Development of a N-type GM-PHD Filter for Multiple Target, Multiple Type Visual Tracking

Nathanael L. Baisa , *Student Member, IEEE,* and Andrew Wallace, *Fellow, IET*

***Abstract*—We propose a new framework that extends the standard Probability Hypothesis Density (PHD) filter for multiple targets having N different types where $N \geq 2$ based on Random Finite Set (RFS) theory, taking into account not only background false positives (clutter), but also confusions among detections of different target types, which are in general different in character from background clutter. Under the assumptions of Gaussianity and linearity, our framework extends the existing Gaussian mixture (GM) implementation of the standard PHD filter to create a N-type GM-PHD filter. The methodology is applied to real video sequences by integrating object detectors' information into this filter for two scenarios. In the first scenario, a tri-GM-PHD filter ($N = 3$) is applied to real video sequences containing three types of multiple targets in the same scene, two football teams and a referee, using separate but confused detections. In the second scenario, we use a dual GM-PHD filter ($N = 2$) for tracking pedestrians and vehicles in the same scene handling their detectors' confusions. For both cases, Munkres's variant of the Hungarian assignment algorithm is used to associate tracked target identities between frames. This approach is evaluated and compared to both raw detection and independent GM-PHD filters using the Optimal Sub-pattern Assignment (OSPA) metric and the discrimination rate. This shows the improved performance of our strategy on real video sequences.**

***Index Terms*—Visual tracking, Random finite set, Multiple target filtering, N-type GM-PHD filter, Gaussian mixture, OSPA metric.**

I. Introduction

Visual detection, tracking and association of multiple targets at each frame in a video sequence is an active research field which has many applications including intelligent surveillance, visual servoing and control, human-computer and human-robot interaction. In some cases, for example for situational awareness, driver assistance and vehicle autonomy, there is also a necessity to distinguish between different target types, e.g. between vehicles and more vulnerable road users such as pedestrians and bicycles to select the best sensor focus and course of action [1]. For sports analysis we often want to track and discriminate sub-groups of the same target type such as players in opposing teams [2]. In this and many other examples, confusion between target types is common; a standard histogram-based detection strategy [3] in an urban environment may provide confused detections between pedestrians and cyclists, and even small cars.

Traditionally, multi-target trackers are based on the concept of finding associations between targets and measurements including Global Nearest Neighbor (GNN) [4], Joint Probabilistic Data Association Filter (JPDAF) [5], and Multiple Hypothesis Tracking (MHT) [6]. However, these approaches have faced challenges not only in the uncertainty caused by data association but also in algorithmic complexity that increases exponentially with the number of targets and measurements. For instance, the MHT has an exponential complexity with time and cubic with the number of targets.

To address the problems of increasing complexity, a unified framework which directly extends single to multiple target tracking by representing multi-target states and observations as random finite sets (RFS) was developed by Mahler [7]. This estimates the states and cardinality of an unknown and time varying number of targets in the scene, and allows for target birth, death, handling clutter (false alarms), and missing detections. Mahler [7] proposed to propagate the first-order moment of the multi-target posterior, called the Probability Hypothesis Density (PHD), rather than the full multi-target posterior.

There are two popular implementation schemes for the PHD filter, the Gaussian mixture (GM-PHD) [8] and the Sequential Monte Carlo (SMC) or particle-PHD filter [9]. The GM-PHD filter is preferred for linear (and by extension mildly non-linear) dynamic and observation models and assumes a Gaussian stochastic process [8]. However, for highly non-linear dynamic and observation models and non-Gaussian stochastic process, the SMC-PHD filter is the better implementation scheme [9]. For example, the GM-PHD filter is used in [10] for tracking pedestrians in video sequences but there is only one type of target and the motion model is fixed, and in [11] for selective tracking in dense environments. As an extension, a GM-PHD Filter was also developed in [12] for maneuvering targets but this employed a Jump Markov System (JMS) that switched between several motion models. In contrast, a particle-PHD filter was applied in [13] to allow for more complex motion models, and to cope with variation of scale, which has significant effects not just on object motion but also on the detection process. It was also used in [14] by treating high-confidence (strong) and low-confidence (weak) detections separately for better performance.

Considering extensions to different target types, Yan et al. [15] developed detection, tracking and classification (JDTC) of multiple targets in clutter which jointly estimates the number of targets, their kinematic states, and types of targets (classes) from a sequence of noisy and cluttered observation sets using a SMC-PHD filter. The dynamics of each target type (class) was modeled as a class-dependent model set and the signal amplitude was included in the multi-target likelihood to enhance the discrimination between targets from different classes and false alarms. Similarly, a joint target tracking and classification (JTC) algorithm was developed in [16] using RFS which takes into account extraneous target-originated

N. L. Baisa and A. Wallace are with the Department of Electrical, Electronic and Computer Engineering, Heriot Watt University, Edinburgh EH14 4AS, United Kingdom. E-mail: {nb30, a.m.wallace}@hw.ac.uk.

measurements (of the same type), i.e. multiple measurements that originated from a target which can be modeled as a Poisson RFS using linear and Gaussian assumptions. In these approaches, the augmented state vector of a target comprises the target kinematic state and class label, i.e. the target type (class) is put into the target state vector. Traditionally, simultaneous multi-object tracking and classification was proposed in [17] using a graphical probabilistic model and an inference procedure and then solving using a variational approximation, however, it suffers from class switching. Although multiple target types were considered, no account was taken of the effect of target confusions between types at the detection stage, as is the case in our work.

The main contributions of this paper are as follows.

- We model the RFS filtering of N different types of multiple targets with separate but confused detections where $N \geq 2$.
- The Gaussian mixture implementation of the standard PHD filter is extended for the proposed N-type PHD filter.
- We train and apply object detectors to video sequences of a soccer game, using players from each team and the referee as three distinct target types, and urban scene, using pedestrians and vehicles as two distinct target types. Compared to ground truth we extract the detection, confusion and background clutter probabilities and integrate these into the N-type GM-PHD filter.
- We integrate Munkres's variant of the Hungarian assignment algorithm to the typed results from the N-type GM-PHD filter to determine individual targets of each type between consecutive frames.
- We compare our approach to both repeated detection and the standard N independent GM-PHD filters to show that our approach yields improved performance in both target identification and location.

We presented preliminary ideas in [18] (simulation under varying probabilities of confusion for N=2) and [19] (on video for N=3). In this work, we further develop our approach. We extend from a tri-PHD filter to a N-type PHD filter as well as conducting an experiment on tracking vehicles and pedestrians as two different target types in addition to tracking two football teams and a referee as three different target types on video sequences. The remainder of this paper is organized as follows. Multiple type, multiple target recursive Bayes filtering with RFS is described in section II. A probability generating functional for deriving a N-type PHD filter and the N-type PHD filtering strategy are given in sections III and IV, respectively. In section V, a Gaussian mixture implementation of the N-type PHD filter is described in detail. The experimental results are analyzed and compared in section VI. The main conclusions and suggestions for future work are summarized in section VII.

II. Multiple Target, Multiple Type Recursive Bayes Filtering with RFS

A RFS represents a varying number of non-ordered target states and observations, analogous to a random vector for single target tracking. More precisely, a RFS is a finite-set-valued random variable i.e. a random variable which is random in both the number of elements and the values of the elements themselves. Finite Set Statistics (FISST), the study of the statistical properties of RFS, is a systematic treatment of multi-sensor multi-target filtering as a unified Bayesian framework using random set theory [7].

When different detectors run on the same scene to detect different target types there is no guarantee that these detectors only detect their own type. It is possible to run an independent PHD filter for each target type, but this will not be correct in most cases, as the likelihood of a positive response to a target of the wrong type will in general be different from, usually higher than, the likelihood of a positive response to the scene background. In this paper, we account for this difference between background clutter and target type confusion. This is equivalent to a single sensor (e.g. a smart camera) that has N different detection modes, each with its own probability of detection and a measurement density for N different target types.

To derive the N-type PHD filter, we define a RFS representation that extends from a single type, single-target Bayes framework to a multiple type, multiple target Bayes framework. Let the multi-target state space $\mathcal{F}(\mathcal{X})$ and the multi-target observation space $\mathcal{F}(\mathcal{Z})$ be the respective collections of all the finite subsets of the state space $\mathcal{X}$ and observation space $\mathcal{Z}$, respectively. If $L_i(k)$ is the number of targets of target type i in the scene at time k, then the multiple states for target type i, $X_{i,k}$, is the set

$$X_{i,k} = \{x_{i,k,1}, ...x_{i,k,L_i(k)}\} \in \mathcal{F}(\mathcal{X}) \tag{1}$$

where $i \in \{1, ..., N\}$. Similarly, if $M_i(k)$ is the number of received observations for target type i, then the corresponding multiple target measurements for that target type is the set

$$Z_{i,k} = \{z_{i,k,1}, ...z_{i,k,M_i(k)}\} \in \mathcal{F}(\mathcal{Z}) \tag{2}$$

where $i \in \{1, ..., N\}$. As stated above, some of these observations may be false, i.e. due to clutter (background) or confusion (response due to another target type).

The uncertainty in the state and measurement is introduced by modeling the multi-target state and the multi-target measurement using RFS. Let $\Xi_{i,k}$ be the RFS associated with the multi-target state of target type i, then

$$\Xi_{i,k} = S_{i,k}(X_{i,k-1}) \cup \Gamma_{i,k}, \tag{3}$$

where $S_{i,k}(X_{i,k-1})$ denotes the RFS of surviving targets of target type i, and $\Gamma_{i,k}$ is the RFS of new-born targets of target type i. We do not consider spawned targets as these have no meaning in our context, discussed below. Further, the RFS $\Omega_{i,k}$ associated with the multi-target measurements of target type i is

$$\Omega_{i,k} = \Theta_{i,k}(X_{i,k}) \cup C_{s_{i,k}} \cup C_{t_{iJ},k}, \tag{4}$$

where $J = \{1, ..., N\} \setminus i$ and $\Theta_{i,k}(X_{i,k})$ is the RFS modeling the measurements generated by the targets $X_{i,k}$, and $C_{s_{i,k}}$ models the RFS associated with the clutter (false alarms) for target type i which comes from the scene background.

However, we also include $C_{t_{iJ,k}}$ which is the RFS associated with all target types $J = \{1, ..., N\} \setminus i$, that is confusions while filtering target type i.

Analogous to the single-target case, the dynamics of $\Xi_{i,k}$ are described by the multi-target transition density $y_{i,k|k-1}(X_{i,k}|X_{i,k-1})$, while $\Omega_{i,k}$ is described by the multi-target likelihood $f_{ji,k}(Z_{i,k}|X_{j,k})$ for target type $i \in \{1, ..., N\}$ from detector $j \in \{1, ..., N\}$. The recursive equations are

$$p_{i,k|k-1}(X_{i,k}|Z_{i,1:k-1}) = \int y_{i,k|k-1}(X_{i,k}|X) p_{i,k-1|k-1}(X|Z_{i,1:k-1}) \mu(dX) \tag{5}$$

$$p_{i,k|k}(X_{i,k}|Z_{i,1:k}) = \frac{f_{ji,k}(Z_{i,k}|X_{j,k}) p_{i,k|k-1}(X_{i,k}|Z_{i,1:k-1})}{\int f_{ji,k}(Z_{i,k}|X) p_{i,k|k-1}(X|Z_{i,1:k-1}) \mu(dX)} \tag{6}$$

where μ is an appropriate dominating measure on $\mathcal{F}(\mathcal{X})$ [7]. Though a Monte Carlo approximation of this optimal multi-target types Bayes recursion is possible according to multi-target for single type [9], the number of particles required is exponentially related to the number of targets and their types in the scene. To make it computationally tractable, we extend Mahler's method of propagating the first-order moment (PHD) of the multi-target posterior instead of the full multi-target posterior for $N \geq 2$ types of multiple targets by deriving the updated PHDs from Probability Generating Functionals (PGFLs) starting from the standard predicted PHDs of each target type for our new filter termed as N-type PHD filter.

