Multiple Target, Multiple Type Visual Tracking using a Tri-GM-PHD Filter

Nathanael L. Baisa and Andrew Wallace
Department of Electrical, Electronic and Computer Engineering, Heriot Watt University, Edinburgh, U.K.
{*nb30, a.m.wallace*}*@hw.ac.uk*

Keywords: Visual Tracking, Random Finite Set, Multiple Target Filtering, Gaussian Mixture, Tri-GM-PHD Filter, OSPA Metric.

Abstract: We propose a new framework that extends the standard Probability Hypothesis Density (PHD) filter for multiple targets having three different types, taking into account not only background false positives (clutter), but also confusion between detections of different target types, which are in general different in character from background clutter. Our framework extends the existing Gaussian Mixture (GM) implementation of the PHD filter to create a tri-GM-PHD filter based on Random Finite Set (RFS) theory. The methodology is applied to real video sequences containing three types of multiple targets in the same scene, two football teams and a referee, using separate detections. Subsequently, 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 detections and independent GM-PHD filters using the Optimal Sub-pattern Assignment (OSPA) metric and discrimination rate. This shows the improved performance of our strategy on real video sequences.

1 INTRODUCTION

Visual detection, tracking and association of multiple targets at each frame in a video sequence is an active research field. 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 (Matzka et al., 2012). For sports analysis we often want to track and discriminate sub-groups of the same target type such as the players in opposing teams (Liu and Carr, 2014). In this and many other examples, confusion between target types is common; a standard histogram-based detection strategy (Dollar et al., 2014) in an urban environment may provide confused detections between pedestrians and cyclists, and even small cars.

Traditional multi-target trackers have been based on finding associations between targets and measurements. These include Global Nearest Neighbor (GNN) (Cai et al., 2006), Joint Probabilistic Data Association Filter (JPDAF) (Rasmussen and Hager, 2001), and Multiple Hypothesis Tracking (MHT) (Cham and Rehg, 1999). 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.

To address the problem 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 (Mahler, 2003). This estimates the states and cardinality of an unknown and time varying number of targets in the scene, and allows for target birth, death, clutter (false alarms), and missing detections. Mahler (Mahler, 2003) 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 implementations for the PHD filter, the Gaussian Mixture (GM-PHD) (Vo and Ma, 2006) and the Sequential Monte Carlo (SMC) or particle-PHD filter (Vo et al., 2005). The GM-PHD filter is used in (Zhou et al., 2014) for tracking pedestrians in video sequences but there is only one type of target and the motion model is fixed. As an extension, a GM-PHD Filter was also developed in (Pasha et al., 2009) 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 (Maggio et al., 2008) to allow for

L. Baisa N. and Wallace A.
Multiple Target, Multiple Type Visual Tracking using a Tri-GM-PHD Filter.
DOI: 10.5220/0006145704670477
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 467-477
ISBN: 978-989-758-227-1

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.

Considering extensions to different target types, Yan et al. (Wei et al., 2012) 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 is 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 (Yang et al., 2014) using RFS which takes into account extraneous target-originated 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. However, although multiple target types were considered, no account was taken of the effect of confusion between target types at the detection stage, as is the case in our work.

We make the following four contributions. First, we model the RFS filtering of three different types of multiple targets with separate but confused detections. Second, the Gaussian mixture implementation of the standard PHD filter is extended for the proposed tri-PHD filter. Third, we extract object detectors' information including the probabilities of detection, confusion detection probabilities among target types and background clutter from receiver operating characteristic (ROC) curves of each of the detectors and then integrate them into tri-GM-PHD filter to apply for visual tracking on real video sequences. Finally, we integrate Munkres's variant of the Hungarian assignment algorithm to the typed results from the tri-GM-PHD filter to determine individual targets of each type between consecutive frames.

2 RANDOM FINITE SET, MULTIPLE TARGET FILTERING FOR THREE TYPES

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 (Mahler, 2003).

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. In this paper we set $N = 3$.

To derive the tri-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 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 $i \in \{1, ..., 3\}$. 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 $i \in \{1, ..., 3\}$. As stated above, some of these observations will 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 Random Finite Sets (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 $J = \{1,...,3\} \setminus i$ and $\Theta_{i,k}(X_{i,k})$ is the RFS modeling the measurements generated by the target $X_{i,k}$, and $C_{s_{i,k}}$ models the RFS associated with the clutters (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 $J = \{1,...,3\} \setminus i$, 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 $f_{ij,k}(Z_{i,k}|X_{j,k})$ for target type $i \in \{1,...,3\}$ from detector $j \in \{1,...,3\}$. The recursive equations are

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

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

where μ is an appropriate dominating measure on $\mathcal{F}(\mathcal{X})$ (Mahler, 2003). We extend Mahler's method of propagating the first-order moment of the multi-target posterior instead of the full multi-target posterior for $N = 3$ types of multiple targets by deriving the updated PHD from the Probability Generating Functional (PGFL) for our tri-PHD filter.

