Online Multi-target Visual Tracking using a HISP Filter

Nathanael L. Baisa[*]
School of Computer Science, University of Lincoln, Lincoln LN6 7TS, U.K.

Keywords: Visual Tracking, Multiple Target Filtering, MHT, PHD Filter, HISP Filter, MOT Challenge.

Abstract: We propose a new multi-target visual tracker based on the recently developed Hypothesized and Independent Stochastic Population (HISP) filter. The HISP filter combines advantages of traditional tracking approaches like multiple hypothesis tracking (MHT) and point-process-based approaches like probability hypothesis density (PHD) filter, and has a linear complexity while maintaining track identities. We apply this filter for tracking multiple targets in video sequences acquired under varying environmental conditions and targets density using a tracking-by-detection approach. In addition, we alleviate the problem of two or more targets having identical label taking into account the weight propagated with each confirmed hypothesis. Finally, we carry out extensive experiments on Multiple Object Tracking 2016 (MOT16) benchmark dataset and find out that our tracker significantly outperforms several state-of-the-art trackers in terms of tracking accuracy.

1 INTRODUCTION

Multi-target tracking is an active research field in computer vision with a wide variety of applications such as intelligent surveillance, human-computer (robot) interaction, augmented reality, and driver assistance systems. It essentially associates the detections corresponding to the same object over time i.e. it assigns consistent labels to the tracked targets in each video frame to generate a trajectory for each target. These can be performed using online (Sanchez-Matilla et al., 2016)(Song and Jeon, 2016) or offline (Leal-Taix et al., 2016)(Milan et al., 2014)(Pirsiavash et al., 2011) approaches. Online methods estimate the target state at each time instant and depends on predictive models in case of miss-detections to carry on tracking, however, both past and future observations are used in offline (batch) methods to overcome miss-detections. Although offline trackers can generally outperform the online trackers, they are limited for real-time applications.

Traditionally, online multi-target trackers have been developed by finding associations between targets and observations using 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. Recently, 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) which not only addresses the problem of increasing complexity, but also estimates the states and cardinality of an unknown and time varying number of targets in the scene by allowing for target birth, death, clutter (false alarms), and missing detections. It propagates the first-order moment of the multi-target posterior, called the Probability Hypothesis Density (PHD) (Vo and Ma, 2006), rather than the full multi-target posterior. This approach is flexible, for instance, it has been used to find the detection proposal with the maximum weight as the target position estimate for tracking a target of interest in dense environments by removing the other detection proposals as clutter (Baisa et al., 2017). Furthermore, the standard PHD filter was extended to develop a novel N-type PHD filter ($N \geq 2$) for tracking multiple target of different types in the same scene (Baisa and Wallace, 2017)(Baisa and Wallace, 2017). However, this approach does not include target identity in the framework because of the indistinguishability assumption of the point process; additional mechanism is necessary for labelling

[*] This work was done while the author was at the Department of Electrical, Electronic and Computer Engineering, Heriot Watt University, Edinburgh EH14 4AS, United Kingdom.

Baisa, N.
Online Multi-target Visual Tracking using a HISP Filter.
DOI: 10.5220/0006564504290438
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP*, pages 429-438
ISBN: 978-989-758-290-5

each target either at the prediction stage (Sanchez-Matilla et al., 2016) or by post-processing the filter outputs (Baisa and Wallace, 2017).

More recently, a new filter based on stochastic populations has been developed with the concept of partially-distinguishable populations and is termed as Distinguishable and Independent Stochastic Populations (DISP) filter (Delande et al., 2016). This filter can handle an unknown and time varying number of targets in the scene with targets birth, death, miss-detections and false alarms, however, it has a high computational complexity. A low-complexity filter called Hypothesized and Independent Stochastic Population (HISP) filter (Houssineau and Clark, 2016) has been derived from the DISP filter under some intuitive approximations and was adapted for space situational awareness in (Delande et al., 2017). This HISP filter has a linear complexity with both the number of hypotheses and the number of observations similar to the PHD filter, however, unlike the PHD filter, it can preserve the distinct tracks for detected targets.

In this work, we propose an online multi-target visual tracker using tracking-by-detection approach for real-time applications. Accordingly, we make the following three contributions. First, we apply the HISP filter for tracking multiple targets in video sequences acquired under varying environmental conditions and targets density. Second, we alleviate the problem of two or more targets having identical label taking into account the weight propagated with each confirmed hypothesis. Finally, we make extensive experiments on Multiple Object Tracking 2016 (MOT16) benchmark dataset using the public detections provided in the benchmark's test set.

The paper is organized as follows. In section 2, the HISP filter in video tracking context is described in detail. In section 3, the applications and determination of some important variable values are given. The experimental results are analyzed and compared in section 4. The main conclusions and suggestions for future work are summarized in section 5.

2 THE HISP FILTER

The HISP filter is a principled approximation of the DISP filter for practical applications especially for filtering in scenarios involving a large number of targets with moderately ambiguous data association. It combines the advantages of engineering solutions like MHT and point-process-based approaches like PHD filter. It propagates track identities through time similar to MHT, however, it overcomes the drawbacks of MHT such as its strong reliance on heuristics for the appearance and disappearance of targets and a lack a adaptivity by modelling all sources of uncertainties in a unified probabilistic framework. Moreover, it has a linear complexity in the number of hypotheses and in the number of observations, however, the MHT filter has an exponential complexity with time and cubic with the number of targets.

