IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 20, 2017, accepted March 3, 2018, date of publication March 16, 2018, date of current version April 4, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2816805

Particle PHD Filter Based Multiple Human Tracking Using Online Group-Structured Dictionary Learning

ZEYU FU[1], (Student Member, IEEE), PENGMING FENG[2], (Member, IEEE), FEDERICO ANGELINI[1], (Student Member, IEEE), JONATHON CHAMBERS[1,3], (Fellow, IEEE), AND SYED MOHSEN NAQVI[1], (Senior Member, IEEE)

[1]Intelligent Sensing and Communications Research Group, School of Engineering, Newcastle University, Newcastle upon Tyne Ne1 7Ru, U.K.
[2]State Key Laboratory of Space-Ground Integrated Information Technology, Beijing Institute of Satellite Information Engineering, Beijing 100029, China
[3]Department of Engineering, University of Leicester, Leicester LE1 7RH, U.K.

Corresponding author: Zeyu Fu (z.fu2@newcastle.ac.uk)

This work was supported in part by the Engineering and Physical Sciences Research Council under Grant EP/K014307 and in part by MOD University Defence Research Collaboration in Signal Processing.

ABSTRACT An enhanced sequential Monte Carlo probability hypothesis density (PHD) filter-based multiple human tracking system is presented. The proposed system mainly exploits two concepts: a novel adaptive gating technique and an online group-structured dictionary learning strategy. Conventional PHD filtering methods preset the target birth intensity and the gating threshold for selecting real observations for the PHD update. This often yields inefficiency in false positives and missed detections in a cluttered environment. To address this issue, a measurement-driven mechanism based on a novel adaptive gating method is proposed to adaptively update the gating sizes. This yields an accurate approach to discriminate between survival and residual measurements by reducing the clutter inferences. In addition, online group-structured dictionary learning with a maximum voting method is used to robustly estimate the target birth intensity. It enables the new-born targets to be automatically detected from noisy sensor measurements. To improve the adaptability of our group-structured dictionary to appearance and illumination changes, we employ the simultaneous code word optimization algorithm for the dictionary update stage. Experimental results demonstrate our proposed method achieves the best performance amongst state-of-the-art random finite set-based methods, and the second best online tracker ranked on the leaderboard of latest MOT17 challenge.

INDEX TERMS Multiple human tracking, SMC-PHD filter, adaptive gating, group-structured sparsity, birth intensity estimation, dictionary learning.

I. INTRODUCTION

Video-based multiple human tracking has been an emerging technique in the last decade, since it is crucial in many applications such as intelligent video surveillance, behavior analysis, assistive technology and human-computer interactions [2]–[4]. Many researchers have been seeking higher-level tracking systems to locate a number of targets, retrieve their trajectories, and recognise their identities from some video sequences. However, there still exist many challenging problems caused by complicated environments such as the presence of noise, occlusions, background clutter, and illumination changes.

To address these challenges, traditional approaches have involved explicit association between measurements and targets in multiple human tracking such as multiple hypotheses tracking (MHT) and the joint probabilistic data association filter (JPDAF) [3], [5]. Recently, tracking-by-detection with data association driven by the advancements in object detection has become a commonly-used framework for multi-target tracking in video [6]–[15]. These methods can be categorized into online and off-line tracking modes. Off-line trackers [6], [7], [9], [10], [12], [13] process the video sequences using both past and future detection responses, but such non-causal systems are difficult to be applied in real-time applications. However, online trackers [8], [11], [14]–[18] only rely on the detections given up to the present time, which is more suitable for real-time processing. An effective online multi-target

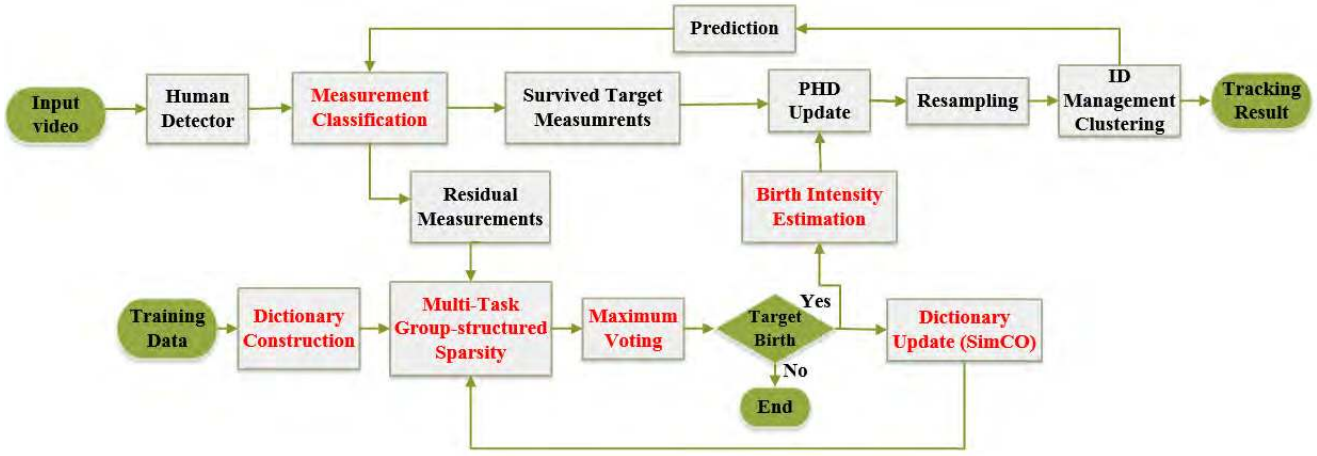


FIGURE 1. Block diagram of the proposed multiple human tracking system: the main contributions are labelled in red colour. Each input video sequence uses a human detector for measurements extraction. The SMC-PHD filtering framework is exploited and adaptive gating based measurement classification is proposed to categorise the raw measurements into survival targets and residual measurements. The residual measurements are further processed via proposed online group-structured dictionary learning which includes dictionary construction based on training data, multi-task group-structured sparsity, maximum voting technique and SimCo based dictionary update steps. For efficient processing, we only perform the birth intensity estimation (birth measurements) and dictionary update, if any new-born target is detected by the maximum voting technique. The both sets of survival targets and new born target(s) measurements are further processed at the PhD update step. The resampling and ID management clustering steps are performed to achieve the final tracking results. The ID management clustering also provides prediction for measurement classification for the following input video, as shown in the diagram.

Bayesian filter has been developed within the signal processing community. Based upon the random finite set (RFS), the probability hypothesis density (PHD) filter [19] is proposed to deal with varying number of targets, reduce missed detections, mitigate spatial noise, and also assist to reduce the computational complexity in data association techniques [5], [17]. The PHD filter is a natural extension of the single-target Bayesian framework to multi-targets; representing the multi-target states and multi-target measurements, as well as recursively propagating the first-order moment of the multi-target posterior [19]. The posterior function has previously been approximated by the Gaussian mixtures method GM-PHD filter [20] or Sequential Monte Carlo method via a set of weighted random particles known as the SMC-PHD filter [21]. The advantages of this technique has enabled it to be extensively applied in video-based multiple human tracking [2], [17], [18], [22]–[26]. Therefore, the focus of this paper is to explore and improve the PHD filter for multiple human tracking in video.

Due to the imperfections in the human detector, two challenging issues in the SMC-PHD filter remain that are accurately discriminating the survival and birth measurements from the original detection results with uncertainty, as well as adaptively determining the birth intensity of the newborn targets. To solve these issues, we present an improved PHD filter based online multiple human tracking system as shown in Fig. 1. The proposed system consists of two main steps: adaptive gating (AG) step and birth intensity estimation driven by online group-structured dictionary learning (Online-GSDL) step to achieve more reliable detected measurements and thereby being able to raise the efficiency and accuracy of implementing the PHD filter update. The AG step aims to adaptively boost the gating threshold fusing spatio-temporal and human information according to the corresponding measurements so as to refine the measurement set of survival targets. This gives the advantages of achieving better measurement classification. In the Online-GSDL step, we employ hierarchical K-means clustering to learn a dictionary with group structure information. The multi-task group-structured sparsity is achieved by exploiting a collaborative hierarchical Lasso (C-HiLasso) model [27] to strengthen the discriminability of the sparse coefficients at the group level. In this way, a maximum voting method based on the sparsity solution is then proposed to eliminate the existing interferences induced by noise or clutter from the measurement set, which leads to increasing the accuracy of birth intensity generalization. Moreover, we use the SimCO algorithm [28] with the proposed structure pattern to carry out the dictionary update stage so as to handle the dynamic variations in appearance and background. Our main contributions are summarized as follows:

1) To enhance the observation model, a novel adaptive gating strategy is developed in the system to aid the classification of measurements.
2) Online group-structured dictionary learning is integrated adaptively in the birth intensity estimation via the proposed maximum voting method. To the best of our knowledge, this is the first work to explore group-structured sparsity at the multi-task level for birth intensity estimation in multiple human tracking.
3) The SimCO algorithm is exploited to update the dictionary, which is devoted to simultaneously updating some active groups of atoms specified by the proposed structure pattern.

4) Experiments demonstrate our proposed method achieves the best performance amongst state-of-the-art RFS based methods (mainly from the signal processing community), and is the second best online tracker ranked on the leaderboard of latest MOT17 Challenge [1].

Part of this work has previously appeared in [29]. In this paper, we present the entire tracking algorithm for the first time, including additional experiments with more benchmark datasets and a detailed study of contributions from individual system component. More importantly, we reformulate the group-structured dictionary learning with maximum voting presented in [29] to improve the birth intensity estimation. Moreover, we propose a newly proposed adaptive gating technique for measurement classification, and also exploit the SimCO algorithm for the dictionary update.

The rest of the paper is organized as follows. We begin by summarizing the related work in the next section. Section III provides a detailed description of the proposed tracking algorithm with background and preliminaries. Experimental results and comparisons between the proposed approach and other state-of-the-art methods are presented in Section IV. Finally, some concluding remarks and future work are discussed in Section V.

II. RELATED WORK

In this section, we briefly discuss some related work that motivates this paper: recent tracking-by-detection approaches, measurement-driven birth intensity estimation and dictionary learning.

Tracking-by-detection has become one of the most successful frameworks to tackle the multiple human tracking problem. The key task of this framework is to associate accurately each detection with an individual target. Online methods only considering information from past and current frames are more suitable for time-critical and real applications [30]. Xiang *et al.* [8] formulated multi-target tracking as decision making with the similarity function learned by reinforcement learning. Yoon *et al.* [11] exploited structural constraints to solve the online multi-target tracking problems with frame-by-frame data association. In [16], relative motion models are built by considering motion context which describes the relative movement between targets, thereby facilitating robust prediction and data association. Most batch methods [6], [7], [9], [10], [13] utilize the whole set of detections to address a global optimization problem. Recently, deep convolutional neural networks (CNNs) have remarkably outperformed heuristic and hand-crafted feature methods in appearance modelling for data association. Leal-Taixe *et al.* [31] explored a so-called Siamese CNN to learn descriptors using image pixel values and optical flow as multi-modal inputs. Son *et al.* [12] proposed a Quadruplet CNN for multi-target tracking, which employs quadruplet losses with appearance and motion cues to associate detections across video frames.