2.1 Tri-PHD Filtering Strategy

The PHDs, $\mathcal{D}_{\Xi_1}(x)$, $\mathcal{D}_{\Xi_2}(x)$, $\mathcal{D}_{\Xi_3}(x)$, are the first-order moments of RFSs, Ξ_1, Ξ_2, Ξ_3, and are intensity functions on a single state space $\mathcal{X}$ whose peaks identify the likely positions of the targets. For any region $R \subseteq \mathcal{X}$

$$E[|(\Xi_1 \cup \Xi_2 \cup \Xi_3) \cap R|] = \sum_{i=1}^{3} \int_R \mathcal{D}_{\Xi_i}(x)dx \tag{7}$$

where$|.|$ is used to denote the cardinality of a set. In practice, Eq. (7) 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

$$\begin{aligned} \mathcal{D}_{i,k|k-1}(x) = \quad & \int p_{s_i,k|k-1}(\zeta) y_{i,k|k-1}(x|\zeta) \mathcal{D}_{i,k-1|k-1}(\zeta)d\zeta \\ & + \gamma_{i,k}(x), \end{aligned} \tag{8}$$

where $\gamma_{i,k}(.)$ is the intensity function of a new target birth RFS $\Gamma_{i,k}$, $p_{s_i,k|k-1}(\zeta)$ 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, the final updated PHD for target type i is obtained by

$$\begin{aligned} &\mathcal{D}_{i,k|k}(x) = \Bigg[1 - p_{ii,D}(x) + \\ &\sum_{z \in Z_{i,k}} \frac{p_{ii,D}(x) f_{ii,k}(z|x)}{c_{s_{i,k}}(z) + c_{t_{i,k}}(z) + \int p_{ii,D}(\xi) f_{ii,k}(z|\xi) \mathcal{D}_{i,k|k-1}(\xi)d\xi} \Bigg] \mathcal{D}_{i,k|k-1(x)}, \end{aligned} \tag{9}$$

The clutter intensity $c_{t_{i,k}}(z)$ due to all types of targets $j = 1,...,3$ except target type i in (9) is given by

$$c_{t_{i,k}}(z) = \sum_{j=1,...,3 \setminus i} \int p_{ji,D}(y) \mathcal{D}_{j,k|k-1}(y) f_{ji,k}(z|y)dy, \tag{10}$$

This means that when filtering target type i, all the other target types are included as confusing detections. (10) 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 the target type j conditioned on state x from detector i taking into account the confusion probability $p_{ji,D}(y)$, when target type j is detected by detector i.

The clutter intensity due to the background i, $c_{s_{i,k}}(z)$, in (9) is given by

$$c_{s_{i,k}}(z) = \lambda_i c_i(z) = \lambda_{c_i} A c_i(z), \tag{11}$$

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 tri-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 ($N = 3$) i.e. computational order of $O(3mn)$.

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 ROC curves of the detection processes. Moreover, the probabilities of detection $p_{ii,D}(x)$ and $p_{ij,D}(x)$ may all be different although assumed constant across both the time and space.

2.2 Tri-PHD Filter Implementation based on Gaussian Mixture

The Gaussian mixture implementation of the standard PHD (GM-PHD) filter (Vo and Ma, 2006) is a closed-form solution of the PHD filter that assumes a linear Gaussian system. In this section, this is extended for the tri-PHD filter by solving (10). Assuming each target follows a linear Gaussian model,

$$y_{i,k|k-1}(x|\zeta) = \mathcal{N}(x; F_{i,k-1}\zeta, Q_{i,k-1}) \quad (12)$$

$$f_{ij,k}(z|x) = \mathcal{N}(z; H_{ij,k}x, R_{ij,k}) \quad (13)$$

where $\mathcal{N}(.;m,P)$ denotes a Gaussian density with mean m and covariance P; $F_{i,k-1}$ and $H_{ij,k}$ are the state transition and measurement matrices, respectively. $Q_{i,k-1}$ and $R_{ij,k}$ are the covariance matrices of the process and the measurement noise, respectively, where $i \in \{1,2,3\}$ and $j \in \{1,2,3\}$. A measurement driven birth intensity, similar in principle to (Ristic et al., 2012), 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) (Ristic et al., 2010) or a-priori birth (Vo and Ma, 2006) and is discussed further in Section 5. The intensity of the spontaneous birth RFS is $\gamma_{i,k}(x)$ for target type i

$$\gamma_{i,k}(x) = \sum_{v=1}^{V_{\gamma_i,k}} w_{i,\gamma,k}^{(v)} \mathcal{N}(x; m_{\gamma_i,k}^{(v)}, P_{\gamma_i,k}^{(v)}) \quad (14)$$

