Efficient 3D Multi-View Multi-Object Tracking with Labeled Multi-Bernoulli Approximation

Tran Thien Dat Nguyen, Hoa Van Nguyen, Changbeom Shim, Viet Nguyen, and Hyunjin Choi

***Abstract*—In this paper, we propose an efficient algorithm for 3D multi-view multi-object tracking. The method is based on labeled multi-Bernoulli approximation of a state-of-the-art 3D visual tracking based on generalized labeled multi-Bernoulli filter. Compare to the baseline, our method demonstrates the improved efficiency on real 3D pedestrian tracking datasets. If the detection quality is high, the tracking accuracy of our method is comparable to the baseline. Our source codes will be published at https://github.com/TranThienDat-Nguyen/3D-VisualTracking.**

***Index Terms*—multi-view multi-object tracking, 3D visual tracking, multi-sensor data fusion.**

I. Introduction

3D multi-object tracking (MOT) is crucial for many applications including surveillance, sports analytics and aged care. 3D MOT can be performed by fusing information from different cameras that observe the scenes from multiple viewpoints. Further, multi-view information could also help resolve object occlusion and improve tracking results since an object that is occluded in one view can be seen in the other views. Nevertheless, performing 3D MOT with multi-view detection is not a straightforward task. Recent techniques rely on 3D object detectors and standard trackers to obtain 3D tracks [1], although locating objects in 3D is challenging [2]. State-of-the-art algorithms are mostly based on deep learning and have demonstrated accurate 3D detection results [3], [4], [5]. Nevertheless, training 3D object detectors is typically expensive and requires 3D training data. These factors limit the practicability of the learning-based methods in real-world scenarios. Further, most state-of-the-art 3D detectors [4], [5], [6] only report object ground plane positions, not their extents.

Alternative approaches for 3D multi-view MOT use 2D bounding box detection (from single-view cameras) to estimate 3D tracks, reporting both object positions and extents (shapes) [7], [8]. Compared to the learning-based methods, these approaches pose several advantages. First, they leverage 2D object detectors which have demonstrated robust performance in various tracking scenarios, without the need of re-training prior to system deployment. Second, no 3D training data are required. Further, these methods do not require 2D training data to have the same camera configurations to the test data. Lastly, they can also estimate object extents, not only the ground plane position. Nevertheless, these methods process multiple multi-sensor association hypotheses, resulting in relatively high computational complexity.

In this paper, we propose an approximation of the filter in [8] that results in a low complexity 3D multi-view MOT algorithm. We reduce the computation time by approximating the generalized labeled multi-Bernoulli (GLMB) density [9] by a labeled multi-Bernoulli (LMB) density [9] at each time step, following the approach in [10]. Our contribution includes:

- A single-term approximation of the multi-sensor GLMB filter for 3D visual MOT; and
- Extensive experiments to verify the improved efficiency of the proposed method on real 3D MOT datasets.

The paper is organized as follows. Section II presents backgrounds on 2D/3D visual tracking/detection. In Section III, we propose an efficient 3D multi-view MOT algorithm based on the LMB approximation of the GLMB-based filter. Section IV shows the experimental results, demonstrating the efficiency of the proposed algorithm in different camera configurations. Finally, Section V concludes the paper.

II. Backgrounds

Visual MOT is an important area of research in computer vision. The SORT algorithm [11] is a popular method for 2D visual MOT owing to its low complexity. This algorithm considers the best data association hypothesis (per time step) and uses a track management module to handle track initialization and termination. Since only the best data association hypothesis is considered, the SORT algorithm is fast and ready for real-time applications. Nevertheless, while its performance is reasonable in scenarios with good detection, its accuracy is compromised in complex scenes. Recent developments lead to other efficient algorithms with improved performance such as DeepSORT [12] which uses object appearance features in the data association, ByteTrack [13] which additionally processes low confidence detection, or StrongSORT [14] that uses track interpolation.

3D visual MOT is usually performed with multi-view camera settings. Early approaches track human heads or feet in world coordinates using homography transformation [15], [16]. In

T.T.D. Nguyen, H.V. Nguyen and C. Shim are with the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Australia (email: {t.nguyen1, hoa.v.nguyen, changbeom.shim}@curtin.edu.au). V. Nguyen is an independent researcher (email: viet.ng@proton.me). H. Choi is with AI R&D Center, AIBIZ Co. Ltd., Seoul, Republic of Korea (email: hjchoi@ai-biz.net). Corresponding author: Tran Thien Dat Nguyen.

This work was supported by grant funding from the AKF of the Department of Foreign Affairs and Trade. The work of H.V. Nguyen was supported by the Australian Research Council under Grant LP200301507.