III. Probability Generating Functional (PGFL)

A probability generating functional is a convenient representation for stochastic modelling with a point process [7], a type of random process for which any one realisation consists of a set of isolated points either in time or space. Now, we model joint (probability generating) functionals which take into account the clutter due to the other target types (confusion) in addition to the background clutter for deriving the updated PHDs. Starting from the standard proved predicted PHDs [7], we derive novel extensions for the updated PHDs of a N-type PHD filter from PGFLs of each target type, handling confusions among target types.

The joint functional for target type i treating all other target types as clutter is given by

$$F_i[g, h] = G_{T_i}(hG_{L_{i,i}}(g|.)) G_{c_i}(g) \prod_{j=1\setminus i}^{N} G_{T_j}(G_{L_{j,i}}(g|.)), \tag{7}$$

where $i \in \{1, ..., N\}$ denotes target types.

$$G_{c_i}(g) = \exp(\lambda_i(c_i[g] - 1)), \tag{8}$$

where $G_{c_i}(g)$ is the Poisson PGFL [7] for false alarms where λ_i is the average number of false alarms for target type i and the functional $c_i[g] = \int g(z) c_i(z) dz$ where $c_i(.)$ is the uniform density over the surveillance region;

$$G_{T_i}(h) = \exp(\mu_i(s_i[h] - 1)), \tag{9}$$

where $G_{T_i}(h)$ is the prior PGFL and μ_i is the average number of targets, each of which is distributed according to $s_i(x)$ for target type i; and

$$G_{L_{j,i}}(g|x) = 1 - p_{ji,D}(x) + p_{ji,D}(x) \int g(z) f_{ji}(z|x) dz, \tag{10}$$

where $G_{L_{j,i}}(g|x)$ is the Bernoulli detection process for each target of target type i using detector j with probability of detection for target type i by detector j, $p_{ji,D}$, and $f_{ji}(z|x)$ is a likelihood defining the probability that z is generated by the target type i conditioned on state x from detector j [7]. Expanding $s[hG_{L_{j,i}}(g|x)]$ and $s[G_{L_{j,i}}(g|x)]$ as

$$s[hG_{L_{j,i}}(g|x)] = \int s(x) h(x) \big(1 - p_{ji,D}(x) + p_{ji,D}(x) \int g(z) f_{ji}(z|x) dz\big) dx, \tag{11}$$

and

$$s[G_{L_{j,i}}(g|x)] = \int s(x) \big(1 - p_{ji,D}(x) + p_{ji,D}(x) \int g(z) f_{ji}(z|x) dz\big) dx, \tag{12}$$

Accordingly, $F_i[g, h]$ is expanded as

$$F_i[g, h] = \exp\Big(\lambda_i(\int g(z) c_i(z) dz - 1) + \textstyle\sum_{j=1\setminus i}^{N} \mu_j [\int s_j(x)(1 - p_{ji,D}(x) + p_{ji,D}(x) \int g(z) f_{ji}(z|x) dz) dx - 1] + \mu_i [\int s_i(x) h(x)(1 - p_{ii,D}(x) + p_{ii,D}(x) \int g(z) f_{ii}(z|x) dz) dx - 1]\Big), \tag{13}$$

The updated PGFL $G_i(h|z_1, ...z_{M_i})$ for target type i is obtained by finding the M_i^{th} functional derivative of $F_i[g, h]$ [7] and is given by

$$G_i(h|z_1, ..., z_{M_i}) = \frac{\frac{\delta^{M_i}}{\delta_{\varphi_{z1}} ... \delta_{\varphi_{z_{M_i}}}} F_i[g, h]|_{g=0}}{\frac{\delta^{M_i}}{\delta_{\varphi_{z1}} ... \delta_{\varphi_{z_{M_i}}}} F_i[g, 1]|_{g=0}}, \tag{14}$$

The updated PHD for target type i treating all other target types as clutter can be obtained by taking the first-order moment (mean) [7] of Eq. (14) and setting $h = 1$,

$$\begin{aligned}\mathcal{D}_i(x|z_1,...,z_{M_i}) &= \tfrac{\delta}{\delta_{\varphi_x}} G_i(h|z_1,...z_{M_i})|_{h=1},\\ &= \mu_i s_i(x)(1-p_{ii,D}(x)) + \textstyle\sum_{m=1}^{M_i} \frac{\mu_i s_i(x) p_{ii,D}(x) f_{ii}(z_m|x)}{\lambda_i c_i(z_m) s_{i,k} + \sum_{j=1\setminus i}^{N} \mu_j \int s_j(x) p_{ji,D}(x) f_{ji}(z_m|x)dx + \mu_i \int s_i(x) p_{ii,D}(x) f_{ii}(z_m|x)dx},\end{aligned} \tag{15}$$

This, $\mathcal{D}_i(x|z_1,...,z_{M_i})$ in Eq. (15), is the updated PHD for target type i treating all other target types as clutter in the case of the N-type PHD filter. The term $\mu_i s_i(x)$ in Eq. (15) is the predicted PHD for target type i.

IV. N-type PHD Filtering Strategy

The PHDs, $\mathcal{D}_{\Xi_1}(x)$, $\mathcal{D}_{\Xi_2}(x)$,..., $\mathcal{D}_{\Xi_N}(x)$, are the first-order moments of RFSs, Ξ_1, Ξ_2,..., Ξ_N, and are intensity functions on a single state space $\mathcal{X}$ whose peaks identify the likely positions of the targets. For any region $R \subseteq \mathcal{X}$

$$E[|(\Xi_1 \cup \Xi_2 ... \cup \Xi_N) \cap R|] = \sum_{i=1}^{N} \int_R \mathcal{D}_{\Xi_i}(x)dx \tag{16}$$

where$|.|$ is used to denote the cardinality of a set. In practice, Eq. (16) means that by integrating the PHDs on any region R of the state space, it is possible to obtain the expected number of targets (cardinality) in R.

At any time step, k, new targets may appear (births) and are added to those targets that persist and have moved position from the previous time step. Consequently, the PHD *prediction* for target type i at time k is

$$\begin{aligned}\mathcal{D}_{i,k|k-1}(x) = &\int p_{i,S,k|k-1}(\zeta) y_{i,k|k-1}(x|\zeta) \mathcal{D}_{i,k-1|k-1}(\zeta)d\zeta\\ &+\gamma_{i,k}(x),\end{aligned} \tag{17}$$

where $\gamma_{i,k}(.)$ is the intensity function of a new target birth RFS $\Gamma_{i,k}$, $p_{i,S,k|k-1}(\zeta)$ is the probability that a target still exists at time k, $y_{i,k|k-1}(.|\zeta)$ is the single target state transition density at time k given the previous state ζ for target type i.

Thus, following Eq. (15), the final updated PHD for target type i is obtained by setting $\mu_i s_i(x) = \mathcal{D}_{i,k|k-1}(x)$

$$\mathcal{D}_{i,k|k}(x) = \left[1 - p_{ii,D}(x) + \textstyle\sum_{z \in Z_{i,k}} \frac{p_{ii,D}(x) f_{ii,k}(z|x)}{c_{s_{i,k}}(z) + c_{t_{i,k}}(z) + \int p_{ii,D}(\xi) f_{ii,k}(z|\xi) \mathcal{D}_{i,k|k-1}(\xi)d\xi}\right] \mathcal{D}_{i,k|k-1(x)}, \tag{18}$$

The clutter intensity $c_{t_{i,k}}(z)$ due to all types of targets $j \in \{1,...,N\}$ except target type i in Eq. (18) is given by

$$c_{t_{i,k}}(z) = \textstyle\sum_{j=1,...,N\setminus i} \int p_{ji,D}(y) \mathcal{D}_{j,k|k-1}(y) f_{ji,k}(z|y)dy, \tag{19}$$

This means that when filtering target type i, all the other target types are included as confusing detections. Eq. (19) converts state space to observation space by integrating the PHD estimator $\mathcal{D}_{j,k|k-1}(y)$ and likelihood $f_{ji,k}(z|y)$ which defines the probability that z is generated by detector j conditioned on state x of the target type i taking into account the confusion probability $p_{ji,D}(y)$, when target type i is detected by detector j.

The clutter intensity due to the background for target type i, $c_{s_{i,k}}(z)$, in Eq. (18) is given by

$$c_{s_{i,k}}(z) = \lambda_i c_i(z) = \lambda_{c_i} A c_i(z), \tag{20}$$

where $c_i(.)$ is the uniform density over the surveillance region A, and λ_{c_i} is the average number of clutter returns per unit volume for target type i i.e. $\lambda_i = \lambda_{ci} A$. While the standard PHD filter has linear complexity with the current number of measurements (m) and with the current number of targets (n) i.e. computational order of $O(mn)$, the N-type PHD filter has linear complexity with the current number of measurements (m), with the current number of targets (n) and with the total number of target types (N) i.e. computational order of $O(mnN)$.

In general, the clutter intensities due to the background for each target type i, $c_{s_{i,k}}(z)$, can be different as they depend on the receiver operating characteristic (ROC) curves of the detection processes. Moreover, the probabilities of detection $p_{ii,D}(x)$ and $p_{ji,D}(x)$ may all be different although assumed constant across both the time and space continua.

V. N-type PHD Filter Implementation based on Gaussian Mixture

The Gaussian mixture implementation of the standard PHD (GM-PHD) filter [8] is a closed-form solution of the PHD filter that assumes a linear Gaussian system. In this section, this is extended for the N-type PHD filter by solving Eq. (19). Assuming each target follows a linear Gaussian model,

$$y_{i,k|k-1}(x|\zeta) = \mathcal{N}(x; F_{i,k-1}\zeta, Q_{i,k-1}) \tag{21}$$

$$f_{ji,k}(z|x) = \mathcal{N}(z; H_{ji,k}x, R_{ji,k}) \tag{22}$$

where $\mathcal{N}(.; m, P)$ denotes a Gaussian density with mean m and covariance P; $F_{i,k-1}$ and $H_{ji,k}$ are the state transition and measurement matrices, respectively. $Q_{i,k-1}$ and $R_{ji,k}$ are the covariance matrices of the process and the measurement noise, respectively, where $i \in \{1,...,N\}$ and $j \in \{1,...,N\}$. A measurement driven birth intensity, similar in principle to [20], is introduced at each time step, with a non-informative zero initial target velocity. This choice is preferred to the options of covering the whole state space (random) [21] or

a-priori birth [8] and is discussed further in Section VI. The intensity of the spontaneous birth RFS is $\gamma_{i,k}(x)$ for target type i

$$\gamma_{i,k}(x) = \sum_{v=1}^{V_{\gamma_i,k}} w_{i,\gamma,k}^{(v)} \mathcal{N}(x; m_{\gamma_i,k}^{(v)}, P_{\gamma_i,k}^{(v)}) \tag{23}$$

where $V_{\gamma_i,k}$ is the number of birth Gaussian components for target type i where $i \in \{1, ..., N\}$, $m_{\gamma_i,k}^{(v)}$ is the current measurement and zero initial velocity used as mean and $P_{\gamma_i,k}^{(v)}$ is the birth covariance for Gaussian component v of target type i.