Another popular direction in multi-target tracking is to use recursive Bayesian filtering techniques such as Kalman filtering, particle filtering and RFS based filtering. The RFS based filtering method has the advantage of estimating the time-varying number of targets and their trajectories with a relatively low computational cost. Our proposed method therefore falls in the online multi-target tracking using RFS based filtering framework. However, a particular issue for RFS based filtering trackers is to select valid measurements from detection responses and track a number of new-born targets that are highly dependent on the birth intensity function. Conventional PHD and multi-Bernoulli filtering methods [26], [21], [19] empirically preset the birth intensity to cover the whole state space of interest. This requires prior knowledge of the scene information, and might have limitations in real world scenarios. To avoid the necessity of prior knowledge in estimating birth intensity, Ristic *et al.* [32] built the adaptive target birth model based on the new-born particles with high measurement likelihood. Alternatively, existing gating methods in the data-driven mechanism have been widely explored to select valid measurements and reduce the computation time in the PHD update step by designing a validation region with spatial relation. For instance, [33] and [34] proposed the gating technique for the measurements classification between the birth and survival measurements originated from detection results. Si *et al.* [35] pre-defined a confidence level to perform the gating strategy to adaptively generate the target birth intensity. However, all of the above methods may be no longer applicable to achieve effective measurements without the assumption: the initial distribution of the new-born target is known, at most one new target can enter to the scene at one time instant, and target births and deaths can be generated at arbitrary positions at any time.

On the other hand, the measurement set of new-born targets separated from the original measurements is still likely to contain some clutter or noise, which means that the target birth intensity directly determined by the measurements originating from the newborn targets will result in many false positives. Zhou *et al.* [22] proposed an entropy distribution-based algorithm to automatically estimate birth intensity. Wu *et al.* [24] improved the GM-PHD filter combined with an iterative random sample consensus (I-RANSAC) method to estimate the target birth intensity from uncertain measurements, whereas they approximated the trajectory of a new-born target as a straight line by means of regressing a line model with a given measurement set, which is not always feasible to track targets with nonlinear movements especially in video surveillance. Besides, Feng *et al.* [2] recently adopted a one-class support vector machine (OCSVM) to remove background noise. Different from existing measurement dependent birth intensity estimation algorithms that eliminate the interference by noise, we proposed an improved two-stage measurement dependent birth intensity estimation algorithm. Inspired by the data-driven methods of [33]–[35], our adaptive gating method takes the advantage of the

fusion of spatio-temporal and human size information to dynamically control the validation area for measurements partitioning. The second stage aims to build a robust classifier via online group-structured dictionary learning to separate the target candidates from noise to promote the birth intensity estimation.

Recent literature shows structured sparsity has been proved to provide better efficiency and robustness than simple sparsity in image classification and object tracking applications [36]–[39], the success of which is attributed to exploiting the block structure in the learning process and considering prior information in the predefined structure of the dictionary. Here, we present an online group-structured dictionary learning algorithm along with a voting method to improve the PHD filter when used for multiple human tracking in video. Numerous approaches have been proposed to cope with the target appearance variations over time. In [40], Liu *et al.* proposed a K-selection algorithm with a locally constrained sparse representation for dictionary update to provide stronger discriminative ability. Both methods in [41] and [36] adopted the block coordinate descent to render the dictionary updated in an online manner. In [42], Ma *et al.* established an occlusion-aware dictionary update scheme which is capable of handling appearance changes during tracking especially caused by occlusion. To avoid false samples accumulating in the dictionary, we exploit the SimCO algorithm [28] following the proposed structure pattern to selectively update the dictionary, in order to simultaneously achieve good efficiency of implementation and improve the robustness of the dictionary.

III. THE PROPOSED TRACKING SYSTEM

A. THE SMC-PHD FILTER

Based upon the framework of random finite set (RFS), a multiple target state and a multiple target measurement at time k can be represented by two finite sets: $\mathbf{X}_k = \{\mathbf{x}_k^1, \ldots, \mathbf{x}_k^{M_k}\} \in \mathcal{F}(\mathcal{X})$ and $\mathbf{Z}_k = \{\mathbf{z}_k^1, \ldots, \mathbf{z}_k^{N_k}\} \in \mathcal{F}(\mathcal{Z})$, where M_k and N_k denote the number of targets and the number of measurements at time k respectively. Here $\mathcal{F}(\mathcal{X})$ and $\mathcal{F}(\mathcal{Z})$ are the finite subsets of $\mathcal{X}$ and $\mathcal{Z}$ respectively [32]. For all $m = 1, \ldots, M_k$, the state of a target m is $\mathbf{x}_k^m = [p_{x,k}^m, p_{y,k}^m, v_{x,k}^m, v_{y,k}^m, w_k^m, h_k^m]^T$ and contains the actual 2D image location, velocity and the size of the target. The observed measurement vector $\mathbf{z}_k^n = [\bar{p}_{x,k}^n, \bar{p}_{y,k}^n, \bar{w}_k^n, \bar{h}_k^n]^T$, where $n = 1, \ldots, N_k$, typically contains the n-th target location and size information. The PHD filter proposed by Mahler [19] aims to recursively propagate the first-order moment of the multi-target posterior $p_{k|k}(\mathbf{X}_k|\mathbf{Z}_{1:k})$, referred to as the intensity function $v_{k|k}(\mathbf{x}|\mathbf{Z}_{1:k})$ abbreviated by $v_{k|k}(\mathbf{x})$. We use the decomposed form of PHD filter [32] that intends to distinguish the survival targets $v_{k|k}(\mathbf{x}, 0)$ and new-born targets $v_{k|k}(\mathbf{x}, 1)$ in both the prediction and update steps. Hence, the PHD prediction equation is given by,

$$v_{k|k-1}(\mathbf{x}, 0) = \int e_{k|k-1}(\xi) f_{k|k-1}(\mathbf{x}|\xi) v_{k-1|k-1}(\xi) d(\xi) \quad (1)$$

$$v_{k|k-1}(\mathbf{x}, 1) = \gamma_{k|k-1}(\mathbf{x}) \quad (2)$$

where $\gamma_{k|k-1}(\mathbf{x})$ is the intensity function of the new-born target, $f_{k|k-1}(\cdot)$ is the single-target transition density, $e_{k|k-1}(\xi)$ is the probability that a target state ξ at time $k-1$ will exist until time k. The PHD update step [32] can be defined with the available measurements from survival targets $\mathbf{Z}_{k,s}$ and new-born targets $\mathbf{Z}_{k,b}$ as:

$$v_{k|k}(\mathbf{x}, 0) = \sum_{\mathbf{z} \in \mathbf{Z}_{k,s}} \frac{\psi_k(\mathbf{x}) v_{k|k-1}(\mathbf{x}, 0)}{\kappa_k(\mathbf{z}) + \langle \psi_k(\mathbf{x}), v_{k|k-1}(\mathbf{x}, 0) \rangle} + v_{k|k-1}(\mathbf{x}, 0) p_M(\mathbf{x}) \quad (3)$$

$$v_{k|k}(\mathbf{x}, 1) = \sum_{\mathbf{z} \in \mathbf{Z}_{k,b}} \frac{\psi_k(\mathbf{x}) v_{k|k-1}(\mathbf{x}, 1)}{\kappa_k(\mathbf{z}) + \langle \psi_k(\mathbf{x}), v_{k|k-1}(\mathbf{x}, 0) \rangle} + v_{k|k-1}(\mathbf{x}, 1) p_M(\mathbf{x}) \quad (4)$$

where $p_M(\cdot)$ is the missed detection probability, $\psi_k(\mathbf{x}) = (1 - p_M(\mathbf{x})) g_k(\mathbf{z}|\mathbf{x})$, $g_k(\mathbf{z}|\mathbf{x})$ is the measurement likelihood of an individual target, $\kappa_k(\mathbf{z})$ is the clutter intensity, and $\langle f, g \rangle = \int f(x) g(x) dx$.

In our work, we adopt the sequential Monte Carlo method to approximate the PHD filter with a set of weighted random samples $\{\tilde{\omega}_{k-1}^i, \tilde{\mathbf{x}}_{k-1}^i\}_{i=1}^{i=(M_{k-1}) \times \mathcal{N}}$, where $\mathcal{N}$ is the number of particles used to represent each target. The PHD prediction at time k can be represented with a set of weighted particles including both survived targets and new-born targets,

$$\{\tilde{\omega}_{k|k-1}^i, \tilde{\mathbf{x}}_k^i\}_{i=1}^{i=(M_{k-1}+J_k) \times \mathcal{N}} \quad (5)$$

where J_k denotes the number of new-born targets at time k. Hence, the predicted weights are given as,

$$\tilde{\omega}_{k|k-1}^i = \begin{cases} f_{k|k-1}(\tilde{\mathbf{x}}_k^i) \tilde{\omega}_{k-1}^i, i = 1, \ldots, M_{k-1} \times \mathcal{N}. \\ \frac{\gamma_{k|k-1}(\mathbf{x})}{J_k}, i = M_{k-1} \times \mathcal{N} + 1, \ldots, (M_{k-1} + J_k) \times \mathcal{N}. \end{cases} \quad (6)$$

Once the new set of observations is available, we can substitute the approximation of $v_{k|k-1}(\mathbf{x})$ into (3) and (4), the weights of each particle are updated as,

$$\tilde{\omega}_k^i = \left[p_M(\tilde{\mathbf{x}}_k^i) + \sum_{\mathbf{z} \in \mathbf{Z}_k} \frac{\psi_k(\tilde{\mathbf{x}}_k^i)}{\kappa_k(\mathbf{z}) + C_k(\mathbf{z})} \right] \tilde{\omega}_{k|k-1}^i \quad (7)$$

where

$$C_k(\mathbf{z}) = \sum_{i=1}^{(M_{k-1}+J_k) \times \mathcal{N}} \psi_k(\tilde{\mathbf{x}}_k^i) \tilde{\omega}_{k|k-1}^i. \quad (8)$$

The likelihood function for each particle is given by,

$$g_k(\mathbf{z}|\tilde{\mathbf{x}}_k^i) = \frac{1}{(2\pi\sigma_g)^{1/2}} \exp\left(- \frac{(\mathbf{z} - \mathbf{H}\tilde{\mathbf{x}}_k^i)^T (\mathbf{z} - \mathbf{H}\tilde{\mathbf{x}}_k^i)}{2\sigma_g^2} \right) \quad (9)$$

where $\mathbf{H}$ is the observation matrix and σ_g^2 denotes the variance for likelihood function. The expected number of targets M_k at time k can be estimated by the total mass of the

weights from (7), $M_k = \sum_{i=1}^{(M_{k-1}+J_k)\times\mathcal{N}} \tilde{\omega}_k^i$. Furthermore, a resampling step will be performed with normalized weights $\tilde{\omega}_k^i = \tilde{\omega}_k^i / M_k$ after the update step, aiming to eliminate particles with low importance weight and avoid the degeneracy problem [5].