Let the time be indexed by the set $\mathbb{T} \doteq \mathbb{N}$. For any $t \in \mathbb{T}$, the target state space of interest and the observation space of interest are given by $\mathbf{X}_t^\bullet \subseteq \mathbb{R}^d$ and $\mathbf{Z}_t^\bullet \subseteq \mathbb{R}^{d'}$, respectively. They are augmented with the empty state ψ which describes the state of targets outside of the scene of interest and the empty observation ϕ which describes missed detections, respectively, to form the *(full) target state space* $\mathbf{X}_t = \mathbf{X}_t^\bullet \bigcup \{\psi\}$ and the *(full) observation space* $\mathbf{Z}_t = \mathbf{Z}_t^\bullet \bigcup \{\phi\}$. The set of collected observations is represented by $\bar{Z}_t = Z_t \bigcup \{\phi\}$; Z_t for detected observations.

At any time $t \in \mathbb{T}$, the HISP filter is basically based on the following modelling assumptions: 1) a target produces at most one observation (if not, a miss detection occurs), 2) an observation originates from at most one target (if not, a false alarm occurs), 3) targets evolve independently of each other, and 4) observations resulting from target detections are produced independently from each other.

For tracking applications, targets are distinguished by considering their observation histories. Let the space $\mathbb{O}_t$ be

$$\bar{\mathbb{O}}_t = \bar{Z}_0 \times ... \times \bar{Z}_t, \tag{1}$$

so that $\mathbf{o}_t \in \mathbb{O}_t$ takes the form $\mathbf{o}_t = (\phi, ..., \phi, z_{t_+}, ..., z_{t_-}, \phi, ..., \phi)$ with t_+ and t_- the time of appearance and disappearance of the considered track in the scene of interest, and with $z_t \in \bar{Z}_t$ for any $t_+ \leq t \leq t_-$. The observation history $\mathbf{o}_t$ can also be referred to as the observation path and the empty observation path $(\phi, ..., \phi) \in \mathbb{O}_t$ is denoted by ϕ_t.

Each target is identified by some index $\boldsymbol{i}$ in a set $\mathbb{I}$. A track $\boldsymbol{i}$ associated to an observation path with at least one detection (i.e. $\mathbf{o}_t^{\boldsymbol{i}} \neq \phi_t$) cannot have a multiplicity $n^{\boldsymbol{i}}$ greater than one since it cannot represent more than one target, hence, the *previously-detected* target represented by the track $\boldsymbol{i}$ is then *distinguishable*. However, a track $\boldsymbol{i}$ associated to the empty observation path $\mathbf{o}_t^{\boldsymbol{i}} = \phi_t$ represents a sub-population of *yet-to-be-detected* (undetected) targets that are *indistinguishable* from one another, and may have a multiplicity $n^{\boldsymbol{i}}$ greater than one. The tracks cover all the possible combinations of non-empty observation paths representing the previously-detected targets, and one (or possibly several) track(s) representing sub-population(s) of yet-to-be-detected targets.

Each subset of pairwise compatible tracks $H \subseteq \mathbb{I}_t \setminus \{u\}$ which represents the previously-detected targets is called an hypothesis, and the set of all the hypotheses is represented by $\mathbf{H}_t$ whereas the undetected track u, with multiplicity $n^u \in \mathbb{N}$, denotes a subpopulations of n^u yet-to-be-detected targets. Each element in the set $\mathbb{I}_t^u$ is denoted by $\boldsymbol{i}_t^u$. In the HISP filter, hypotheses are assumed to be independent of each other.

Accordingly, a target is indexed by a pair (t,**o**), i.e. $\boldsymbol{i}$ = (t,**o**), where t is the last epoch where the target was known to be in the scene, and the observation path **o** stores its detections across time. Thus, at any time $t \in \mathbb{T}$, the representation of targets after the prediction and after the update steps can be indexed by the sets $\mathbb{I}_{t|t-1} = \{(t,\mathbf{o})|\mathbf{o} \in \bar{\mathbb{O}}_{t-1}\}$ and $\mathbb{I}_t = \{(t,\mathbf{o})|\mathbf{o} \in \bar{\mathbb{O}}_t\}$, respectively.

Using the aforementioned notations and concept, the HISP filter can be expressed via a set of hypotheses. For instance, after the observation (data) update step at time t (see section 2.2), it can be expressed by set of triples of the form $\mathcal{P}_t = \{p_t^{\boldsymbol{i}}, w_t^{\boldsymbol{i}}, n_t^{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{I}_t}$, where $p_t^{\boldsymbol{i}}$ is the probability density corresponding to the index $\boldsymbol{i} \in \mathbb{I}_t$, $w_t^{\boldsymbol{i}} \in [0,1]$ is the weight (or probability of existence) of the hypothesis, and $n_t^{\boldsymbol{i}}$ is the multiplicity of the hypothesis. Each hypothesis maintained by the HISP filter corresponds to a track (a confirmed hypothesis, see section 2.4) and is described by its own probability of existence.

The important steps of the HISP filter are briefly described as follows.

2.1 Time Prediction

The motion of a target from time $t-1$ to time t is modelled by a Markov transition q_t^π verifying for any $x' \in \mathbf{X}_{t-1}^\bullet$

$$q_t^\pi(\psi,\psi) = 1 \quad \text{and} \quad q_t^\pi(x',\psi) = 0, \tag{2}$$

The transition q_t^π models propagation in the scene only excluding target appearance and disappearance of the scene. The probability that a target at point x at time $t-1$ does not disappear is given by the function $p_t^\pi(x) = \int q_t^\pi(x,x')dx'$. The disappearance of a target between time $t-1$ and time t is modelled separately by a transition q_t^ω verifying for any $x' \in \mathbf{X}_{t-1}^\bullet$

$$\int_{\mathbf{X}_t^\bullet} q_t^\omega(x',x)dx = 0 \quad \text{and} \quad \int q_t^\omega(\psi,x)dx = 0, \tag{3}$$

It is assumed that the transition q_t^π and q_t^ω are complementary in the sense that $q_t^\omega(x,\psi) + p_t^\pi(x) = 1$, i.e. either the target disappear or it does not. Hence, the probability of survival of target with state x is given by the scalar $p_t^\pi(x) = 1 - q_t^\omega(x,\psi)$. Besides, there are n_t^α targets potentially appearing at time t, modelled by a probability density q_t^α on $\mathbf{X}_t$ and by a scalar w_t^α.