It is assumed that the posterior intensity for target type i at time $k-1$ is a Gaussian mixture of the form

$$\mathcal{D}_{i,k-1}(x) = \sum_{v=1}^{V_{i,k-1}} w_{i,k-1}^{(v)} \mathcal{N}(x; m_{i,k-1}^{(v)}, P_{i,k-1}^{(v)}), \tag{24}$$

where $i \in \{1, ..., N\}$ and $V_{i,k-1}$ is the number of Gaussian components of $\mathcal{D}_{i,k-1}(x)$. Under these assumptions, the predicted intensity at time k for target type i is given following Eq. (17) by

$$\mathcal{D}_{i,k|k-1}(x) = \mathcal{D}_{i,S,k|k}(x) + \gamma_{i,k}(x), \tag{25}$$

where

$$\begin{aligned}\mathcal{D}_{i,S,k|k-1}(x) = \; & p_{i,S,k} \textstyle\sum_{v=1}^{V_{i,k-1}} w_{i,k-1}^{(v)} \mathcal{N}(x; \\ & m_{i,S,k|k-1}^{(v)}, P_{i,S,k|k-1}^{(v)}),\end{aligned}$$

$$m_{i,S,k|k-1}^{(v)} = F_{i,k-1} m_{1,k-1}^{(v)},$$

$$P_{i,S,k|k-1}^{(v)} = Q_{i,k-1} + F_{i,k-1} P_{i,k-1}^{(v)} F_{1,k-1}^T,$$

where $p_{i,S,k}$ is the survival rate for target type i and $\gamma_{i,k}(x)$ is given by Eq. (23).

Since $\mathcal{D}_{i,S,k|k-1}(x)$ and $\gamma_{i,k}(x)$ are Gaussian mixtures, $\mathcal{D}_{i,k|k-1}(x)$ can be expressed as a Gaussian mixture of the form

$$\mathcal{D}_{i,k|k-1}(x) = \sum_{v=1}^{V_{i,k|k-1}} w_{i,k|k-1}^{(v)} \mathcal{N}(x; m_{i,k|k-1}^{(v)}, P_{i,k|k-1}^{(v)}), \tag{26}$$

where $w_{i,k|k-1}^{(v)}$ is the weight accompanying the predicted Gaussian component v for target type i and $V_{i,k|k-1}$ is the number of predicted Gaussian components for target type i where $i \in \{1, ..., N\}$.

Assuming the probabilities of detection to be constant, the posterior intensity for target type i at time k (updated PHD), considering incorrect detection of target types as confusion is also a Gaussian mixture which corresponds to Eq. (18), and is given by

$$\mathcal{D}_{i,k|k}(x) = (1 - p_{ii,D,k}) \mathcal{D}_{i,k|k-1}(x) + \sum_{z \in Z_{i,k}} \mathcal{D}_{i,D,k}(x; z), \tag{27}$$

where

$$\mathcal{D}_{i,D,k}(x; z) = \sum_{v=1}^{V_{i,k|k-1}} w_{i,k}^{(v)}(z) \mathcal{N}(x; m_{i,k|k}^{(v)}(z), P_{i,k|k}^{(v)}),$$

$$w_{i,k}^{(v)}(z) = \frac{p_{ii,D,k} w_{i,k|k-1}^{(v)} q_{i,k}^{(v)}(z)}{c_{s_{i,k}}(z) + c_{t_{i,k}}(z) + p_{ii,D,k} \sum_{l=1}^{V_{i,k|k-1}} w_{i,k|k-1}^{(l)} q_{i,k}^{(l)}(z)},$$

$$q_{i,k}^{(v)}(z) = \mathcal{N}(z; H_{ii,k} m_{i,k|k-1}^{(v)}, R_{ii,k} + H_{ii,k} P_{i,k|k-1}^{(v)} H_{ii,k}^T),$$

$$m_{i,k|k}^{(v)}(z) = m_{i,k|k-1}^{(v)} + K_{i,k}^{(v)}(z - H_{ii,k} m_{i,k|k-1}^{(v)}),$$

$$P_{i,k|k}^{(v)} = [I - K_{i,k}^{(v)} H_{ii,k}] P_{i,k|k-1}^{(v)},$$

$$K_{i,k}^{(v)} = P_{i,k|k-1}^{(v)} H_{ii,k}^T [H_{ii,k} P_{i,k|k-1}^{(v)} H_{ii,k}^T + R_{ii,k}]^{-1},$$

$c_{s_{i,k}}(z)$ is given in Eq. (20). Therefore, all that is left is to formulate the implementation scheme for $c_{t_{i,k}}(z)$ which is given in Eq. (19) and is given again as

$$c_{t_{i,k}}(z) = \textstyle\sum_{j=1,...,N \setminus i} \int p_{ji,D}(y) \mathcal{D}_{j,k|k-1}(y) f_{ji,k}(z|y) dy, \tag{28}$$

where $\mathcal{D}_{j,k|k-1}(y)$ is given in Eq. (26), $f_{ji,k}(z|y)$ is given in Eq. (22) and $p_{ji,D}(y)$ is assumed constant. Since $w_{j,k|k-1}^{(i)}$ is independent of the integrable variable y, Eq. (28) becomes

$$\begin{aligned}c_{t_{i,k}}(z) = \; & \textstyle\sum_{j=1,...,N \setminus i} \sum_{v=1}^{V_{j,k|k-1}} p_{ji,D} w_{j,k|k-1}^{(v)} \int \mathcal{N}(y; \\ & m_{j,k|k-1}^{(v)}, P_{j,k|k-1}^{(v)}) \mathcal{N}(z; H_{ji,k} y, R_{ji,k}) dy,\end{aligned} \tag{29}$$

This can be simplified further using the following equality given that P_1 and P_2 are positive definite

$$\int \mathcal{N}(y; m_1 \zeta, P_1) \mathcal{N}(\zeta; m_2, P_2) d\zeta = \mathcal{N}(y; m_1 m_2, P_1 + m_1 P_2 m_2^T). \tag{30}$$

Therefore, (29) becomes,

$$\begin{aligned}c_{t_{i,k}}(z) = \; & \textstyle\sum_{j=1,...,N \setminus i} \sum_{v=1}^{V_{j,k|k-1}} p_{ji,D} w_{j,k|k-1}^{(v)} \mathcal{N}(z; \\ & H_{ji,k} m_{j,k|k-1}^{(v)}, R_{ji,k} + H_{ji,k} P_{j,k|k-1}^{(v)} H_{ji,k}^T),\end{aligned} \tag{31}$$

where $i \in \{1, ..., N\}$.

The key steps of the N-type GM-PHD filter are summarised in Algorithms 1 and 2. These are expressed in terms of frames k and $k-1$; for the first frame, $k = 1$, of a sequence there is only detection and target birth, but no prediction and update for existing targets. For subsequent frames, we have chosen measurement driven target birth, rather than a random or a-priori birth model, inspired by but not identical to [20]. Maggio et al. [13] also assume that targets are born in a limited volume around measurements. The advantage of random birth is in the potential detection of weak target signatures, but in these examples the presence of a human should, in general, generate a strong probability of detection provided the target is in view. This is borne out by experiments and parameter setting in Section VI. A further disadvantage of random birth is the increased complexity of processing

Algorithm 1 Pseudocode for the N-type GM-PHD filter

given $\{w_{i,k-1}^{(v)}, m_{i,k-1}^{(v)}, P_{i,k-1}^{(v)}\}_{v=1}^{V_{i,k-1}}$, and the measurement set $Z_{i,k}$ for target type $i \in \{1, ..., N\}$
step 1. (prediction for birth targets)
for $i = 1, ..., N$ **do** ▷ for all target type i
  $e_i = 0$
  for $u = 1, ..., V_{\gamma_i,k}$ **do**
    $e_i := e_i + 1$
    $w_{i,k|k-1}^{(e_i)} = w_{i,\gamma,k}^{(u)}$
    $m_{i,k|k-1}^{(e_i)} = m_{i,\gamma,k}^{(u)}$
    $P_{i,k|k-1}^{(e_i)} = P_{i,\gamma,k}^{(u)}$
  end for
end for
step 2. (prediction for existing targets)
for $i = 1, ..., N$ **do** ▷ for all target type i
  for $u = 1, ..., V_{i,k-1}$ **do**
    $e_i := e_i + 1$
    $w_{i,k|k-1}^{(e_i)} = p_{i,S,k} w_{i,k-1}^{(u)}$
    $m_{i,k|k-1}^{(e_i)} = F_{i,k-1} m_{i,k-1}^{(u)}$
    $P_{i,k|k-1}^{(e_i)} = Q_{i,k-1} + F_{i,k-1} P_{i,k-1}^{(u)} F_{i,k-1}^T$
  end for
end for
$V_{i,k|k-1} = e_i$
step 3. (Construction of PHD update components)
for $i = 1, ..., N$ **do** ▷ for all target type i
  for $u = 1, ..., V_{i,k|k-1}$ **do**
    $\eta_{i,k|k-1}^{(u)} = H_{ii,k} m_{i,k|k-1}^{(u)}$
    $S_{i,k}^{(u)} = R_{ii,k} + H_{ii,k} P_{i,k|k-1}^{(u)} H_{ii,k}^T$
    $K_{i,k}^{(u)} = P_{i,k|k-1}^{(u)} H_{ii,k}^T [S_{i,k}^{(u)}]^{-1}$
    $P_{i,k|k}^{(u)} = [I - K_{i,k}^{(u)} H_{ii,k}] P_{i,k|k-1}^{(u)}$
  end for
end for
step 4. (Update)
for $i = 1, ..., N$ **do** ▷ for all target type i
  for $u = 1, ..., V_{i,k|k-1}$ **do**
    $w_{i,k}^{(u)} = (1 - p_{ii,D,k}) w_{i,k|k-1}^{(u)}$
    $m_{i,k}^{(u)} = m_{i,k|k-1}^{(u)}$
    $P_{i,k}^{(u)} = P_{i,k|k-1}^{(u)}$
  end for
  $l_i := 0$
  for each $z \in Z_{i,k}$ **do**
    $l_i := l_i + 1$
    for $u = 1, ..., V_{i,k|k-1}$ **do**
      $w_{i,k}^{(l_i V_{i,k|k-1}+u)} = p_{ii,D,k} w_{i,k|k-1}^{(u)} \mathcal{N}(z; \eta_{i,k|k-1}^{(u)}, S_{i,k}^{(u)})$
      $m_{i,k}^{(l_i V_{i,k|k-1}+u)} = m_{i,k|k-1}^{(u)} + K_{i,k}^{(u)} (z - \eta_{i,k|k-1}^{(u)})$
      $P_{i,k}^{(l_i V_{i,k|k-1}+u)} = P_{i,k|k}^{(u)}$
    end for
    for $u = 1, ..., V_{i,k|k-1}$ **do**
      $c_{s_{i,k}}(z) = \lambda_{c_i} A c_i(z)$
      $c_{t_{i,k}}(z) = \sum_{j=1,...,N\setminus i} \sum_{e=1}^{V_{j,k|k-1}} p_{ji,D} w_{j,k|k-1}^{(e)} \mathcal{N}(z; H_{ji,k} m_{j,k|k-1}^{(e)}, R_{ji,k} + H_{ji,k} P_{j,k|k-1}^{(e)} H_{ji,k}^T)$
      $c_{i,k}(z) = c_{s_{i,k}}(z) + c_{t_{i,k}}(z)$
      $w_{i,k,N} = \sum_{e=1}^{V_{i,k|k-1}} w_{i,k}^{(l_i V_{i,k|k-1}+e)}$
      $w_{i,k}^{(l_i V_{i,k|k-1}+u)} = \frac{w_{i,k}^{(l_i V_{i,k|k-1}+u)}}{c_{i,k}(z) + w_{i,k,N}}$
    end for
  end for
  $V_{i,k} = l_i V_{i,k|k-1} + V_{i,k|k-1}$
end for
output $\{w_{i,k}^{(v)}, m_{i,k}^{(v)}, P_{i,k}^{(v)}\}_{v=1}^{V_{i,k}}$

