP^+H^T + \hat{X}$,

$K = P^+H^T(S)^{-1}$,

$N = \hat{X}^{1/2}S^{-1/2}\epsilon\epsilon^T S^{-T/2}\hat{X}^{T/2}$.

Predicted likelihood, where $\Gamma(\cdot)$ is the Gamma function, and $\Gamma_d(\cdot)$ is the multivariate Gamma function,

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

Algorithm 3 GGIW Prediction

Input: $\zeta_{k|k}$

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

$$\zeta_{k+1|k} = \begin{cases} \alpha_{k+1|k} = \alpha_{k|k}\eta_k, \\ \beta_{k+1|k} = \beta_{k|k}\eta_k, \\ m_{k+1|k} = f(m_{k|k}), \\ P_{k+1|k} = F_{k|k}P_{k|k}F_{k|k}^T + Q, \\ v_{k+1|k} = 2d + 2 + e^{-Ts/\tau}\frac{v_{k|k}-2d-2}{v_{k|k}-2d-2}, \\ V_{k+1|k} = \left(\frac{v_{k+1|k}-2d-2}{v_{k|k}-2d-2}\right)^{-1} \times M_{k|k}V_{k|k}M_{k|k}^T \end{cases}$$

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

V. RESULTS

Figures 1, 4, 7 and 10 show reference images of different scenarios at different times. Figures 2, 5, 8 and 11 show Laser data measurements in blue. Figures 3, 6, 9 and 12 show 3D-INSEG detections in yellow. In these figures, red denotes the target location and shape estimates (a cross for the center position and an ellipse for the shape). The figures highlight the advantages of employing 3D-INSEG for MEOT using a PMBM extended filter for target estimation. When using laser data, even in less dense outdoor scenarios, the proximity between objects and the prevalence of obstacles complicate the estimation task.

We explore a densely populated scenario involving extended targets. The experiment showcases the goal of estimating

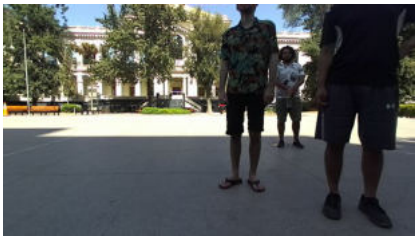

Fig. 1: Reference Image of scenario 1 at $t = 0.7s$.

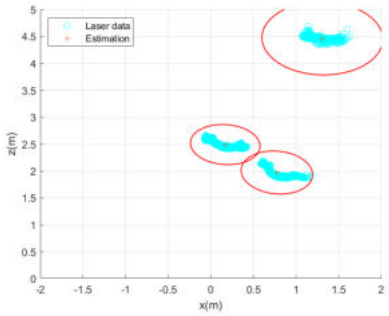

Fig. 2: Target location and shape estimation (scenario 1, $t = 0.7s$): Laser data.

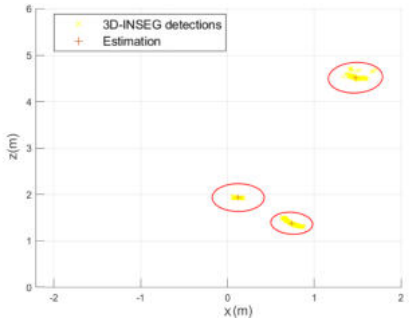

Fig. 3: Target location and shape estimation (scenario 1, $t = 0.7s$): 3D-INSEG.

Fig. 4: Reference Image of scenario 1 at $t = 8.07$.

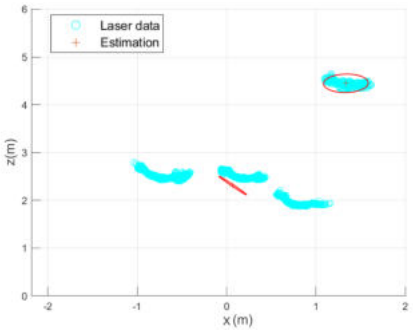

Fig. 5: Target location and shape estimation (scenario 1, $t = 8.07s$): Laser data.