where $V_{\gamma_i,k}$ is the number of birth Gaussian components for target type i where $i \in \{1,2,3\}$, $m_{\gamma_i,k}^{(v)}$ is the current measurement and zero initial velocity used as mean and $P_{\gamma_i,k}^{(v)}$ is the birth covariance for 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

$$\mathcal{D}_{i,k-1}(x) = \sum_{v=1}^{V_{i,k-1}} w_{i,k-1}^{(v)} \mathcal{N}(x; m_{i,k-1}^{(v)}, P_{i,k-1}^{(v)}), \quad (15)$$

where $i \in \{1,2,3\}$ 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 (8) by

$$\mathcal{D}_{i,k|k-1}(x) = \mathcal{D}_{i,S,k|k}(x) + \gamma_{i,k}(x), \quad (16)$$

where

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

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

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

where $p_{i,s,k}$ is the survival rate for target type i and $\gamma_{i,k}(x)$ is given by (14).

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

$$\mathcal{D}_{i,k|k-1}(x) = \sum_{v=1}^{V_{i,k|k-1}} w_{i,k|k-1}^{(v)} \mathcal{N}(x; m_{i,k|k-1}^{(v)}, P_{i,k|k-1}^{(v)}), \quad (17)$$

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

Assuming the probabilities of detection are 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 (9), and is given by

$$\mathcal{D}_{i,k|k}(x) = (1 - p_{ii,D,k}) \mathcal{D}_{i,k|k-1}(x) + \sum_{z \in Z_{i,k}} \mathcal{D}_{i,D,k}(x;z), \quad (18)$$

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. (11). Finally, the implementation scheme for $c_{t_{i,k}}(z)$ is formulated in (10) and is given again as

$$c_{t_{i,k}}(z) = \sum_{j=1,\ldots,3\setminus i} \int p_{ji,D}(y)\mathcal{D}_{j,k|k-1}(y) f_{ji,k}(z|y)dy, \tag{19}$$

where $\mathcal{D}_{j,k|k-1}(y)$ is given in (17), $f_{ji,k}(z|y)$ is given in (13) and $p_{ji,D}(y)$ is assumed constant. Since $w^{(i)}_{j,k|k-1}$ is independent of the integrable variable y, (19) becomes

$$\begin{aligned} c_{t_{i,k}}(z) = \quad & \sum_{j=1,\ldots,3\setminus i} \sum_{v=1}^{V_{j,k|k-1}} p_{ji,D} w^{(v)}_{j,k|k-1} \int \mathcal{N}(y; \\ & m^{(v)}_{j,k|k-1}, P^{(v)}_{j,k|k-1}) \mathcal{N}(z; H_{ji,k}y, R_{ji,k})dy, \end{aligned} \tag{20}$$

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

$$\begin{aligned} & \int \mathcal{N}(y; m_1\zeta, P_1)\mathcal{N}(\zeta; m_2, P_2)d\zeta \\ & = \mathcal{N}(y; m_1 m_2, P_1 + m_1 P_2 m_2^T). \end{aligned} \tag{21}$$

Therefore, (20) becomes,

$$\begin{aligned} c_{t_{i,k}}(z) = \quad & \sum_{j=1,\ldots,3\setminus i} \sum_{v=1}^{V_{j,k|k-1}} p_{ji,D} w^{(v)}_{j,k|k-1} \mathcal{N}(z; \\ & H_{ji,k} m^{(v)}_{j,k|k-1}, R_{ji,k} + H_{ji,k} P^{(v)}_{j,k|k-1} H^T_{ji,k}), \end{aligned} \tag{22}$$

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

The key steps of the tri-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 (Ristic et al., 2012). Maggio et al. (Maggio et al., 2008) 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 5. A further disadvantage of random birth is the increased complexity of processing a large number of incorrect targets. For humans 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. Reemerging 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 GM-PHD filter (Vo and Ma, 2006) but are extended to take into account the three detection processes and the subsequent confusion between detections. In the proposed algorithm, birth and prediction both precede

Algorithm 1: Pseudocode for the tri-GM-PHD filter.

