Novel Data Association Methods for Online Multiple Human Tracking

by

Zeyu Fu

A doctoral thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD), from Newcastle University.

July 2019

Intelligent Sensing and Communications (ISC) Research Group,
School of Engineering,
Newcastle University,
Newcastle upon Tyne, UK, NE1 7RU.

CERTIFICATE OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgements or in footnotes, and that neither the thesis nor the original work contained herein has been submitted to this or any other institution for a degree.

..... (Signed)

..... (candidate)

I dedicate this thesis to my loving family.

Abstract

Video-based multiple human tracking has played a crucial role in many applications such as intelligent video surveillance, human behavior analysis, and health-care systems. The detection based tracking framework has become the dominant paradigm in this research field, and the major task is to accurately perform the data association between detections across the frames. However, online multiple human tracking, which merely relies on the detections given up to the present time for the data association, becomes more challenging with noisy detections, missed detections, and occlusions. To address these challenging problems, there are three novel data association methods for online multiple human tracking are presented in this thesis, which are online group-structured dictionary learning, enhanced detection reliability and multi-level cooperative fusion.

The first proposed method aims to address the noisy detections and occlusions. In this method, sequential Monte Carlo probability hypothesis density (SMC-PHD) filtering is the core element for accomplishing the tracking task, where the measurements are produced by the detection based tracking framework. To enhance the measurement model, a novel adaptive gating strategy is developed to aid the classification of measurements. In addition, online group-structured dictionary learning with a maximum voting method is proposed to estimate robustly the target birth intensity. It enables the new-born targets in the tracking process to be accurately initialized from noisy sensor measurements. To improve the adaptability of the

group-structured dictionary to target appearance changes, the simultaneous codeword optimization (SimCO) algorithm is employed for the dictionary update.

The second proposed method relates to accurate measurement selection of detections, which is further to refine the noisy detections prior to the tracking pipeline. In order to achieve more reliable measurements in the Gaussian mixture (GM)-PHD filtering process, a global-to-local enhanced confidence rescoring strategy is proposed by exploiting the classification power of a mask region-convolutional neural network (R-CNN). Then, an improved pruning algorithm namely soft-aggregated non-maximal suppression (Soft-ANMS) is devised to further enhance the selection step. In addition, to avoid the misuse of ambiguous measurements in the tracking process, person re-identification (ReID) features driven by convolutional neural networks (CNNs) are integrated to model the target appearances.

The third proposed method focuses on addressing the issues of missed detections and occlusions. This method integrates two human detectors with different characteristics (full-body and body-parts) in the GM-PHD filter, and investigates their complementary benefits for tracking multiple targets. For each detector domain, a novel discriminative correlation matching (DCM) model for integration in the feature-level fusion is proposed, and together with spatio-temporal information is used to reduce the ambiguous identity associations in the GM-PHD filter. Moreover, a robust fusion center is proposed within the decision-level fusion to mitigate the sensitivity of missed detections in the fusion process, thereby improving the fusion performance and tracking consistency.

The effectiveness of these proposed methods are investigated using the MOTChallenge benchmark, which is a framework for the standardized evaluation of multiple object tracking methods. Detailed evaluations on challenging video datasets, as well as comparisons with recent state-of-the-art

techniques, confirm the improved multiple human tracking performance.

Contents

Statement of Originality

The contributions of this thesis are mainly to improve measurement driven filtering for online multiple human tracking. The following international journal and conference papers verify the novelty of the contributions.

In Chapter 3, a novel adaptive gating method is proposed to achieve better measurement classification for the filtering process. Then an online group-structured dictionary learning strategy is developed for robust target birth intensity estimation. Additionally, the SimCO algorithm is exploited for the dictionary update, which is devoted to efficiently dealing with the target appearance changes. These research outputs have been published in:

1. 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 International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4376-4380, 2017.

2. Z. Fu, P. Feng, F. Angelini, J. A. Chambers, and S. M. Naqvi, 'Particle PHD Filter based Multiple Human Tracking using Online Group-Structured Dictionary Learning', *IEEE Access*, vol. 6, pp. 14764-14778, 2018.

In Chapter 4, a novel two-stage measurement selection technique based on enhanced detection reliability is proposed in the tracking pipeline. In the first stage, a global-to-local enhanced confidence rescoring strategy is proposed to mitigate the misalignment between the given confidence scores and true human detections. In the second stage, a Soft-ANMS algorithm is developed to improve the robustness in the suppression process. Moreover, CNN features based on the person re-identification are employed to build target appearance models for improving the target association. The outputs of these three solutions are presented in:

3. Z. Fu, S. M. Naqvi, and J. A. Chambers, 'Enhanced GM-PHD Filter Using CNN-Based Weight Penalization for Multi-Target Tracking', in *Proc. Sensor Signal Processing for Defence Conference* (*SSPD*), pp. 1-5, 2017.

4. Z. Fu, Xin Lai, and S. M. Naqvi, 'Enhanced Detection Reliability for Human Tracking Based Video Analytics', in *Proc. International Conference on Information Fusion* (*FUSION*), pp. 1-7, 2019.

In Chapter 5, novel multi-level (feature-level and decision-level) cooperative fusion within the measurement driven GM-PHD filter framework is presented for online multiple human tracking. For the feature-level fusion, a novel DCM model is proposed and fused with spatio-temporal information to mitigate the ambiguities in the identity associations. For the decision-level fusion, a robust fusion center with virtual and real zones is proposed to improve the fusion process and tracking consistency. The contributions of this tracking method appear in:

5. Z. Fu, F. Angelini, S. M. Naqvi, and J. A. Chambers, 'GM-PHD Filter Based Online Multiple Human Tracking Using Deep Discriminative Correlation Matching', in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4299-4303, 2018.

6. Z. Fu, S. M. Naqvi, and J. A. Chambers, 'Collaborative Detector Fusion of Data-Driven PHD Filter for Online Multiple Human Tracking', in *Proc. International Conference on Information Fusion* (*FUSION*), pp. 1976-1981, 2018.

7. Z. Fu, F. Angelini, J. A. Chambers, and S. M. Naqvi, 'Multi-Level Cooperative Fusion of GM-PHD Filters for Online Multiple Human Tracking', *IEEE Transactions on Multimedia*, 2019. In Press.

Acknowledgements

First and foremost I would like to express the deepest gratitude to my supervisors, Dr Mohsen Naqvi and Professor Jonathon Chambers for their constant support of my PhD study, for their patience, motivation and kind advice throughout the past four years. I have benefited immensely from their useful guidance and exceptional knowledge for all the related research and writing up this thesis. It would not have ever been possible to accomplish this thesis without their continuous encouragement. It has been my great honour to be one of their research students. I do wish that we could work together again in the future.

I am grateful to acknowledge the fee contributions from the School of Engineering, Newcastle University, as well as my parental financial support, which made my PhD research possible.

I would also like to thank my friends and colleagues Pengming Feng, Yang Sun, Jiachen Yin, Federico Angelini, Yang Xian, Jiawei Yan, and Yuxing Yang who have contributed enormously and provided great help to my PhD life professionally and personally during the past four years in the United Kingdom.

Lastly, but most importantly, I would like to thank my family, in particular my loving parents for raising me up with much love and supporting me all the time, as well as my loving wife for giving me faithful love, support and encouragement throughout my PhD life.

Zeyu Fu

July, 2019

List of Acronyms

CNN	Convolutional Neural Network
DCF	Discriminative Correlation Filter
DCM	Discriminative Correlation Matching
DPM	Deformable Part-based Model
EDR	Enhanced Detection Reliability
EKF	Extended Kalman Filter
EMD	Exponential Mixture Densities
FAF	False Alarms per Frame
FFT	Fast Fourier Transform
FISST	Finite Set Statistics
FN	False Negative
FOV	Field of View
FP	False Positive
GCI	Generalized Covariance Intersection
GM	Gaussian Mixture
GSDL	Group-Structured Dictionary Learning

GT	Ground Truth
HOG	Histogram of Oriented Gradient
IDS	Identity Switches
IOA	Intersection Over Area
IOU	Intersection Over Union
I-RANSAC	Iterative Random Sample Consensus
JPDAF	Joint Probabilistic Data Association Filter
KCF	Kernelized Correlation Filter
LSTM	Long-Short Term Memory
MCMC	Markov Chain Monte Carlo
MHT	Multiple Hypothesis Tracking
ML	Mostly Lost
MOTA	Multiple Object Tracking Accuracy
MOTP	Multiple Object Tracking Precision
MSE	Mean Square Error
MT	Mostly Tracked
NMS	Non-Maximal Suppression
OCSVM	One Class Support Vector Machine
OSPA	Optimal Sub-Pattern Assignment
PDF	Probability Density Function
PSR	Peak to Sidelobe Ratio