In the estimation framework for stochastic population, the appearing targets and the yet-to-be detected (undetected) targets are mixed in a single subpopulation. Using "u" in place of the indices $\boldsymbol{i}_{t-1}^u$ and $\boldsymbol{i}_{t|t-1}^u$ when there is no possible ambiguity, the newborn and the undetected targets are represented together after time prediction by

$$p_{t|t-1}^u(x) = \frac{n_{t-1}^u \int q_t^\pi(x',x)p_{t-1}^u(x')dx' + n_t^\alpha p_t^\alpha(x)}{n_{t-1}^u + n_t^\alpha}, \tag{4a}$$

$$(w_{t|t-1}^u, n_{t|t-1}^u) = \left(\frac{n_{t-1}^u w_{t-1}^u + n_t^\alpha w_t^\alpha}{n_{t-1}^u + n_t^\alpha}, n_{t-1}^u + n_t^\alpha\right), \tag{4b}$$

The targets that have already been observed at least once in the past and which have prior indices in $\mathbb{I}_{t-1}$ of the form $\kappa = (t-1,\mathbf{o})$, with $\mathbf{o} \neq \phi_{t-1}$, can either be propagated (kernel q_t^π) or disappear (kernel q_t^ω), and they are characterized after time prediction by

$$p_{t|t-1}^{\boldsymbol{i}}(x) = \int q_t^\iota(x',x)p_{t-1}^\kappa(x')dx', \tag{5a}$$

$$(w_{t|t-1}^{\boldsymbol{i}}, n_{t|t-1}^{\boldsymbol{i}}) = (w_{t-1}^{\boldsymbol{i}}, 1), \tag{5b}$$

with $\iota \in \{\pi,\omega\}$ and with $\boldsymbol{i}$ equals to $(t,\mathbf{o})$ if $\iota = \pi$ (the target is still in the scene at epoch t) and $(t-1,\mathbf{o})$ otherwise (the target has left the scene since last epoch $t-1$). The hypotheses corresponding to disappeared targets are not indexed in the set $\mathbb{I}_{t|t-1}$ since they are not considered for the following observation update. Though they are ignored for the purpose of filtering, they need to be stored as they will be useful for track extraction (see section 2.4).

The approximated multi-target configuration $\mathcal{P}_{t|t-1}$ after prediction from time $t-1$ to time t is then given by $\mathcal{P}_{t|t-1} = \{p_{t|t-1}^{\boldsymbol{i}}, w_{t|t-1}^{\boldsymbol{i}}, n_{t|t-1}^{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathbb{I}_{t|t-1}}$. The time prediction step applies independently to each hypothesis as seen in the prediction equations (4) and (5) due to the modelling assumption on the independence of the targets making it have a linear complexity with respect to the number of hypotheses.

2.2 Observation Update

The observation process at time t is modelled by a potential ℓ_t^z on $\mathbf{X}_t$ defined for any $z \in \bar{Z}_t$ and verifying $\ell_t^\phi(\psi) = 1$ as no observation can be generated from targets that are not present in the scene. For any $x \in \mathbf{X}_t^\bullet$, the potential ℓ_t^z can be given by

$$\ell_t^z = p_{d,t}(x) l_t^z(x), \;\; z \in Z_t \quad \text{and} \quad \ell_t^{\phi}(x) = 1 - p_{d,t}(x), \tag{6}$$

where $p_{d,t}$ is the probability of detection and the dimensionless potential l_t^z is the likelihood of association with measurement z, and is given in the one-dimensional, linear Gaussian case as

$$l_t^z(x) = \exp\left(- \frac{(Hx - z)^2}{2\sigma^2} \right), \tag{7}$$

where H is the observation matrix and σ^2 is the variance of the observation noise.

To maintain a low computational cost for the HISP filter, all the terms in the observation update can be computed with a linear complexity by making an assumption on the term

$$\breve{w}_t^{\kappa,z} = w_{t|t-1}^{\kappa} \int \ell_t^z(x) p_{t|t-1}^{\kappa} dx \tag{8}$$

which corresponds to the association of the target with index $\kappa \in \mathbb{I}_{t|t-1}$ with the observation $z \in Z_t$.

For any $\kappa = (t, \mathbf{o}) \in \mathbb{I}_{t|t-1}$ and any $z \in \bar{Z}_t$, define $\boldsymbol{i}$ as the index $(t, \mathbf{o} \times z)$, with $(\mathbf{o} \times z)$ being the concatenation of $\mathbf{o}$ and z, and define $p_t^{\boldsymbol{i}}$ as the probability density function on $\mathbf{X}_t$ characterized by

$$p_t^{\boldsymbol{i}} = \frac{\ell_t^z(x) p_{t|t-1}^{\kappa}(x)}{\int \ell_t^z(x') p_{t|t-1}^{\kappa}(x') dx'} \tag{9}$$

for any $x \in \mathbf{X}_t$ and let the weights be characterized equivalently by

$$w_t^{\boldsymbol{i}} = \frac{w_{ex}^{\kappa,z} \breve{w}_t^{\kappa,z}}{\sum_{z' \in \bar{Z}_t} w_{ex}^{\kappa,z'} \breve{w}_t^{\kappa,z'}} \quad \text{or} \quad w_t^{\boldsymbol{i}} = \frac{w_{ex}^{\kappa,z} \breve{w}_t^{\kappa,z}}{\sum_{\kappa' \in \mathbb{I}_{t|t-1}} w_{ex}^{\kappa',z} \breve{w}_t^{\kappa',z}} \tag{10}$$