a large number of incorrect targets. For targets moving in video sequences there is no spawn process, but occlusions do result anywhere in the field of view, and may be caused either by other targets or other obstacles. Re-emerging targets are detected and constitute births, are not spawned because they may be occluded by obstacles other than targets, and have no a-priori location.

The prediction and update, steps 2 to 4, follow the standard procedures for the GM-PHD filter [8] but are extended to take into account the N detection processes and the subsequent confusion between detections. In the proposed algorithm, birth and prediction both precede the construction and update of the PHD components, so the total number at the conclusion of step 4 is the sum of the persistent and birthed components. The number of Gaussian components in the posterior intensities may increase without bound as time progresses, particularly as a birth at this stage may be due to an existing target that has moved from the previous frame and then is re-detected in the current frame. Therefore, it is necessary to prune weak and duplicated components in Algorithm 2. First, weak components with weight $w_{i,k}^{(v)} < T = 10^{-5}$ are pruned. Further, Gaussian components with Mahalanobis distance less than $U = 4$ pixels from each other are merged. These pruned and merged Gaussian components, output of Algorithm 2, are predicted as existing targets in the next iteration. Finally, Gaussian components of the posterior intensity, output of Algorithm 1, with means corresponding to weights greater than 0.5 as a threshold are selected as multi-target state estimates.

VI. Experimental Results

We apply the N-type GM-PHD filter to video sequences by integrating the object detectors' information such as the probabilities of detections for each target type and the confusion detection probabilities among target types at a specific background clutter rate. Accordingly, we consider two scenarios as follows.

A. Multiple Target, Implicit Multiple Type Tracking using a Tri-GM-PHD Filter

In this part, we consider tracking of football teams and a referee in the same scene handling their confusions using a

Algorithm 2 Pruning and merging for the N-type GM-PHD filter

given $\{w_{i,k}^{(v)}, m_{i,k}^{(v)}, P_{i,k}^{(v)}\}_{v=1}^{V_{i,k}}$ for target type $i \in \{1, ..., N\}$, a pruning weight threshold T, and a merging distance threshold U.
for $i = 1, ..., N$ **do** ▷ for all target type i
  Set $\ell_i = 0$, and $I_i = \{v = 1, ..., V_{i,k} | w_{i,k}^{(v)} > T\}$
  repeat
  $\ell_i := \ell_i + 1$
  $u := \arg\max_{v \in I_i} w_{i,k}^{(v)}$
  $L_i := \left\{ v \in I_i \middle| (m_{i,k}^{(v)} - m_{i,k}^{(u)})^T (P_{i,k}^{(v)})^{-1} (m_{i,k}^{(v)} - m_{i,k}^{(u)}) \leq U \right\}$
  $\tilde{w}_{i,k}^{(\ell_i)} = \sum_{v \in L_i} w_{i,k}^{(v)}$
  $\tilde{m}_{i,k}^{(\ell_i)} = \frac{1}{\tilde{w}_{i,k}^{(\ell_i)}} \sum_{v \in L_i} w_{i,k}^{(v)} m_{i,k}^{(v)}$
  $\tilde{P}_{i,k}^{(\ell_i)} = \frac{1}{\tilde{w}_{i,k}^{(\ell_i)}} \sum_{v \in L_i} w_{i,k}^{(v)} (P_{i,k}^{(v)} + (\tilde{m}_{i,k}^{(\ell_i)} - m_{i,k}^{(v)})(\tilde{m}_{i,k}^{(\ell_i)} - m_{i,k}^{(v)})^T)$
  $I_i := I_i \setminus L_i$
  until $I_i = \emptyset$
end for
output $\{\tilde{w}_{i,k}^{(v)}, \tilde{m}_{i,k}^{(v)}, \tilde{P}_{i,k}^{(v)}\}_{v=1}^{\ell_i}$ as pruned and merged Gaussian components for target type i.

tri-GM-PHD filter ($N = 3$). We call it implicit multiple type since the multiple types we are dealing with are fundamentally the same target type but grouped into sub-groups which we try to track and discriminate by handling their confusions.

1) Object Detection, Training and Evaluation: The RFS methodology post-processes a set of detections with parameters defining the probabilities of detection and clutter (false alarms). For the tri-PHD filter, we also need parameters for confusion. We employ the existing, state-of-the-art, Aggregated Channel Features (ACF) pedestrian detector [3] adapted to our data set due to its computational efficiency and ease of use. This uses three different kinds of features in 10 channels: normalized gradient magnitude (1 channel), histograms of oriented gradients (6 channels), and LUV color (3 channels). It is applied to detect the actors (football teams and a referee) using a sliding window at multiple scales. The Adaboost classifier [22] is used to learn and classify the feature vectors acquired by the ACF detector.

For training, evaluation and parameter setting we use the VS-PETS'2003 football video data[1]. This consists of 2500 frames which have players from the red and white teams and the referee. We trained 3 separate detectors for each target type (red, white, referee). We used every 10'th frame, i.e. 240 frames taken from the last 2400 frames, including 2000 positive samples for each footballer type, 240 samples for the referee, and 5000 randomly selected negative samples. This captures the appearance variation of players due to articulated motion. The correct player type or referee positions and windows were labeled manually for training as positive samples. The first 100 frames (video) are used to evaluate and test the tri-GM-PHD filter in comparison with repeated detection and three separate GM-PHD filters in section VI-A3.

The RFS methodology assumes point detections and a Gaussian error distribution on locations accuracy. However, humans in a video sequence are extended targets and the ACF detector has a bounding box that encloses the target. Therefore, overlapping detections are merged using a greedy non-maximum suppression (NMS) overlap threshold (intersection over union of two detections) of 0.05 (we made the overlap threshold very tight to ignore multiple bounding boxes on the same object). However, when evaluating the detectors, an overlap threshold (intersection over union of detection and ground truth bounding boxes) of 0.5 is used to identify true positives vs false positives. The receiver operating characteristic (ROC) curves for each of the detectors are given in Fig. 1.

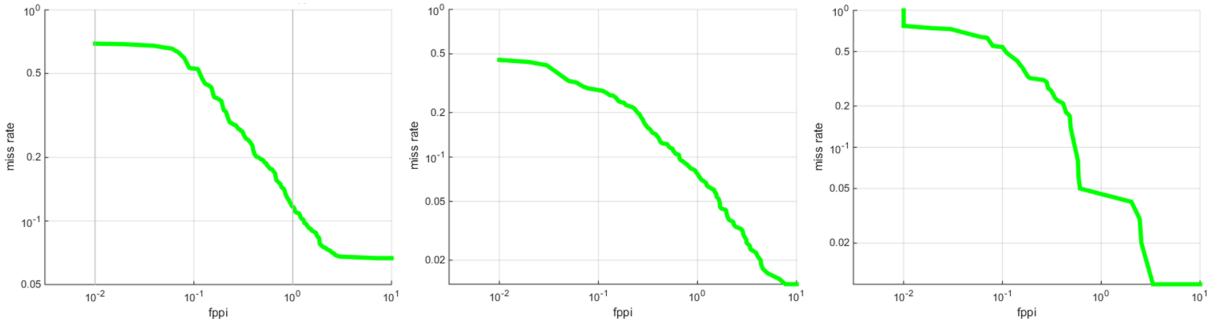
For the tri-GM-PHD strategy, we must set the thresholds on detection from the ROC curves in Fig. 1, taking into account the probabilities of confusion that arise from the corresponding ROC curves (not shown) of each detector applied to targets of a confusing type. From our own simulations and the published literature, e.g. [8], [20], we know that the RFS methodology is most effective when applied with a high probability of detection, albeit with a higher clutter rate, and in our case a higher confusion rate. Obviously, for a target detection to be useful, the probability of true detection must be higher than the probability of confusion. Therefore, from Fig. 1, we standardise a clutter rate of 10 false positive per image (fppi), which gives probabilities of detection of 0.93 (p_{11}), 0.99 (p_{22}) and 0.99 (p_{33}) for red, white and referee, respectively. With these values, the corresponding confusion parameters are 0.24 (white footballer detected as red, p_{21}), 0.5 (referee as red, p_{31}), 0.24 (red as white, p_{12}), 0.18 (referee as white, p_{32}), 0.19 (red as referee, p_{13}) and 0.17 (white as referee, p_{23}).

2) Data Association: The tri-GM-PHD filter handles sensor noise, clutter and distinguishes between true and false targets of each type. However, this does not distinguish between two different targets of the same type, so an additional step can be applied if we wish to identify different targets of the same type between consecutive frames. Although not part of the tri-GM-PHD strategy, this is commonly required so we include results from this post-labeling process for completeness in section VI-A3. It does not affect our error metrics but is a post-process to label individuals from frame to frame. For data association, the Euclidean distance between each previous filtered centroid (track) and the current filtered centroids is computed and we compute an assignment which minimizes the total cost returning assigned tracks to current filtered outputs. This assignment problem represented by the cost matrix is solved using Munkres's variant of the Hungarian algorithm [23].

This also returns the unassigned tracks and unassigned current filtered results. The unassigned tracks are deleted and the unassigned current filtered outputs create new tracks if the targets are not created earlier. If some targets are miss-detected and incorrectly labeled, labels are uniquely re-assigned by re-identifying them using the approach in [24].