PHD	Probability Hypothesis Density
R-CNN	Region-Convolutional Neural Network
RFS	Random Finite Set
RNN	Recurrent Neural Network
ROI	Region of Interest
RPN	Region Proposal Network
SFM	Social Force Model
SMC	Sequential Monte Carlo

List of Symbols

$\odot$	Hadamard (element-wise) product
$(\cdot)^{\dagger}$	Complex conjugation
$\mathscr{F}^{-1}$	Inverse fast Fourier transform
Λ	Association cost
$\mathbf{\Delta}$	Cost matrix
e	Target survival probability
f	State transition function
$\mathbf{F}$	State transition matrix
$\mathcal{F}$	Finite subsets
g	Target likelihood function
$\mathbf{h}$	Measurement transition function
$\mathbf{H}$	Observation matrix
$\mathbf{x}$	State vector
$\mathbf{z}$	Measurement vector
$\|\|.\|\|$	Euclidean distance
$(.)^T$	Transpose operator

T	The threshold of validation gate
p	Probability distribution
ν	intensity function
$\sum$	Summation
$\int$	Integral
$\mathcal{N}$	Gaussian distribution
min	Minimum value
$argmin$	Argument of the minimum
max	Maximum value
$\|\cdot\|_1$	$l1$ norm
$\|\cdot\|_2$	$l2$ norm
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|$	Euclidean distance
$\setminus$	Set subtraction
$\emptyset$	Empty set
σ	Standard deviation
$\|\cdot\|$	Cardinality of a set
Υ	Intensity function of new-born target birth RFS
Γ	Gamma function
$\in$	Belonging to
$\subseteq$	Subset

∇	Gradient
$\bigcap$	Intersection
$\bigcup$	Union
$\approx$	Approximately equal
det	Determinant of a matrix
$\triangleq$	Determinant of a matrix
∞	Infinity
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
N	Number of measurements
M	Number of targets
J	Number of new-born targets
exp	Exponential
φ	Fusing parameter

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Motivation

With the rapid advancement in industrial technologies, surveillance cameras have become ubiquitous [7]. The importance of video surveillance continues to rise in contemporary society, as the utility of video data has been rapidly evolving through the development of autonomous video analytics tools which are able to analyze the visual world. In fact, video analytics plays a crucial role in the development of intelligent video surveillance systems, the ultimate aim of which is to automatically perceive and understand visual content from video camera recordings [5].

Detecting and tracking multiple humans [6,8–13] which represents a fundamental and important processing step has enormously advanced many video analytics applications, particularly in crime or terrorism investigation, outdoor environment surveillance, assisted living, indoor environment surveillance, homeland security and sports analysis [7], [14]. Example applications of multiple human tracking are represented in Fig. 1.1. Of particular interest in multiple human tracking is to facilitate the autonomous system to better understand human-involved events occurring in a surveillance region [7]. During the last two decades, many researchers have been seeking reliable multiple human trackers which enable the surveillance system to acquire a richer human description from the scene, thereby facilitating holistic scene interpretation. Major tasks of multiple human tracking are to localize

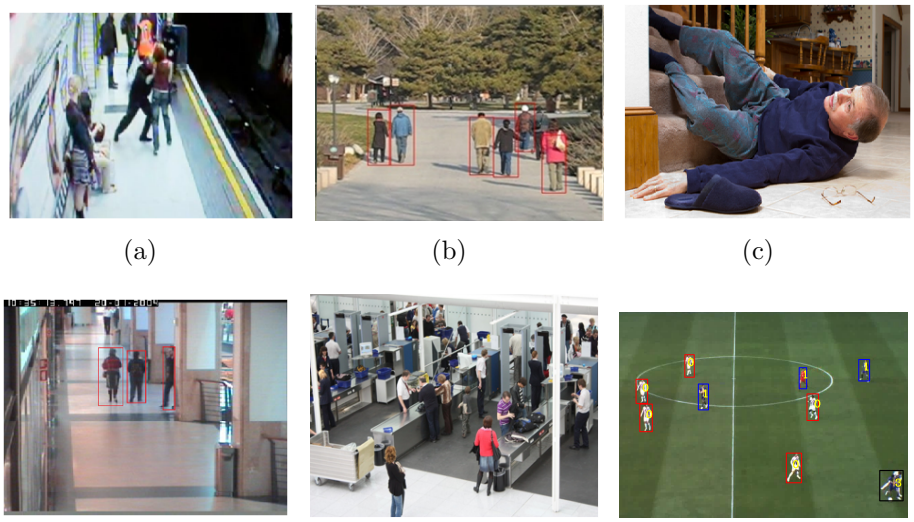


Figure 1.1. Examples of multiple human tracking application. Common racking scenarios include (a) crime or terrorism investigation, (b) outdoor environment surveillance, (c) assisted living, (d) indoor environment surveillance, (e) homeland security, and (f) sports analysis.

accurately varying number of targets, retrieve their trajectories, as well as maintaining their identities in a given video sequence [6]. However, there still exist many challenging problems needed to be further tackled. These issues are particularly caused by complicated environments such as the varying number of targets, noisy detections or clutter, missed detections and occlusions.

To deal with these challenging issues, there are several approaches that have been established to execute this tracking task. 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) [8], [15]. With the rapid advancements of human detection [2, 12, 16], recent video-based multiple human tracking methods have benefited significantly from the tracking-by-detection framework in either online or offline processing. In this framework, it necessitates detections that are accurately associated and linked

across frames, so as to achieve a higher tracking performance. Off-line trackers [17–22] process the video sequences using both past and future detection responses, but such non-causal systems are difficult to apply in time critical video surveillance. This thesis focuses on online multiple human tracking, by only relying on the detections given up to the present time.

Recursive Bayesian filtering based on Bayes theory has proved to be effective and appropriate mathematical tools to efficiently perform the inference task in different tracking domains, e.g. radar tracking and sensor fusion [23]. This framework performs dynamic state estimation by combining target motion and observation, which is appropriate for the focused online tracking methods. Established Bayesian tracking schemes such as Kalman filtering [24] and particle filtering [25] have been adapted in various studies so that they are better usable in video processing applications. However, most of these approaches notably focused on single target tracking; when the tracking scenario becomes complicated especially with increasing number of targets, they may not be able to demonstrate promising performance.

More recent multi-target Bayesian filtering methods are developed based on the probability hypothesis density (PHD) filters which are built upon random finite set (RFS) theory [26]. The key advantage of this approach is that it can naturally deal with a varying number of targets, which allows it to be well integrated with the detection based tracking framework. This approach is also suitable for providing the tracking estimates in both cardinality and localization with relatively low computational cost [9]. Due to the aforementioned merits, the PHD filtering technique has been successfully developed from its original radar/sonar tracking domain to be applicable in vision tasks [1, 9, 27–31]. Typically, Gaussian mixture (GM)-PHD and sequential Monte Carlo (SMC)-PHD filters are two commonly used resulting implementations from this theory.

However, applying conventional PHD filters directly with the tracking-

by-detection framework is inadequate to achieve the desired tracking performance, since there are several key points within this approach which need particular attention to expand its capability in video tracking applications. Firstly, the entire tracking process is closely dependent on the quality of the measurements, so it is necessary to perform initial preprocessing on the obtained measurements e.g. detection analysis, so the PHD updating step can be better realized without being contaminated by clutter or noise. The tracker must also be able to decide the number of targets at each time frame as human targets may appear and disappear from the scene, which is important for target initialization and termination. In the context of video tracking, visual similarities between frames should be well exploited for target matching, which could be also useful to handle the complex occlusion issues. Moreover, the tracker should include a track management mechanism to deduce preliminary tracking estimates of the pixel-level position of each human target, and therefore ultimately output the bounding boxes with unique identities around tracked targets at each time frame.

1.2 Aims and objectives

This thesis aims to exploit fully the measurements obtained from the human detection, by different machine learning or fusion techniques to drive the PHD filtering together with the detection based tracking framework for online multiple human tracking. The main goal is to address the aforementioned issues existing in multiple human tracking by improving several primary components in the tracking pipeline. The particular objectives are:

- Objective 1: Strengthening the measurement model in the particle PHD filter by classifying the raw measurements achieved from the human detector.

In Chapter 3, an adaptive gating step which exploits the spatio-temporal

and human size information is developed to distinguish the measurements originated from survival and new-born targets.

- Objective 2: Elevating the birth intensity estimation by applying online group-structured dictionary learning.