[17], different cues are combined with the transformed 2D detection via a composition model to estimate 3D tracks. Leverage the advancement of deep learning, the methods reported in [1] detect objects with learned occlusion reasoning [6] and apply standard trackers to process 3D detection results. Indeed, detecting and tracking 3D objects using deep learning is an active topic in computer vision [4], [5], [18]. However, deep-learning-based techniques require 3D training data which is, in general, more difficult to obtain compared to 2D training data. Neglecting 3D detection, the method in [7] applies the multi-sensor GLMB filter [19] to fuse 2D bounding box detection from multi-view cameras and estimate object 3D shapes and positions. Nonetheless, it needs prior information on the initial locations of objects and does not exploit object appearance features. On the other hand, the recent method in [8] can automatically initialize tracks and handle track re-ID, considering object appearance features in the tracking models, as well as handle partial occlusion.

The 3D visual MOT algorithms proposed in [7], [8] are developed based on the random finite set framework [20]. In fact, this approach has also been applied in various visual tracking applications. In [21], numerous objects are tracked using multi-Bernoulli filter and 2D image observations, while in [22], [23], cell motions and lineages are estimated from 2D microscopy images with the GLMB-based filters. The GLMB filter [24] has also been applied to track pedestrians in [25], [26], [27], and the LMB filter [10] has been applied to track vehicles in [28], [29].

The performance of track-by-detect 2D visual MOT and the 3D visual MOT algorithms in [7], [8] depends on the quality of the 2D detection. Detecting objects in 2D images is an important problem in computer vision with applications beyond MOT task. Early 2D detectors rely on region-proposal algorithms and features from convolutional neural networks to detect objects [30]. Omitting the region-proposal step, YOLO algorithms [31] demonstrate remarkable improvements in efficiency, making them suitable for real-time applications. Recently, the transformer-based object detector [32] has gained significant attention due to its accuracy and efficiency. On the other hand, some detectors report object appearance features along with the 2D detection. Indeed, the 3D MOT algorithm in [8] uses these features for re-ID and improving the data association. State-of-the-art 2D detectors with feature extraction capability include ones proposed in [33]. However, since detection and feature extraction are interrelated, the balance (in terms of accuracy) between the two tasks must be considered [33]. In general, compared to 3D object detectors, 2D object detectors are relatively more robust given the wide availability of 2D training data.

III. Approximate 3D Visual MOT Filter

In this section, we propose an efficient 3D visual MOT algorithm by approximating the multi-object density computed with the method in [8] by an LMB density at each time step. Subsections III-A and III-B present the dynamic and observation models, respectively. In Subsection III-C, we present the approximate filtering recursion that yields efficient implementation.

A. 3D Objects Dynamic Model

Following [7], we model each (3D) object with a labeled state $\boldsymbol{x} = (x, \ell)$ with x belongs to an attribute space $\mathbb{X}$, and ℓ belongs to a label space $\mathbb{L}$. The object attribute includes the 3D position, velocity, and shape parameters of the object (i.e., logarithms of the half-axis lengths of the 3D ellipsoid enclosing the objects). A set of such objects is denoted as $\boldsymbol{X}$. At time step k, there will be a new set of objects that appear in the scene. The statistics of such set of objects are described by an LMB with parameters $\{r_B^{(\ell)}, p_B^{(\ell)}\}_{\ell\in\mathbb{B}}$, where $r_B^{(\ell)}$ is the existence probability of the object ℓ and $p_B^{(\ell)}$ is a probability distribution on object attribute. In addition to the new labels at time k, to handle track re-ID, the birth label set $\mathbb{B}$ might also include labels that appear at some time step k' before k. These (reappeared) labels are not in the current label set, i.e., $\mathbb{B} \cap \mathbb{L} = \emptyset$, because of the multi-object density truncation. The parameters of the LMB birth object density are computed from the measurements at $k-1$ using Algorithm 1 in [8].

For an object with attribute x at the current time step, it will survive to the next time step with the probability P_S or disappear with the probability $1 - P_S$. If it survives, its kinematic state follows a constant velocity model while its half-axis lengths are randomly scaled with the mean scaling factor of 1, i.e., [7]

$$f_{S,+}(x_+|x,\ell) = \mathcal{N}\left(x_+; Fx + b, Q\right), \tag{1}$$

where $\mathcal{N}(\cdot, \mu, P)$ is a Gaussian density with mean μ and covariance matrix P,

$$F = \begin{bmatrix} I_3(T) & 0_{6\times 3} \\ 0_{3\times 6} & I_3 \end{bmatrix}, I_3(T) = I_3 \otimes \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix},$$

$$b = \begin{bmatrix} 0_{6\times 1} \\ -\upsilon^{(\varsigma)}/2 \end{bmatrix}, Q = \begin{bmatrix} V(\upsilon^{(\zeta)}, T) & 0_{6\times 3} \\ 0_{3\times 6} & \text{diag}(\upsilon^{(\varsigma)}) \end{bmatrix},$$

$$V(\upsilon^{(\zeta)}, T) = \text{diag}(\upsilon^{(\zeta)}) \otimes \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \begin{bmatrix} \frac{T^2}{2} & T \end{bmatrix},$$

T is the sampling period, $\upsilon^{(\zeta)}$ and $\upsilon^{(\varsigma)}$ are respectively the noise variances for the 3D position and shape parameters[1], I_n is an n-dimension identity matrix, and $0_{m\times n}$ is an m by n matrix of zeros. This model is a direct-discrete type of discretization, see [34] for details.