The above work underpins the decomposed form of SMC-PHD filter, which has been used extensively in multiple human tracking [32]–[34]. Besides, it is necessary for the PHD filter to add an additional mechanism to provide target identity information. For instance, an ID management clustering algorithm [25] can be utilized to extract the current states from all the particles.

B. ADAPTIVE GATING BASED MEASUREMENT CLASSIFICATION

To implement the measurement-driven particle PHD filter, the measurements obtained from the detector are required to be classified as survival target measurements, new-born target measurements and background clutter. To obtain more available real measurements, we propose a novel adaptive gating method which is designed to extract the survival targets from the entire measurement set, and also to discard false positives within the survival measurements set.

For our video-based multiple human tracking system, the proposed measurement-driven mechanism is not limited to employing the positions and velocities, but an elliptical human shape with height and width is also included to develop this gating technology. The validation gate threshold T_k is designed to reduce the number of candidate measurements which can be reasonably associated with the predicted target state as,

$$T_k = \left[\left(\frac{1}{N_k} \sum_{n=1}^{N_k} \|\bar{\mathbf{s}}_k^n\|_1 \right)^2 + \left(\frac{1}{M_k} \sum_{m=1}^{M_k} \|\mathbf{s}_{k|k-1}^m\|_1 \right)^2 \right]^{\frac{1}{2}} \quad (10)$$

where $\bar{\mathbf{s}}_k^n = [\bar{w}_k^n, \bar{h}_k^n] \in \mathbf{z}_k^n$ offers the size information for the n-th measurement, $\mathbf{s}_{k|k-1}^m = [w_{k|k-1}^m, h_{k|k-1}^m] \in \mathbf{x}_{k|k-1}^m$ denotes the width and height of the m-th predicted target state obtained by $\mathbf{x}_{k|k-1}^m = \mathbf{H}\mathbf{F}\mathbf{x}_{k-1}^m$, $\mathbf{F}$ denotes the transition matrix and $\|\cdot\|_1$ is the $l1$ norm.

However, the size of human target may alter with different views from a physically static camera especially when targets appear and disappear in the monitored area. To address the uncertainties in the validation region, we propose an alternative method is mainly to update the gating threshold with an adaptive step [23] associated with previous measurements,

$$T_k = (1 - \lambda_k) T_{k-1} + \lambda_k T_k \quad (11)$$

$$\lambda_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \sum_{j=1}^{N_{k-1}} \exp\left(- \frac{(\mathbf{z}_k^n - \mathbf{z}_{k-1}^j)^T (\mathbf{z}_k^n - \mathbf{z}_{k-1}^j)}{2\sigma_\lambda^2} \right). \quad (12)$$

The adaptive parameter λ_k is computed by the similarity between the measurement set $\mathbf{Z}_k$ at time k and $\mathbf{Z}_{k-1}$ at time $k-1$, where σ_λ represents the standard deviation. As a consequence, the n-th survival measurement $\mathbf{z}_{k,s}^n$ can be effectively distinguished from available measurement $\mathbf{z}_k^n \in \mathbf{Z}_k$, $n = 1, 2, \ldots, N_k$ as follows,

$$\mathbf{z}_{k,s}^n = \{ \mathbf{z}_k^n : \min_n \|\bar{\mathbf{p}}_k^n - \mathbf{p}_{k|k-1}^m\| < T_k \} \quad (13)$$

where the n-th measurement location is $\bar{\mathbf{p}}_k^n = [\bar{p}_{x,k}^n, \bar{p}_{y,k}^n] \in \mathbf{z}_k^n$, $\mathbf{p}_{k|k-1}^m = [p_{x,k|k-1}^m, p_{y,k|k-1}^m] \in \mathbf{x}_{k|k-1}^m$, $m = 1, \ldots, M_k$ denotes the location of the m-th predicted state, and $\|\cdot\|$ denotes the Euclidean distance. Furthermore, we discard some duplicate detections that have the same Euclidean distance to a predicted state $\mathbf{x}_{k|k-1}^m$ correspondingly, which would implicitly reduce the amount of false alarms within the survival measurement set. The measurement set of survival targets is defined as the union of all survival measurements,

$$\mathbf{Z}_{k,s} = \bigcup_{n=1}^{N_k} \mathbf{z}_{k,s}^n \quad (14)$$

and then the residual measurement set $\mathbf{Z}_{k,r}$ is defined as,

$$\mathbf{Z}_{k,r} = \mathbf{Z}_k \setminus \mathbf{Z}_{k,s}. \quad (15)$$

The complementary set $\mathbf{Z}_{k,r}$ comprising the new-born targets, clutter and false detections will be further distinguished via online group-structured dictionary learning in the following section.

Algorithm 1 Adaptive Gating Measurement Classification (at time $k > 1$)

Input : $\mathbf{Z}_k$, $\mathbf{Z}_{k-1}$, and $\mathbf{X}_{k|k-1}$.
Output: $\mathbf{Z}_{k,s}$ and $\mathbf{Z}_{k,r}$.
Initialization: Set the gating threshold T_1 and standard deviation σ_λ.
Set $\mathbf{Z}_{k,s} = \emptyset$, and $\mathbf{Z}_{k,r} = \emptyset$.
Compute T_k using Eq. (10).
Compute the adaptive parameter λ_k with Eq. (12)
Update T_k with Eq. (11)
for *each* $\mathbf{x}_{k|k-1}^m \in \mathbf{X}_{k|k-1}$, $m = 1, \ldots, M_k$ **do**
  for *each* $\mathbf{z}_k^n \in \mathbf{Z}_k$, $n = 1, 2, \ldots, N_k$ **do**
    Obtain each survival measurement $\mathbf{z}_{k,s}^n$ with Eq. (13).
  end
end
Compute $\mathbf{Z}_{k,s}$ using Eq. (14) and $N_{k,s} = |\mathbf{Z}_{k,s}|$.
Compute $\mathbf{Z}_{k,r}$ with Eq. (15).

The example pseudo-code in Algorithm 1 summarizes the proposed adaptive gating method. It is noteworthy that the adaptive step in [23] combining the forward and backward processing (batch method) is developed to reduce the approximation error induced by delayed measurements. While our approach is different from the previous solution, it utilizes this adaptive step as a part of our proposed method to enhance the gating technique for measurement classification. Specifically the adaptive parameter λ_k achieved by only using past and current inputs, can be considered as a forgetting process that weights the contribution of the updating gating threshold to

the previous threshold value, thereby handling the uncertainties in the validation region and increasing the robustness to parameter changes. Moreover, the proposed technique can be interpreted as a reliable approach that fuses temporal, spatial and human size information and provides flexibility to dynamically control the validation region, and thereby enhancing the ability to reduce the false measurements without causing any further impact on the system's performance.

C. DICTIONARY CONSTRUCTION

Prior to commencing the study of group-structured dictionary learning for birth intensity estimation, feature extraction is a necessary step for target appearance modelling to be applied in the training and testing process. The training phase is processed using data with higher confidence score from the MOTChallenge Benchmark. Hand-crafted features are extracted with training data from each image in the target region $S = (x, y, w, h)$, including the RGB colour histogram with 8 bins for each channel and the grey-scale histogram of oriented gradients (HOGs) with 9 orientation bins [43]. Typically, we employ a feature vector $\mathbf{c}_n \in \mathbb{R}^{d_c}$ that consists of transformed coefficients of the RGB colour histogram for characterizing an image patch, where d_c is the dimensionality of the RGB colour feature. Then, we use these RGB colour features to form an feature set $\mathbf{F}_c = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n] \in \mathbb{R}^{d_c \times n}$, where n denotes the total number of feature vectors in the training data. Likewise the vectorized HOG features $\mathbf{h}_n \in \mathbb{R}^{d_h}$ are represented by a matrix $\mathbf{F}_h = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n] \in \mathbb{R}^{d_h \times n}$, where d_h is the dimensionality of HOG features. For simplicity, the HOG and RGB colour features are concatenated to a combined feature set, $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n] \in \mathbb{R}^{(d_c+d_h)\times n}$.

Different from imposing data directly to a dictionary, we employ an unsupervised learning method - hierarchical K-means clustering algorithm [4] to fuse the group structure information into a dictionary from the combined feature template $\mathbf{F} \in \mathbb{R}^{d\times n}$, which allows the dictionary atoms in each class to be well clustered, and results in a large within-class similarity. For example, the same tracked target in different image frames under different illumination and pose conditions can be clustered into the same group (class). Furthermore, the learned dictionary with group structure enforces the label consistency between sub-dictionaries and training data [36]. As a consequence, this dictionary $\mathbf{D} \in \mathbb{R}^{d\times n}$ is learned off-line with the pre-defined group structure $\mathcal{G} = \{1, \ldots, q\}$ having q groups with the same l sub-dictionary atoms in each group, which is a column matrix concatenation $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_g, \ldots, \mathbf{D}_q]$ consisting of the number of q independent sub-dictionaries, where $\mathbf{D}_g \in \mathbb{R}^{d\times l}$, $g \in \mathcal{G}$ represents the sub-dictionary with l atoms, as shown in Fig. 2. In addition, we crop the observed target $\mathbf{z}_k \in \mathbf{Z}_{k,r}$ from the residual measurement set at the current image frame as well as extracting the hand-crafted features to constitute an observed target vector $\mathbf{y} \in \mathbb{R}^d$. In fact, learning the representation for each measurement can be viewed as an individual task in the feature space, while we intend to exploit similarities between observed signals and the group-structured dictionary in a multi-task approach, which yields an observation matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{N_{k,r}}] \in \mathbb{R}^{d\times N_{k,r}}$, where $N_{k,r}$ denotes the cardinality of $\mathbf{Z}_{k,r}$.