In Chapter 3, target appearance models constructed by group-structured dictionary learning are used to retrieve more accurately the new-born targets. The SimCO algorithm is utilized to update the dictionary with the aim of handling the appearance variations.

- Objective 3: Enhancing detection reliability prior to the tracking model by applying a two-stage measurement selection of detections.

In Chapter 4, the establishment of enhanced detection reliability to aid the earlier tracking process is divided into two stages: rescoring the detection confidence by taking advantage of the classification power of the Mask R-CNN and further to suppress false alarms by using the Soft-ANMS algorithm.

- Objective 4: Developing novel target appearance modeling by using CNN features based on the person re-identification.

In Chapter 4, instead of utilizing hand-crafted features, CNN features driven by the person re-identification are used to establish the target appearance models with the aim of mitigating the target ambiguity during the measurement grouping process.

- Objective 5: Investigating the ambiguous identity associations in the GM-PHD filtering process by performing the feature-level fusion.

In Chapter 5, target appearance models are further enhanced via the developed DCM model, which are jointly used with spatial-temporal constraints to penalize the ambiguous target associations.

- Objective 6: Addressing missed detections by collaboratively fusing two human detectors in a designed fusion center.

Missed detections can easily affect the performance of multiple human tracking, therefore, in Chapter 5, a robust fusion center which resorts to the complementary strengths of two human detectors is developed at the decision level to mitigate the issue of missed detection.

To sum up, this thesis systematically contributes to the multiple human tracking in the following aspects: enhancing the acquisition of newborn targets via online group-structured dictionary learning, improving the limitations of measurement model by establishing the enhanced detection reliability, achieving better target association by developing robust target appearance models and dealing with the missed detection by proposing a multi-level fusion approach.

1.3 Thesis outline

The remainder of this thesis is structured as follows:

Chapter 2, in general, gives a relevant literature review of multiple human tracking, and also explains the background preliminaries that are helpful to derive and evaluate the proposed techniques in the thesis. The important relevance of previous works to this thesis is firstly described in three main aspects. Human detection as an essential step in the detection based tracking framework is discussed. Different challenges associated with multiple human tracking are described, including various solutions to address them. Moreover, related developed multiple human tracking methods are categorized by carrying through the primary components in the tracking system, where the application of different categories as well as their advantages and disadvantages are given. Then, the problem formulation of multiple human tracking is presented by a probabilistic Bayesian filtering framework, with

a particular interest on the RFS based PHD filtering. Benchmark datasets and performance measures are also given with details for the tracking evaluations.

Major technical contributions are divided into the following three chapters. Chapter 3 pertains to the first and second objectives, and provides the contributions to improving the measurement model and birth intensity estimation in the measurement-driven filtering framework via the exploitation of online group-structured dictionary learning. This chapter is mostly based on the published works in [10] and [28]. Chapter 4 presents an enhanced detection reliability (EDR) module signifying a stand-alone algorithm to improve the measurement selection of detections, which possesses an essential step in the early stage of the entire tracking framework. This chapter also deals with mitigating the target ambiguity by leveraging CNN features based upon target visual similarity measures. The majority of the technical parts in this chapter were previously published in [31] and [32]. Chapter 5 continues with alleviating ambiguous target associations in the tracking framework by introducing a novel DCM model and performing feature fusions. Furthermore, this chapter provides a collaborative detector fusion perspective to address the missed detections by compensating for the missing information from either detector domain. Parts of this approach were presented in [33]. In the final chapter, contributions are drawn and suggestions for future work are given.

Chapter 2

RELEVANT LITERATURE REVIEW AND PRELIMINARIES

2.1 Introduction

As introduced before, multiple human tracking plays a fundamental role in many video analytics applications, which requires the tracking performance to be more accurate and robust in complex environments. To this end, many researchers have been seeking relevant solutions to improve tracking performance in the last decade. Thus, it may not be feasible to conduct a complete review within this research topic, instead the goal of this chapter is to summarize recent progress regarding multiple human tracking associated with essential techniques that are mostly close to the proposed solutions in this thesis.

In fact, it is not easy to categorize tracking methods with a universal criterion, which means that a particular tracking method must be classified into multiple categories in the tracking family [34, 35]. Therefore, this chapter starts with introducing the dominant detection based tracking framework, which generally categorizes most tracking methods into online and offline depending upon the processing mode. Then more discussions are made on the

recent tracking methods by following two major aspects: current tracking challenges and primary components of a tracking system. The main challenges of multiple human tracking are firstly presented, followed by relevant established tracking methods for each of them. Then the key components included in modelling multiple human tracking are reviewed. This thesis focuses on recursive Bayesian filtering methods to realize the tracking inference in the detection based tracking framework. Therefore, the background preliminaries related to Bayesian tracking models are presented.

The rest of this chapter is presented as follows: Section 2.2 presents the recent development of human detection which advances multiple human tracking. Section 2.3 describes the main difficulties for designing a multiple human tracking system as well as relevant existing solutions. Section 2.4 reviews the established tracking methods based on recent advancements of primary components in the tracking process. Section 2.5 details the systematic problem formulation based on Bayesian filtering. Finally, evaluation sequences from three different datasets as well as the performance evaluation criteria employed in this thesis for making comparison between different tracking algorithms are presented in Sections 2.6 and 2.7.

2.2 Human detection

With the advancement of object detection, the detection based tracking framework as shown in Fig. 2.1, has become a commonly-used paradigm to track multiple humans in video [6, 36–39], since it has better ability to deal automatically with how the targets are initialized and terminated. As introduced before, this thesis also focuses on the detection based tracking framework, which acts as a pre-processing step independent from the tracking pipeline. Specifically, a pre-trained human detector is initially employed in each frame of an input video sequence to capture a set of candidate bound-

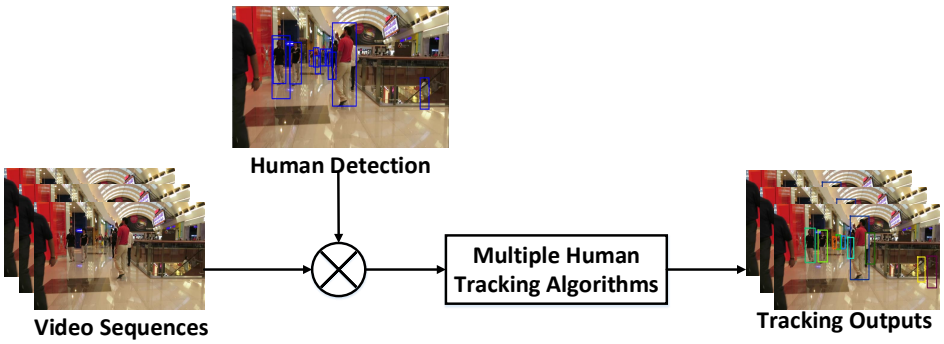


Figure 2.1. An overview of detection based tracking systems. A human detector is firstly performed on the given video sequence to achieve detection responses. Based on the obtained detections, tracking algorithms can be established to construct the target trajectories using either online or offline processing.

ing boxes.

Early tracking works [1], [28], [40] often employed background subtraction methods which mainly investigate the pixel-wise differences between the trained background model and foreground objects to obtain the regions of blobs. As can be seen from Fig. 2.2, obtained foreground blobs usually need a post-processing step to reduce the effect of small patches of noise as well as handling the merging targets, so that they can be applied as input to multiple human tracking. This kind of technique to capture the region of interest also requires the background in the scene to be less altered, which limits its application only to tracking scenarios with static cameras.