where the scalar $w_t^{\kappa,z} = \breve{w}_t^{\kappa,z} + \mathbf{1}_{\phi}(z)(1 - w_{t|t-1}^{\kappa})$ is the probability mass attributed to the association between κ and z including the possibility that the target does not actually exist in the case of detection failure. The probability that a false alarm will be generated for $z \in \bar{Z}_t$ is denoted by v_t^z. The posterior probability for an observation $z \in Z_t$ to be a false alarm is also obtained via (10) when $\kappa = z$, by setting $w_t^{z,z} = \breve{w}_t^{z,z} = v_t^z$, $w_t^{z,\phi} = 1 - v_t^z$, and $w_t^{z,z'} = 0$ if $z \neq z'$. For any $z \in \bar{Z}_t$ and any $\kappa \in \mathbb{I}_{t|t-1}$ or $\kappa = z$, the scalar $w_{ex}^{\kappa,z}$ is the weight corresponding to the association of the observations in $Z_t \setminus \{z\}$ with false alarms, any of the remaining undetected individuals, or any remaining hypotheses in $\mathbb{I}_{t|t-1} \setminus \{\kappa\}$. This scalar can be expressed as

$$w_{ex}^{\kappa,z} = C'_t(\kappa, z) \prod_{\kappa' \in \mathbb{I}_{t|t-1} \setminus \{\kappa\}} \left[w_t^{\kappa',\phi} + \sum_{z' \in Z_t \setminus \{z\}} \frac{w_t^{\kappa',z'}}{C_t(z')} \right] \tag{11}$$

where $C_t(z) = w_t^{u,z} / w_t^{u,\phi} + v_t^z / (1 - v_t^z)$ and where

$$C'_t(\kappa, z) = \; [w_t^{u,\phi}]^{n_{t|t-1}^u - \mathbf{1}_u(\kappa)} \left[\prod\nolimits_{z' \in Z_t \setminus Z'} (1 - v_t^{z'}) \right] \left[\prod\nolimits_{z' \in Z_t \setminus \{z\}} C_t(z') \right] \tag{12}$$

with $Z' = \emptyset$ when $\kappa \in \mathbb{I}_{t|t-1}$ and $Z' = \{z\}$ when κ corresponds to a false alarm ($\kappa = z$). The hypotheses corresponding to false alarms are not indexed in the set $\mathbb{I}_t$ since they are not considered for the next time step. Though they are ignored for the purpose of filtering, they need to be stored as they will be useful for track extraction (see section 2.4).

The approximated multi-target configuration $\mathcal{P}_t$ after the data update at time t is then given by $\mathcal{P}_t = \{p_t^{\boldsymbol{i}}, w_t^{\boldsymbol{i}}, n_t^{\boldsymbol{i}}\}_{\boldsymbol{i} \in \mathbb{I}_t}$ where $n_t^{\boldsymbol{i}} = n_{t|t-1}^u$ if $\boldsymbol{i} = u$ and $n_t^{\boldsymbol{i}} = 1$ otherwise. There are two assumptions that lead to the structure of the posterior weights (10), (11) of the hypotheses. The first one is for any $\kappa, \kappa' \in \mathbb{I}_{t|t-1}$ such that $\kappa \neq \kappa'$ and any $z \in Z_t$, it holds that $\breve{w}_t^{\kappa,z} \breve{w}_t^{\kappa',z} \approx 0$. This implies that the data association is moderately ambiguous. The second assumption is that hypotheses are independent of each other. Particularly, the computation of the weight $w_{ex}^{\kappa,z}$ does not involve combinatorial operations on the subsets of observations and/or hypotheses making the observation update step have a linear complexity with respect to the number of hypotheses and the number of observations.

2.3 Pruning and Merging

Although the HISP filter has a linear complexity in the number of hypotheses and in the number of observations, reducing the computational cost by limiting the number of propagated hypotheses without a reasonable information loss is crucial while ensuring a meaningful track extraction. From the output of the HISP filter at time t with a multi-target configuration $\mathcal{P}_t = \{p_t^{\boldsymbol{i}}, w_t^{\boldsymbol{i}}, n_t^{\boldsymbol{i}}\}_{\boldsymbol{i} \in \mathbb{I}_t}$, the pruning and merging steps are given as follows:

1. A hypothesis $\boldsymbol{i} \in \mathbb{I}_t$ may have a negligible weight $w_t^{\boldsymbol{i}}$. Such hypothesis can be pruned by retaining the subset of hypotheses having a weight greater than a threshold of τ_p.
2. Some hypotheses $I \subseteq \mathbb{I}_t$ may have probability densities $p_t^{\boldsymbol{i}}$, $\boldsymbol{i} \in I$, that are very close to each other. Such probability densities can be merged since they represent very similar information. Thus, the Mahalanobis distance between the given probability distributions with less than a threshold of τ_m is used as a merging metric.
3. Some hypotheses $I \subseteq \mathbb{I}_t$ may have the same observation path over the extraction window T so that

they can be assumed to represent the same potential target. Such hypotheses can be merged into a single hypothesis with $w = \sum_{\boldsymbol{i}\in I} w_t^{\boldsymbol{i}}$ if $w \leq 1$ since hypotheses cannot have a weight strictly greater than 1.

After the pruning and merging steps, the multi-target configuration will be $\tilde{\mathcal{P}}_t = \{\tilde{p}_t^{\boldsymbol{i}}, \tilde{w}_t^{\boldsymbol{i}}, \tilde{n}_t^{\boldsymbol{i}}\}_{\boldsymbol{i}\in\tilde{\mathbb{I}}_t}$, and is used in the next time step.