B. 3D Objects Observation Model

1) Single-View Single-Object Measurement Likelihood: In a camera view c, an object $\boldsymbol{x} = (x, \ell)$ could generate a 2D bounding box and a feature vector, i.e., $z^{(c)} = (z_b^{(c)}, z_f^{(c)})$,

[1]It is equivalent to modeling the half-lengths with log-normal distribution of mean 1 and variances $\exp(\upsilon_i^{(\varsigma)}) - 1$, $i = 1, 2, 3$.

where $z_b^{(c)}$ is the 2D box center and extent, and $z_f^{(c)}$ is the feature vector. The detected bounding box and the feature vector are assumed to be independent. Hence, the measurement likelihood function is given by [8]

$$g^{(c)}(z_b^{(c)}, z_f^{(c)}|\boldsymbol{x}) = g_b^{(c)}(z_b^{(c)}|x,\ell)g_f^{(c)}(z_f^{(c)}|\ell), \tag{2}$$

where $g_b^{(c)}$ and $g_f^{(c)}$ are, respectively, the bounding box and feature measurement likelihoods.

The observed bounding box measurement $z_b^{(c)}$ is modeled by the box $\Phi^{(c)}(x)$, bounding the image of object $\boldsymbol{x}$ in camera c's image plane, corrupted by some additional Gaussian noise. Noting that $\Phi^{(c)}(x)$ is computed using the camera projection matrix described in [35]. Hence, the likelihood of observing $z_b^{(c)}$ is given by [7]

$$g_b^{(c)}(z_b^{(c)}|x,\ell) = \mathcal{N}\left(z_b^{(c)}; \Phi^{(c)}(x), \upsilon_b^{(c)}\right), \tag{3}$$

where $\upsilon_b^{(c)}$ is the noise variance matrix for the center and the extent of the box.

On the other hand, the visual feature has two modes, one is for the stable condition and the other is for changes in appearance (due to lightning etc.), i.e, [8]

$$g_f^{(c)}(z_f^{(c)}|\ell) \propto \sigma s_f(z_f^{(c)}, \alpha^{(\ell,c)}) + \bar{\sigma} s_f(z_f^{(c)}, \bar{\alpha}^{(\ell,c)}). \tag{4}$$

In (4), s_f is non-negative and increases monotonically as the similarity between its arguments grows, $\alpha^{(\ell,c)}$ and $\bar{\alpha}^{(\ell,c)}$ represent the nominal feature vectors for the stable and unstable modes, respectively, with associated weights σ and $\bar{\sigma}$. Given the observed feature $z_f^{(c)}$, the object feature is updated as

$$\alpha_+^{(\ell,c)} = \vartheta_0 \alpha^{(\ell,c)} + (1-\vartheta_0) z_f^{(c)}. \tag{5}$$

The nominal feature vector $\alpha^{(\ell,c)}$ when track ℓ first appears in the scene is computed from the measurements from which ℓ is initialized.

2) Multi-View Multi-Object Measurement Likelihood: The measurement set $Z^{(c)}$ from camera c includes measurements generated by objects and clutter measurements. The number of clutter measurements is assumed Poisson distributed with mean $\langle \kappa^{(c)}, 1 \rangle$, where $\kappa^{(c)}$ is the clutter intensity function and $\langle f, g \rangle$ denotes $\int f(x)g(x)dx$. At each time step, for a camera c, we define an *association map* $\gamma^{(c)} : \mathbb{L} \to \{-1 : |Z^{(c)}|\}$ such that: if $\gamma^{(c)}(\ell) = -1$ the object with label ℓ does not exist; if $\gamma^{(c)}(\ell) = 0$, the object is miss-detected; and if $\gamma^{(c)}(\ell) > 0$ the object is associated with measurement $z^{(c)}_{\gamma^{(c)}(\ell)}$ [19]. The single-view multi-object likelihood function is defined as [9]:

$$\boldsymbol{g}^{(c)}(Z^{(c)}|\boldsymbol{X}) \propto \sum_{\gamma^{(c)} \in \Gamma^{(c)}} \delta_{\mathcal{L}(\gamma^{(c)})}[\mathcal{L}(\boldsymbol{X})] \left[\psi_{Z^{(c)},\boldsymbol{X}}^{(c,\gamma^{(c)}(\mathcal{L}(\cdot)))}(\cdot)\right]^{\boldsymbol{X}}, \tag{6}$$

where $\mathcal{L}_\gamma = \{\ell : \gamma^{(1)}(\ell), ..., \gamma^{(C)}(\ell) \geq 0\}$, $\delta_a[b]$ is a generalized Kronecker delta function that takes 1 if $a = b$ and 0 otherwise, and

$$\psi^{(c,j)}_{\{z^{(c)}_{1:|Z^{(c)}|}\},\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} 1 - P_D^{(c)}(\boldsymbol{x};\boldsymbol{X}), & j = 0 \\ \frac{P_D^{(c)}(\boldsymbol{x};\boldsymbol{X}) g^{(c)}(z_j^{(c)}|\boldsymbol{x})}{\kappa^{(c)}(z_j^{(c)})}, & j > 0 \end{cases}. \tag{7}$$