Recent developments on object detection particularly have allowed the detector to robustly find human localization in more dynamic environments with different camera motion and lighting conditions. The deformable part-based model (DPM) proposed by Felzenszwalb et al. [2] has become one of the commonly-used human detectors from the MOTChallenge benchmark [4], since this model well discovered the connection between different individual parts in the entire target bounding box, and therefore provides better detection accuracy. Fig. 2.3 shows a few exemplar detection results

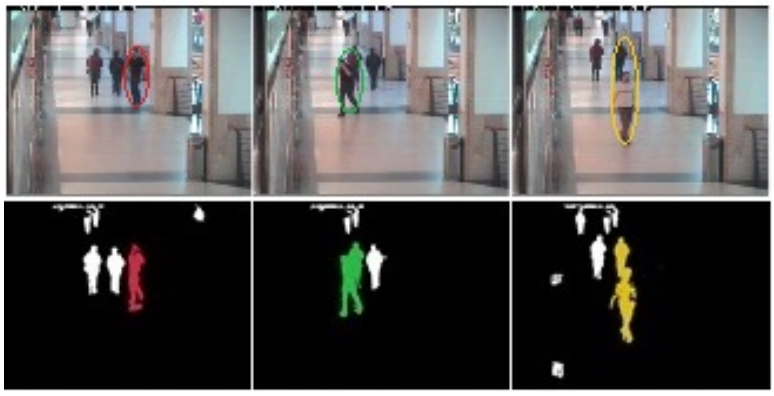


Figure 2.2. Exemplar results of human detections using background subtraction method [1]. Some failure cases are given in the second row, where small patches of noise falsely detected in all three results, and merged measurements appear in the latter two results.

using the DPM detector. Owing to the recent success in deep learning, a unified region-based convolutional neural network object detector namely Faster R-CNN [16] has been developed, due to its superior detection performance, this deep learning based detector has been adopted for human detections in the MOTChallenge benchmark [4]. In this approach, a region proposal network (RPN) which shares convolutional features with the detection network is introduced. This RPN can provide highly reliable region proposal, thus guiding the detection network to pay more attention on potential object regions for detections. After obtaining a set of candidate detections given a video sequence, then a tracking system begins with the collected measurements to perform the subsequent tracking process. Under the detection based tracking framework, there are essentially two processing modes for multiple human tracking: online and offline tracking. The major difference is whether measurements from future frames are used or not when performing the current tracking processing. This is quite similar to the notations of causal and non-causal systems within the signal processing theory. In practice, existing offline tracking approaches [36, 38, 41] employ both past and future detections to address the data association problem by

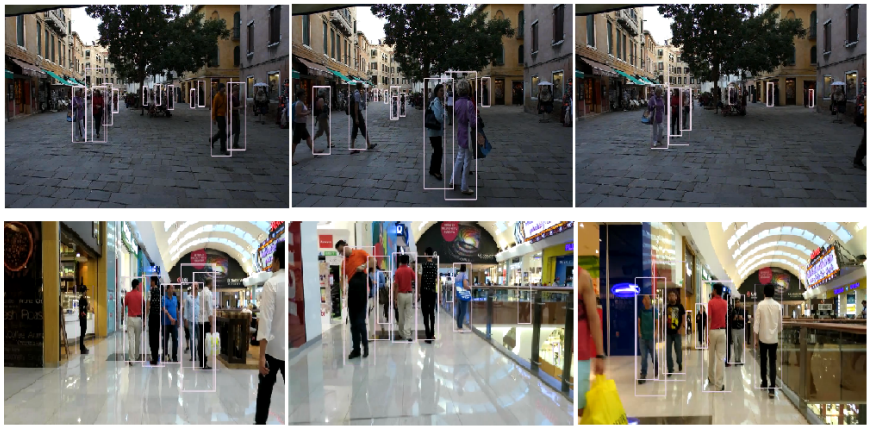


Figure 2.3. Example results of human detection using DPM detector [2] for multiple human tracking. There are some failure cases given by this method, such as false alarms which mostly appear in the first row of results, whereas missed detections which are clearly shown in the second row.

globally formulating an optimization problem. These offline trackers often perform more promisingly than online counterparts, however they may not be suitable for automated real-time surveillance systems. Online tracking approaches [6, 8, 10, 37, 39, 42] achieve the tracking estimates only relying on detections from past and current time.

Although this framework has become the dominant initializing step in the entire tracking pipeline, its imperfection may cause several critical issues in the subsequent tracking process. The following section will summarize and discuss the existing solutions for the challenging problems in multiple human tracking.

2.3 The challenges of multiple human tracking

Different from single target tracking, multiple human tracking becomes more challenging especially with complex scene conditions, including unknown number of targets, noisy detections or clutter, missed detections, and occlusions. Fig. 2.4 represents some example challenges within video-based

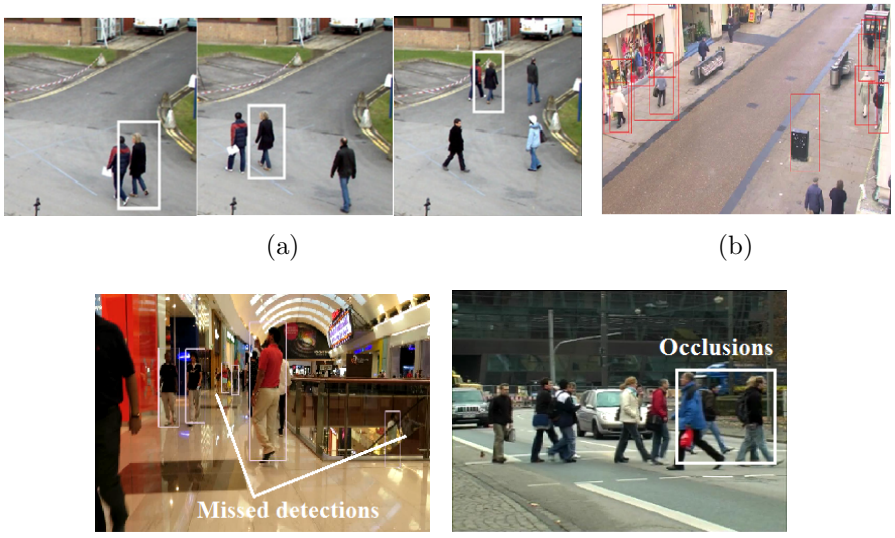


Figure 2.4. Visual illustration of challenging issues in multiple human tracking. (a) Varying number of targets: the number of targets varies between frames. (b) Noisy detections: objects other than humans as well as background noise are detected; (c) Missed detections: targets are missed by the detector. (d) Occlusions: partial or full occlusions often occur in the crowded environment. Image frames shown in this figure are obtained from the MOTChallenge benchmark [3] [4].

multiple human tracking work. More video demonstrations to explain the tracking challenges are publicly available in the website[1]. In the following subsections, numerous tracking methods with related solutions to these challenges are discussed.

2.3.1 Unknown number of targets

The first issue of multiple target tracking is how to cope with unknown number of targets, which is due to targets appearing and disappearing from the scene. To address this issue, traditional approaches have involved explicit association between measurements and targets in multiple human tracking such as multiple hypotheses tracking (MHT) [43] and the joint probabilistic

[1]http://www.intellsensing.com/research/information-processing/

data association filter (JPDAF) [15], but both of these suffer from being computationally expensive with increasing number of targets. The RFS based PHD filter [26] provides a promising alternative to handle variable dimension state and measurement vector by representing them as finite sets, as well as recursively propagating the first-order moment of the multiple target posterior [44], [45]. In addition to the PHD filter, the multi-Bernoulli filter [46] [47], as one of the resulting filters from RFS theory, was also proposed for multiple human tracking [48].

However, conventional PHD and multi-Bernoulli filtering methods [26], [44], [46] empirically preset the birth intensity to cover the entire region of interest. This requires prior knowledge of the scene information, and has limitations in real world scenarios. To avoid the necessity of prior knowledge in estimating birth intensity, Ristic et al. [49] built the adaptive target birth model based on new-born particles with high measurement likelihood. Likewise, Maggio et al. [50] attempted to model the birth intensity function within a limited volume around the observations to generate new-born particles from a mixture of Gaussians centered on the components of original measurements. However, these approaches are still unable to automatically deal with target birth and deaths that are highly random and unpredictable.

Detection based tracking has become the most commonly used initialization method in most tracking systems, since this paradigm can automatically deal with the targets appearing or disappearing from a practical perspective. Therefore, the PHD filtering based tracking methods driven by the detection based tracking framework can better handle the tracking scenario where the number of targets changes. In this thesis, the aim is to develop new measurement-driven filtering algorithmic solutions for multiple human tracking in video, where measurements are captured by the detection based tracking framework.

2.3.2 Noisy detections

In detection based tracking framework, a human detector is likely to be confused by other classes of objects or the background. This results in clutter or noisy detections which can deteriorate the tracking accuracy. Offline trackers [36, 38, 51] usually consider global data association by exploring future detection responses to remove noisy detections. As for online tracking methods, conventional Bayesian filtering approaches often used spatial and temporal constraints to filter the interferences in the measurement set. For example, existing gating techniques [52], [53] have been widely explored to select valid measurements and reduce the computation time in the update step by designing a validation region with spatial relation. Si et al. [54] pre-defined a confidence level to perform the gating strategy to generate adaptively the target birth intensity. However, the above traditional 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. Wu et al. [55] improved the measurement model by introducing 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. Zhou et al. [56] proposed an algorithm based on entropy distribution and coverage rate to eliminate irrelevant measurements e.g. background noises for accurate birth intensity estimation.