2.4 Track Extraction

Tracks, the subset of hypotheses that is the likeliest candidate to represent the population of targets in the scene, are extracted as follows from the multi-target configuration propagated by the HISP filter. The track extraction process has no effect on the filtering process and thus the set of hypotheses is not modified; it is merely for output. The simplest and efficient track extraction method is to select the subset of hypotheses with the highest possible weights and whose observation paths agree with the observations collected during some sliding time window T. The posterior probabilities for each observation produced during this time window to be false alarms need to be computed and stored as hypotheses along with hypotheses corresponding to targets that disappeared during the time window. This is important to know all the observations collected in this time window for the purpose of track extraction. Given the temporary set of hypotheses $\hat{\mathbb{I}}_t$ resulting from these modifications, the track extraction can be solved through the following optimization problem

$$\underset{I\subseteq\hat{\mathbb{I}}_t}{\operatorname{argmax}} \prod_{\boldsymbol{i}\in I} \tilde{w}_t^{\boldsymbol{i}} \tag{13}$$

subject to 1) the union of all observation paths over the time window $T \subseteq \mathbb{T} \cap [0,t]$ must contain all the observations over this window, and 2) the observation paths in I must be pairwise compatible i.e. each observation cannot be used more than once. The solution to this problem is the same as the one for

$$\underset{I\subseteq\hat{\mathbb{I}}_t}{\operatorname{argmax}} \sum_{\boldsymbol{i}\in I} \log \tilde{w}_t^{\boldsymbol{i}} \tag{14}$$

with the same constraints since all $\tilde{w}_t^{\boldsymbol{i}}$ are strictly positive. Taking this way helps us to solve it using integer programming, for instance, using the GNU Linear Programming Kit (GLPK). Hypotheses that are not associated to any observations during the time window are considered as non-conflicting and are extracted on an individual basis. This track extraction approach is only one among many possible. It is one of the simplest that uses the structure of the filter instead of selecting hypotheses individually based on their weight, for example.

In video tracking context, specially when targets density is very high, two or more nearby targets can be detected as a single bounding box due to their extended nature. When these targets start to move apart, they might be detected by their own bounding boxes. This situation is similar to spawning of targets from the original target. However, spawning targets are currently not modelled in the HISP filter. Therefore, when tracks are extracted according to the above procedure, there are cases when the spawning targets take the same label as the original target. These cause difficulty to identify them as they share the same label. In this work, we use the weight propagated with each track (confirmed hypothesis) to discriminate them with the assumption that the original target has a maximum weight, after track extraction process. Thus, if two or more tracks with the same label are confirmed at the same time, we give new label(s) to those spawned target(s) except the original target with the assumption that the original target has a maximum weight and needs to retain the original label. This approach solves the problem of having the same label, however, it is rarely prone to identity switches since the spawned target(s) can have weight(s) greater than the original target violating our assumption. Though this approach overall alleviates the problem, using appearance model might give better results. Note that this process is merely for output purpose as it does not affect the filtering process.

3 THE APPLICATIONS AND DETERMINATION OF THE VARIABLE VALUES

The HISP filter can easily be implemented using any Bayesian filtering technique for each hypothesis, for instance, sequential Monte Carlo (SMC) (Houssineau et al., 2015) or Kalman filtering. In this work, we use the Kalman filter implementation of the HISP filter referred to as KF-HISP filter with the assumption of a linear Gaussian model. In this implementation scheme, a probability density, for instance $p_t^{\boldsymbol{i}}$, is characterized by multivariate normal distribution $\mathcal{N}(m_t^{\boldsymbol{i}}, P_t^{\boldsymbol{i}})$ where $m_t^{\boldsymbol{i}}$ is the mean and $P_t^{\boldsymbol{i}}$ is the covariance for $\boldsymbol{i} \in \mathbb{I}_t$.

Our state vector includes the centroid positions, velocities, width and height of the bounding boxes, i.e. $x_t = [p_{cx,xt}, p_{cy,xt}, \dot{p}_{x,xt}, \dot{p}_{y,xt}, w_{xt}, h_{xt}]^T$. Similarly, the measurement is the noisy version of the target area in the image plane approximated with a w x h rectangle centered at $(p_{cx,xt}, p_{cy,xt})$ i.e. $z_t =$

$[p_{cx,zt}, p_{cy,zt}, w_{zt}, h_{zt}]^T$.

A target state evolves from time $t-1$ to time t through the Markov transition kernel q_t^{π} with matrices taking into account the box width and height at the given scale.

$$F_{t-1} = \begin{bmatrix} I_2 & \Delta I_2 & 0_2 \\ 0_2 & I_2 & 0_2 \\ 0_2 & 0_2 & I_2 \end{bmatrix},$$
$$Q_{t-1} = \sigma_v^2 \begin{bmatrix} \frac{\Delta^4}{4} I_2 & \frac{\Delta^3}{2} I_2 & 0_2 \\ \frac{\Delta^3}{2} I_2 & \Delta^2 I_2 & 0_2 \\ 0_2 & 0_2 & \Delta^2 I_2 \end{bmatrix}, \qquad (15)$$

where F and Q denote the state transition matrix and process noise covariance, respectively; I_n and 0_n denote the n x n identity and zero matrices, respectively, and $\Delta = 1$ second is the sampling period defined by the time between frames. $\sigma_v = 5$ pixels$/s^2$ is the standard deviation of the process noise. The disappearance kernel q_t^{ω} is assumed constant and verifies, for any $x \in \mathbf{X}_t^{\bullet}$, $q_t^{\omega}(x, \psi) = 10^{-2}$ (i.e. the probability of survival p_t^{π} of the targets is 0.99). The HISP filter is sensitive to p^{π}: $p^{\pi} = 1$ implies that if an hypothesis is present almost surely then it will be displayed at all following time steps, alternatively, if $p_t^{\pi} \leq p_d$ then hypotheses stop to be considered as tracks as soon as a detection failure happens. Thus, it is preferable to set the value of p^{π} greater than the value of the probability of detection p_d to handle some miss-detections.