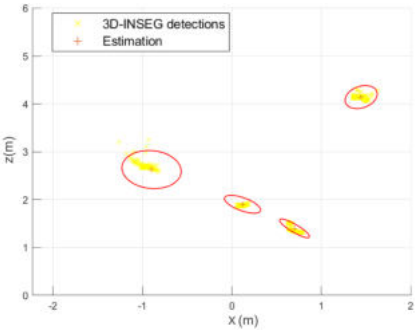

Fig. 6: Target location and shape estimation (scenario 1, $t = 8.07s$): 3D-INSEG.

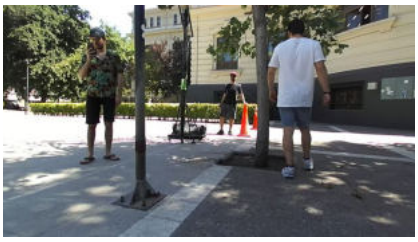

Fig. 7: Reference Image of scenario 2 at $t = 8.59s$.

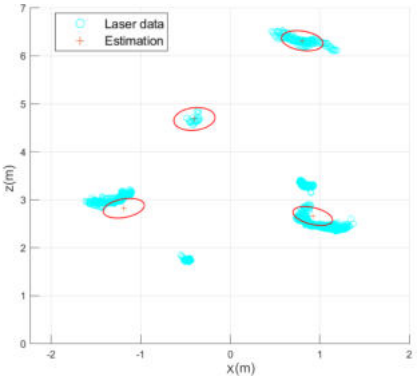

Fig. 8: Target location and shape estimation (scenario 2, $t = 8.59s$): Laser data.

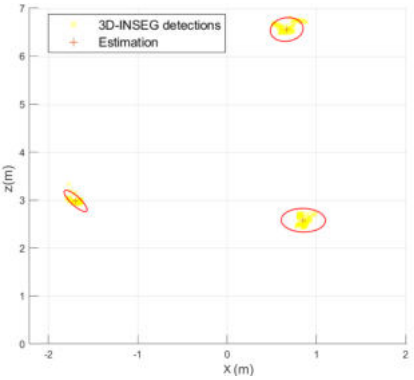

Fig. 9: Target location and shape estimation (scenario 2, $t = 8.59s$): 3D-INSEG.

the state of two humans in an indoor environment for 11 seconds (160 frames) as they navigate their surroundings and eventually intersect paths. 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 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 [8]. Initially, we applied DB-SCAN [9] 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 [10] to identify the $\lceil w^a_{k|k} \rceil^2$ 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 $x = (\gamma, \xi, X)$

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

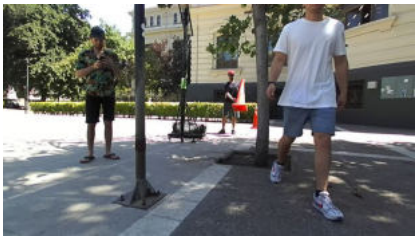

Fig. 10: Reference Image of scenario 2 at $t = 17.31s$.

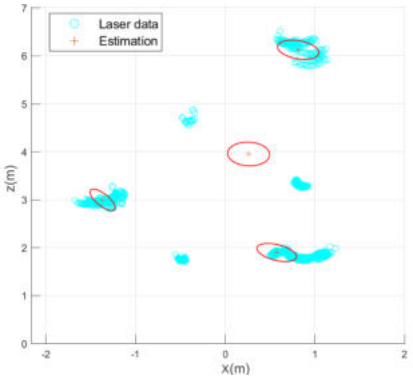

Fig. 11: Target location and shape estimation (scenario 2, $t = 17.31s$): Laser data.

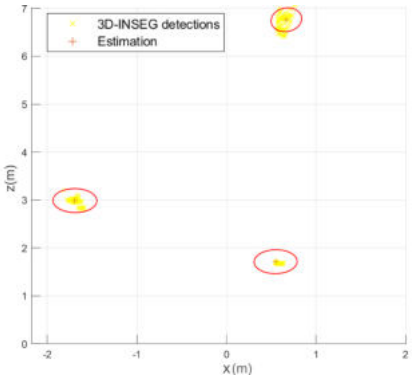

Fig. 12: Target location and shape estimation (scenario 2, $t = 17.31s$): 3D-INSEG.