In addition to employing the spatial and temporal constraints, leveraging a classifier based on the target appearances would better discriminate desired targets from the background. Feng et al. [9] employed colour features

to learn a one-class support vector machine (OCSVM) to discriminate the valid targets and background clutter. Breitenstein et al. [57] used online-learned target specific classifiers to enhance the observation model. Furthermore, Chu et al. [58] devised a dynamic convolutional neural network (CNN) based model which combines spatial-temporal attention scheme with CNN features to prevent the target drifting to background. However, the increased computational cost of coupling appearance models may not be avoided.

Exploiting detection confidence scores has become an important step to acquire more reliable detections in some tracking systems, due to its less computational load. Tang et al. [38] incorporated the detection confidence into the pairwise affinity model for better clustering to tracks. Bae and Yoon [59] devised a tracklet confidence measure via detectability and continuity to improve tracklet association. However, the confidence score may not be always reliable, high-scored false detections can still exist, and thus worsen the tracking performance. Sanchez-Matilla et al. [29] exploited strong and weak detections in multiple detectors to improve the detection model. Chen et al. [39] made full advantage of deep neural networks to enhance the detection scores for candidate selection. Alternatively, contextual information from the scene structure has been also incorporated to promote the tracking problem. For instance, Chen et al. [51] exploited scene understanding and mutual detection correlation to improve the detection model, and thus avoid false detections.

2.3.3 Missed detections

Considering more complex or congested scenes, missed detections (false negatives) are often generated by a human detector with limited ability. This challenge could degrade the performance of the established methods, especially in PHD filtering based approaches which are prone to missed detections. According to the recent literature, offline trackers often attempt to

use target interpolation or stitching to fill the gap between two tracks, and then recover the missed detections via tracklet association [60] [38], however these approaches must gain access to future image frames. For online trackers, most of them relied on the target motion patterns from the past and present to predict target movement in the next frames when it has been missed from the detector, so linear [61] [57] and non-nonlinear motion models [62], [63] have been investigated. Recent learning based methods introduced recurrent neural networks (RNNs) [37], [64] for the non-linear motion prediction.

Although the above methods have been able to mitigate the problem of missed detections, these single detector based tracking approaches are limited and used to explore the image context more comprehensively. Multi-detector fusion provides an effective solution to reinforce the tracking process. To leverage the advantages of fusing multiple input sources, Ma et al. [65] proposed to combine multiple detectors with different modalities (RGB and depth) at detection level, and model the deformable spatial relationship to overcome the multiple human tracking problems. In [27], a multiple detector approach which is more suitable for computer vision task is presented by developing a novel likelihood model averaging a sum of individual likelihoods. Kutschbach et al. [66] extended the GM-PHD filter with kernelized correlation filters (KCFs) in a sequential fusion approach for refining target predictions. Khalid et al. [67] proposed to fuse a variety of trackers at decision level by hierarchically clustering the trackers based on spatio-temporal agreement to achieve final tracking estimates. Moreover, Henschel et al. [36] globally formulated the tracking task using multiple detectors as a weighted graph labeling problem, which is solved by Binary Quadratic optimization.

Generalized covariance intersection (GCI) [68] fusion has become a commonly-used approach to implement multi-detector fusion in the PHD filtering frame-

work. According to the systematic formulations in [69], Uney et al. [70] presented a distributed fusion of SMC-CPHD filters via exponential mixture densities (EMD). Then a further realization developed by Battistelli et al. [71] is to use GM implementation for GCI distributed fusion. For the labelled RFS filters, Wang et al. [72] has analyzed the *label space mismatching phenomenon*, and meanwhile proposed a robust fusing strategy which is to perform GCI fusion with the unlabelled versions of posteriors transformed from labelled counterparts. However, the aforementioned methods using the original GCI product rule are prone to missed detections [73]. To address this, a proposed multi-level cooperative fusion approach will be presented in Chapter 5.

2.3.4 Occlusions

Long or short term occlusion can be considered as a special type of missed detection for the occluded target, since it often occurs when the measurement of the occluded target is absent. This problem leads to less convincing tracking results with increased number of identity switches and missed detections. To mitigate this problem, many researchers have developed different solutions in the last decade. Intuitively, the use of multi-cameras [74] in video tracking giving multiple views of targets can provide better tracking visibility when targets are occluded in some point of view. However, fusion of different cameras requires extra processing time in communicating and exchanging information, and the camera calibration may be also considered.

In addition, methods with other modalities different from the camera sensor have also been developed for occlusion handling. For instance, depth information [75] [65] has been demonstrated to be helpful for separating the occluded targets during tracking, as it allows targets to be spatially visible. Audio aided tracking methods [11] [76] [77] usually employ audio data as an extra source to localize the speakers due to its independence to visual

occlusions. However, introducing these modalities limits the tracker only to be applicable for the tracking scenarios with audio recording or depth information, but this may not be suitable for real world tracking applications such as video surveillance.

Although the aforementioned approaches can benefit the tracking task, this thesis focuses on 2D multiple human tracking with a single RGB camera. To handle the occlusion problem in this research direction, appearance and interaction models have been investigated as two important cues to accurately identity or separate the targets within the occlusion region. Zhou et al. [56] proposed a game theory based algorithm with an improved spatial colour appearance model for mutual occlusion handling. Zhou et al. [78] explored the fusion of multiple features to penalize the ambiguous targets in the occlusion area, so as to improve the tracking accuracy. Interaction models such as social force model [79] aim to constrain the target motions by capturing the interactions and forces (repulsion or attraction) between different targets. For instance, Ata et al. [8] developed a variational Baeysian clustering combined with social force model to address complex inter-target occlusions. Feng et al. [9] proposed a novel exponential function based social model combined with Markov chain Monte Carlo (MCMC) resampling to improve the target prediction step when interactions occur. In contrast to model the interactions between targets, Lan et al. [41] exploited close and distant interactions between target tracklets to facilitate multiple object tracking. Following the recent success in deep learning, human interactions can be learned through a data-driven approach instead of hand-crafted functions. Amir et al. [37] utilized a long-short term memory (LSTM) network with occupancy grids to learn an interaction model. To address this, a proposed multi-level cooperative fusion approach will be presented in Chapter 5.

In this section, the main challenges as well as the existing solutions have

been reviewed. In the following, primary components for designing a multiple human tracking system will be introduced, and also recent developments regarding these components will be reviewed.

2.4 Primary components in multiple human tracking

The previous section has grouped existing multiple human tracking works based on different challenging problems. In this section, related tracking methods will be categorized based on the recent advancements of primary components in the tracking process. Fig. 2.5 illustrates the taxonomy of multiple human tracking algorithms based on two crucial parts: target representation and tracking inference. For each part, important milestones associated with different established techniques on multiple human tracking are discussed.