Similarly, the measurement follows the observation model (6) with matrices taking into account the box width and height,

$$H_t = \begin{bmatrix} I_2 & 0_2 & 0_2 \\ 0_2 & 0_2 & I_2 \end{bmatrix},$$
$$R_t = \sigma_r^2 \begin{bmatrix} I_2 & 0_2 \\ 0_2 & I_2 \end{bmatrix}, \qquad (16)$$

where H_t and R_t denote the observation matrix and the observation noise covariance, respectively, and $\sigma_r = 6$ pixels is the measurement standard deviation. The probability of detection is assumed to be constant across the state space and through time and is set to a value of $p_d = 0.90$. The false positives are independently and identically distributed (i.i.d), and the number of false positives per frame is Poisson-distributed with mean 10 (false alarm rate of $v_t^z = 4.8 \times 10^{-6}$; dividing the mean 10 by frame resolution).

The average number of appearing targets per frame n_t^{α} is set to 0.1. This number is then divided uniformly across frame resolution to give the probability w_t^{α} that any potential observation represents an appearing target. The distribution p_t^{α} is uninformative since nothing is known about the appearing targets before the first observation. The distribution after the observation is determined by the current measurement and zero initial velocity used as a mean of the Gaussian distribution and using a predetermined initial covariance given in (17) for birthing of targets.

$$P_t^{\alpha} = diag([100, 100, 25, 25, 20, 20]). \qquad (17)$$

To reduce the computational cost, the pruning threshold τ_p is set to 10^{-3} and the merging threshold τ_m is set to 4 pixels, and are used on the collection of individual posterior laws (probability densities). For track extraction, the sliding time window T is set to 5. We set the maximum number of hypotheses to 10^7.

4 EXPERIMENTAL RESULTS

We validate our proposed tracker, HISP-T, and compare it against state-of-the-art online and offline tracking methods (GM-PHD-MA (Song and Jeon, 2016), DP-NMS (Pirsiavash et al., 2011), SMOT (Dicle et al., 2013), CEM (Milan et al., 2014) and JPDA-m (Rezatofighi et al., 2015)) on the MOT16 benchmark datasets (Milan et al., 2016). We use the *public detections* provided by the MOT benchmark. We use the following evaluation measures: Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP) (Kasturi et al., 2009), Mostly Tracked targets (MT), Mostly Lost targets (ML) (Li et al., 2009), Fragmented trajectories (Frag), False Positives (FP), False negatives (FN) and Identity Switches (IDS). For detailed description of each metric, please refer to (Milan et al., 2016).

Quantitative evaluation of our proposed method with other trackers is compared in Table 1. The Table shows that HISP-T outperforms both online and offline trackers listed in the table in terms of MOTA and MT. In terms of MOTP, our tracker outperforms the online tracker(s) and the offline trackers such as CEM and SMOT. The number of ML and FN percentage are overall lower than the other online and offline trackers except one offline tracker (i.e. second to SMOT). The higher number of IDS and Frag compared to the other online tracker and some of the offline trackers is due to the fact that our tracker relies only on the position and size of the bounding box of the detections; we are not using any appearance models to discriminate nearby targets. Spawning targets are also currently not modelled in the HISP filter, therefore, identity switches are more likely to occur in such crowded scenes. Our tracker runs about 4.8 frames per second (fps). The computational costs arise from experiments on a i7 2.30 GHz core processor with 8 GB RAM using Matlab (not well optimized).