given $\{w^{(v)}_{i,k-1}, m^{(v)}_{i,k-1}, P^{(v)}_{i,k-1}\}_{v=1}^{V_{i,k-1}}$, and the measurement set $Z_{i,k}$ for target type $i \in \{1,2,3\}$
step 1. (prediction for birth targets)
for $i = 1,\ldots,3$ **do** ▷ for all target type i
  $e_i = 0$
  for $u = 1,\ldots,V_{\gamma_i,k}$ **do**
    $e_i := e_i + 1$
    $w^{(e_i)}_{i,k|k-1} = w^{(u)}_{i,\gamma,k}$
    $m^{(e_i)}_{i,k|k-1} = m^{(u)}_{i,\gamma,k}$
    $P^{(e_i)}_{i,k|k-1} = P^{(u)}_{i,\gamma,k}$
  end for
end for
step 2. (prediction for existing targets)
for $i = 1,\ldots,3$ **do** ▷ for all target type i
  for $u = 1,\ldots,V_{i,k-1}$ **do**
    $e_i := e_i + 1$
    $w^{(e_i)}_{i,k|k-1} = p_{i,s,k} w^{(u)}_{i,k-1}$
    $m^{(e_i)}_{i,k|k-1} = F_{i,k-1} m^{(u)}_{i,k-1}$
    $P^{(e_i)}_{i,k|k-1} = Q_{i,k-1} + F_{i,k-1} P^{(u)}_{i,k-1} F^T_{i,k-1}$
  end for
end for
$V_{i,k|k-1} = e_i$
step 3. (Construction of PHD update components)
for $i = 1,\ldots,3$ **do** ▷ for all target type i
  for $u = 1,\ldots,V_{i,k|k-1}$ **do**
    $\eta^{(u)}_{i,k|k-1} = H_{ii,k} m^{(u)}_{i,k|k-1}$
    $S^{(u)}_{i,k} = R_{ii,k} + H_{ii,k} P^{(u)}_{i,k|k-1} H^T_{ii,k}$
    $K^{(u)}_{i,k} = P^{(u)}_{i,k|k-1} H^T_{ii,k} [S^{(u)}_{i,k}]^{-1}$
    $P^{(u)}_{i,k|k} = [I - K^{(u)}_{i,k} H_{ii,k}] P^{(u)}_{i,k|k-1}$
  end for
end for
step 4. (Update)
for $i = 1,\ldots,3$ **do** ▷ for all target type i
  for $u = 1,\ldots,V_{i,k|k-1}$ **do**
    $w^{(u)}_{i,k} = (1 - p_{ii,D,k}) w^{(u)}_{i,k|k-1}$
    $m^{(u)}_{i,k} = m^{(u)}_{i,k|k-1}$
    $P^{(u)}_{i,k} = P^{(u)}_{i,k|k-1}$
  end for
  $l_i := 0$
  for each $z \in Z_{i,k}$ **do**
    $l_i := l_i + 1$
    for $u = 1,\ldots,V_{i,k|k-1}$ **do**
      $w^{(l_i V_{i,k|k-1}+u)}_{i,k} = p_{ii,D,k} w^{(u)}_{i,k|k-1} \mathcal{N}(z; \eta^{(u)}_{i,k|k-1}, S^{(u)}_{i,k})$
      $m^{(l_i V_{i,k|k-1}+u)}_{i,k} = m^{(u)}_{i,k|k-1} + K^{(u)}_{i,k}(z - \eta^{(u)}_{i,k|k-1})$
      $P^{(l_i V_{i,k|k-1}+u)}_{i,k} = P^{(u)}_{i,k|k}$
    end for

for $u = 1, ..., V_{i,k|k-1}$ **do**
$c_{s_{i,k}}(z) = \lambda_{c_i} A c_i(z)$
$c_{t_{i,k}}(z) = \sum_{j=1,2,3\setminus i} \sum_{e=1}^{V_{j,k|k-1}} p_{ji,D}$ $w_{j,k|k-1}^{(e)} \mathcal{N}(z; H_{ji,k} m_{j,k|k-1}^{(e)}, R_{ji,k} + H_{ji,k} P_{j,k|k-1}^{(e)} H_{ji,k}^T)$
$c_{i,k}(z) = c_{s_{i,k}}(z) + c_{t_{i,k}}(z)$
$w_{i,k,N} = \sum_{e=1}^{V_{i,k|k-1}} w_{i,k}^{(l_i V_{i,k|k-1}+e)}$
$w_{i,k}^{(l_i V_{i,k|k-1}+u)} = \frac{w_{i,k}^{(l_i V_{i,k|k-1}+u)}}{c_{i,k}(z)+w_{i,k,N}}$
end for
end for
$V_{i,k} = l_i V_{i,k|k-1} + V_{i,k|k-1}$
end for
output $\{w_{i,k}^{(v)}, m_{i,k}^{(v)}, P_{i,k}^{(v)}\}_{v=1}^{V_{i,k}}$

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 $w_{i,k}^{(v)} < 10^{-5}$ 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.

3 OBJECT DETECTION, TRAINING AND EVALUATION

For the tri-PHD filter, we need parameters for the probabilities of detection, confusion and clutter. We employ the existing, state-of-the-art, Aggregated Channel Features (ACF) pedestrian detection algorithm (Dollar et al., 2014) although any detector can be used. 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 (Appel et al., 2013) is used to learn and classify

Algorithm 2: Pruning and merging for the tri-GM-PHD filter.