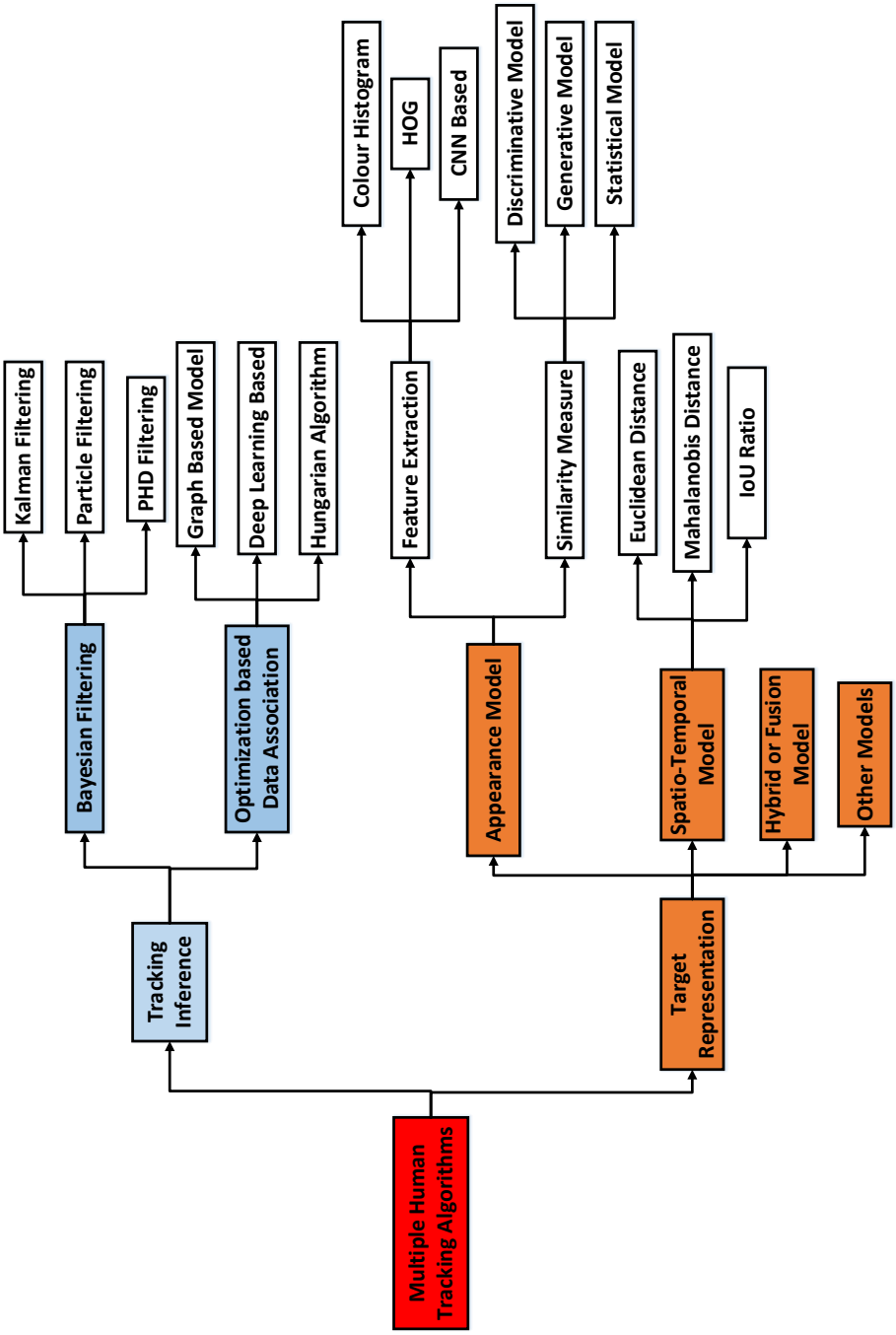


Figure 2.5. Taxonomy of multiple human tracking algorithms.

2.4.1 Target representation

Once the candidate detections are obtained from a human detector, the next important step is how to measure the target similarity between frames. The similarity measurement between targets is built upon target representation models which can describe the characteristics of targets, such as appearance, position, movement and gait in the human tracking task [34]. Ideally, the representation of targets can be modelled with unique features, which can distinguish individuals from the desired target, background and other targets in the same category. In fact, there are several features that have been applied to represent human targets in the tracking literature, such as, motion, colour, shape, gradient, depth, occupancy map, and optical flow [35]. These features can be used to construct different representation or measurement models.

In this section, recent developments of appearance model, spatio-temporal model, and hybrid model will be mainly discussed, since these are closely related to the proposed works in the thesis.

Appearance Model

In the context of video-based multiple human tracking, exploiting human appearances is crucial to compute the target similarity or affinity. For human beings, it is feasible to recognize the uniqueness of targets by grasping their appearances, but this is much more difficult for an ordinary camera to optimize this objective. The surveillance system needs to learn how to visually represent or describe what it sees from the targets. In essence, extracting descriptive features has become an importantly prior step to build up the appearance model.

In the last decade, many researchers have been seeking for better feature engineering techniques for the visual target representation. In general, feature extraction is performed on the region of interest e.g. a bounding box,

where various appearance-related features can be employed for the visual representation, such as colour histograms [80], histogram of oriented gradients (HOG) [81], and convolutional features [82]. The colour histograms [80] have been extensively investigated due to its robustness to shape and orientation variations, whereas neglecting the spatial information about the target region limits its ability. Gradient based features such as HOG [81], which are robust to geometric transformation and lighting changes have been also well studied and successfully used for different vision applications, but the accuracy can be affected by occlusions. Recently, with the advancements of deep learning [83], convolutional neural networks (CNNs) have been shown to outperform the hand-crafted feature extractors as introduced above in a wide range of computer vision tasks. CNN features with the aid of person re-identification have been prevalently used to encode the target appearances for single or multi-target tracking [39], [84], [85]. However, only relying on one type of feature extractor may be limited to fully characterize the target appearance, integration between several types of features can enables the appearance model to be more robust to different challenges. Techniques such as feature concatenation [1], and product [40] can be utilized for the feature fusion. For instance, Zhou et al. [78] developed a multi-feature fusion approach, which combines spatial-colour appearance, histogram of oriented gradient and Gaussian spatial constraints to realize the weight penalization, thereby improving the tracking accuracy.

After the feature extraction, similarity measure or feature matching between targets can be made the extracted features by means of statistical computation, generative and discriminative models. For the statistical modelling, comparing the distance between feature vectors has been broadly applied to efficiently compute the similarity across targets, such as Bhattacharyya distance [40] and Cosine distance [61]. The calculated distance is then transformed to likelihood function via Gaussian kernels like [86].

For the generative models, a typical method, sparse representation has been successfully applied to robust face recognition [87] and visual tracking [88] To achieve the similarity measure, this method usually employs the with minimal reconstruction error [89] or maximal sparsity coefficient [90], which is computed between the candidate target and an over-complete dictionary learned from the training targets. Recent advancements of structured sparsity has been proved to provide better efficiency and robustness than simple sparsity in image classification and object tracking applications [91–94], 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. Discriminative models are designed as a classifier using the extracted features to discriminate the desired target from the background or other targets, such as conventional approaches of boosting [95] and support vector machines [9]. In recent years, deep neural networks with different architectures have been used for similarity measure. Typical CNN based models are siamese network [85] with two symmetrical CNN and quadruplet network [18] with four identical branches. Furthermore, correlation filters have been effectively used as a discriminative classified in single object tracking applications [84, 96]. This kind of technique can achieve high computational efficiency due to the use of fast Fourier transforms, and its collaborating with convolutional features gives better robustness for target specific appearance modelling. To investigate how correlation filters can improve multiple human tracking, [97] and [98] have been proposed to apply multiple single object trackers based on the KCFs in parallel for fast tracking.

Spatio-temporal Information

Apart from the appearance model, spatial-temporal information can be also useful to describe the characterises of the targets, when the motion un-

certainty is low. This model which simplifies the measurement model can largely relax the computation cost on the tracking algorithm without using sophisticated visual information. The idea is to usually used as a constraint to overlapping measure between frames such as Mahalanobis distance, Euclidean distance and intersection-over-union (IOU) ratio.

Early tracking methods [52], [53] often exploit the spatia-temporal constraints to collect valid measurements by comparing the Mahalanobis distance between the predicted states and current measurements. Additionally, Sanchez et al. [29] proposed an early association, in which, the Euclidean distance is used to measure the difference on the position and size between a pair of bounding boxes. Bochinski et al. [99] assumed that same targets in consecutive frames return a relatively high overlapping ratio, so they proposed to use IOU ratio to measure the similarity for target matching. Moreover, Milan et al. [64] presented an end-to-end learning approach to investigate spatio-temporal correlation between targets instead of the use of hand-designed distance measure. This is particularly use helpful to cope with the realistic scenarios where targets have complex movements.

Hybrid model

In reality, the performance of methods in the previous section may degrade drastically when targets move in close proximity, because relying solely on spatio-temporal model is not robust enough to address the target ambiguity in the image plane. Likewise, the use of appearance model alone may not be adequate to give distinct cues to separate targets, especially when they have similar appearances or occlusions occur to them. Therefore, fusing different models to represent targets can complement each other and handle the uncertainties in different aspects.

Following this, Wojke et al. [61] combined the motion information using Mahalanobis distance with the CNN based appearance descriptor for better

target assignment. Feng et al. [9] combined a social force based interaction model with target motion to better handle the target ambiguity in the occlusion region. Park et al. [62] proposed to integrate correlation filters and a confidence-based relative motion network to perform a two-step data association to track multiple objects, where Correlation filters are employed as a verifying step to confirm the target estimates. To incorporate more cues, Tang et al. [38] gathered a set of matching keypoints, spatio-temporal relation and detection confidence to measure the affinity between a pair of targets. With regard to the deep learning based fusion approach, Sadeghian et al. [37] presented a structure of RNN to encode long-term dependencies across motion, appearance and interaction models.

The above reviews underpin the major target representation models which are closely related to the proposed works in the thesis. However, there are other models that have been used to address the challenges and thus improve the tracking accuracy, such as interaction and exclusion models [8], [79].

2.4.2 Tracking inference

Until this stage, detections (measurements) which are obtained from a human detector have been richly characterized with different feature models. Then the real tracking part is to establish the link between these transformed detections with tracked targets in the previous time [35]. In the following subsections, two mainstreams for implementing tracking are discussed. The one is Bayesian filtering based tracking using the Bayes theory. The other is data association based tracking via an optimization solution.