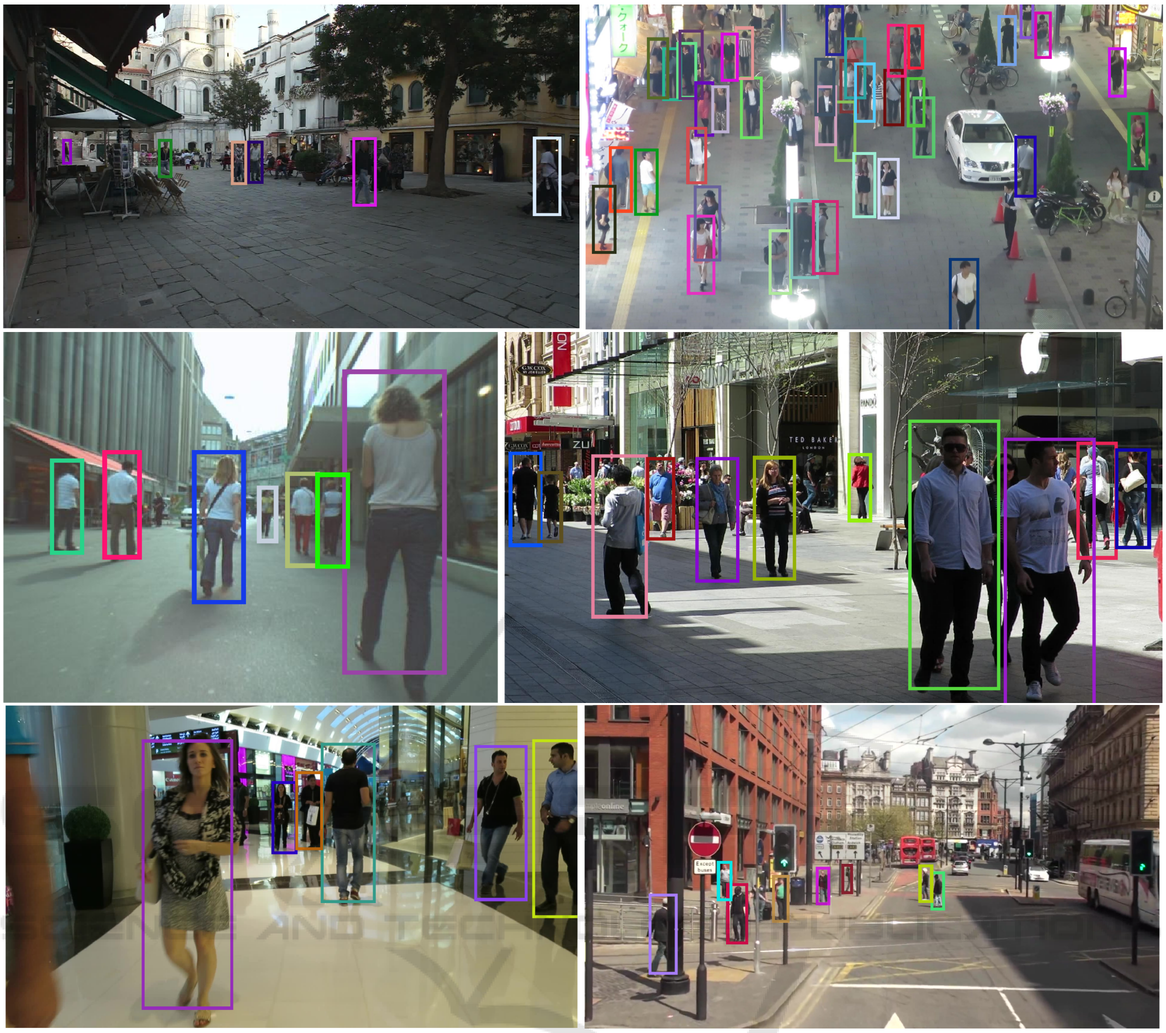

Figure 1: Sample results on several sequences of MOT16 datasets, bounding boxes represents the tracking results with their color-coded identities. From left to right: MOT16-01, MOT16-03 (top row), MOT16-06, MOT16-08 (middle row),and MOT16-12, MOT16-14 (bottom row).

Examples of tracking results of all MOT16 test sequences except MOT16-07 are shown in Figure 1; from left to right: MOT16-01, MOT16-03 (top row), MOT16-06, MOT16-08 (middle row), and MOT16-12, MOT16-14 (bottom row). Three frames from MOT16-07 are shown in Figure 2. In all figures, the bounding boxes represent the tracking results with their color-coded identities. The MOT16-07 shown in Figure 2 contains 54 tracks recorded by a moving camera in a sequence of 500 frames. Tracking in this sequence is a very challenging task, not only because the density of pedestrians is quite high, but also because significant camera motion makes the person trajectories to be both rough and discontinuous. Our tracker reasonably performs even on this sequence though some identity switches occur due to significant camera motion, detection failures and lack of appearance model in our approach.

5 CONCLUSIONS

We have developed a novel multi-target visual tracker based on the recently developed Hypothesized and Independent Stochastic Population (HISP) filter. We apply this filter for tracking multiple targets in video sequences acquired under varying environmental conditions and targets density. We followed a tracking-by-detection approach using the public detections provided in the Multiple Object Tracking 2016 (MOT16) benchmark datasets. We also allevi-