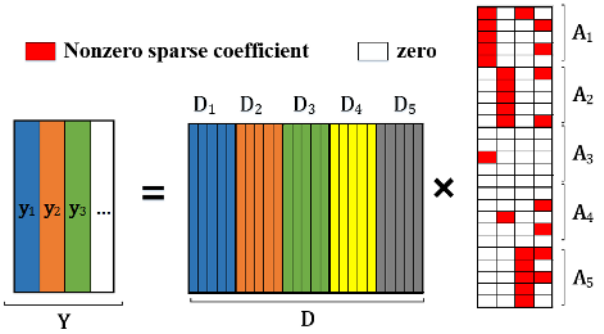


FIGURE 2. Example illustration of multi-task structured sparsity solution induced by the C-HiLasso model. The dictionary $\mathbf{D}$ consists of sub-dictionaries for five different groups, $\mathbf{D}_1, \ldots, \mathbf{D}_5$, with five atoms in each group. Input signals $\mathbf{Y}$ contains different measurements in feature space. All input signals within the same class are forced to reveal the group-sparsity structure $\mathbf{A}_1, \ldots, \mathbf{A}_5$.

D. GROUP-STRUCTURED DICTIONARY LEARNING FOR BIRTH INTENSITY ESTIMATION

Based on the analysis in Section III-B, the entire measurement set $\mathbf{Z}_k$ has been divided into the set of survival measurements $\mathbf{Z}_{k,s}$ in (14) and the residual measurement set $\mathbf{Z}_{k,r}$ in (15). Considering the number of new-born targets is unknown, and any initialization or prior information is unavailable for generating birth measurements, this residual measurement set $\mathbf{Z}_{k,r}$ potentially containing the new-born targets $\mathbf{Z}_{k,b}$ and other different kinds of false detections Γ_k can be updated as,

$$\mathbf{Z}_{k,r} = \Gamma_k \cup \mathbf{Z}_{k,b} \tag{16}$$

Hence, it is essential for estimating birth intensity before the PHD prediction, to remove the false alarms from the remaining measurements. To achieve this, we attempt to employ the online group-structured dictionary learning to discriminate the new-born targets from false alarms or background clutter, and thus correctly estimate the birth intensity. It is known that seeking the sparsity solution $\mathbf{A}$ is NP-hard in Fig. 2. Traditionally, the sparse coding solution $\mathbf{a}_i$ for each test target $\mathbf{y}_i$ is performed separately via *Lasso* or *Basis pursuit* [44], because different tasks choose the dictionary atoms independently. However, the dictionary atoms of proposed approach have been grouped with pre-defined structure in the learning process instead of being treated as singletons, which allows multiple test targets to be represented by a few active groups of atoms, and a few atoms of each group are selected to be active at a time. For this study, we introduce a C-HiLasso method [27] to acquire group structured sparsity at the multi-task level. Specifically, an over-complete dictionary $\mathbf{D} \in \mathbb{R}^{d\times n}$ and input signals $\mathbf{Y} \in \mathbb{R}^{d\times N_{k,r}}$ are effectively taken from the learning process in Section III-C. The sparse coefficient matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{N_{k,r}}] \in \mathbb{R}^{n\times N_{k,r}}$ can be

accomplished by the following multi-task C-HiLasso model [27],

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times N_{k,r}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2 + \lambda_2 \sum_{g\in\mathcal{G}} \|\mathbf{A}_g\|_F + \lambda_1 \sum_{j=1}^{N_{k,r}} \|\mathbf{a}_j\|_1 \tag{17}$$

where $\mathbf{A}_g$ is the sub-matrix consisting of l rows belonging to the g-th group, and $\|\cdot\|_F$ denotes the Frobenius norm. In addition, the selection of λ_1 and λ_2 is dependent on the application and data, such parameters can be obtained by cross validation. The sparsity pattern is shown in Fig. 2, which is effective and suitable to perform classification for multi-target tracking, since using the group structure of this sparsity solution could enforce the sparse coefficients for different classes to deal with different subspaces, so the sparse coefficients in our system would be further strengthened to discriminate the new-born targets from the uncertain measurements. In general, the nonzero sparse codes for each measurement are gathered within a group g, as depicted in Fig. 2. However, when candidate targets are outliers and out of the dictionary, the nonzero coefficients tend to scatter among groups instead of centralizing in some single group [44].

According to the sparsity solution, a maximum voting method is developed to acquire all the new-born targets from the residual measurements $\mathbf{Z}_{k,r}$. The i-th residual measurement $\mathbf{z}_{k,r}^i$ is clustered to a birth measurement $\mathbf{z}_{k,b}^i$ with the following equation,

$$\mathbf{z}_{k,b}^i = \left\{ \mathbf{z}_{k,r}^i : \frac{\max_{g\in\mathcal{G}} \|\mathbf{A}_{g,i}\|_1}{\|\mathbf{A}_{:,i}\|_1} \geq \varepsilon \right\} \tag{18}$$

where $\mathbf{A}_{g,i}$ denotes the i-th column in $\mathbf{A}_g$, and ε is a predefined threshold value. The measurement set of new-born targets is effectively defined as the union of all actual birth measurements,

$$\mathbf{Z}_{k,b} = \bigcup_{i=1}^{N_{k,r}} \mathbf{z}_{k,b}^i. \tag{19}$$

Then, the false measurements can be removed as follows,

$$\mathbf{Z}_{k,r} \setminus \mathbf{Z}_{k,b} = \emptyset. \tag{20}$$

Let α_u record each valid index i, where $u = 1, \ldots, J_k$ ($J_k > 0$) and J_k is the cardinality of $\mathbf{Z}_{k,b}$, and $\Omega = \{\alpha_1, \ldots, \alpha_{J_k}\}$ be a set comprised of each α_u. The voting index θ_u for each new-born target $\mathbf{z}_{k,b}^u$ can be calculated by,

$$\theta_u = \arg\max_{g\in\mathcal{G}} \|\mathbf{A}_{g,\alpha_u}\|_1, \tag{21}$$

then each obtained θ_u consists of a set $Q = \{\theta_1, \ldots, \theta_{J_k}\}$. It is worth noting that the Q and Ω are defined as the structure pattern which will be further used in the dictionary update. The above voting index is used to compute the average of the selected group sparse codes for the birth intensity estimation,

$$\eta_u = \frac{1}{l}\|\mathbf{A}_{\theta_u,\alpha_u}\|_1. \tag{22}$$

Once all the birth measurements are obtained, the birth intensity function can be formulated as,

$$\gamma_{k|k-1}(\mathbf{x}) = \frac{1}{J_k}\sum_{u=1}^{J_k} \frac{1}{(2\pi\sigma_b)^{1/2}} \exp\left(-\frac{\eta_u}{2\sigma_b^2}\right). \tag{23}$$

The details of the proposed maximum voting method for birth intensity estimation are summarized in Algorithm 2. The obtained birth intensity function can be finally taken as the input to the prediction step (6), and meanwhile both survival measurement set $\mathbf{Z}_{k,s}$ and birth measurement set $\mathbf{Z}_{k,b}$ distinguished by the proposed method are offered to realize the weights update of (7) (8) in Section III-A.

Algorithm 2 Birth Intensity Estimation by Maximum Voting (at Time $k > 1$)

Input : The residual measurement set $\mathbf{Z}_{k,r}$; The sparse coefficients matrix $\mathbf{A} \in \mathbb{R}^{n\times h}$; The group structure $\mathcal{G} = \{1, \ldots, q\}$, and each group g consists of same l columns.

Output: The birth intensity function $\gamma_{k|k-1}(\mathbf{x})$, Q and Ω

Initialization: Set the threshold to ε and $\mathbf{Z}_{k,b} = \emptyset$.
for *each* $\mathbf{z}_{k,r}^i$, $i = 1, 2, \ldots, N_{k,r}$ **do**
  Obtain each birth measurement $\mathbf{z}_{k,b}^i$ with Eq. (18).
end
Compute the measurement set of new born targets $\mathbf{Z}_{k,b}$ with Eq. (19).
Remove the false measurements with Eq. (20).
if $J_k > 0$ **then**
  for *each* $\mathbf{z}_{k,b}^u$, $u = 1, 2, \ldots, J_k$ **do**
    Compute the voting index θ_u with Eq. (21).
    Calculate the average of selected sparse codes η_u with Eq. (22).
  end
  Compute the Birth intensity $\gamma_{k|k-1}(\mathbf{x})$ with Eq. (23).
  Compute the structure pattern Q and Ω.
end

E. DICTIONARY UPDATE WITH SIMCO ALGORITHM

The dictionary $\mathbf{D}$ mentioned earlier is a pre-trained dictionary that only fuses the information of a few relevant frames, while using such off-line dictionary may not adapt to the tracking scenario or robustly cope with the occlusions because of the appearance variation in both the object target and the background [41]. In order to improve the robustness of our dictionary, the SimCO algorithm proposed by Dai *et al.* [28] is utilized for our dictionary update stage, since the key characteristics of this approach can achieve our goal of updating an arbitrary subset of atoms in $\mathbf{D}$. In general, the dictionary update problem with SimCO algorithm can be written as [28],

$$\arg\min_{\mathbf{D}\in\mathbb{R}^{d\times n}} f(\mathbf{D}) = \arg\min_{\mathbf{D}\in\mathbb{R}^{d\times n}} \left(\min_{\mathbf{A}\in\mathbb{R}^{n\times N_{k,r}}} \|\mathbf{Y}-\mathbf{D}\mathbf{A}\|_F^2 \right) \tag{24}$$

where the dictionary matrix $\mathbf{D}$ contains unit ℓ_2-norm columns, and sparse coding matrix $\mathbf{A}$ is obtained from (17).

Algorithm 3 Dictionary Update (at Time $k > 1$)

Input : $\mathbf{D}^k$,$\mathbf{A}^k$,$\mathbf{Y}$, J_k, Q, Ω
Output: $\mathbf{D}^{k+1}$
if $J_k > 0$ **then**
  Extract the matrix of $\mathbf{B}^k \leftarrow \mathbf{D}^k$ with the Q structure pattern.
  Find a proper step size δ with the method of golden section search [28].
  Compute the search direction, with (28) and (29).
  Update $\mathbf{B}^k \rightarrow \mathbf{B}^{k+1}$ with (30).
  Update the dictionary $\mathbf{D}^{k+1} \leftarrow \mathbf{B}^{k+1}$.
end

Instead of updating all the atoms of $\mathbf{D}$, our dictionary update can be carried through an efficient way of following the structure pattern Q and Ω determined by proposed maximum voting method to update the selected groups of atoms. To be specific, let $Q \subseteq \mathcal{G}$ with $|Q| = J_k$, $0 < J_k \leq q$ be the index set of sub-dictionaries to be updated. Then a new matrix concatenation $\mathbf{B} = [\mathbf{D}_{\theta_1}, \ldots, \mathbf{D}_{\theta_u}, \ldots, \mathbf{D}_{\theta_{J_k}}] \in \mathbb{R}^{d\times(J_k\times l)}$ indexed by Q is to be updated, where $\theta_u \in Q$ and $\mathbf{D}_{\theta_u} \in \mathbb{R}^{d\times l}$ denotes the sub-matrix of $\mathbf{D}$ formed by l columns of $\mathbf{D}$, while we remain $\mathbf{B}^c$ including other sub-matrices of $\mathbf{D}$ indexed by Q^c to be constant, where Q^c is a set complementary to Q over $\mathcal{G}$. Similarly, define $\mathbf{K} = [\mathbf{A}^T_{\theta_1}, \ldots, \mathbf{A}^T_{\theta_u}, \ldots, \mathbf{A}^T_{\theta_{J_k}}]^T \in \mathbb{R}^{(J_k\times l)\times h}$, where $\mathbf{A}_{\theta_u}$ is the sub-matrix of $\mathbf{A}$ containing l rows of the sparse coding matrix $\mathbf{A}$, and let $\mathbf{K}^c$ be the composite of remaining sub-matrices of $\mathbf{A}$ indexed by Q^c.