given $\{w_{i,k}^{(v)}, m_{i,k}^{(v)}, P_{i,k}^{(v)}\}_{v=1}^{V_{i,k}}$ for target type $i \in \{1,2,3\}$, a pruning weight threshold T, and a merging distance threshold U.
for $i = 1, ..., 3$ **do** ▷ for all target type i
Set $\ell_i = 0$, and $I_i = \{v = 1, ..., V_{i,k} | w_{i,k}^{(v)} > T\}$
repeat
$\ell_i := \ell_i + 1$
$u := arg\ max_{v \in I_i} w_{i,k}^{(v)}$
$L_i := \left\{ v \in I_i \middle| (m_{i,k}^{(v)} - m_{i,k}^{(u)})^T (P_{i,k}^{(v)})^{-1} (m_{i,k}^{(v)} - m_{i,k}^{(u)}) \leq U \right\}$
$\tilde{w}_{i,k}^{(\ell_i)} = \sum_{v \in L_i} w_{i,k}^{(v)}$
$\tilde{m}_{i,k}^{(\ell_i)} = \frac{1}{\tilde{w}_{i,k}^{(\ell_i)}} \sum_{v \in L_i} w_{i,k}^{(v)} x_{i,k}^{(v)}$
$\tilde{P}_{i,k}^{(\ell_i)} = \frac{1}{\tilde{w}_{i,k}^{(\ell_i)}} \sum_{v \in L_i} w_{i,k}^{(v)} (P_{i,k}^{(v)} + (\tilde{m}_{i,k}^{(\ell_i)} - m_{i,k}^{(v)})(\tilde{m}_{i,k}^{(\ell_i)} - m_{i,k}^{(v)})^T)$
$I_i := I_i \setminus L_i$
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.

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 random 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 filtering process in comparison with repeated detection and three separate GM-PHD filters in Section 5.

The RFS methodology assumes point detections and a Gaussian error distribution on location. However, humans in a video sequence are extended targets and the ACF detector employs a bounding box. Therefore, overlapping detections are merged using a greedy non-maximum suppression (NMS) overlap threshold (intersection over union of two detections)

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

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 box) 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 Figure 1.

For the tri-GM-PHD strategy, we set the thresholds on detection from the ROC curves in Figure 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. (Vo and Ma, 2006; Ristic et al., 2012), 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 Figure 1, we standardise a clutter rate of 10 false positive per image (fppi), which gives probabilities of detection of 0.93, 0.99 and 0.99 for red, white and referee respectively. With these values, the corresponding confusion parameters are 0.24 (white footballer detected as red), 0.5 (referee as red), 0.24 (red as white), 0.18 (referee as white), 0.19 (red as referee) and 0.17 (red as referee).

4 DATA ASSOCIATION

The tri-GM-PHD filter 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 5. 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 (Bourgeois and Lassalle, 1971).

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 mis-detected and incorrectly labeled, labels are uniquely re-assigned by re-identifying them using the approach in (Ahmed et al., 2015).

5 EXPERIMENTAL RESULTS

Referring to (1), our state vector includes the centroid positions, velocities, 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 $(p_{cx,xk}, p_{cy,xk})$ 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 (12) with matrices taking into account the box width and height at the given scale.

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

$$Q_{i,k-1} = \sigma_{v_i}^2 \begin{bmatrix} \frac{\Delta^4}{4} I_2 & \frac{\Delta^3}{2} I_2 & 0_2 \\ \frac{\Delta^3}{2} I_2 & \Delta^2 I_2 & 0_2 \\ 0_2 & 0_2 & \Delta^2 I_2 \end{bmatrix}, \quad (23)$$

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. $\sigma_{v_i} = 5$ 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 team), target type 2 (white team) and target type 3 (referee).

Similarly, the measurement follows the observation models of (13) with matrices taking into account the box width and height,

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

$$R_{ij,k} = \sigma_{r_i j}^2 \begin{bmatrix} I_2 & 0_2 \\ 0_2 & I_2 \end{bmatrix}, \quad (24)$$

where $\sigma_{r_i j}$ 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 (25). The current measurement and

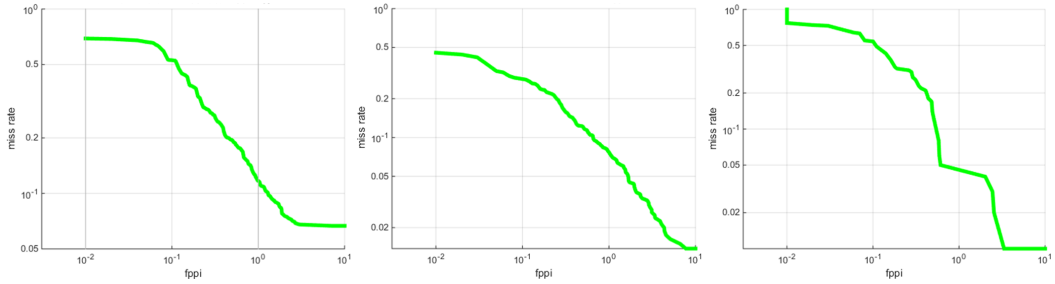

Figure 1: Extracting detection probabilities for three target types (red, white and referee) from ROCs of 3 detectors: red team detector (left), white team detector (middle) and referee detector (right) when tested on red team instances, white team instances and referee instances, respectively.