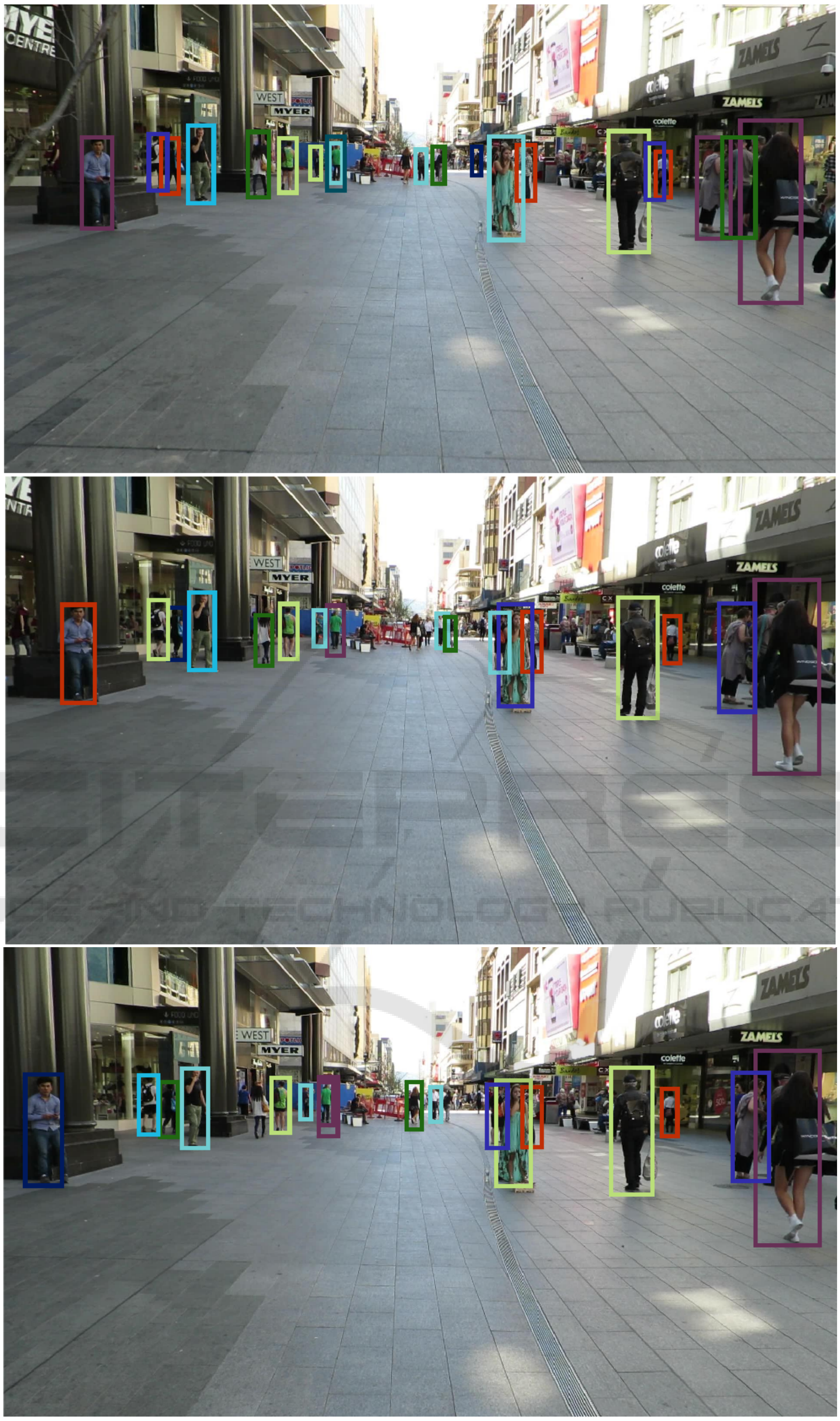

Figure 2: Sample results on the sequence MOT16-07, bounding boxes represents the tracking results with their color-coded identities, for frames 354, 368 and 380 from top to bottom.

ate the problem of identical labels that two or more nearby targets share through the employed track extraction approach by using the weight of the confirmed tracks which is very crucial in the case of video tracking. Results show that our method outperforms state-of-the-art trackers developed using both online and offline approaches on the MOT16 benchmark datasets in terms of tracking accuracy.

Table 1: Tracking performance of representative trackers developed using both online and offline methods. All trackers are evaluated on the test dataset of the MOT16 (Milan et al., 2016) benchmark using public detections. The first and second highest values are highlighted by bold and underline.

Tracker	Tracking Mode	MOTA↑	MOTP↑	MT (%)↑	ML (%)↓	FP↓	FN↓	IDS↓	Frag↓
CEM (Milan et al., 2014)	offline	<u>33.2</u>	75.8	<u>7.8</u>	54.4	6,837	114,322	642	<u>731</u>
DP-NMS (Pirsiavash et al., 2011)	offline	32.2	**76.4**	5.4	62.1	**1,123**	121,579	972	944
SMOT (Dicle et al., 2013)	offline	29.7	75.2	5.3	**47.7**	17,426	**107,552**	3,108	4,483
JPDF-m (Rezatofighi et al., 2015)	offline	26.2	<u>76.3</u>	4.1	67.5	<u>3,689</u>	130,549	**365**	**638**
GM-PHD-MA (Song and Jeon, 2016)	**online**	30.5	75.4	4.6	59.7	5,169	120,970	<u>539</u>	<u>731</u>
HISP-T (ours)	**online**	**35.9**	76.1	**7.8**	<u>50.1</u>	6,406	<u>107,905</u>	2,592	2,299

The tracker works at an average speed of 4.8 fps. In the future work, we will use appearance features, either hand-engineered or deep learning, to alleviate identity switches and trajectory fragmentation.

REFERENCES

Baisa, N. L., Bhowmik, D., and Wallace, A. (2017). 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*.

Baisa, N. L. and Wallace, A. (2017). Multiple Target, Multiple Type Filtering in RFS Framework. *ArXiv e-prints*.

Baisa, N. L. and Wallace, A. (2017). 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*.

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

Delande, E., Houssineau, J., and Clark, D. (2016). Multi-object filtering with stochastic populations. *arXiv*, 1501.04671v2.

Delande, E., Houssineau, J., Franco, J., Frh, C., and Clark, D. (2017). A new multi-target tracking algorithm for a large number of orbiting objects. *27th AAS/AIAA Space Flight Mechanics Meeting*.

Dicle, C., Camps, O. I., and Sznaier, M. (2013). The way they move: Tracking multiple targets with similar appearance. In *2013 IEEE International Conference on Computer Vision*, pages 2304–2311.

Houssineau, J. and Clark, D. (2016). Multi-target filtering with linearised complexity. *arXiv*, 1404.7408v2.

Houssineau, J., Clark, D. E., and Del Moral, P. (2015). A sequential monte carlo approximation of the HISP filter. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1251–1255. IEEE.

Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., and Zhang, J. (2009). Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336.

Leal-Taix, L., Canton-Ferrer, C., and Schindler, K. (2016). Learning by tracking: Siamese CNN for robust target association. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR). DeepVision: Deep Learning for Computer Vision.*

Li, Y., Huang, C., and Nevatia, R. (2009). Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *In CVPR*.

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

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*. arXiv: 1603.00831.

Milan, A., Roth, S., and Schindler, K. (2014). Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72.

Pirsiavash, H., Ramanan, D., and Fowlkes, C. C. (2011). Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR 2011*, pages 1201–1208.

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.

Rezatofighi, S. H., Milan, A., Zhang, Z., Shi, Q., Dick, A., and Reid, I. (2015). Joint probabilistic data association revisited. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3047–3055.

Sanchez-Matilla, R., Poiesi, F., and Cavallaro, A. (2016). Online multi-target tracking with strong and weak detections. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 84–99.

Song, Y. and Jeon, M. (2016). Online multiple object tracking with the hierarchically adopted GM-PHD filter using motion and appearance. In *IEEE/IEIE The International Conference on Consumer Electronics (ICCE) Asia*.

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