In accordance with the method in [28], we can define the following equation,

$$\mathbf{Y}_r = \mathbf{Y} - \mathbf{B}^c\mathbf{K}^c \tag{25}$$

Since $\mathbf{Y}-\mathbf{DA} = \mathbf{Y}_r-\mathbf{BK}$, then the dictionary update problem in (24) can be written as,

$$\underset{\mathbf{B}\in\mathbb{R}^{d\times(J_k\times l)}}{\arg\min}\ f(\mathbf{B}) = \underset{\mathbf{B}\in\mathbb{R}^{d\times(J_k\times l)}}{\arg\min}\left(\min_{\mathbf{K}\in\mathbb{R}^{(J_k\times l)\times h}} \|\mathbf{Y}_r - \mathbf{BK}\|_F^2\right) \tag{26}$$

Assuming that appearance variation would possibly happen when targets with new IDs being detected or existing targets re-entering to the scene, only valid input signals $\mathbf{Y}_\Omega = [(\mathbf{Y}_r)_{\alpha_1}, \ldots, (\mathbf{Y}_r)_{\alpha_u}, \ldots, (\mathbf{Y}_r)_{\alpha_{J_k}}] \in \mathbb{R}^{d\times J_k}$, $\alpha_u \in \Omega$ can be considered to be implemented in the dictionary update with the corresponding $\mathbf{K}_\Omega = [\mathbf{K}_{\alpha_1}, \ldots, \mathbf{K}_{\alpha_u}, \ldots, \mathbf{K}_{\alpha_{J_k}}] \in \mathbb{R}^{(J_k\times l)\times J_k}$, so the objective function $f(\mathbf{B})$ is given by,

$$f(\mathbf{B}) = \min_{(\mathbf{K}_\Omega)_{:,u}} \sum_{u=1}^{J_k} \|(\mathbf{Y}_\Omega)_{:,u} - \mathbf{B}(\mathbf{K}_\Omega)_{:,u}\|_2^2 \tag{27}$$

where $(\mathbf{Y}_\Omega)_{:,u}$ is the u-th column of $\mathbf{Y}_\Omega$, and $(\mathbf{K}_\Omega)_{:,u}$ denotes the u-th column of $\mathbf{K}_\Omega$. Here, the gradient descent line search method can be applied for this stage. Firstly, the search direction $\mathbf{E}$ is defined as follows,

$$\begin{aligned}\mathbf{E} &= -\nabla f(\mathbf{B}) \\ &= -2(\mathbf{Y}_\Omega - \mathbf{BK}_\Omega)\mathbf{K}_\Omega^T\end{aligned} \tag{28}$$

The line search path for this study was prepared using the product of Grassmann manifolds that was detailed in [28]. Let $\mathbf{e}_j$ be the j-th column of $\mathbf{E}$. We define

$$\bar{\mathbf{e}}_j = \mathbf{e}_j - \mathbf{B}_{:,j}\mathbf{B}_{:,j}^T\mathbf{e}_j, \tag{29}$$

where $\mathbf{B}_{:,j}$ denotes the j-th column of $\mathbf{B}$ to be updated. Therefore, the line search path for dictionary update $\mathbf{B}(\delta)$ can be written as,

$$\mathbf{B}_{:,j}(\delta) = \begin{cases} \mathbf{B}_{:,j} & \text{if } \|\bar{\mathbf{e}}_j\|_2 = 0, \\ \mathbf{B}_{:,j}\cos(\|\bar{\mathbf{e}}_j\|_2\delta) + \left(\frac{\bar{\mathbf{e}}_j}{\|\bar{\mathbf{e}}_j\|_2}\right)\sin(\|\bar{\mathbf{e}}_j\|_2\delta) & \\ & \text{if } \|\bar{\mathbf{e}}_j\|_2 \neq 0. \end{cases} \tag{30}$$

where the step size $\delta \in \mathbb{R}^+$ is properly chosen via the method of golden section search [28]. Besides, the dictionary update stage is summarized in Algorithm 3.

IV. EXPERIMENTS

In this section, we initially introduce the datasets used in our experiment and parameter settings, then explain the widely used evaluation metrics. To evaluate both the effectiveness and strength of our proposed tracking method, we investigate the individual contribution of each component in our tracking system by testing it on five commonly used datasets, as well as showing quantitative comparisons in a fair way with a range of state-of-the-art tracking methods using the MOTChallenge benchmark. We additionally discuss the runtime performance of our method.

A. DATASETS

We firstly validate the proposed tracking method on five commonly used video sequences: CAVIAR-EnterExitCrossing Paths1cor (CAVIAR) [45], PETS2009-View001-S2L1 (PETS2009) [46], TUD-Stadtmitte [47], TUD-Campus [47], and TUD-Crossing [47], so as to conduct the analysis of different contribution components. The CAVIAR dataset shows the scene of two people entering and exiting from a shopping mall, and a couple walking in the corridor, whilst including partial occlusions and complete occlusions and similar appearances. The PETS2009 dataset shows a large variable number of walking pedestrians with highly dynamic movement in an outdoor environment. It contains frequent occlusions with moving targets or static objects as well as illumination changes. In TUD datasets, there are more complex mutual occlusions because of low camera viewpoint and closely moving pedestrians with similar speeds. In addition, we further use the MOTChallenge Benchmark dataset[1] to

[1] https://motchallenge.net/

TABLE 1. Parameter values used in the Experiments .

p_M : missed detection probability	0.01
e : survival probability	0.99
κ : the clutter intensity	0.0001
$\mathcal{N}$: the number of particles for each target	100
T_1 : initial gating threshold	60
σ_g^2 : variance for the measurement likelihood function	25
σ_λ^2 : variance for the adaptive parameter function	25
σ_b^2 : variance for the birth intensity function	10
λ_1 : first regularization parameter	0.06
λ_2 : second regularization parameter	0.03
ε : maximum voting threshold	0.52

evaluate the tracking performance of the proposed tracking system. This benchmark collects a set of video sequences from other datasets and some new challenging sequences, as well as providing public object detections for fair comparisons. The 2D MOTChallenge 2015 dataset [48], consists of 11 training and 11 testing video sequences captured by both static and dynamic cameras. MOT17 dataset contains 7 training and 7 testing sequences, in which each sequence is provided with 3 sets of public detections. All the video sequences are only suitable for pedestrian tracking. Multi-camera tracking is out of the scope of this paper. The testing video sequences are used for performance evaluation in comparisons with other recent trackers.

B. PARAMETER SETTINGS

For this study, we apply the state transition model with constant velocity $\mathbf{F} = [\mathbf{I}_2, \Delta t \times \mathbf{I}_2, \mathbf{0}_2; \mathbf{0}_2, \mathbf{I}_2, \mathbf{0}_2; \mathbf{0}_2, \mathbf{0}_2, \mathbf{I}_2]$ and the observation model $\mathbf{H} = [\mathbf{I}_2, \mathbf{0}_2, \mathbf{0}_2; \mathbf{0}_2, \mathbf{0}_2, \mathbf{I}_2]$ from previous work in [22], where $\mathbf{I}_2$ and $\mathbf{0}_2$ are the 2×2 identity and zero matrices respectively, and Δt is the time interval between frame k and $k+1$. In accordance with [2], each concatenated feature vector $\mathbf{f} \in \mathbb{R}^d$ from the training data contains 512 elements from the colour histogram and 81 from the oriented gradient histogram. The principal component analysis (PCA) method is used for dimensionality reduction of the feature template. The dictionary size for all sequences is 5 dictionary atoms for each group. The MOTChallenge benchmark provides the ground truth of training sequences, thus we use the training video sequences for fine-tuning system parameters which remain fixed for all testing sequences. The system parameters used for all testing video sequences are summarized in Table 1.

C. EVALUATION METRICS

To achieve a fair comparison with previous state-of-the-art tracking algorithms, there are two performance measures, the optimal subpattern assignment (OSPA) [49] from signal processing community and the MOTChallenge Benchmark metrics [48] in this article being utilized for evaluation. Let $\mathbf{O}_k = \{\mathbf{o}_k^1, \ldots, \mathbf{o}_k^i, \ldots, \mathbf{o}_k^m\}$ be the ground truth with m targets at time k, where $\mathbf{o}_k^i = \{\mathbf{p}_k^i, I_k^i\}$ contains the actual 2D positions and identity information. Likewise, $\hat{\mathbf{O}}_k = \{\hat{\mathbf{o}}_k^1, \ldots, \hat{\mathbf{o}}_k^j, \ldots, \hat{\mathbf{o}}_k^n\}$ gives tracking results at time k with n targets, where each $\hat{\mathbf{o}}_k^j = \{\hat{\mathbf{p}}_k^j, \hat{I}_k^j\}$ represents the estimated target positions and the corresponding target identity [3].

1) THE OSPA METRIC

The miss-distance has generally played an essential part in the formulation and evaluation of filtering and control algorithms [49]. For single target tracking, the performance measures including the Euclidean errors and mean squared errors are based on the concept of miss-distance [49]. However, those errors seem to be unsuitable for the case in multi-target tracking. A performance metric for evaluating the multi-target tracking was developed in [49], which is intended to capture both cardinality and localization errors. This OSPA metric has been widely applied by several state-of-the-art methods in video-based multiple human tracking [2], [24], [29], [50], [51] to examine the tracking performance. Let $d_k^{(c)}(\mathbf{o}_k^i, \hat{\mathbf{o}}_k^j) = \min(c, d(\mathbf{o}_k^i, \hat{\mathbf{o}}_k^j))$ be the distance between $\mathbf{o}_k^i$ and $\hat{\mathbf{o}}_k^j$ at time k, where $d(\mathbf{o}_k^i, \hat{\mathbf{o}}_k^j)$ is the Euclidean distance, and c denotes the cut off parameter. For $1 \leq p \leq \infty$, and $c > 0$, we define

$$d_{k,p}^{(c)}(\mathbf{O}_k, \hat{\mathbf{O}}_k) := \left(\frac{1}{n}\left(\min_{\pi \in \Pi_n} \sum_{i=1}^{m} d_k^{(c)}(\mathbf{o}_k^i, \hat{\mathbf{o}}_k^{\pi(i)})^p + c^p(n-m)\right)\right)^{\frac{1}{p}} \tag{31}$$

for $m \leq n$; $d_{k,p}^{(c)}(\mathbf{O}_k, \hat{\mathbf{O}}_k) = d_{k,p}^{(c)}(\hat{\mathbf{O}}_k, \mathbf{O}_k)$ if $m > n$, and where Π_n is the set of permutations on $\{1, 2, \ldots, n\}$ for any $n \in \mathbb{N} = \{1, 2, \ldots\}$. The function of $d_{k,p}^{(c)}$ is called as the OSPA metric of order p with cut-off c at time k step.