Algorithm 4 GGIW PMBM prediction

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

$$D^{u+}(x) = \sum_{n=1}^{N_b} w_{b,n}\text{GGIW}(x;\zeta_{b,n}) + \sum_{n=1}^{N_u} w_{u,n}p_S(\hat{x}_{u,n})\text{GGIW}(x;\zeta_{u,n})$$

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

and $\mathcal{W}_j^+ = \mathcal{W}_j$, where $\zeta_{u,n}^+$ and $\zeta_{j,i}^+$ are computed as in algorithm 3.

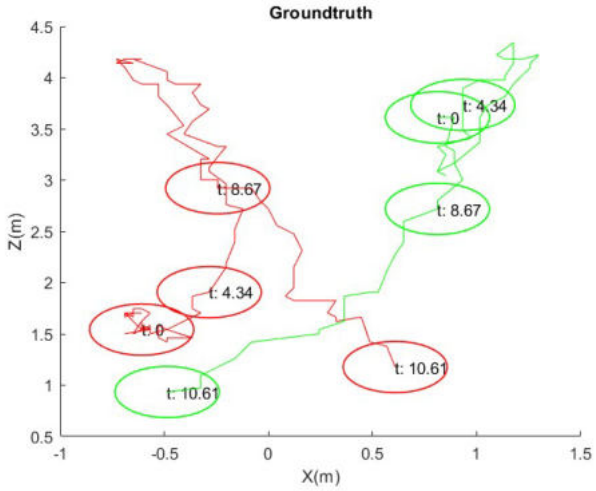

Fig. 13: Ground truth for the extended target scenario.

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 constant velocity, therefore the extent transformation function M is an identity matrix $M(x_k) = \mathbf{I}_2$. 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 $w_b^k = 0.1$ 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 $\bar{x}_b^k = [0\,\text{m}, 0\,\text{m/s}, 0\,\text{m}, 0\,\text{m/s}]^T$, and a covariance matrix $P_b^k = \text{diag}([50^2\,(\text{m}^2), 1\,(\text{m}^2/\text{s}^2), 50^2\,(\text{m}^2), 1\,(\text{m}^2/\text{s}^2)])$, along with an inverse-Wishart distribution with a mean of $\text{diag}([2, 2])\,(\text{m}^2)$ and 100 degrees of freedom.

The ground truth trajectories were manually marked and are visualized in Fig. 13, 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.

To evaluate the performance of extended object estimates with ellipsoidal extents, [21] showed that the Gaussian Wasserstein Distance (GWD) is a suitable choice. The GWD, defined as:

$$d_{GW}(x,\hat{x}) = \|H\xi - H\hat{\xi}\|_2^2 + \text{Tr}\left(X + \hat{X} - 2\sqrt{X}\sqrt{\hat{X}}\right)^{1/2}, \quad (14)$$

where H corresponds to H_k in Eq. 10. 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 [22]. The GOSPA metric, formulated as:

$$d_{(c,\alpha)}^p(X,\hat{X}) = \min_{\theta\in\Theta(|X|,|\hat{X}|)} \sum_{(i,j)\in\theta} d_{(c)}^{GW}(x_i,\hat{x}_j)^p + \frac{c^p}{\alpha}\left(|X| - |\theta| + |\hat{X}| - |\theta|\right)^{1/p}, \quad (15)$$

which incorporates $d_{(c)}^{GW}(x_i,\hat{x}_j) = \min(c, d_{GW}(x_i,\hat{x}_j))$, where $\Theta(|X|,|\hat{X}|)$ is the set of all possible 2D assignments, c denotes the base distance cut-off distance 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 [22] as an extension of the OSPA metric [23], and allows for the decomposition of multi-object error into three components: 1) state estimation error, 2) missed targets,