Note that the dependence of the detection probability $P_D^{(c)}(\boldsymbol{x};\boldsymbol{X})$ on the other objects $\boldsymbol{X}$ is due to object occlusion. We use the occlusion model proposed in [8] that can handle partial occlusion.

For a multi-view detection $Z \triangleq (Z^{(1:C)})$, the multi-view association map is defined as $\gamma \triangleq (\gamma^{(1:C)})$. Note that if ℓ does not exist, it implies $\gamma^{(1)}(\ell) = ... = \gamma^{(C)}(\ell) = -1$. Thus, $\gamma : \mathbb{L} \to \{-1\}^C \uplus (\mathbb{J}^{(1)} \times \cdots \times \mathbb{J}^{(C)})$, where $\mathbb{J}^{(c)} \triangleq \{0 : |Z^{(c)}|\}$. The space of such γ maps is denoted as Γ. The *multi-view multi-object measurement likelihood* is given by [19]

$$\boldsymbol{g}(Z|\boldsymbol{X}) \propto \sum_{\gamma \in \Gamma} \delta_{\mathcal{L}_\gamma}[\mathcal{L}(\boldsymbol{X})] \left[\psi_{Z,\boldsymbol{X}}^{(\gamma(\mathcal{L}(\cdot)))}(\cdot)\right]^{\boldsymbol{X}}, \tag{8}$$

where $\mathcal{L}_\gamma \triangleq \{\ell : \gamma^{(1)}(\ell), ..., \gamma^{(C)}(\ell) \geq 0\}$ and

$$\psi_{Z,\boldsymbol{X}}^{(j^{(1:C)})}(\boldsymbol{x}) \triangleq \prod_{c=1}^{C} \psi_{Z^{(c)},\boldsymbol{X}}^{(c,j^{(c)})}(\boldsymbol{x}). \tag{9}$$

C. Approximate 3D Visual Tracking Filter

Given an initial LMB density with parameters $\{r^{(\ell)}, p^{(\ell)}\}_{\ell \in \mathbb{L}}$, applying the recursion in Algorithm 2 of [7] (treating the LMB density as a special GLMB density that only has one term) yields a GLMB density with parameters

$$\widehat{\boldsymbol{\pi}}_+ = \left\{(\omega_+^{(I_+,\gamma_+)}, p_+^{(\gamma_+)}) : (I_+,\gamma_+) \in \mathcal{F}(\mathbb{L}_+) \times \Gamma_+\right\},$$

where $\omega_+^{(I_+,\gamma_+)}$ is a non-negative weight satisfying $\sum_{(I_+,\gamma_+)\in\mathcal{F}(\mathbb{L}_+)\times\Gamma_+} \omega_+^{(I_+,\gamma_+)} = 1$, $\mathcal{F}(\mathbb{L}_+)$ is all finite subsets of $\mathbb{L}_+$, $p_+^{(\gamma_+)}$ is a distribution on $\mathbb{X}$. We then approximate the GLMB $\widehat{\boldsymbol{\pi}}_+$ by an LMB $\tilde{\boldsymbol{\pi}}_+$ with parameters $\{\tilde{r}_+^{(\ell)}, \tilde{p}_+^{(\ell)}\}_{\ell \in \mathbb{L}_+}$ [10] such that

$$\tilde{r}_+^{(\ell)} = \sum_{(I_+,\gamma_+)\in\mathcal{F}(\mathbb{L}_+)\times\Gamma_+} 1_\ell(I_+)\omega_+^{(I_+,\gamma_+)}, \tag{10}$$

$$\tilde{p}_+^{(\ell)}(x) = \sum_{(I_+,\gamma_+)\in\mathcal{F}(\mathbb{L}_+)\times\Gamma_+} 1_\ell(I_+) p_+^{(\gamma_+)}(x,\ell). \tag{11}$$

Noting that $\tilde{\boldsymbol{\pi}}_+$ matches the first moment and the mean cardinality of $\widehat{\boldsymbol{\pi}}_+$ [10]. If the detection quality is high, we can further approximate the LMB distribution by setting

$$\tilde{p}_+^{(\ell)}(x) \approx p_+^{(\hat{\gamma}_\ell)}(x,\ell), \tag{12}$$

where

$$(\cdot, \hat{\gamma}_\ell) = \arg \max_{(I_+,\gamma_+)\in\mathcal{F}(\mathbb{L}_+)\times\Gamma_+} 1_\ell(I_+)\omega_+^{(I_+,\gamma_+)}. \tag{13}$$

The schematic of the algorithm is given in Figure 1.