2) THE MOTCHALLENGE BENCHMARK METRICS

We also employ the widely accepted evaluating tool MOTChallenge Benchmark metrics including the standard CLEAR MOT metrics [52] and the metrics defined in [53] to examine the performance of our proposed method. The CLEAR MOT metrics [52] mainly entail Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA). The MOTP as a precision score, is designed to measure the average position errors in 2D image plane between estimated tracking results and ground truth. The MOTA as an accuracy score, is comprised of the total number of false negatives (FN), the total number of false positives (FP), the total number of identity switches (IDS). The metrics from [53] include Mostly Track targets (MT, the ratio of ground truth objects whose trajectories are covered by a tracking result more than 80%), Mostly Lost targets (ML, the ratio of ground truth objects whose trajectories are covered by a tracking result less than 20%), and the total number of times a trajectory is fragmented (Frag). In addition, the average number of false alarms per frame (FAF) and the runtime performance in frames per second (Hz) are also included in benchmark evaluation metrics.

TABLE 2. Average OSPA (pixel) performance comparison of different system component on five video sequences. The best results are shown in bold.

Dataset	SMC-PHD [21]	SMC-PHD -AG	SMC-PHD -Online GSDL	Combined method
CAVIAR	48.26	34.76	22.25	**15.16**
PETS2009	33.08	27.65	21.98	**15.35**
TUD-stadtmitte	34.07	26.98	21.44	**15.74**
TUD-crossing	31.84	24.84	21.97	**16.74**
TUD-campus	39.01	27.37	22.88	**17.63**

D. EFFECTIVENESS EVALUATION OF PROPOSED CONTRIBUTIONS

In this paper, the adaptive gating method (SMC-PHD-AG) and online group-structured dictionary learning method (SMC-PHD-Online GSDL) have been proposed, and integrated with the particle PHD filter (SMC-PHD) framework to handle the challenging issues of multi-target tracking. To achieve a better understanding of the contribution of the individual system component, we typically check the OSPA performance measure to compare and evaluate the different stage tracking performance of our approach on five commonly used video datasets.

Table 2 shows the average OSPA performance comparison computed on all five sequences individually using different stage methods, so that we can observe the performance improvement from both SMC-PHD-AG and SMC-PHD-Online GSDL methods. Overall, the highest performance which returns the smallest OSPA error is reported when using both proposed terms regardless of different scenarios, in the meantime the performance can be effectively decreased by removing each individual term. By comparing the individual improvement of each proposed contribution, from the third and fourth columns in Table 2, we can understand the performance is relatively more improved by employing SMC-PHD-Online GSDL than in SMC-PHD-AG. This can be explained by the fact that the SMC-PHD-AG approach is fundamentally capable of offering prior information for measurement classification and handling the missed detections, whereas the SMC-PHD-Online GSDL method intends to further strengthen the ability of discriminating the targets from noisy environment, as well as resolving the occlusions.

For CAVIAR dataset, our combined method achieves the highest improvement of 68.58% over 5 benchmark sequences, where the average OSPA value is reduced from 48.26 to 15.16. This is because our proposed tracker can effectively eliminate a large number of false alarms caused by the raw background subtraction results in the less crowded CAVIAR dataset. Since most targets in the TUD-Crossing having similar sizes walk with similar speeds, it is likely to generate more errors in the case of heavy occlusions or long-term interactions using proposed method without occlusion reasoning and detection confidence. Nonetheless the tracking accuracy in the TUD-Crossing dataset is still increased by 47.42% compared with the SMC-PHD method.

On the other hand, the effectiveness evaluation can be also visually seen from Fig. 3, where each curve demonstrates

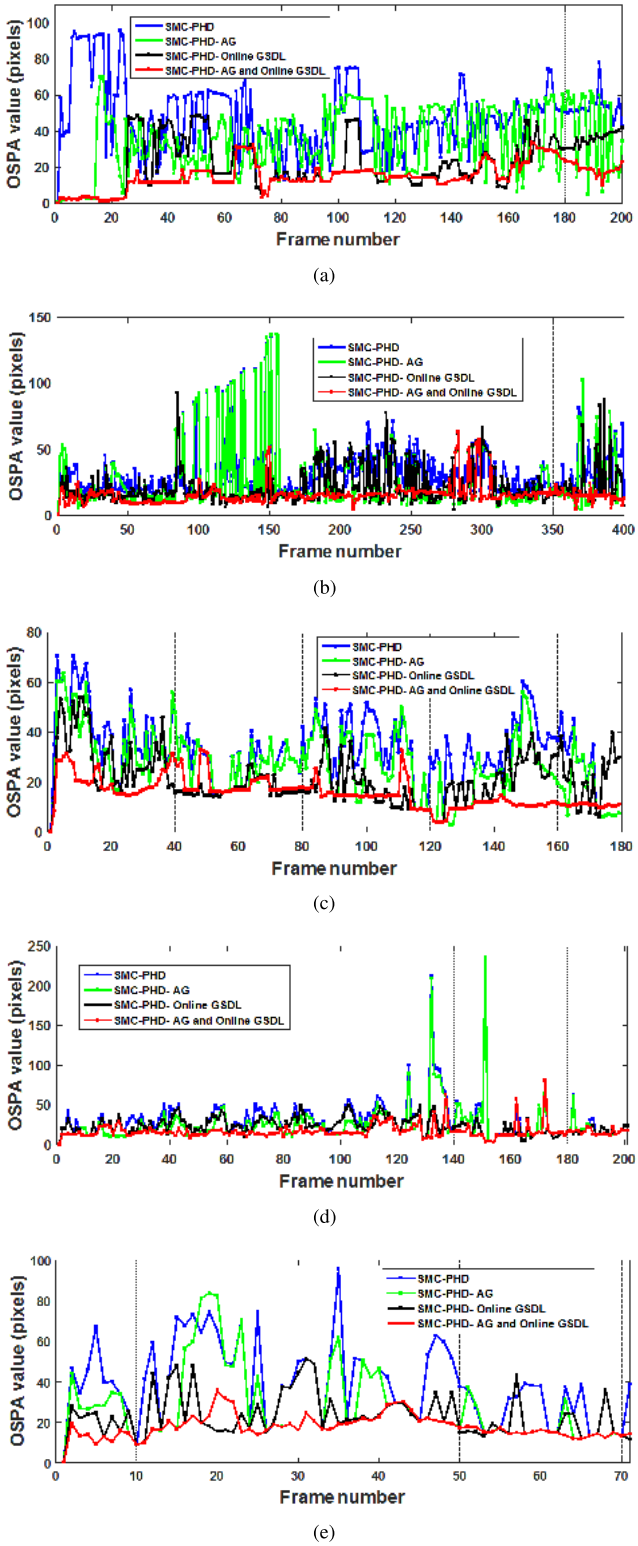


FIGURE 3. Effectiveness evaluation for different stages of our proposed tracking system on five video datasets. The performance is examined with OSPA metric.

the change of tracking performance over the entire sequence. From the results, we can see the OSPA performance using SMC-PHD-AG method (green curve) includes instability,

FIGURE 4. Qualitative performance of our proposed method on the test video sequences of the 2D MOTChallenge 2015. Different colors of the bounding boxes and trajectories demonstrate the identities of tracked targets.

which can be attributed to the inefficiency of dealing with the complex targets interactions especially within a crowded scene such as in the PETS2009-S2L1 dataset. As can be seen from the black curve of Fig. 3, the SMC-PHD-Online GSDL method apparently performs better and more robust than that in SMC-PHD-AG, which to some extent benefits from using both maximum voting to remove the false detections and adaptively estimating birth intensity to enhance the birth and death targets processing. More importantly, it is apparent to see that the OSPA value of our combined approach (red curve) is generally shown to be much lower than other baseline methods in most frames, and also the steady performance proves the robustness of the combined method.

Table 3 shows the comparison in terms of average OSPA measure between the proposed method and three recent state-of-the-art algorithms on five datasets. These three trackers are all reliant on the RFS-based Bayesian filtering method, including conventional particle PHD filter (SMC-PHD) [21], background subtraction based multi-Bernoulli filter (MB) [26], and social force model based particle PHD filter (SFM-PHD) [2]. We implemented the original MATLAB codes provided by the authors. The comparable results above emphasise the fact that our proposed tracker reports the best OSPA performance over five sequences. Specifically, our tracker outperforms the tracker in [26] with the improvement between approximately 31.8% and 60.2%. Moreover, the SFM-PHD method performs worse than our proposed method, which means that the group-structured sparsity of our tracker shows clear advantage over

TABLE 3. Average OSPA (pixel) comparison between proposed method and different state-of-the-art methods on five video sequences. The best results are shown in bold.

Dataset	SMC-PHD method [21]	MB method [26]	SFM-PHD method [2]	Proposed method
CAVIAR	48.26	29.38	17.11	**15.16**
PETS2009	33.08	36.94	17.63	**15.35**
TUD-stadtmitte	34.07	39.54	23.10	**15.74**
TUD-crossing	31.84	39.08	21.81	**16.74**
TUD-campus	39.01	25.85	22.70	**17.63**

TABLE 4. Quantitative comparison with other state-of-the-art methods on the 2D MOTChallenge 2015 benchmark with public detections. Our method is PHD_GSDL. The results are sorted as tracking mode and MOTA score. The best results are shown in bold, the second best are underlined. (Last accessed on 06/08/2017).