3) Tracking Results: Referring to Eq. (1), our state vector includes the centroid positions, velocities, and the

[1]http://www.cvg.reading.ac.uk/slides/pets.html



(a) ROC for red team detector on red football instances (b) ROC for white team detector on white football instances (c) ROC for referee detector on referee instances

Fig. 1: Extracting detection probabilities for three target types (red, white and referee) from ROCs of 3 detectors: red team detector, white team detector and referee detector when tested on red team instances, white team instances and referee instances.

width and height of the bounding boxes, i.e. $x_k = [p_{cx,xk}, p_{cy,xk}, \dot{p}_{x,xk}, \dot{p}_{y,xk}, w_{xk}, h_{xk}]^T$. Similarly, the measurement is the noisy version of the target area in the image plane approximated with a w x h rectangle centered at $(p_{cx,zk}, p_{cy,zk})$ i.e. $z_k = [p_{cx,zk}, p_{cy,zk}, w_{zk}, h_{zk}]^T$.

As stated above, the detection and confusion probabilities are set by experimental evaluation of the ACF detection processes. Additional parameters are set from simulation and previous experience. For each target type, we set survival probabilities $p_{1,S} = p_{2,S} = p_{3,S} = 0.99$, and we assume the linear Gaussian dynamic model of Eq. (21) with matrices taking into account the box width and height at the given scale.

$$F_{i,k-1} = \begin{bmatrix} I_2 & \Delta I_2 & 0_2 \\ 0_2 & I_2 & 0_2 \\ 0_2 & 0_2 & I_2 \end{bmatrix},$$

$$Q_{i,k-1} = \sigma_{v_i}^2 \begin{bmatrix} \frac{\Delta^4}{4} I_2 & \frac{\Delta^3}{2} I_2 & 0_2 \\ \frac{\Delta^3}{2} I_2 & \Delta^2 I_2 & 0_2 \\ 0_2 & 0_2 & \Delta^2 I_2 \end{bmatrix}, \quad (32)$$

where I_n and 0_n denote the n x n identity and zero matrices, respectively and Δ is the sampling period defined by the time between frames (we use 1 second). $\sigma_{v_i} = 5$ pixels$/s^2$ are the standard deviations of the process noise for target type i where $i \in \{1, 2, 3\}$ i.e. type 1 (red football team), target type 2 (white football team) and target type 3 (a referee).

Similarly, the measurement follows the observation model of Eq. (22) with matrices taking into account the box width and height,

$$H_{ii,k} = H_{ji,k} = \begin{bmatrix} I_2 & 0_2 & 0_2 \\ 0_2 & 0_2 & I_2 \end{bmatrix},$$

$$R_{ii,k} = \sigma_{r_{ii}}^2 \begin{bmatrix} I_2 & 0_2 \\ 0_2 & I_2 \end{bmatrix},$$

$$R_{ji,k} = \sigma_{r_{ji}}^2 \begin{bmatrix} I_2 & 0_2 \\ 0_2 & I_2 \end{bmatrix}, \quad (33)$$

where $i \in \{1, 2, 3\}$, $j \in \{1, 2, 3\}$, and $\sigma_{r_{ii}}$ and $\sigma_{r_{ji}}$ are the measurement standard deviations taken from the distribution of distance errors of the centroids from ground truth in the evaluation of the detection process, effectively 6 pixels.

Accordingly, in our approach, positive detections specify the possible birth locations with the initial covariance given in Eq. (34). The current measurement and zero initial velocity are used as a mean of the Gaussian distribution using a pre-determined initial covariance for birthing of targets, i.e. new targets are born in the region of the state space for which the likelihood will have high values. Precisely, the birthing of targets is completely automatic using the very recent measurements obtained from object detectors. Very small initial weight (e.g. 10^{-4}) is assigned to the Gaussian components for new births as this is effective for high clutter rates. This is basically equivalent to the average number of appearing (birth) targets per frame (n_b) divided uniformly across the frame resolution (A).

$$P_{i,\gamma,k} = diag([100, 100, 25, 25, 20, 20]). \quad (34)$$

where $i \in \{1, 2, 3\}$.

We evaluate the tracking methodology of the tri-GM-PHD tracker in comparison with first, repeated independent detection on each frame, and second, with three independent GM-PHD trackers. Using the football video sequence, the examples shown in Fig. 2, Fig. 3 and Fig. 4 are for repeated detection (no tracking), three independent GM-PHD trackers, and the tri-GM-PHD tracker for frames 25, 57 and 73, respectively. Hence, Fig. 3a designates detections in which the red footballers, white footballers and the referee are detected both correctly and incorrectly, i.e. one object may be detected by many detectors. In this example the referee is detected 3 times: by the red team detector (red), by the white team detector (yellow) and the referee detector (black). Moreover, there are many background false positives (clutter) in the scene that arise from our choice to set the detection probability high at the expense of higher clutter as this is the detection scenario that is favored by the PHD process. Using the three independent GM-PHD trackers to effectively eliminate false positives, confused detections are not resolved as shown in Fig. 3b. However, our proposed tri-GM-PHD tracker effectively eliminates the false positives as well as confused detections as shown in Fig. 3c.

The tri-GM-PHD filter is evaluated quantitatively for the whole test sequence and compared with three independent GM-PHD filters and repeated detection using cardinality, OSPA metric [25], discrimination rate and time taken. We use the OSPA metric which is designed for evaluating RFS-based filters rather than multi-object tracking accuracy (MOTA) [26] which is widely used for evaluating other traditional multi-target tracking algorithms [27], [28]. Our algorithm is developed not only for tracking but also for discriminating different target types overcoming their confusions unlike algorithms such as [27], [28]. Therefore, the OSPA is the right evaluation metric to compare our approach with repeated raw detection and three independent GM-PHD trackers. The computational figures arise from experiments on a i5 2.50 GHz core processor with 6 GB RAM laptop using MATLAB and we acknowledge that these are not definitive and give a rough guide only to implementation costs. Though labeling of the targets using Munkres's variant of the Hungarian assignment algorithm works well as shown in Figs. 2c, 3c and 4c, we did not include this in our evaluation as it is not part of the quantitative comparison of the filtering and type labeling of either the detection or distinct GM-PHD filters. We present the cardinality and OSPA error plots in Fig. 5a and Fig. 5b respectively, in red for ground truth (cardinality), green for the tri-GM-PHD filter, blue for the three independent GM-PHD filters and magenta for repeated detection. As summarised in Table I the average absolute cardinality error using detection only is 10.22, reduced to 5.76 using the standard GM-PHD filters and to 0.11 using the tri-GM-PHD filter. The overall frame-averaged value of OSPA error for the tri-GM-PHD filter is 10.59 pixels, compared to three independent GM-PHD filters of 30.86 pixels, and repeated detections of 37.61 pixels. The proposed approach reduces the cardinality and OSPA errors by a large margin over three independent GM-PHD filters and repeated detection, although this has more computational cost as also shown in Table I.

Independent GM-PHD trackers do not take confusion into account, so treat such confusion as 'background' clutter; the problem is that such confused detections are not likely to be accurately modeled by random detections distributed uniformly in space as is commonly the case. Our approach can effectively discriminate true positives from clutter, while eliminating confused detections with a discrimination rate of 99.20%. The mis-discrimination rate of 0.80% occurs primarily during the initial frames (e.g. the first 7 frames) until the prediction-update process stabilises and the true detections are confirmed by the motion between adjacent frames.

Fig. 6 shows another example in which the individual footballers are detected, filtered, tracked and labeled for 100 frames. The image has been cropped as the action is confined to the top half of the image, and immediately follows a throw-in as the players move away left from the touchline. The examples also show the individual tracks and labels of the footballers and referee as small numbers over the targets. From this sequence, we see for example that the red player number 6 and the white player number 10, and several others, are consistently tracked through the sequence. However the labeling does occasionally make mistakes, for example red player 3 who starts near the touchline is finally labeled as red player number 49 in frame 293. In this instance the mislabeling is due to occlusion and lack of persistence in the detection and tracking as it uses successive frames only, so that if a player disappears then re-appears after several frames he is treated as a new target. Nevertheless, although this evaluation is not part of the Tri-GM-PHD filter, the labeling that we apply has good performance with a mean label switch error of only 0.43%.

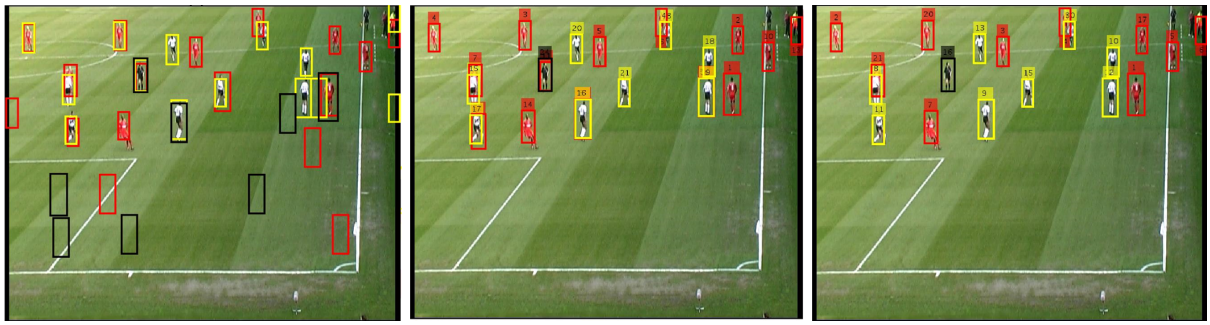
B. Multiple Target, Explicit multiple Type Tracking using a Dual GM-PHD Filter

In this part, we consider tracking of pedestrians and vehicles in the same scene handling their confusions using a dual GM-PHD filter ($N = 2$). We call it explicit multiple type since the multiple type we are dealing with are fundamentally different target types which we try to track and discriminate handling their confusions. The main difference between implicit and explicit target types is that in the latter case they can have different aspect ratios which can affect the means to differentiate the true and false positives in evaluating probabilities of confusion.

1) Pedestrian and Vehicle Detection: We adapted the ACF pedestrian detector [3] by considering appearance variations. Similarly, we adapted the vehicle detector in [29] which uses the same type of features as the ACF detector considering additional geometrical features such as truncation level, occlusion level and occlusion type features in addition to 3D orientation which depends on the ground truth information. However, we only consider 3D geometric orientation from the ground truth information available in the KITTI dataset [30]. Similar to [29], we also consider visual features to capture appearance variations due to varying orientation, truncation and occlusion degree for detecting both pedestrians and vehicles.

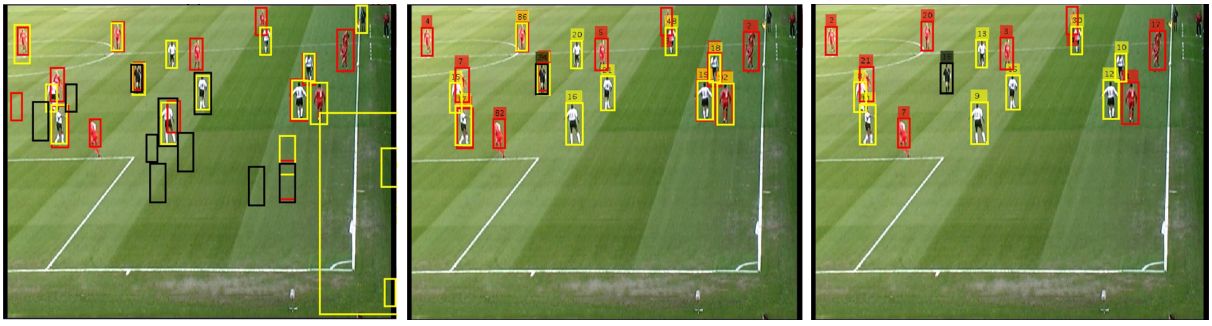
3D orientation: Appearance variation due to observation angle is common when detecting vehicles in different driving settings. Accordingly, the observation angle i.e. relative orientation of the object with respect to the camera is used by considering the angle of the vector joining the camera center in 3D and an object which takes into account the ego-vehicle. This 3D geometric orientation is available in KITTI dataset ranging from -π (-3.14) to π (3.14) and is quantized into L labels (L = 5 for pedestrians and L = 20 for vehicles). The mean of the aspect ratios of the image instances (samples) with the specific quantized label is used as an aspect ratio for which a specific detector model is trained on that specific image instances.