IV. Experimental Results

In this section, we evaluate the proposed algorithm compared to the baseline [8] in real datasets. We use Curtin multi camera (CMC) [7] and WILDTRACK (WT) [1] datasets for our evaluation. The numbers of objects and cameras for different sequences are given in Table I. For the CMC dataset, we

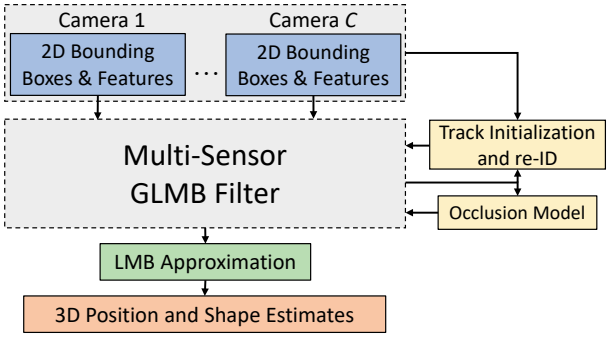

Fig. 1: Schematic of the proposed method.

evaluate the tracking accuracy against 3D bounding boxes enclosing the true objects. For the WT dataset, we only use ground plane position for evaluation since there is no information on the true object extent.

TABLE I: Number of objects and cameras.

Sequence	No. Objects	No. Cameras
CMC1	3	4
CMC2	10	4
CMC3	15	4
CMC4	3	4
CMC5	7	4
WT	24	7

For comparison, we use multi-object accuracy (MOTA) score in the CLEARMOT criteria [36], IDF1 score in the identity criteria [37], and OSPA$^{(2)}$ distance [38], [39] to evaluate the tracking accuracy. For 3D bounding boxes, we use 3D generalized intersection over union (GIoU) distance [40] between 3D bounding boxes as the base distance. For ground plane evaluation, we use Euclidean distance as the base distance. To compute MOTA and IDF1 scores, the threshold distance is 0.5 for GIoU base distance and 1m for Euclidean base distance. To compute OSPA$^{(2)}$ distance, we use the cut-off of 1 for the GIoU base distance and 1m for the Euclidean base distance. We use the frame rate per second (FPS) to evaluate the efficiency of the algorithms.

We consider three other 3D MOT algorithms in our evaluation: MV-GLMB-AB [8], MV-GLMB [7], and MS-GLMB [19] filters. To compare the speed, MV-GLMB-AB filter and our algorithm are implemented in C++ and tested on the same computer. We exclude the other two methods in the speed comparison since they are not implemented in C++ and, thus, the comparison including them would not be fair. To take into account the probabilistic nature of the filters, we test the algorithms over 25 Monte Carlo runs and report the mean and standard deviation over all trials. We use FairMOT detector [33] in our experiment.

The parameters are set up the same as in [8]. Object modes include "upright" and "fallen". The dynamic model parameters are the same for the two modes. In particular, the variance in velocity is 0.0012 m/s^2 for the CMC dataset, and 0.0225 m/s^2 for the WT dataset (for all three dimensions). For the 3D extents, the variance is 0.0036 for the (logarithm of the) width and length of the extents and 0.0004 for the (logarithm of the) height. These parameters are the same for both CMC and WT datasets. The observation noise variance of the 2D bounding box centroid is 400 squared pixels for both two object modes. For the "upright" mode, the noise variance of the logarithm of the width and the height of the box is 0.00995 and 0.002, respectively. For the "fallen" mode, it is vice versa.

A. Standard Camera Configuration

In this experiment, all cameras are in normal operation. In Table II, we report the tracking results on different sequences (of CMC and WT datasets). In general the MV-GLMB-AB filter and ours outperform others in terms of tracking accuracy, although the MV-GLMB-AB filter exhibits higher accuracy than our method. In CMC1 and CMC4 sequences, our filter performance is only slightly below the MV-GLMB-AB filter. This can be explained due to the high-quality detection in these two sequences (because of the low numbers of objects in the scenes). The performance gap between ours and the MV-GLMB-AB is considerable in CMC5 sequence due to the high object intensity and the complexity of the scenes with objects change their modes between "fallen" and "upright" frequently. Nevertheless, in WT sequence, while the MV-GLMB-AB filter outperforms ours in terms of OSPA$^{(2)}$ metric, ours is better in terms of MOTA score and comparable in terms of IDF1 score. In terms of efficiency, our method is significantly faster than the MV-GLMB-AB filter in all tested sequences.