Optimization based data association

In contrast to Bayesian filtering based tracking approaches, most of data association based methods categorized to offline tracking generally utilize the

entire set of detections to address a global optimization problem, where the cost function is needed to be well designed. [17, 19–22]. In [100], tracking multiple targets was formulated as a submodular maximization problem to globally find the most related tracklets for trajectory generation. In [41], authors exploited the interactions between non-associable tracklets to facilitate multi-target tracking, and addressed the binary labeling problem using efficient quadratic pseudo-Boolean optimization. For graph based approaches, they associate and cluster the tracking hypotheses as different nodes on the directed graphs, such as conditional random field [42], energy minimization [22] and subgraph multi-cut [38]. Deep learning based data association has attracted more attention recently. Kim et al. [101] developed a bilinear LSTM which jointly encodes both appearance and motion information for target tracks, in order to make the full use of past target appearance Son et al. [18] proposed a Quadruplet CNN for multi-target tracking, which employs quadruplet losses with appearance and motion cues to associate detections across video frames. However, some of the techniques such as notably Hungarian algorithm which is usually involved with data association based tracking can be also integrated into Bayesian tracking approaches for better tracking accuracy.

Bayesian filtering

Recursive Bayesian filtering based on Bayes theory effectively provides a general framework for dynamic state estimation, which is closely related to the focused online tracking methods. There are generally two stages in Bayesian recursive estimation: prediction and update. In prediction stage, the system model is used to predict the state posterior probability density. The objective of the update step is to employ the new available measurements to refine the estimated posterior probability density function (PDF) [25]. The Kalman filter [24] is widely known as the optimal algorithm addressing

the linear-Gaussian estimation problem, relating the mean and covariance of posterior distribution. The extended versions of the Kalman filter have been developed, such as the extended Kalman filter and the unscented Kalman filter. Regarding the scenarios in nonlinear or non-Gaussian filtering, particle filters [102] were developed to provide a better way for estimation via a set of randomly weighted samples. Although these fundamental algorithms have been shown to be effective in situations where the number of targets are small, they are not always feasible to deal with high dimensional state spaces and varying number of targets.

In recent years, random finite set theory [26] has become an increasing trend in multi-target Bayesian tracking with the employment of finite set statistics (FST). The PHD filter as one of the resulting filters from this theory [26] provides a promising alternative to the previous conventional Bayesian methods. This is because the PHD filtering method has the advantage of estimating the time-varying number of targets and also gives estimates in both cardinality and localization with relatively low computational cost [15]. Two main implementations of approximating the PHD intensity function have been made by a Gaussian mixture as in the GM-PHD filter [103] or the sequential monte carlo method via a set of weighted random particles known as the SMC-PHD filter [44].

Driven by the detection based tracking framework, a practical version of this filtering technique has been introduced from its original radar/sonar tracking domain to be applicable in vision tasks [1,29]. Several recent strategies have been made to enhance the performance of the PHD filtering based video tracker for multiple human tracking. In [29], an early association step has been presented to be helpful for better handling the missed detection and clutter problems. Besides, the development of the GM-PHD filter has been extended to track multiple targets with different types in [104]. Therefore, in this thesis, this kind of filtering framework becomes the baseline track-

ing inference to develop different algorithmic solutions to multiple human tracking.

2.5 Background preliminaries of Bayesian tracking approaches

The previous section presented the taxonomy of multiple human tracking algorithms in accordance with two crucial components: target representation and tracking inference, and reviewed recent tracking methods and associated technologies of based on these two components. Recent developments of target representation, such as appearance model, spatio-temporal model, and hybrid model have been reviewed. In addition, Bayesian filtering and optimization based tracking as two main approaches for implementing the tracking inference are discussed. Since this thesis focuses on online multiple human tracking, which typically requires only past and present detection responses for the estimation. Probabilistic tracking approaches particularly with the recursive Bayesian filtering methods are largely suitable for the tracking process in an online mode. Therefore, the following subsections are aimed at demonstrating the systematic formulation of Bayesian tracking approaches, including the details of multi-target Bayesian filtering, RFS theory, and PHD filter.

2.5.1 Multiple target Bayesian filtering

Multiple human tracking can be generally considered as an online multiple state estimation problem. Given a video sequence, the state for the i-th target in the k-th frame is defined as $\mathbf{x}_k^i$. Let $\mathbf{X}_k = (\mathbf{x}_k^1, ..., \mathbf{x}_k^{M_k})$ denote the states of all present targets in a monitored scene in the k-th frame, where M_k is the number of targets at time k. Thus, all the sequential states of targets from the initial frame to the k-th frame are represented as $\mathbf{X}_{1:k} = \{\mathbf{X}_1, ..., \mathbf{X}_k\}$. It should be noted that the number of targets may be

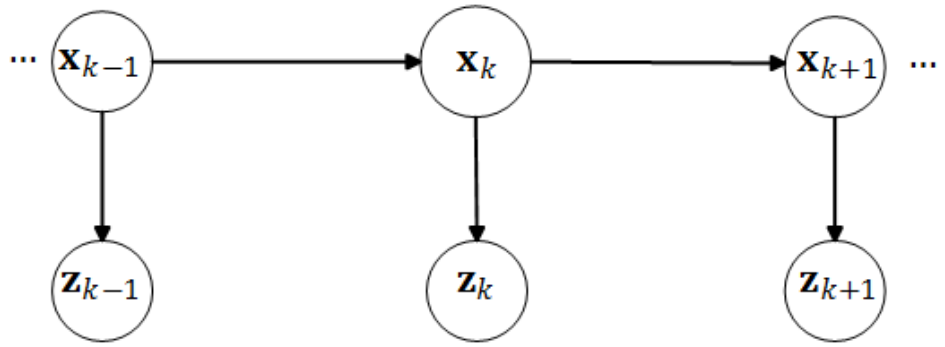


Figure 2.6. Graphical representation of a hidden Markov model within Bayesian filtering. [5]

varied at each frame. Assuming that all the target states follow a first-order Markov process, the current target state only relies on the very last state, which yields the target dynamic or transition model $f_{k|k-1}$,

$$f_{k|k-1}(\mathbf{X}_k|\mathbf{X}_{1:k-1}) = f_{k|k-1}(\mathbf{X}_k|\mathbf{X}_{k-1}) \tag{2.5.1}$$

where $k|k-1$ and $k|k$ denote the prediction and filtering process respectively. A graphical representation of the hidden Markov model within the Bayesian filtering is described in Fig. 2.6, where the target state evolves from left to right.

This Markov process also contains a measurement model to provide observation measurements for the target states. Since the developed tracking systems in this thesis follow the detection based tracking paradigm, detection responses are viewed as the viable measurements. The obtained measurements for all the target states at the k-th frame are denoted as $\mathbf{Z}_k = (\mathbf{z}_k^1, ..., \mathbf{z}_k^{N_k})$, where N_k is the number of measurements. Thus all the obtained sequential measurements of all the targets for the entire sequence are denoted as, $\mathbf{Z}_{1:k} = \{\mathbf{Z}_1, ..., \mathbf{Z}_k\}$. In the Markov process, the measurement of a target is only associated with its target state, which means that all the measurements are conditionally independent. Thus, the measurement model

or likelihood function can be formulated as,

$$p(\mathbf{Z}_k|\mathbf{X}_k) \tag{2.5.2}$$

In the Bayesian filtering tracking model, the goal is to estimate the multi-target posterior density $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ at time k, which is conditioned on the given measurements $\mathbf{Z}_{1:k} = \{\mathbf{Z}_1, ..., \mathbf{Z}_k\}$ up to time k [103]. There are two main steps in this tracking model: prediction and updating. The prediction step is to predict the current state by combining all the previous obtained measurements with the previous state through the state transition model $f_{k|k-1}(\mathbf{X}_k|\mathbf{X}_{k-1})$. The update step aims to update the predicted state with the currently available measurements via the measurement model $p(\mathbf{Z}_k|\mathbf{X}_k)$. Then the optimal multiple target Bayesian filtering recursion can be shown as [25] [103],

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k-1}) = \int f_{k|k-1}(\mathbf{X}_k|\mathbf{X}_{k-1})p(\mathbf{X}_{k-1}|\mathbf{Z}_{1:k-1})d\mathbf{X}_{k-1} \tag{2.5.3}$$

where

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k}) = \frac{p(\mathbf{Z}_k|\mathbf{X}_k)p(\mathbf{X}_k|\mathbf{Z}_{1:k-1})}{\int p(\mathbf{Z}_k|\mathbf{X}_k)p(\mathbf{X}_k|\mathbf{Z}_{1:k-1})d\mathbf{X}_k} \tag{2.5.4}$$

Various fundamental methods have been proposed in the literature to give an approximate solution to computing the integral of the above state distribution, such as Kalman filtering [24], extended Kalman filtering [105], and particle filtering [25]. However, issues of variable number of targets and target states with non-fixed dimensions to some extent limit their capacity to be applied in more complex tracking scenarios. In addition, computationally intensive associations in these methods may not be avoided with the increased number of targets. To address these issues, the concept of random finite set based filtering techniques using finite statistics (FISST) will be introduced in the following section.