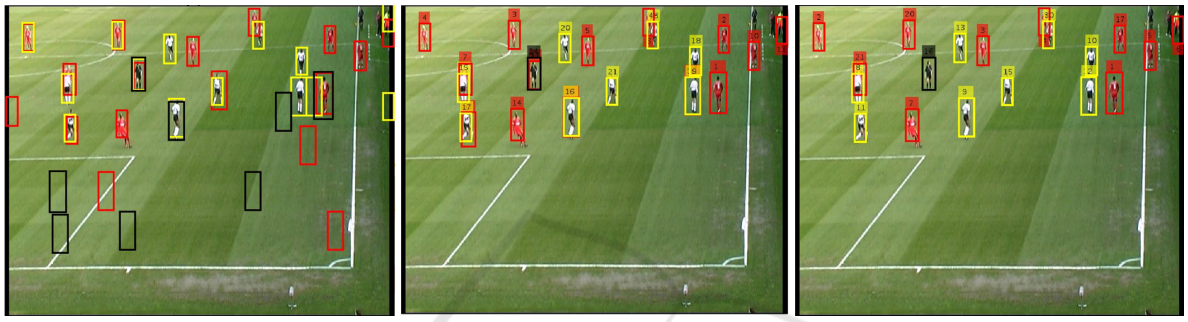

Figure 2: Results of detections (left), three independent GM-PHD trackers (middle) and tri-GM-PHD tracker (right), for frame 25.

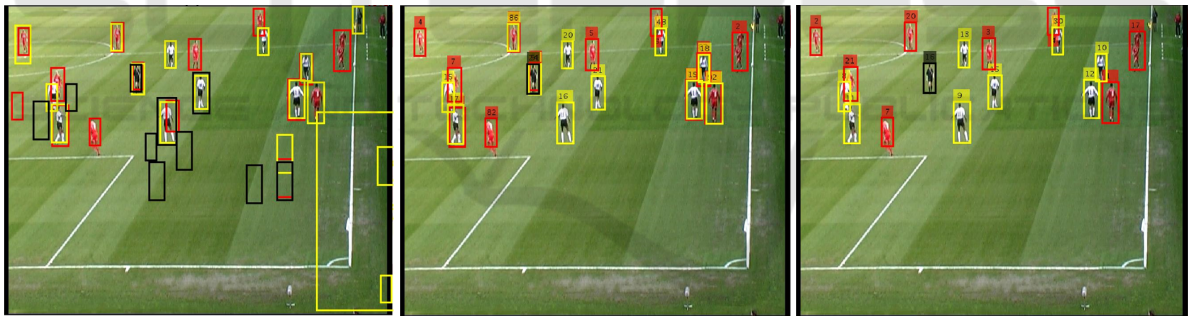

Figure 3: Results of detections (left), three independent GM-PHD trackers (middle) and tri-GM-PHD tracker (right), for frame 57.

zero initial velocity are used as a mean of the Gaussian distribution using a predetermined 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. 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.

$$P_{1,\gamma,k} = P_{2,\gamma,k} = P_{3,\gamma,k} = diag([100,100,25,25,20,20]). \quad (25)$$

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 Figures 2, 3 and 4 are for repeated detection (left), three independent GM-PHD trackers (middle), and the tri-GM-PHD tracker (right) for frames 25, 57 and 73, respectively. Hence, Figure 3 (left) designates detections in which the red, white footballers and the referee are detected both correctly and incorrectly, i.e. one object may be detected by many detectors. For 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 that arise from our choice to set the detection probability high at the expense of higher clutter. Using the three independent

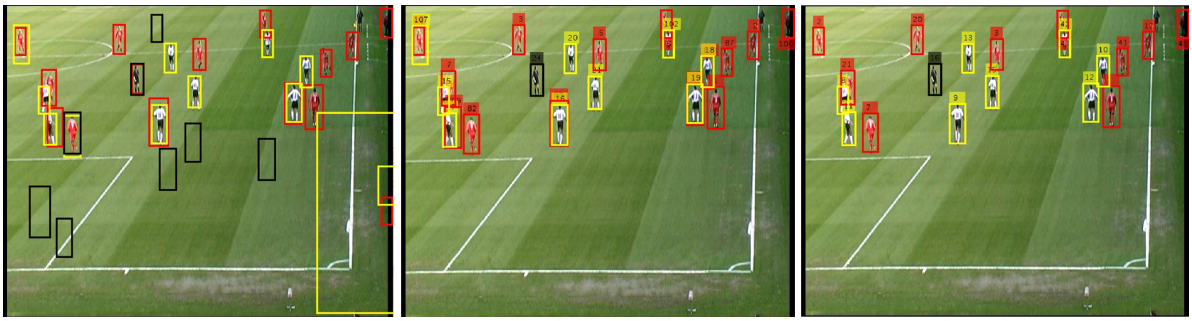

Figure 4: Results of detections (left), three independent GM-PHD trackers (middle) and tri-GM-PHD tracker (right), for frame 73.