B. On-Line Camera Re-Configuration

In this experiment, we split the sequences into five intervals (indicated in Figure 2) with different camera settings that change on-line. The camera settings for these intervals are: (1) all cameras are all; (2) only three fixed cameras are on; (3) three random cameras are on at random times; (4) two fixed cameras are on; and (5) two cameras which are off in interval (4) are turned back on and the other two are turned off. In Figure 2, we plot OSPA$^{(2)}$ distance at different time steps. Note that the OSPA$^{(2)}$ distance at time k considers the tracking outputs from the initial time step up to k. Ours and the MV-GLMB-AB filter perform similarly in CMC1-4 sequences. Nevertheless, in CMC5 our filter accuracy is considerably worse than MV-GLMB-AB filter. This observation is also confirmed by the results reported in Table III. However, the results from this table demonstrate that ours exhibits significantly higher speed compared to the baseline MV-GLMB-AB filter.

V. Conclusion

We have proposed an efficient algorithm for 3D visual MOT using 2D detection from multiple monocular cameras. Our method shows significant improvement in terms of efficiency and demonstrates similar tracking accuracy to the baseline if the

TABLE II: Tracking performance on the CMC and WT datasets: MC means and 1 standard deviation (shown in parenthesis). The best result for each sequence is **highlighted**.

Sequence	Filter	MOTA↑	IDF1↑	OSPA$^{(2)}$↓	FPS↑
CMC1	Ours	98.90(0.00)	99.50(0.00)	**0.3(0.0)**	**41.02(3.02)**
	MV-GLMB-AB [8]	**99.20(0.00)**	**99.60(0.00)**	**0.3(0.0)**	7.19(0.59)
	MV-GLMB [7]	92.1(2.86)	96.0(1.49)	0.83(0.02)	-
	MS-GLMB [19]	96.4(1.97)	98.1(1.11)	0.82(0.01)	-
CMC2	Ours	93.87(3.29)	78.46(7.11)	0.46(0.04)	**10.51(0.60)**
	MV-GLMB-AB [8]	**97.62(0.76)**	**87.78(4.47)**	**0.41(0.02)**	1.81(0.06)
	MV-GLMB [7]	78.3(2.76)	52.4(6.79)	0.88(0.02)	-
	MS-GLMB [19]	83.3(2.71)	47.8(5.06)	0.88(0.02)	-
CMC3	Ours	91.26(1.15)	72.30(5.31)	0.5(0.04)	**7.11(0.32)**
	MV-GLMB-AB [8]	**94.20(0.57)**	**78.98(3.69)**	**0.47(0.02)**	1.21(0.02)
	MV-GLMB [7]	71.2(3.61)	43.8(3.76)	0.86(0.02)	-
	MS-GLMB [19]	72.2(4.94)	38.7(2.59)	0.89(0.01)	-
CMC4	Ours	94.60(0.00)	97.20(0.00)	**0.26(0.0)**	**24.11(2.83)**
	MV-GLMB-AB [8]	**94.61(0.37)**	**97.24(0.21)**	**0.26(0.0)**	0.93(0.05)
	MV-GLMB [7]	70.9(4.00)	66.6(3.31)	0.71(0.05)	-
	MS-GLMB [19]	56.2(15.93)	74.6(6.90)	0.66(0.06)	-
CMC5	Ours	84.25(0.56)	37.07(2.63)	0.94(0.01)	**7.10(0.19)**
	MV-GLMB-AB [8]	**86.39(0.67)**	**47.62(3.12)**	**0.88(0.01)**	0.69(0.02)
	MV-GLMB [7]	65.0(7.06)	24.7(2.47)	0.93(0.01)	-
	MS-GLMB [19]	65.0(6.53)	20.3(3.32)	0.96(0.01)	-
WT	Ours	**37.87(2.61)**	**57.75(3.15)**	0.76(0.02)	**0.19(0.00)**
	MV-GLMB-AB [8]	35.89(2.33)	57.61(2.89)	**0.73(0.01)**	0.01(0.00)
	MV-GLMB [7]	16.2(1.79)	31.8(0.77)	0.86(0.00)	-
	MS-GLMB [19]	15.5(1.48)	31.4(0.8)	0.87(0.00)	-

TABLE III: Tracking performance on the CMC dataset in camera re-configuration experiment. MC means and 1 standard deviation (shown in parenthesis). The best result for each sequence is **highlighted**.

Sequence	Filter	MOTA↑	IDF1↑	OSPA$^{(2)}$↓	FPS↑
CMC1	Ours	**99.10(0.00)**	**99.50(0.00)**	**0.31(0.0)**	**42.54(4.31)**
	MV-GLMB-AB [8]	**99.10(0.00)**	**99.50(0.00)**	**0.31(0.0)**	9.13(0.79)
CMC2	Ours	91.54(2.30)	72.94(6.44)	0.5(0.03)	**12.11(0.71)**
	MV-GLMB-AB [8]	**95.08(0.73)**	**85.58(2.18)**	**0.44(0.01)**	2.10(0.04)
CMC3	Ours	81.00(3.77)	**62.54(5.62)**	0.55(0.05)	**8.61(0.47)**
	MV-GLMB-AB [8]	**81.28(3.10)**	61.56(3.55)	**0.55(0.03)**	1.21(0.06)
CMC4	Ours	**94.92(0.99)**	96.99(2.29)	0.29(0.05)	**27.72(2.12)**
	MV-GLMB-AB [8]	94.42(1.37)	**97.16(0.73)**	**0.28(0.0)**	2.43(0.28)
CMC5	Ours	61.60(2.08)	27.98(1.31)	0.95(0.01)	**11.90(0.37)**
	MV-GLMB-AB [8]	**67.81(2.79)**	**37.36(4.31)**	**0.90(0.01)**	1.00(0.06)