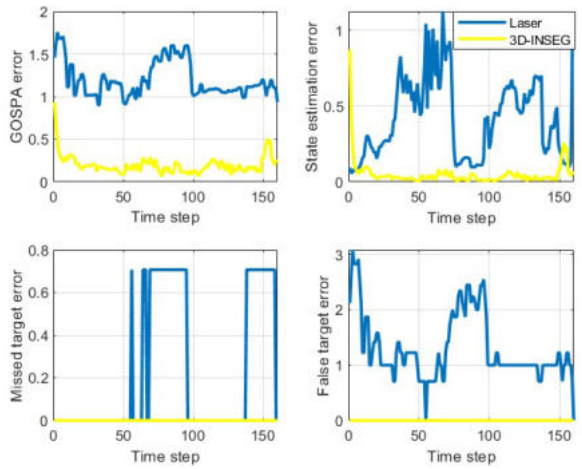

Fig. 14: GOSPA errors and their decomposition against time for the extended target scenario using the the extended PMBM with laser data and the 3D-INSEG data.

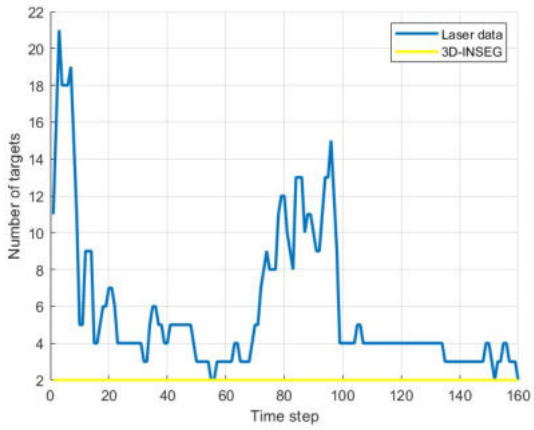

Fig. 15: Target cardinality estimated by the extended PMBM filter using the laser data and the 3D-INSEG data.

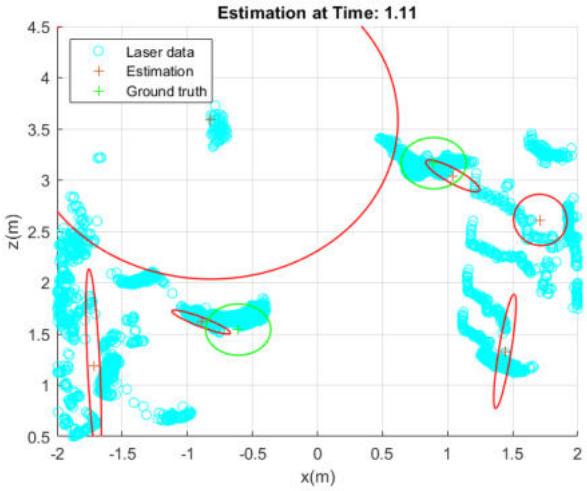

Fig. 16: Estimates at time $t = 1.11s$ using laser data.

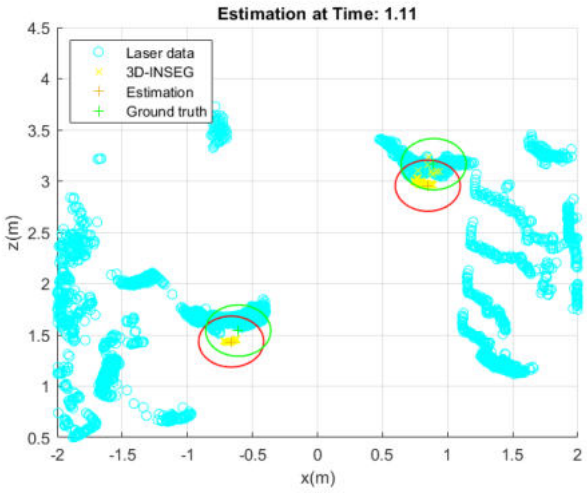

Fig. 17: Estimates at time $t = 1.11s$ using 3D-INSEG data.

and 3) false targets. Fig. 14 illustrates the GOSPA metric and its components over time, while Fig. 15 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 for the laser data, and 0.75 for the 3D-INSEG data. The processing time of the 3D-INSEG algorithm was 84 seconds for RAFT-Stereo and 32 seconds for Mask R-CNN on an NVIDIA Corporation TU102 GeForce RTX 2080 GPU, giving a total time of 116 seconds. However, this process is parallelizable, so the time could be reduced.