Visual features: We use visual features by clustering them as a means of capturing appearance variations of objects to detect them under challenging appearance changes. This approach is very generic as it does not depend on the availability of ground truth orientation though it gives slightly less accuracy when compared to 3D geometrical orientation. Though color and gradient features (HOG, LUV color and normalized gradient magnitude) can also be used [29], in our case, high quality convolutional neural network (CNN) features from a R-CNN object detector [31] are used to learn appearance variations of objects. The R-CNN object detector model is then used to



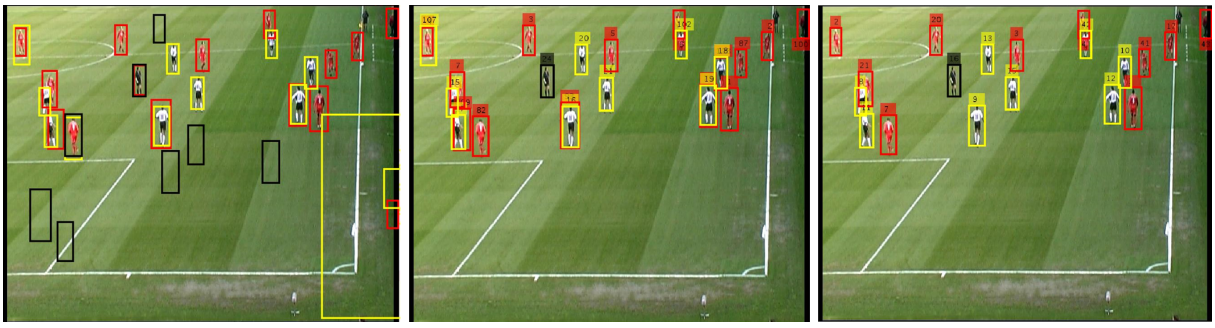
(a) Detections, frame 25 (b) Three independent GM-PHD trackers, frame 25 (c) Tri-GM-PHD Tracker, frame 25

Fig. 2: Results of detections, three independent GM-PHD trackers and tri-GM-PHD tracker, respectively, for frame 25.



(a) Detections, frame 57 (b) Three independent GM-PHD trackers, frame 57 (c) Tri-GM-PHD Tracker, frame 57

Fig. 3: Results of detections, three independent GM-PHD trackers and tri-GM-PHD tracker, respectively, for frame 57.



(a) Detections, frame 73 (b) Three independent GM-PHD trackers, frame 73 (c) Tri-GM-PHD Tracker, frame 73

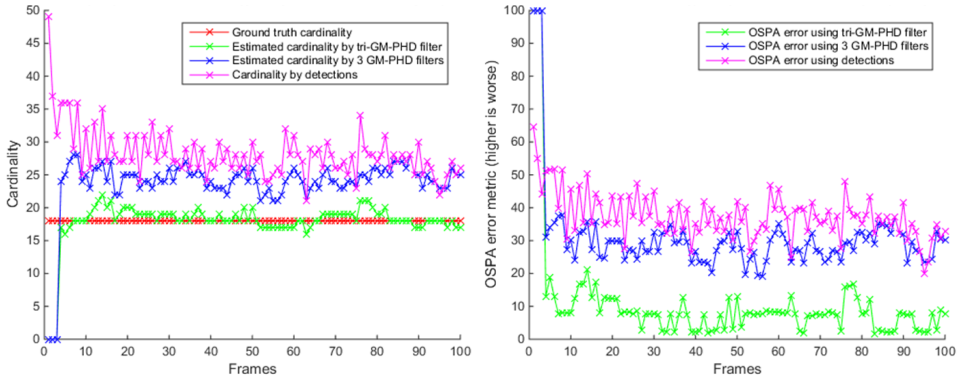
Fig. 4: Results of detections, three independent GM-PHD trackers and tri-GM-PHD tracker, respectively, for frame 73.

Method	frame averaged cardinality	frame-averaged OSPA error	time taken	discrimination rate
Detections	10.22	37.61 pixels	0.59 seconds/frame	0%
3 GM-PHDs	5.76	30.86 pixels	0.80 seconds/frame	0%
Tri-GM-PHD	0.11	10.59 pixels	3.00 seconds/frame	99.20%

TABLE I: Cardinality and OSPA errors, time taken and discrimination rate at the extracted detection probabilities for tri-GM-PHD filter, three independent GM-PHD filters and Detections.

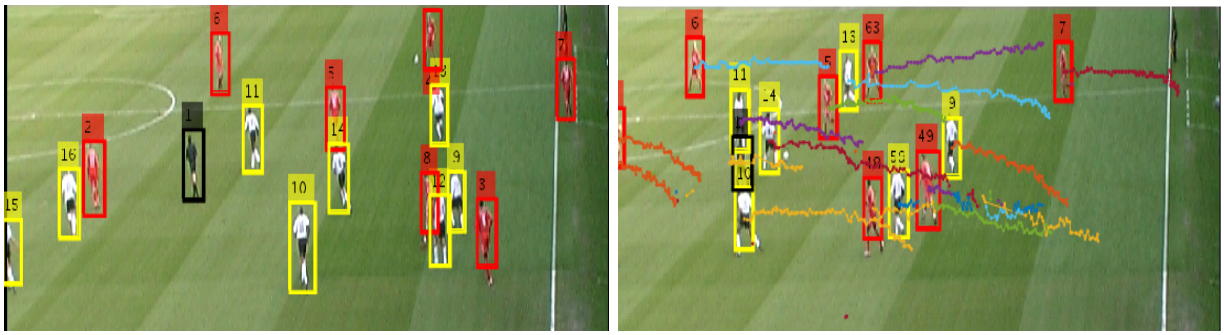
extract 4096-dimensional CNN features from cropped KITTI image samples which is then reduced dimensionally using PCA. This dimensionally reduced features are then clustered using k-means clustering giving cluster labels. The mean of the aspect ratios of the image instances assigned the same label is used to learn a specific detector on the image instances with that specific label.

Pedestrian Detection: The ACF pedestrian detector [3] detects pedestrians at multiple scales using the Adaboost classifier. However, unlike the original ACF [3], we consider the



(a) Cardinality (b) OSPA error

Fig. 5: Cardinality and OSPA error: Ground truth (red for cardinality only), tri-GM-PHD filter (green), three independent GM-PHD filters (blue), detections (magenta).



(a) Frame 193 (b) Tracking both teams and referee, frame 293

Fig. 6: Tracking the red and white teams, and referee from frame 193 to frame 293

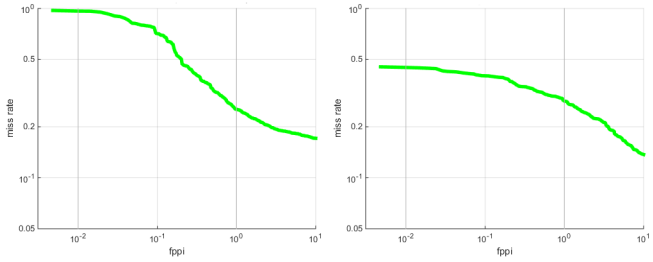
appearance variations of pedestrians due to varying orientation, truncation and occlusion levels. The KITTI benchmark [30] consists of 7481 training frames from which around 3583 pedestrian instances are extracted in moderate setting. From these 7481 training frames, 6501 frames are used as the training set and the rest (980 frames) as the testing or validation set. To capture the appearance variations, we trained 5 geometrical orientations-based detection models and 5 visual CNN features clustering-based detection models which perform better than using only one model as in [3]. This is necessary as using only one model, it is not possible to get the required detection performance for extracting detection probabilities on this data set. Overlapping detections are merged using a greedy NMS overlap threshold of 0.1. However, when evaluating the detector, an overlap threshold of 0.5 is used to identify true positives vs false positives.

Vehicle Detection: Capturing appearance variations of vehicles due to changing observation angle, illumination variability, vehicle shape and type, truncation, out of camera view, different occlusion levels, etc is very important for developing a vehicle detector [29]. Unlike the approach considered in [29], we consider only 3D orientation rather than other geometrical features such as truncation level, occlusion level and occlusion type features from the ground truth available in KITTI. Moreover, we used visual features which can also capture appearance variations due to varying orientation, truncation and occlusion. The KITTI benchmark [30] consists of 7481 training frames from which around 16105 car instances are extracted in moderate setting. From these 7481 training frames, 6501 frames are used as the training set and the rest (980 frames) as the testing or validation set. Accordingly, we trained 20 geometrical orientations-based detection models and 20 visual CNN features clustering-based detection models which perform better than using only one model. We use a greedy NMS overlap threshold of 0.2 to merge overlapping detections, and an overlap threshold of 0.5 is used to identify true positives vs false positives when evaluating the detector.

Detection Parameters Extraction: The detection parameters, detection probabilities ($p_{11,D}$, $p_{22,D}$) and confusion detection probabilities ($p_{12,D}$, $p_{21,D}$), are extracted as follows. The detection probability for pedestrians by a pedestrian detector, $p_{11,D}$, can be extracted from the ROC curve of the pedestrian detector when it is tested on pedestrian instances. Similarly, the detection probability for vehicles by a vehicle detector, $p_{22,D}$, is obtained from the ROC curve of the vehicle detector when it is tested on vehicle instances. The confusion detection probabilities, detection probability for pedestrians by a vehicle detector, $p_{12,D}$, and detection probability for vehicles by a

pedestrian detector, $p_{21,D}$, can also be obtained when the vehicle detector is evaluated on pedestrian instances and when the pedestrian detector is evaluated on vehicle instances, respectively.

Accordingly, when we want to track on video sequences, we first run both pedestrian and vehicle detectors on that specific video sequence to obtain their ROC curves. We used a combination of CNN-based and 3D orientation-based detectors. Thus, the ROC curve of the pedestrian detector when applied to pedestrian instances in the KITTI video tracking sequence 16 is shown in Fig. 7a from which $p_{11,D}$ of 0.83 is obtained at clutter rate (false positive per image - fppi) of 10. Similarly, $p_{22,D}$ of 0.86 is obtained at fppi of 10 from the ROC curve of the vehicle detector when it is applied to vehicle instances as shown in Fig. 7b. However, the values of $p_{12,D}$ and $p_{21,D}$ are very low, around 0.03 for $p_{12,D}$ and 0.01 for $p_{21,D}$, this happens because even if the vehicle detector detects pedestrian instances, for example, the intersection of the detected bounding boxes by vehicle detector on pedestrian instances and the ground truth of the pedestrian instances is very low as the two bounding boxes have very much different aspect ratios, therefore, it can be classified as false positive though it is detected. Hence, we try to fine-tune values of $p_{12,D}$ and $p_{21,D}$ to some higher values e.g. $p_{12,D} = 0.3$ and $p_{21,D} = 0.1$.