quality of the detection is high. Future works will exploit the efficient data association algorithms such as the tempered Gibbs sampler in [41], and estimate tracks across multiple frames with smoothing algorithms based on the multi-scan GLMB [42], [43], or partial smoothing [44] in 3D visual MOT.

REFERENCES

[1] T. Chavdarova, P. Baqué, S. Bouquet *et al.*, "WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5030–5039, 2018.

[2] C. Rubino, M. Crocco, and A. D. Bue, "3D object localisation from multi-view image detections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1281–1294, 2017.

[3] T. Chavdarova and F. Fleuret, "Deep multi-camera people detection," in *IEEE Int. Conf. Mach. Learning and Appl.*, 2017, pp. 848–853.

[4] D. M. H. Nguyen, R. Henschel, B. Rosenhahn *et al.*, "LMGP: Lifted multicut meets geometry projections for multi-camera multi-object tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 8866–8875.

[5] R. Qiu, M. Xu, Y. Yan *et al.*, "3D random occlusion and multi-layer projection for deep multi-camera pedestrian localization," in *Eur. Conf. Comput. Vis.*, 2022.

[6] P. Baqué, F. Fleuret, and P. V. Fua, "Deep occlusion reasoning for multi-camera multi-target detection," in *IEEE Int. Conf. Comput. Vis.*, 2017, pp. 271–279.

[7] J. Ong, B.-T. Vo, B.-N. Vo *et al.*, "A Bayesian filter for multi-view 3D multi-object tracking with occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2246–2263, 2022.

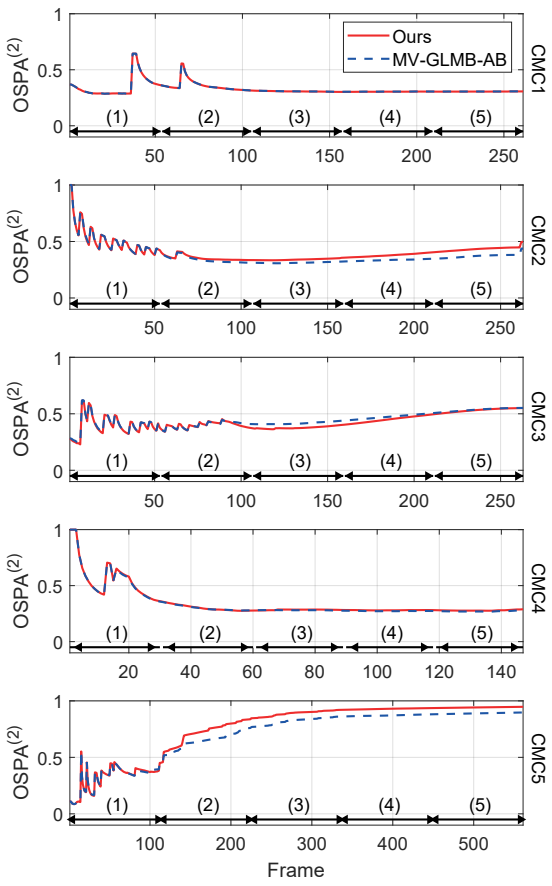

Fig. 2: OSPA$^{(2)}$ errors at different time intervals.

[8] L. V. Ma, T. T. D. Nguyen, B.-N. Vo *et al.*, "Track initialization and re-identification for 3D multi-view multi-object tracking," *Inf. Fusion*, vol. 111, p. 102496, 2024.

[9] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3460–3475, 2013.

[10] S. Reuter, B.-T. Vo, B.-N. Vo *et al.*, "The labeled multi-bernoulli filter," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3246–3260, 2014.

[11] A. Bewley, Z. Ge, L. Ott *et al.*, "Simple online and realtime tracking," in *IEEE Int. Conf. Image Process.*, 2016, pp. 3464–3468.

[12] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *IEEE Int. Conf. Image Process.*, pp. 3645–3649, 2017.

[13] Y. Zhang, P. Sun, Y. Jiang *et al.*, "ByteTrack: Multi-object tracking by associating every detection box," in *Eur. Conf. Comput. Vis.*, 2022, pp. 1–21.

[14] Y. Du, Y. Song, B. Yang *et al.*, "StrongSORT: Make DeepSORT great again," *IEEE Trans. Multimedia*, 2023.