Method	Mode	MOTA (↑)	MOTP (↑)	FAF (↓)	MT (↑)	ML (↓)	FP (↓)	FN (↓)	IDS (↓)	Frag (↓)	Hz (↑)
HybridDaT [14]	Online	**35.0**	72.6	1.5	11.4%	42.2%	8,455	31,140	**358**	1,267	4.6
TDAM [15]	Online	33.0	**72.8**	1.7	**13.3%**	39.1%	10,064	**30,617**	464	1,506	5.9
PHD_GSDL	Online	30.5	71.2	1.1	7.6%	41.2%	6,534	35,284	879	2,208	8.2
MDP [8]	Online	30.3	71.3	1.7	13.0%	**38.4%**	9,717	32,422	680	1,500	1.1
SCEA [11]	Online	29.1	71.1	**1.0**	8.9%	47.3%	**6,060**	36,912	604	**1,182**	6.8
oICF [54]	Online	27.1	70.0	1.3	6.4%	48.7%	7,594	36,757	454	1,660	1.4
EAMTTPub [17]	Online	22.3	70.8	1.4	5.4%	52.7%	7,924	38,982	833	1,485	12.2
RMOT [16]	Online	18.6	69.6	2.2	5.3%	53.3%	12,473	36,835	684	1,282	7.9
GMPHD_15 [18]	Online	18.5	70.9	1.4	3.9%	55.3%	7,864	41,766	459	1,266	**19.8**
JointMC [13]	Batch	**35.6**	71.9	1.8	**23.2%**	39.3%	10,580	**28,508**	457	969	0.6
QuadMOT [12]	Batch	33.8	**73.4**	1.4	12.9%	**36.9%**	7,898	32,061	703	1,430	3.7
NOMT [9]	Batch	33.7	71.9	1.3	12.2%	44.0%	7,762	32,547	**442**	823	11.5
SiameseCNN [31]	Batch	29.0	71.2	**0.9**	8.5%	48.4%	**5,160**	37,798	639	1,316	**52.8**
DCO_X [7]	Batch	19.6	71.4	1.8	5.1%	54.9%	10,652	38,232	521	**819**	0.3

the OCSVM classifier in the SFM-PHD tracker regarding the ability of mitigating background noise and false positives, so as to achieve better performance in both localization and cardinality.

E. EVALUATIONS ON MOTCHALLENGE BENCHMARK

In this section, we test the proposed method denoted by PHD-GSDL on the test set of the 2D MOTChallenge 2015 Benchmark [48] and MOT17 Challenge [1]. In order to achieve a fair comparison between methods, we use the same public detections for all sequences and the centralized evaluation tool provided by the website of the MOTChallenge Benchmark. Tables 4 and 5 show the quantitative comparisons with a number of state-of-the-art tracking methods. These include online tracking approaches: MDP [8], SCEA [11], RMOT [16], GMPHD_15 [18], EAMTTPub [17], oICF [54], GM_PHD [55], and GMPHD_KCF [56], and also other offline (batch) approaches NOMT [9], QuadMOT [12], JointMC [13], SiameseCNN [31], DCO_X [7], jCC [13], EDMT17 [57], IOU17 [58] and DP_NMS [6]. Evaluation measures with (↑) indicate that higher is better, and with (↓) denote lower is better.

As illustrated in Table 4, the proposed method achieves better or competitive performance as compared to other state-of-the-art methods on most evaluation measures, and even outperforms most offline methods using the entire set of future outputs. In fact, off-line methods based on global association techniques usually perform better than online counterparts. Furthermore, Table 5 demonstrates the proposed method reports the highest MOTA score which indicates the most important metric for performance analysis, and also achieves the second best online tracker ranked on the leaderboard of MOT17 Challenge. The justification for the improved performance in MOTA is because many false alarms and missed detections are effectively mitigated by the proposed group-structured sparsity based classifier. In turn, our method also performs well in terms of tracking precision (high MOTP), fewer targets lost (low ML) and more targets tracked (high MT). This is mainly due to the proposed method being able to accurately estimate positions of varying number of targets, as well as robustly maintain the tracking consistency. Fig. 4 depicts some selected qualitative tracking results produced by our method on the test video sequences of the 2D MOTChallenge 2015. We can observe that some pedestrians with similar appearances that are partially or even almost fully occluded are successfully tracked through our tracker. This can be attributed by the online update mechanism with SimCO algorithm gives the benefits of dealing with the target appearance changes and the environment changes.

In order to further demonstrate the advantages of our RFS-based tracker, we compare the proposed method with other recent PHD filter based methods EAMTTPub [17], GMPHD_15 [18], GM_PHD [55] and GMPHD_KCF [56]. As compared to [17] and [18] on MOT15 dataset, our MOTA is improved by 8.2% and 12% respectively. More importantly, the proposed method outperforms the algorithms in [55] and [56] with large margins on MOT17 dataset. All above evaluations indicate that our proposed method within the PHD filter framework can achieve higher tracking performance in dynamic scenarios thereby verifying the robustness of the proposed method. More detailed tracking results and videos produced by our tracker can be found in the website of the MOTChallenge Benchmark.[2][3]

On the other hand, the number of ID switches and fragments of our approach are relatively higher than other methods. In fact, these two challenges are more likely to happen with a large number of targets and higher level difficulty of occlusions. In such a case, it is possible to utilize higher level features such as image textures to better identify targets instead of only using colour cues. Moreover, re-identification problems and contextual information can be explored in future work to further improve the tracking performance.

[2]https://motchallenge.net/results/2D_MOT_2015/
[3]https://motchallenge.net/results/MOT17/

TABLE 5. Quantitative comparison with other state-of-the-art methods presented in the MOT2017 Challenge benchmark using public detections. Our method is PHD_GSDL17. The best results are shown in bold. (Last accessed on 14/12/2017) .

Method	Mode	MOTA (↑)	MOTP (↑)	FAF (↓)	MT (↑)	ML (↓)	FP (↓)	FN (↓)	IDS (↓)	Frag (↓)	Hz (↑)
PHD_GSDL17	Online	**48.0**	**77.2**	**1.3**	**17.1%**	**35.6%**	**23,199**	**265,954**	**3,998**	8,886	6.7
GM_PHD [55]	Online	36.2	76.1	1.3	4.2%	56.6%	23,682	328,526	8,025	11,972	**38.4**
GMPHD_KCF [56]	Online	30.5	74.3	6.1	9.6%	41.8%	107,802	277,542	6,774	**7,833**	3.3
jCC [13]	Offline	**51.2**	75.9	1.5	20.9%	37.0%	25,937	247,822	**1,802**	**2,984**	1.8
EDMT17 [57]	Offline	50.0	**77.3**	1.8	**21.6%**	**36.3%**	32,279	**247,297**	2,264	3,260	0.6
IOU17 [58]	Offline	45.5	76.9	1.1	15.7%	40.5%	19,993	281,643	5,988	7,404	**1522.9**
DP_NMS [6]	Offline	43.7	76.9	**0.6**	12.6%	46.5%	**10,048**	302,728	4,942	5,342	137.7

F. RUNTIME PERFORMANCE

Our experiments were implemented on a desktop with an Intel i5 CPU with 3.5GHz and 16GB of memory without parallel processing. The code was written in MATLAB without any optimization. We found that most of the running time of our approach is consumed in two major steps: Online GSDL and PHD update, both of which are dependent on the number of detections. Average runtime performance (Hz) comparisons with other approaches for the MOTChallenge Benchmark are listed in Tables 4 and 5, where the runtime of our method is approximately 8.2 and 6.7 Hz on MOT15 and MOT17 benchmark respectively. Hence, the proposed system is well-suited for online applications. Although our tracker runs slower than the methods in [17], [18], and [55] using same the PHD filter framework, it returns a significant improvement regarding the tracking accuracy.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel multiple human tracking system that incorporated an adaptive gating based particle PHD filter with online group-structured dictionary learning. We developed a novel adaptive-gating strategy: adaptively updating the gating threshold using human information to refine the measurement set of survival targets thereby strengthening the measurement-driven mechanism. To further improve measurements classification and birth intensity estimation, we firstly explored the properties of group-structured dictionary learning to improve the discriminative power of sparse coding, in this sense, proposing a maximum voting method to distinguish the birth measurements from noisy measurements. Additionally, the SimCO algorithm with the structure pattern was feasible to efficiently implement the dictionary update stage for further robustness. Experimental results were shown to demonstrate the effectiveness and stability of the proposed tracker compared to some state-of-the-art methods. Future work will introduce an occlusion reasoning method to further tackle the occlusions and missed detections.

REFERENCES

[1] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler. (2016). "MOT16: A benchmark for multi-object tracking." [Online]. Available: https://arxiv.org/abs/1603.00831
[2] P. Feng, W. Wang, S. Dlay, S. M. Naqvi, and J. Chambers, "Social force model-based MCMC-OCSVM particle PHD filter for multiple human tracking," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 725–739, Apr. 2017.
[3] A. Ur-Rehman, S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-target tracking and occlusion handling with learned variational Bayesian clusters and a social force model," *IEEE Trans. Signal Process.*, vol. 64, no. 5, pp. 1320–1335, Mar. 2016.
[4] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 864–880, Apr. 2014.
[5] E. Maggio and A. Cavallaro, *Video Tracking: Theory and Practice*. Hoboken, NJ, USA: Wiley, 2011.
[6] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1201–1208.
[7] A. Milan, K. Schindler, and S. Roth, "Multi-target tracking by discrete-continuous energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2054–2068, Oct. 2016.
[8] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4705–4713.
[9] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3029–3037.
[10] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, "Multi-class multi-object tracking using changing point detection," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2016, pp. 68–83.
[11] J. H. Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon, "Online multi-object tracking via structural constraint event aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1392–1400.
[12] J. Son, M. Baek, M. Cho, and B. Han, "Multi-object tracking with quadruplet convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5620–5629.
[13] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, and B. Schiele. (2016). "A multi-cut formulation for joint segmentation and tracking of multiple objects." [Online]. Available: https://arxiv.org/abs/1607.06317
[14] M. Yang, Y. Wu, and Y. Jia, "A hybrid data association framework for robust online multi-object tracking," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5667–5679, Dec. 2017.
[15] M. Yang and Y. Jia, "Temporal dynamic appearance modeling for online multi-person tracking," *Comput. Vis. Image Understand.*, vol. 153, pp. 16–28, Dec. 2016.
[16] J. H. Yoon, M. H. Yang, J. Lim, and K. J. Yoon, "Bayesian multi-object tracking using motion context from multiple objects," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2015, pp. 33–40.
[17] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2016, pp. 84–99.
[18] Y. Song and M. Jeon, "Online multiple object tracking with the hierarchically adopted GM-PHD filter using motion and appearance," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2016, pp. 1–4.
[19] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.
[20] B. N. Vo and W. K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.
[21] B.-N. Vo, S. Singh, and D. Arnaud, "Sequential Monte Carlo methods for multitarget filtering with random finite sets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 4, pp. 1224–1245, Oct. 2005.
[22] X. Zhou, Y. Li, B. He, and T. Bai, "GM-PHD-based multi-target visual tracking using entropy distribution and game theory," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1064–1076, May 2014.
[23] P. Feng, W. Wang, S. M. Naqvi, and J. Chambers, "Adaptive retrodiction particle PHD filter for multiple human tracking," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1592–1596, Nov. 2016.
[24] J. Wu *et al.*, "Iterative RANSAC based adaptive birth intensity estimation in GM-PHD filter for multi-target tracking," *Signal Process.*, vol. 131, pp. 412–421, Feb. 2017.