Figures 16 and 18 show estimates at different times using laser data. Figures 17 and 19 show estimates at different times based on 3D-INSEG generated data. Again, 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.

From the results we can see that when using the 3D-INSEG algorithm generated data, the GGIW-PMBM filter produces less false positives and the estimates have lower error than the estimates from the laser data with the same filter. Also the filter is faster because DB-SCAN clustering is avoided.

VI. Conclusions

This paper has demonstrated the efficacy of using the 3D-INSEG algorithm in densely populated scenarios, where object detection and segmentation play pivotal roles in robust tracking. Specifically, our investigation has underscored the enhanced performance of the extended GGIW-PMBM filter in challenging environments compared to traditional approaches reliant solely on raw laser data. These findings suggest future research which investigates a hybrid approach that integrates lidar measurements with 3D-INSEG detections depending on

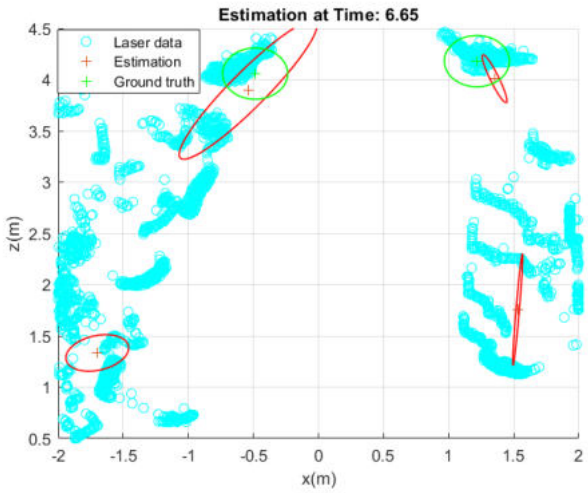

Fig. 18: Estimates at time $t = 6.65s$ using laser data.

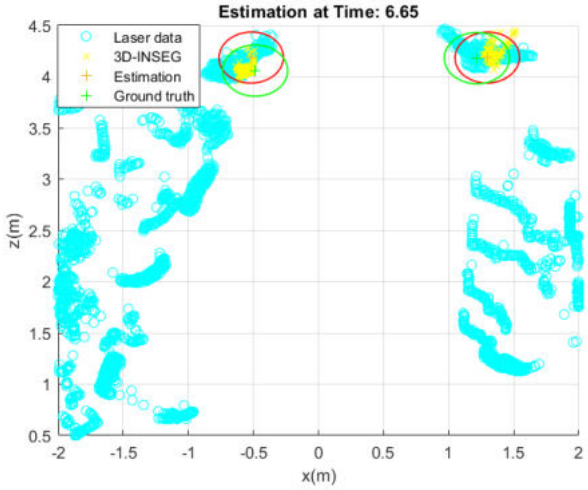

Fig. 19: Estimates at time $t = 6.65s$ using 3D-INSEG data.

the density of the targets in the environment. By combining the long-range capabilities of lidar with the precision in object detection afforded by the 3D-INSEG algorithm, we anticipate achieving superior tracking results across diverse conditions.

Acknowledgment

The authors acknowledge "Agencia Nacional de Investigacion y Desarrollo" (ANID) Fondecyt project 1231658, ANID Master's scholarship ref. no. 22230898 and the Department of Electrical Engineering, Universidad de Chile as well as the US Air Force Office of Scientific Research (AFOSR) Grant 23IOS020.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] L. Cament, J. Correa, M. Adams, and C. Pérez, "The histogram poisson, labeled multi-bernoulli multi-target tracking filter," *Signal Processing*, vol. 176, p. 107714, 2020.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] K. Granström and M. Baum, "A tutorial on multiple extended object tracking," 2022.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.