[15] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Eur. Conf. Comput. Vis.* Springer, 2006, pp. 133–146.

[16] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[17] Y. Xu, X. Liu, Y. Liu *et al.*, "Multi-view people tracking via hierarchical trajectory composition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4256–4265.

[18] S. Aung, H. Park, H. Jung *et al.*, "Enhancing multi-view pedestrian detection through generalized 3d feature pulling," in *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 1196–1205.

[19] B.-N. Vo, B.-T. Vo, and M. Beard, "Multi-sensor multi-object tracking with the generalized labeled multi-Bernoulli filter," *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 5952–5967, 2019.

[20] B.-N. Vo, B.-T. Vo, T. T. D. Nguyen *et al.*, "An overview of multi-object estimation via labeled random finite set," *IEEE Transactions on Signal Processing*, vol. 72, pp. 4888–4917, 2024.

[21] R. Hoseinnezhad, B.-N. Vo, B.-T. Vo *et al.*, "Visual tracking of numerous targets via multi-Bernoulli filtering of image data," *Pattern Recognit.*, vol. 45, no. 10, pp. 3625–3635, 2012.

[22] T. T. D. Nguyen, B.-N. Vo, B.-T. Vo *et al.*, "Tracking cells and their lineages via labeled random finite sets," *IEEE Trans. Signal Process.*, vol. 69, pp. 5611–5626, 2021.

[23] T. T. D. Nguyen, C. Shim, and W. Kim, "Biological cell tracking and lineage inference via random finite sets," in *Int. Symp. Biomed. Imaging*, 2021, pp. 339–343.

[24] B.-N. Vo, B.-T. Vo, and H. G. Hoang, "An efficient implementation of the generalized labeled multi-Bernoulli filter," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1975–1987, 2017.

[25] L. V. Ma, T. T. D. Nguyen, C. Shim *et al.*, "Visual multi-object tracking with re-identification and occlusion handling using labeled random finite sets," *Pattern Recognit.*, vol. 156, p. 110785, 2024.

[26] M. J. Abbaspour and M. A. Masnadi-Shirazi, "Online multi-object tracking with δ-glmb filter based on occlusion and identity switch handling," *Image Vis. Comput.*, vol. 127, p. 104553, 2022.

[27] D. Y. Kim, B.-N. Vo, B.-T. Vo *et al.*, "A labeled random finite set online multi-object tracker for video data," *Pattern Recognit.*, vol. 90, pp. 377–389, 2019.

[28] N. Ishtiaq, A. K. Gostar, A. Bab-Hadiashar *et al.*, "Interaction-aware labeled multi-Bernoulli filter," *IEEE Trans. Intell. Transp. Syst.*, 2023.

[29] ——, "Integration of intelligent driver model with interaction-aware LMB (IA-LMB) filter for vehicle tracking," in *Int. Conf. Control Autom. Inf. Sci.*, 2023, pp. 97–102.

[30] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[31] Z. Ge, S. Liu, F. Wang *et al.*, "YOLOX: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[32] N. Carion, F. Massa, G. Synnaeve *et al.*, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 213–229.

[33] Y. Zhang, C. Wang, X. Wang *et al.*, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3069–3087, 2021.

[34] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[35] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000.

[36] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. on Image and Video Process.*, vol. 2008, pp. 1–10, 2008.

[37] E. Ristani, F. Solera, R. Zou *et al.*, "Performance measures and a data set for multi-target, multi-camera tracking," in *Eur. Conf. Comput. Vis.* Springer, 2016, pp. 17–35.

[38] M. Beard, B.-T. Vo, and B.-N. Vo, "A solution for large-scale multi-object tracking," *IEEE Trans. Signal Process.*, vol. 68, pp. 2754–2769, 2020.

[39] T. T. D. Nguyen, H. Rezatofighi, B.-N. Vo *et al.*, "How trustworthy are the existing performance evaluations for basic vision tasks?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8538–8552, 2023.

[40] H. Rezatofighi, N. Tsoi, J. Gwak *et al.*, "Generalized intersection over union: A metric and a loss for bounding box regression," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 658–666.

[41] C. Shim, B.-T. Vo, B.-N. Vo *et al.*, "Linear complexity gibbs sampling for generalized labeled multi-bernoulli filtering," *IEEE Trans. Signal Process.*, vol. 71, pp. 1981–1994, 2023.

[42] D. Moratuwage, B.-N. Vo, B.-T. Vo *et al.*, "Multi-scan multi-sensor multi-object state estimation," *IEEE Trans. Signal Process.*, vol. 70, pp. 5429–5442, 2022.

[43] B.-N. Vo and B.-T. Vo, "A multi-scan labeled random finite set model for multi-object state estimation," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 4948–4963, 2019.

[44] H. V. Nguyen, T. T. D. Nguyen, C. Shim *et al.*, "The smooth trajectory estimator for LMB filters," in *Int. Conf. Control Autom. Inf. Sci.*, 2023, pp. 115–120.