(a) ROC for pedestrian detector on pedestrian instances (b) ROC for vehicle detector on vehicle instances

Fig. 7: ROCs using 3D orientation and CNN visual features detector models tested on KITTI sequence 16.

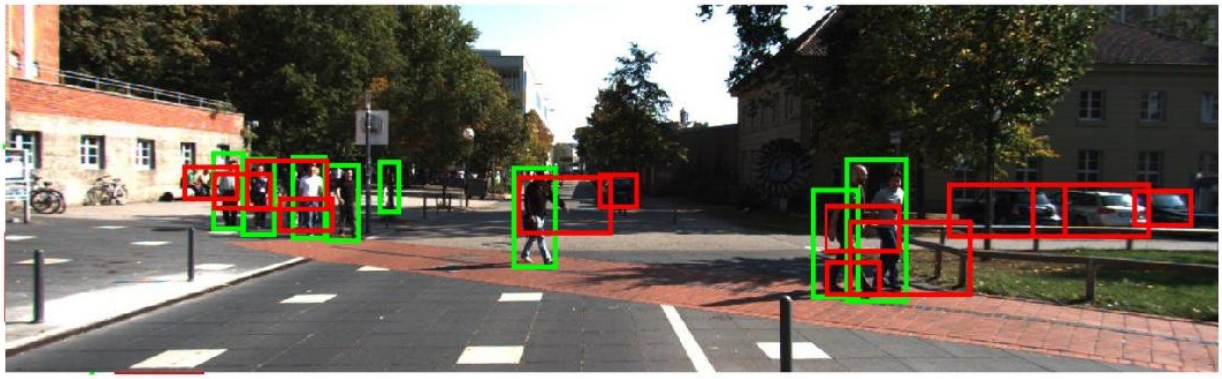
2) Data Association and Tracking Results: Munkres's variant of the Hungarian assignment algorithm [23] is used to associate tracked target identities between two consecutive frames in the same fashion discussed in section VI-A2 with the exception that if a target disappears and then reappears, a new label is given without using any re-identification algorithm.

The state vector includes the centroid positions, velocities, and the width and height of the bounding boxes; the measurement is the noisy version of the target area, in the same fashion as in section VI-A3. A dynamic model, an observation model and a birth covariance follow Eqs. (32), (33) and (34) respectively with the exception of setting $i \in \{1, 2\}$. We set $\sigma_{v_1} = 5$ pixels$/s^2$ and $\sigma_{v_2} = 6$ pixels$/s^2$ for target type 1 (pedestrians) and target type 2 (vehicles), respectively. We also set survival probabilities $p_{1,S} = p_{2,S} = 0.99$ for each target of both types, and the measurement standard deviations $\sigma_{r_{ii}}$ and $\sigma_{r_{ij}}$ ($i \in \{1, 2\}$ and $j \in \{1, 2\}$) are evaluated to 7 pixels.

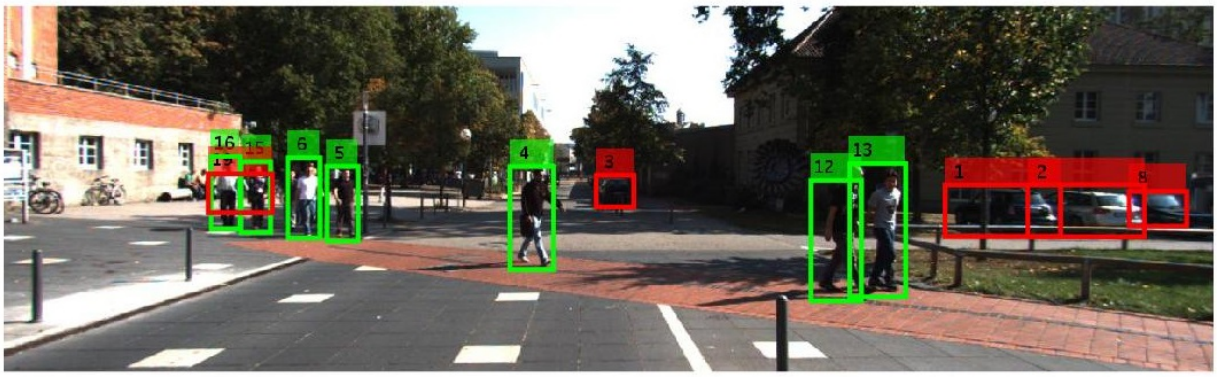
This proposed visual tracking approach is analyzed using the KITTI tracking video sequence 16. A multi-target measurement for pedestrians is obtained using a pedestrian detector and a multi-target measurement for vehicles is obtained using a vehicle detector. The sample frames of results of detections, two independent GM-PHD trackers and dual GM-PHD tracker are shown in Fig. 8 and Fig. 9. For instance, for the sample frame 23 in Fig. 9, many clutter responses from detections in Fig. 9a are removed by 2 independent GM-PHD trackers as shown in Fig. 9b. However, the pedestrian targets with labels 10, 30, 34 and 31 are confused by the vehicle detector and then tracked by standard GM-PHD trackers as shown in Fig. 9b. These are removed by our dual GM-PHD tracker as shown in Fig. 9c. Hence, our approach eliminates the wrong tracking of vehicles or pedestrians which are confused at detection.

The dual GM-PHD filter is evaluated quantitatively and compared with two independent GM-PHD filters and raw detection using the cardinality error, OSPA metric [25], time taken and discrimination rate in Table II. We also show the cardinality and OSPA error plots as shown in Fig. 10a and Fig. 10b, respectively, in red for ground truth (cardinality), green for dual GM-PHD filter, blue for two independent GM-PHD filters and magenta for detections. As shown in Table II, the overall average value of the OSPA error for the dual GM-PHD filter is 20.74 pixels compared to using two independent GM-PHD filters of 35.29 pixels and raw detection of 49.81 pixels. Our proposed approach reduces the OSPA error by a large margin over both using two independent GM-PHD filters and raw detection.

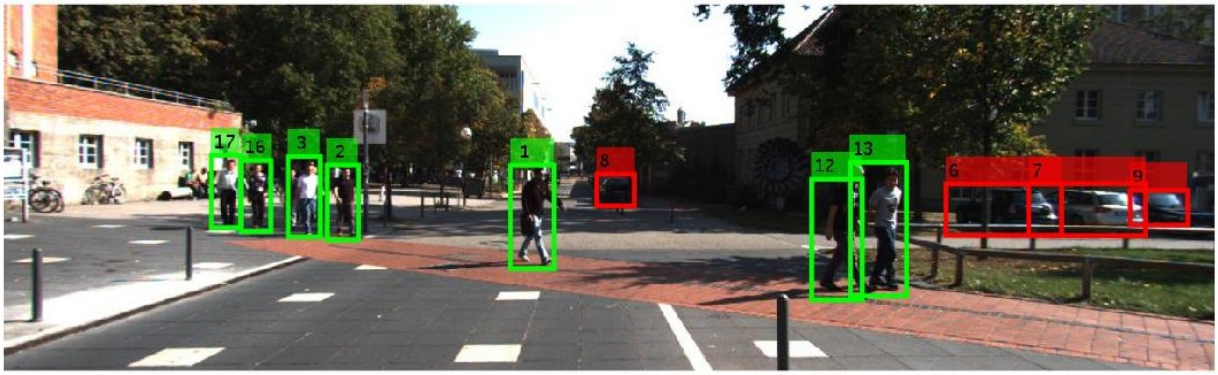
The time taken for the 209 video frames of the KITTI tracking sequence 16 is shown in Table II where the dual GM-PHD tracker takes 4.03 seconds per frame (both detection and tracking), two independent GM-PHD trackers take 1.61 seconds per frame and raw detection takes 1.25 seconds per frame on a i5 2.50 GHz core processor with 6 GB RAM laptop using MATLAB. Since we use many detection models for each actor (10 for pedestrians, 40 for vehicles), it takes more computational time than the scenario considered in section VI-A. As shown in Table II, the dual GM-PHD tracker has only 0.32 (below 1 target) cardinality error and 1.81% discrimination rate error when compared to 3.82 cardinality error and 100% discrimination rate error using 2 independent GM-PHD trackers as well as 9.86 cardinality error and 100% discrimination rate error using raw detection. As can be seen from Fig. 8c and Fig. 9c, labels of some of the actors (cars - labels 6 and 9; pedestrians - labels 1, 2, 16, 17, etc) are consistent from frame to frame. Since we are using two frames to associate the targets, a new label is given to a target which disappears and reappears as well as for a newly appearing target. For example, pedestrians labeled 12 and 13 in frame 13 are re-detected as one target in frame 23 and given a label 26. The car labeled 7 in frame 13 is miss-detected, and then is re-detected in frame 23 and is given a new label, 25. Still, the labeling approach we use has reasonable performance with a mean label switch error of only 1.07%, and it is obviously not part of the dual GM-PHD filter.



(a) Detections, frame 13



(b) Two independent GM-PHD trackers, frame 13



(c) Dual GM-PHD Tracker, frame 13

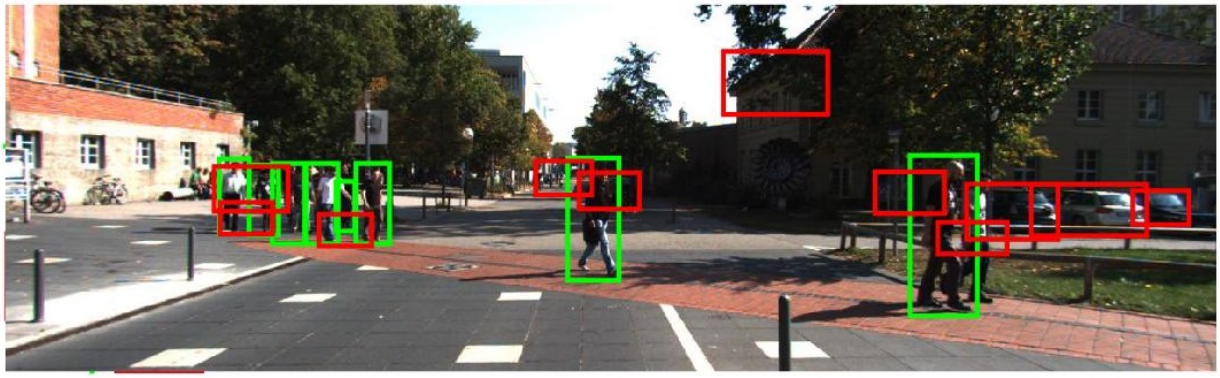
Fig. 8: Results of detections, two independent GM-PHD trackers and dual GM-PHD tracker, respectively, for frame 13.