Table 1: Frame-averaged 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.

Method	Cardinality error	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%

GM-PHD trackers to effectively eliminate false positives, confused detections are not resolved as shown in Figure 3 (middle). However, our proposed tri-GM-PHD tracker effectively eliminates the false positives and confused detections as shown in Figure 3 (right).

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 (Schumacher et al., 2008), discrimination rate and time taken. We use OSPA metric which is designed for evaluating RFS-based filters rather than multi-object tracking accuracy (MOTA) (Bernardin and Stiefelhagen, 2008) which is widely used for evaluating other traditional multi-target tracking algorithms (Yoon et al., 2016; Choi, 2015). Furthermore, our algorithm is developed not only for tracking but also for discriminating different target types overcoming their confusions unlike algorithms such as (Yoon et al., 2016; Choi, 2015). Therefore, OSPA is the right evaluation metric to evaluate our approach. The computational costs arise from experiments on a i5 2.50 GHz core processor with 6 GB RAM using Matlab and we acknowledge that these are not definitive and give a rough guide only. Though labeling of the targets using Munkres's variant of the Hungarian assignment algorithm works well as shown in Figures 2 (right), 3 (right) and 4 (right), we didn't 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 Figure 5 (left) and Figure 5 (right) 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 1 the average absolute cardinality error using detection only is 10.22, reduced to 5.76 using the standard GM-PHD filter 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 1. Overall, this demonstrates that 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.

Figure 6 shows another example in which the individual footballers are detected, filtered, tracked and labeled for 100 frames. The image has been cropped and and immediately follows a throw-in as the players move away left from the touchline. The figures show the individual tracks, the labels of the footballers and the 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 mis-

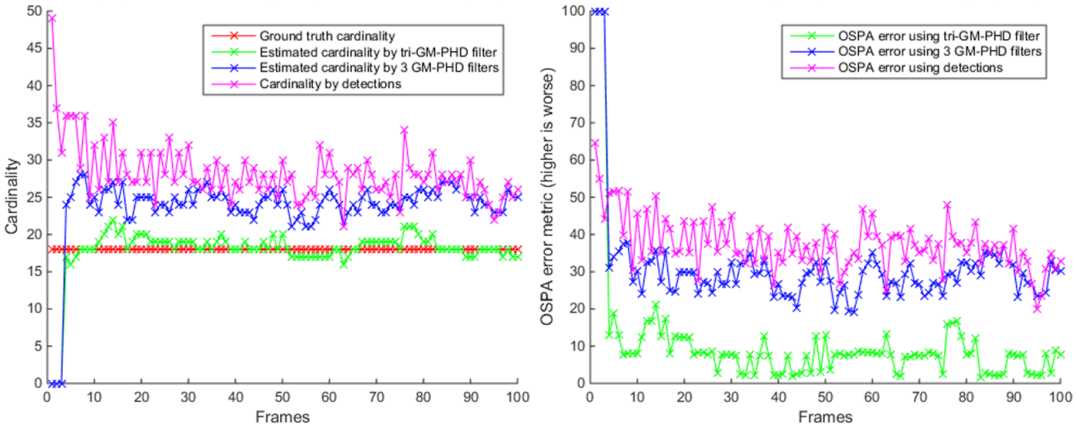

Figure 5: Cardinality error (left) and OSPA error (right): Ground truth (red for cardinality only), tri-GM-PHD filter (green), three independent GM-PHD filters (blue), detections (magenta).

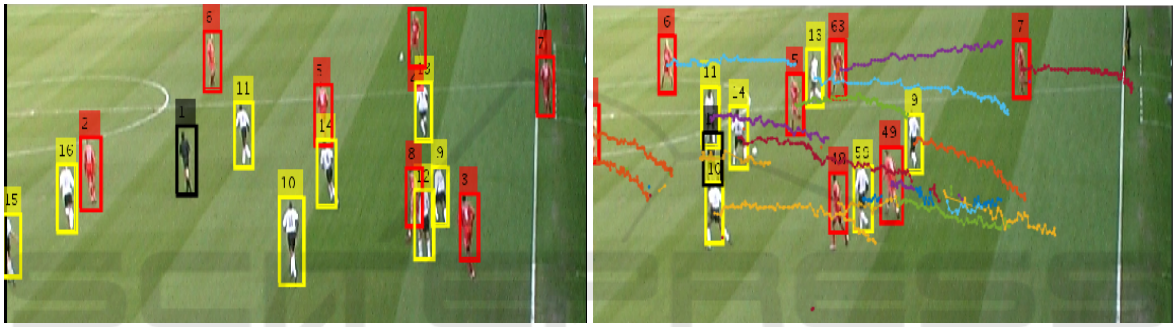

Figure 6: Tracking the red and white teams, and referee from frame 193 (left) to frame 293 (right).

takes, for example red player 3 who starts near the touchline is finally labeled as red player number 49 in frame 293. This 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%.