[25] E. Maggio, M. Taj, and A. Cavallaro, "Efficient multitarget visual tracking using random finite sets," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1016–1027, Aug. 2008.

[26] R. Hoseinnezhad, B.-N. Vo, and B.-T. Vo, "Visual tracking in background subtracted image sequences via multi-Bernoulli filtering," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 392–397, Jan. 2013.

[27] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, Sep. 2011.

[28] W. Dai, T. Xu, and W. Wang, "Simultaneous codeword optimization (SimCO) for dictionary update and learning," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6340–6353, Dec. 2012.

[29] Z. Fu, P. Feng, S. M. Naqvi, and J. A. Chambers, "Particle PHD filter based multi-target tracking using discriminative group-structured dictionary learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4376–4380.

[30] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.

[31] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun./Jul. 2016, pp. 418–425.

[32] B. Ristic, D. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for PHD and CPHD filters," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 2, pp. 1656–1668, Apr. 2012.

[33] Y. D. Wang, J. K. Wu, A. A. Kassim, and W. Huang, "Data-driven probability hypothesis density filter for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1085–1095, Aug. 2008.

[34] Y. Zheng, Z. Shi, R. Lu, S. Hong, and X. Shen, "An efficient data-driven particle PHD filter for multitarget tracking," *IEEE Trans. Ind. Informat.*, vol. 9, no. 4, pp. 2318–2326, Nov. 2013.

[35] W. Si, L. Wang, and Z. Qu, "A measurement-driven adaptive probability hypothesis density filter for multitarget tracking," *Chin. J. Aeronautics*, vol. 28, no. 6, pp. 1689–1698, 2015.

[36] Y. Suo, M. Dao, U. Srinivas, V. Monga, and T. Tran. (2014). "Structured dictionary learning for classification." [Online]. Available: https://arxiv.org/abs/1406.1943

[37] Y. Xu, Y. Sun, Y. Quan, and B. Zheng, "Discriminative structured dictionary learning with hierarchical group sparsity," *Comput. Vis. Image Understand.*, vol. 136, pp. 59–68, Jul. 2015.

[38] T. Zhang *et al.*, "Structural sparse tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 150–158.

[39] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 367–383, Jan. 2013.

[40] B. Liu, J. Huang, C. Kulikowski, and L. Yang, "Robust visual tracking using local sparse appearance model and K-selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2968–2981, Dec. 2013.

[41] Y. Xie, W. Zhang, C. Li, S. Lin, Y. Qu, and Y. Zhang, "Discriminative object tracking via sparse representation and online dictionary learning," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 539–553, Apr. 2014.

[42] B. Ma, J. Shen, Y. Liu, H. Hu, L. Shao, and X. Li, "Visual tracking using strong classifier and structural local sparse descriptors," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1818–1828, Oct. 2015.

[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1. no. 1, pp. 886–893.

[44] W. Z. Lu, C. Bai, K. Kpalma, and J. Ronsin, "Multi-object tracking using sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 2312–2316.

[45] R. Fisher. (2003). *Caviar Test Case Scenarios*. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[46] I. Goldberg and M. J. Atallah. (2009). *Privacy Enhancing Technologies*. [Online]. Available: http://ftp.pets.rdg.ac.uk/pub/PETS2009/Crowd_PETS09_dataset/a_data/a.html

[47] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[48] L. Leal-Taixé, A. Milan, I. D. Reid, S. Roth, and K. Schindler. (Apr. 2015). "MOTChallenge 2015: Towards a benchmark for multi-target tracking." [Online]. Available: http://arxiv.org/abs/1504.01942

[49] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.

[50] N. Chenouard, I. Bloch, and J. Olivo-Marin, "Multiple hypothesis tracking for cluttered biological image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2736–3750, Nov. 2013.

[51] Z. Fu, P. Feng, S. M. Naqvi, and J. A. Chambers, "Robust particle PHD filter with sparse representation for multi-target tracking," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Oct. 2016, pp. 281–285.

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

[53] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2953–2960.

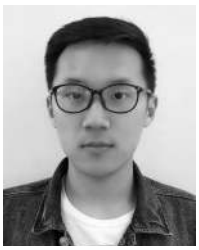
[54] H. Kieritz, S. Becker, W. Hübner, and M. Arens, "Online multi-person tracking using integral channel features," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 122–130.

[55] V. Eiselein, D. Arp, M. Pätzold, and T. Sikora, "Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors," in *Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 325–330.

[56] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora, "Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data," in *Proc. IEEE Int. Workshop Traffic Street Surveill. Safety Secur. (AVSS)*, Aug. 2017, pp. 1–5.

[57] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, "Enhancing detection model for multiple hypothesis tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2143–2152.

[58] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. IEEE Int. Workshop Traffic Street Surveill. Safety Secur. (AVSS)*, Sep. 2017, pp. 1–6.



ZEYU FU (S'16) received the B.Eng. degree (Hons.) in electrical and electronic engineering from Newcastle University, Newcastle Upon Tyne, U.K., in 2015, where he is currently pursuing the Ph.D. degree with the Intelligent Sensing and Communications Research Group, School of Engineering.

He is with the University Defence Research Collaboration sponsored by the U.K. Defence Science and Technology Laboratory and the Engineering and Physical Science Research Council, on the project Signal Processing in Networked Battlespace. His research interests include video-based multi-target tracking, enhanced RFS filtering, dictionary learning, and deep learning.

PENGMING FENG (S'13–M'17) was born in Jilin, China. He received the B.Sc. degree in automatic control from the Beijing University of Chemical Technology, Beijing, China, in 2012, the M.Sc. degree in digital communication systems from Loughborough University, Loughborough, U.K., in 2013, and the Ph.D. degree in intelligent signal processing from Newcastle University, Newcastle Upon Tyne, U.K., in 2016.

During his Ph.D. degree, he was with the University Defence Research Collaboration sponsored by the U.K. Defence Science and Technology Laboratory and the Engineering and Physical Science Research Council, on the project Signal Processing in Networked Battlespace.

He joined China Aerospace Science and Technology Corporation in 2017, as a Doctoral Engineer with the State Key Laboratory of Space-Ground Integrated Information Technology. His research interests include remote sensing, multiple target tracking, machine learning, and sparse representation.

FEDERICO ANGELINI (S'17) received the master's degree in pure and applied mathematics from the University of Rome Tor Vergata, Italy, in 2015. He is currently pursuing the Ph.D. degree with the Intelligent Sensing and Communications Research Group, School of Engineering, Newcastle University, U.K. He collaborated with the National Research Council, Rome, Italy, on the Image Denoising Methods Project from 2015 to 2016.

He is working as part of University Defence Research Collaboration sponsored by the U.K. Defence Science and Technology Laboratory and the Engineering and Physical Science Research Council, U.K., on the Project Multimodal Wide Area Surveillance, co-founded by Newcastle University and the iCase Award, EPSRC, U.K. His research interests include human action and behavior recognition via multimodal sensors, machine learning, and signal processing.

JONATHON CHAMBERS (S'83–M'90–SM'98–F'11) received the Ph.D. and D.Sc. degrees in signal processing from Imperial College London, London, U.K., in 1990 and 2014, respectively. From 1991 to 1994, he was a Research Scientist with the Schlumberger Cambridge Research Center, Cambridge, U.K. In 1994, he returned to Imperial College London as a Lecturer in signal processing and was promoted to a Reader (Associate Professor) in 1998. From 2001 to 2004, he was the Director of the Center for Digital Signal Processing and a Professor of signal processing with the Division of Engineering, King's College London, where he is currently a Visiting Professor. From 2004 to 2007, he was a Cardiff Professorial Research Fellow with the School of Engineering, Cardiff University, Cardiff, U.K. From 2007 to 2014, he led the Advanced Signal Processing Group, School of Electronic, Electrical and Systems Engineering, Loughborough University, where he is also a Visiting Professor. In 2015, he joined the School of Electrical and Electronic Engineering, and he has been with the School of Engineering, Newcastle University, Newcastle upon Tyne, U.K., since 2017. Since 2017, he has also been a Professor in engineering and the Head of Department, University of Leicester, Leicester, U.K. He is also the International Honorary Dean and a Guest Professor with the Department of Automation, Harbin Engineering University, Harbin, China. He has advised almost 80 researchers through to Ph.D. graduation and has published over 500 conference proceedings and journal articles, many of which are in IEEE journals. His research interests include adaptive signal processing and machine learning and their application in communications, defence, and navigation systems.

Dr. Chambers is a fellow of the Royal Academy of Engineering, U.K., the Institution of Engineering and Technology, and the Institute of Mathematics and its Applications. He was a Technical Program Co-Chair of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. He is serving on the Organizing Committees of ICASSP 2019, Brighton, U.K., and ICASSP 2022, Singapore. He has served on the IEEE Signal Processing Theory and Methods Technical Committee for six years, the IEEE Signal Processing Society Awards Board for three years, and the Jack Kilby Medal Committee for three year. He was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING for three terms over the periods 1997–1999, 2004–2007, and a Senior Area Editor from 2011 to 2015.

SYED MOHSEN NAQVI (S'07–M'10–SM'15) received the Ph.D. degree in signal processing from Loughborough University, Loughborough, U.K., in 2009. He was a Post-Doctoral Research Associate with the EPSRC U.K.-funded projects and an REF Lecturer from 2009 to 2015. Prior to his postgraduate studies in Cardiff and Loughborough Universities, U.K., he served the National Engineering and Scientific Commission, Islamabad, Pakistan, from 2002 to 2005. He is currently a Lecturer of signal and information processing with the School of Engineering, Newcastle University, Newcastle upon Tyne, U.K. He has over 90 publications with the main focus of his research being on multimodal (audio video) signal and information processing. He has successfully (co)-supervised 25 Ph.D. and M.Sc. students and is (co)-supervising eight Ph.D. students. He organized special sessions on multi-target tracking in FUSION 2013 and 2014, delivered seminars, and was a speaker at UDRC Summer Schools, from 2015 to 2017. His research interests include multimodal processing for human behavior analysis, multi-target tracking, and source separation, all for machine learning.

Dr. Naqvi is a fellow of the Higher Education Academy U.K., a member of the University Defence Research Collaboration in Signal Processing, and a member of the International Society of Information Fusion.