Method	cardinality error	OSPA error	time taken	discrimination rate
Detections	9.86	49.81 pixels	1.25 sec/frame	0%
2 GM-PHDs	3.82	35.29 pixels	1.61 sec/frame	0%
Dual GM-PHD	0.32	20.74 pixels	4.03 sec/frame	98.19%

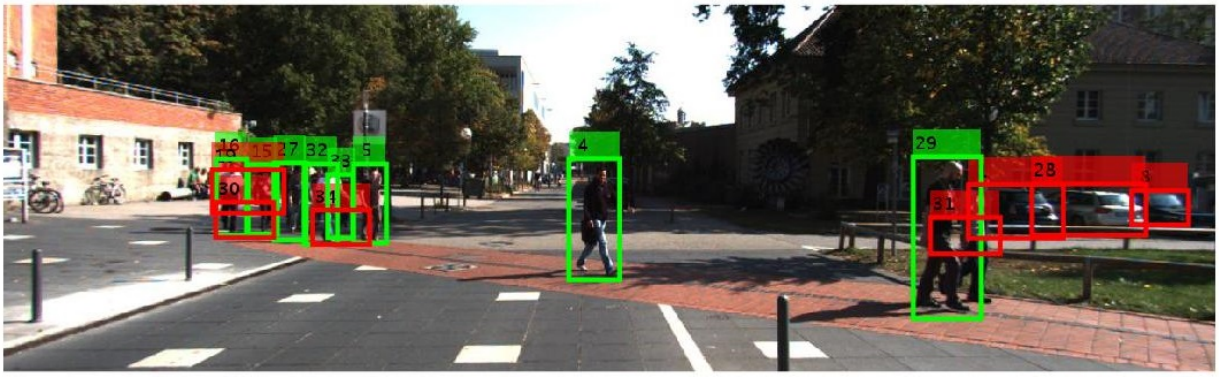
TABLE II: Frame-averaged cardinality and OSPA errors, time taken and discrimination rate at the extracted detection probabilities for dual GM-PHD filter, two independent GM-PHD filters and Detections.

VII. Conclusions

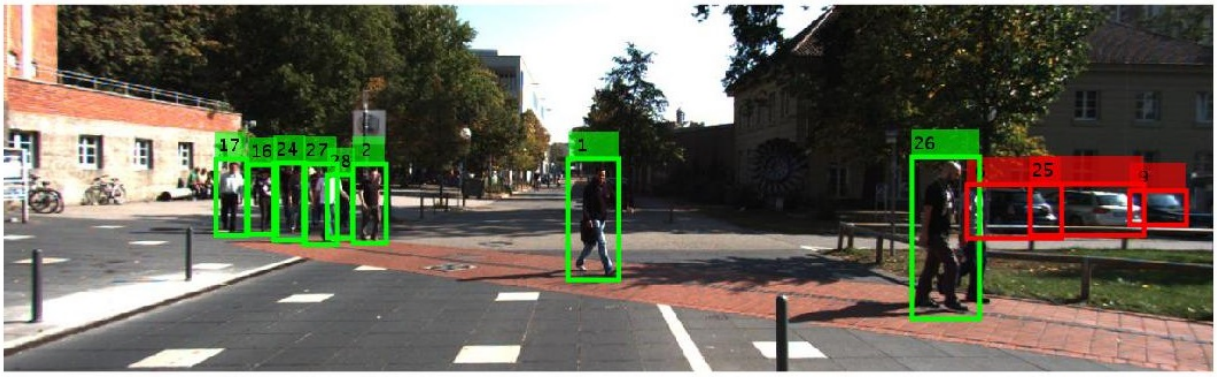
We have developed an extension of the PHD filter in the RFS framework to account for $N \geq 2$ different types of multiple targets with separate observations in the same scene, allowing for different probabilities of detection, scene clutter and possible confusions between targets of different types at the detection stage. This extends the standard GM-PHD filter [8] to a N-type GM-PHD filter. This has been tested and evaluated using video sequences with the separate targets defined as different team players and the referee, and pedes-



(a) Detections, frame 23



(b) Two independent GM-PHD trackers, frame 23

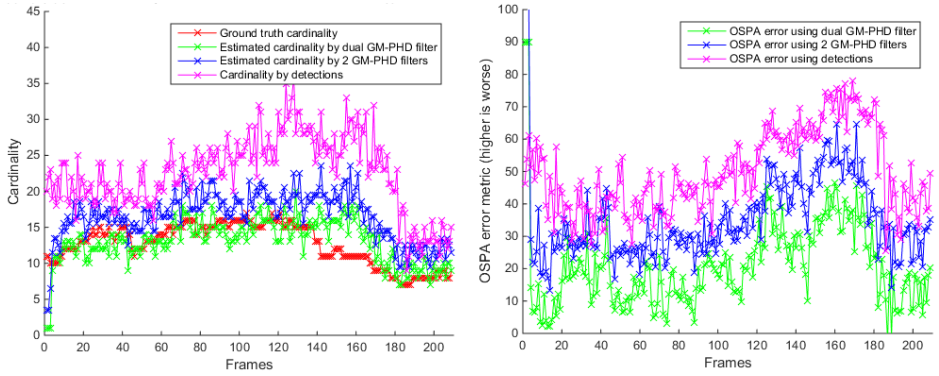


(c) Dual GM-PHD Tracker, frame 23

Fig. 9: Results of detections, two independent GM-PHD trackers and dual GM-PHD tracker, respectively, for frame 23.

trians and vehicles. We have also applied Munkres's variant of the Hungarian assignment algorithm as data association on the filtered results of the filter as a post-process. The key finding of this work is that by considering and modeling confusions between the different types of target and detector we can improve the target discrimination rate, demonstrated by quantitative measurement of cardinality and the OSPA score.

Although the process has been applied here to 3 and 2 types of target, in principle the methodology can be applied to N types of targets where N is a variable, with the caveat that the number of possible confusions may rise as $N(N-1)$. The N-type GM-PHD filter degrades to N GM-PHD filters when we set the probabilities of confusion to 0.0 i.e. no target confusions. However, if each target is regarded as a type, the N-type GM-PHD filter is used as a labeler of each target i.e. it discriminates those targets from frame to frame whether or not confusions between targets exist rather than simply degrading to N standard GM-PHD filters. We also observe that other assumptions about clutter, target location and birth follow the same random models as the standard PHD filter. Hence we assume that our background clutter, and the detection and confusion probabilities are uniform across the image field, which is not unreasonable in the football data, but is less likely to be true when identifying pedestrians in an urban



(a) Cardinality (b) OSPA metric error

Fig. 10: Cardinality and OSPA error: Ground truth (red for cardinality only), dual GM-PHD filter (green), two independent GM-PHD filters (blue), detections (magenta).

environment, where particular street furniture may generate repeated false alarms. As the detections are represented as points (centroids of bounding detection boxes), the filtering process does not explicitly consider scale, and as the boxes and humans/vehicles within the boxes have finite extent, this makes occlusions possible such that targets may disappear for several frames. Notwithstanding these imperfections, the work we have done has shown that the N-type GM-PHD filter has potential both to track targets in video data, and to better address multiple target confusions than the standard method.

ACKNOWLEDGMENT

We would like to acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC), grant references EP/K009931, EP/J015180 and a James Watt Scholarship. We would also like to thank Dr. Daniel Clark for sharing his expertise and understanding of RFS methodology.

REFERENCES

[1] P. Matzka, A. Wallace, and Y. Petillot, "Efficient resource allocation for automotive attentive vision systems," *IEEE Trans. on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 859–872, 2012.

[2] J. Liu and P. Carr, "Detecting and tracking sports players with random forests and context-conditioned motion models," in *Computer Vision in Sports*. Springer, 2014, pp. 113–132.

[3] P. Dollar, R. Appel, P. Perona, and S. Belongie, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, p. 14, 2014.

[4] Y. Cai, N. de Freitas, and L. JJ, "Robust visual tracking for multiple targets," in *IN ECCV*, 2006, pp. 107–118.

[5] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 560–576, 2001.

[6] T.-J. Cham and J. M. Rehg, "A multiple hypothesis approach to figure tracking." in *CVPR*. IEEE Computer Society, 1999, pp. 2239–2245.

[7] R. P. Mahler, "Multitarget bayes filtering via first-order multitarget moments," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, 2003.

[8] B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4091–4104, Nov 2006.

[9] B.-N. Vo, S. Singh, and A. Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1224–1245, 2005.

[10] X. Zhou, Y. Li, B. He, and T. Bai, "GM-PHD-based multi-target visual tracking using entropy distribution and game theory," *Industrial Informatics, IEEE Transactions on*, vol. 10, no. 2, pp. 1064–1076, May 2014.

[11] N. L. Baisa, D. Bhowmik, and A. Wallace, "Long-term correlation tracking using multi-layer hybrid features in dense environments," in *Proceedings of the 12th International Conference on Computer Vision Theory and Applications (VISAPP), VISIGRAPP*, 2017.

[12] S. Pasha, B.-N. Vo, H. D. Tuan, and W.-K. Ma, "A gaussian mixture PHD filter for jump markov system models," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 45, no. 3, pp. 919–936, July 2009.

[13] E. Maggio, M. Taj, and A. Cavallaro, "Efficient multi-target visual tracking using random finite sets," *IEEE Transactions On Circuits And Systems For Video Technology*, pp. 1016–1027, 2008.

[14] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, *Online Multi-target Tracking with Strong and Weak Detections*. Cham: Springer International Publishing, 2016, pp. 84–99.

[15] Y. Wei, F. Yaowen, L. Jianqian, and L. Xiang, "Joint detection, tracking, and classification of multiple targets in clutter using the PHD filter," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 48, no. 4, pp. 3594–3609, October 2012.

[16] W. Yang, Y. Fu, and X. Li, "Joint target tracking and classification via RFS-based multiple model filtering," *Information Fusion*, vol. 18, pp. 101–106, Jul. 2014.

[17] V. Romero-Cano, G. Agamennoni, and J. Nieto, "A variational approach to simultaneous multi-object tracking and classification," *Int. J. Rob. Res.*, vol. 35, no. 6, pp. 654–671, May 2016.

[18] N. L. Baisa and A. Wallace, "Multiple Target, Multiple Type Filtering in RFS Framework," *ArXiv e-prints*, May 2017.

[19] N. L. Baisa and A. Wallace, "Multiple target, multiple type visual tracking using a Tri-GM-PHD Filter," in *Proceedings of the 12th International Conference on Computer Vision Theory and Applications (VISAPP), VISIGRAPP*, 2017.

[20] B. Ristic, D. E. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for PHD and CPHD filters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 2, pp. 1656–1668, 2012.

[21] B. Ristic, D. Clark, and B.-N. Vo, "Improved SMC implementation of the PHD filter," in *Information Fusion (FUSION), 2010 13th Conference on*, 2010, pp. 1–8.

[22] R. Appel, T. Fuchs, P. Dollar, and P. Perona, "Quickly boosting decision trees – pruning underachieving features early," in *ICML*, vol. 28, no. 3, May 2013, pp. 594–602.

[23] F. Bourgeois and J.-C. Lassalle, "An extension of the munkres algorithm for the assignment problem to rectangular matrices," *Commun. ACM*, vol. 14, no. 12, pp. 802–804, Dec. 1971.

[24] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3908–3916.

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

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

[27] J. H. Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon, "Online multi-object tracking via structural constraint event aggregation." in *CVPR*, 2016.

[28] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[29] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Transactions on Intelligent Transportation Systems*, 2015.

[30] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014.