6 CONCLUSIONS

We have developed an extension of the PHD filter in the RFS framework to account for three different types of multiple targets with separate observations in the same scene, allowing for different probabilities of detection, scene clutter and confusion between targets of different types at the detection stage. This extends the standard GM-PHD filter (Vo and Ma, 2006) to a tri-GM-PHD filter. This has been evaluated using video sequences with the separate targets defined as different team players and the referee. We 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 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. In comparison with separate PHD filters, as is usual practice, we can reduce the mean absolute error in cardinality to less than 1 target, with a corresponding reduction in the OSPA location metric to a mean of 10.59 from 30.86 pixels. Application of the Hungarian labeling method shows good data association so that we are able to track individual targets over the sequence with a mean label switch error of only 0.43. The work we have done has shown that the tri-GM-PHD filter has potential both to track targets in video data, and to better address multiple target confusions than the standard method.

ACKNOWLEDGEMENT

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.

REFERENCES

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

Appel, R., Fuchs, T., Dollar, P., and Perona, P. (2013). Quickly boosting decision trees – pruning underachieving features early. In *ICML*, volume 28, pages 594–602.

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The CLEAR MOT metrics. *J. Image Video Process.*, pages 1:1–1:10.

Bourgeois, F. and Lassalle, J.-C. (1971). An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Commun. ACM*, 14(12):802–804.

Cai, Y., de Freitas, N., and JJ, L. (2006). Robust visual tracking for multiple targets. In *IN ECCV*, pages 107–118.

Cham, T.-J. and Rehg, J. M. (1999). A multiple hypothesis approach to figure tracking. In *CVPR*, pages 2239–2245. IEEE Computer Society.

Choi, W. (2015). Near-online multi-target tracking with aggregated local flow descriptor. In *The IEEE International Conference on Computer Vision (ICCV)*.

Dollar, P., Appel, R., Perona, P., and Belongie, S. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99:14.

Liu, J. and Carr, P. (2014). Detecting and tracking sports players with random forests and context-conditioned motion models. In *Computer Vision in Sports*, pages 113–132. Springer.

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

Mahler, R. P. (2003). Multitarget bayes filtering via first-order multitarget moments. *IEEE Trans. on Aerospace and Electronic Systems*, 39(4):1152–1178.

Matzka, P., Wallace, A., and Petillot, Y. (2012). Efficient resource allocation for automotive attentive vision systems. *IEEE Trans. on Intelligent Transportation Systems*, 13(2):859–872.

Pasha, S., Vo, B.-N., Tuan, H. D., and Ma, W.-K. (2009). A gaussian mixture PHD filter for jump markov system models. *Aerospace and Electronic Systems, IEEE Transactions on*, 45(3):919–936.

Rasmussen, C. and Hager, G. D. (2001). Probabilistic data association methods for tracking complex visual objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:560–576.

Ristic, B., Clark, D., and Vo, B.-N. (2010). Improved SMC implementation of the PHD filter. In *Information Fusion (FUSION), 2010 13th Conference on*, pages 1–8.

Ristic, B., Clark, D. E., Vo, B.-N., and Vo, B.-T. (2012). Adaptive target birth intensity for PHD and CPHD filters. *IEEE Transactions on Aerospace and Electronic Systems*, 48(2):1656–1668.

Schumacher, D., Vo, B.-T., and Vo, B.-N. (2008). A consistent metric for performance evaluation of multi-object filters. *Signal Processing, IEEE Transactions on*, 56(8):3447–3457.

Vo, B.-N. and Ma, W.-K. (2006). The Gaussian mixture probability hypothesis density filter. *Signal Processing, IEEE Transactions on*, 54(11):4091–4104.

Vo, B.-N., Singh, S., and Doucet, A. (2005). Sequential monte carlo methods for multitarget filtering with random finite sets. *IEEE Transactions on Aerospace and Electronic Systems*, 41(4):1224–1245.

Wei, Y., Yaowen, F., Jianqian, L., and Xiang, L. (2012). Joint detection, tracking, and classification of multiple targets in clutter using the PHD filter. *Aerospace and Electronic Systems, IEEE Transactions on*, 48(4):3594–3609.

Yang, W., Fu, Y., and Li, X. (2014). Joint target tracking and classification via RFS-based multiple model filtering. *Information Fusion*, 18:101–106.

Yoon, J. H., Lee, C.-R., Yang, M.-H., and Yoon, K.-J. (2016). Online multi-object tracking via structural constraint event aggregation. In *CVPR*.

Zhou, X., Li, Y., He, B., and Bai, T. (2014). GM-PHD-based multi-target visual tracking using entropy distribution and game theory. *Industrial Informatics, IEEE Transactions on*, 10(2):1064–1076.