2.5.2 The random finite set for multiple target filtering

Random finite set [106] filtering has been recently developed in multi-target tracking approaches, particularly to deal with tracking scenarios where the number of targets is varied in time, as well as consistently handling the problem of variable dimension states and measurements. Based upon the RFS framework, a multiple target state and a multiple target measurement at time k can be represented by two finite sets:

$$\mathbf{X}_k = \{\mathbf{x}_k^1, ..., \mathbf{x}_k^{M_k}\} \in \mathcal{F}(\mathcal{X}) \tag{2.5.5}$$

$$\mathbf{Z}_k = \{\mathbf{z}_k^1, ..., \mathbf{z}_k^{N_k}\} \in \mathcal{F}(\mathcal{Z}) \tag{2.5.6}$$

where $\mathcal{F}(\mathcal{X})$ and $\mathcal{F}(\mathcal{Z})$ are the finite subsets of $\mathcal{X}$ and $\mathcal{Z}$ respectively [49]. According to [15], an RFS can be defined as a set where:

- the elements are random stochastic processes
- the set cardinality is also a stochastic process

Another advantage of the RFS framework [106] is that it provides a principled solution to the problem of dynamically estimating multiple targets in the presence of clutter and association uncertainty [103]. The uncertainty in the multiple tracking system is represented by modelling the multiple target state $\mathbf{X}_k$ and multiple target measurement $\mathbf{Z}_k$ as *random finite sets* [103]. Specifically, an RFS model for the time evolution of the multiple target state contains target motion, birth and death. Given a multiple target state $\mathbf{X}_{k-1}$ at time $k-1$, each target $\mathbf{x}_{k-1}$ either survives at time k with a survival probability $e_{k|k-1}(\mathbf{x}_{k-1})$ or disappears with probality $1-e_{k|k-1}(\mathbf{x}_{k-1})$. For a new-born target, it can be either generated spontaneously or spawned from a target at time $k-1$. Therefore, the multi-target state $\mathbf{X}_k$ at time k can be constituted by the union of the surviving targets, spawned targets and

new-born targets [103],

$$\mathbf{X}_k = \left[\bigcup_{\xi \in \mathbf{X}_{k-1}} S_{k|k-1}(\xi) \right] \cup \left[\bigcup_{\xi \in \mathbf{X}_{k-1}} B_{k|k-1}(\xi) \right] \cup \Upsilon_k \tag{2.5.7}$$

where $S_{k|k-1}(\xi)$ denotes the RFS of survived targets, Υ_k is the RFS of spontaneous new-born targets at time k, and $B_{k|k-1}(\xi)$ denotes the RFS of targets spawned at time k from a previous target state ξ [103].

Likewise, the RFS measurement model denoted by $\mathbf{Z}_k$, considers the uncertain detections and clutter [103]. Given a target $\mathbf{x}_k$ at time k, it can be either missed by a sensor or detector with probability of $p_{M,k}(\mathbf{x}_k)$ or detected with probability of $1 - p_{M,k}(\mathbf{x}_k)$. Hence, a measurement $\mathbf{z}_k$ is generated with a probability density by the measurement model (2.5.2). However, the false alarms or clutter can be also unexpectedly generated by the detector, which are modelled as a set Γ_k. Therefore, the multi-target measurement set $\mathbf{Z}_k$ generated by the detector at time k is constituted by the union of target originated measurements and clutter [103],

$$\mathbf{X}_k = \left[\bigcup_{\mathbf{x}_k \in \mathbf{X}_k} \Theta_k(\mathbf{x}_k) \right] \cup \Gamma_k \tag{2.5.8}$$

where $\Theta_k(\mathbf{x}_k)$ defines the measurements generated from the targets $\mathbf{X}_k$, and Γ_k models the false measurements or clutter.

It is similar to the single-target dynamic model and observation, the $\mathbf{X}_k$ and $\mathbf{Z}_k$ are captured by the multi-target transition density $f_{k|k-1}(\mathbf{X}_k|\mathbf{X}_{k-1})$ and the multi-target likelihood $g_k(\mathbf{Z}_k|\mathbf{X}_k)$ respectively. Under the RFS framework, the Bayesian recursive equations of multi-target posterior density $p_{k|k}(\mathbf{X}_k|\mathbf{Z}_{1:k})$ can be expressed as [15],

$$p_{k|k-1}(\mathbf{X}_k|\mathbf{Z}_{1:k-1}) = \int f_{k|k-1}(\mathbf{X}_k|\mathbf{X}_{k-1}) p_{k|k-1}(\mathbf{X}_{k-1}|\mathbf{Z}_{1:k-1}) \mu(d\mathbf{X}_{k-1}) \tag{2.5.9}$$

where

$$p_{k|k}(\mathbf{X}_k|\mathbf{Z}_{1:k}) = \frac{g_k(\mathbf{Z}_k|\mathbf{X}_k)p_{k|k-1}(\mathbf{X}_k|\mathbf{Z}_{1:k-1})}{\int g_k(\mathbf{Z}_k|\mathbf{X}_k)p_{k|k-1}(\mathbf{X}_k|\mathbf{Z}_{1:k-1})\mu(d\mathbf{X}_k)} \tag{2.5.10}$$

and $\mu(\cdot)$ is an appropriate dominating measure on RFS $\mathcal{F}(\mathcal{X})$ [44]. However, there are multiple integrals in (2.5.9) and (2.5.10) in the $\mathcal{F}(\mathcal{X})$, which are computationally intractable. To address this issue, the PHD filter will be introduced in the next section.

2.5.3 The probability hypothesis density filter

The PHD filter proposed by Mahler [26] offers a tractable approximation to mitigate the computational intractability in multiple target Bayesian filtering. It is intended to recursively propagate the first-order statistical moment of the multiple target posterior density $p_{k|k}(\mathbf{X}_k|\mathbf{Z}_{1:k})$, referred to as the intensity function $\nu_{k|k}(\mathbf{x}|\mathbf{Z}_{1:k})$ abbreviated by $\nu_{k|k}(\mathbf{x})$ [107].

There are three important assumptions for the recursions of the PHD filter [103]:

1. Targets evolve and generate observations independently,
2. Clutter is Poisson-distributed and independent of target-originated observations, and
3. the predicted multiple target RFS is Poisson-distributed.

Under the assumptions above, the following PHD recursions for the propagation of the posterior intensity have been demonstrated using FISST [26]. Given a posterior intensity ν_{k-1} at time $k-1$, the prediction of the PHD filter can be defined as [15, 26],

$$\nu_{k|k-1}(\mathbf{x}) = \int e_{k|k-1}(\xi)f_{k|k-1}(\mathbf{x}|\xi)\nu_{k-1}(\xi)d\xi + \int \beta_{k|k-1}(\mathbf{x}|\xi)\nu_{k-1}(\xi)d\xi + \gamma_k(\mathbf{x}) \tag{2.5.11}$$

where $\gamma_k(\cdot)$ is the intensity of the birth RFS Υ_k at time k, $f_{k|k-1}(\cdot)$ denotes the state transition model, and $e_{k|k-1}(\xi)$ is the probability that a target still exists at time k given its previous state of ξ. $\beta_{k|k-1}(\cdot|\xi)$ is the intensity of the RFS $B_k(\xi)$ spawned from the previous state ξ. When the new set of measurements is available, the PHD update step is given as [103] [15],

$$\nu_{k|k}(\mathbf{x}) = \Big[p_{M,k}(\mathbf{x}) + \sum_{\mathbf{z}\in\mathbf{Z}_k} \frac{(1-p_{M,k})(\mathbf{x})g_k(\mathbf{z}|\mathbf{x})}{\kappa_k(\mathbf{z}) + \int(1-p_{M,k})(\mathbf{x})g_k(\mathbf{z}|\mathbf{x})\nu_{k|k-1}}\Big]\nu_{k|k-1}(\mathbf{x}) \tag{2.5.12}$$

where $p_{M,k}(\mathbf{x})$ is the probability of missed detection given a state $\mathbf{x}$ at time k, $g_k(\mathbf{z}|\mathbf{x})$ is the single-target likelihood which defines the probability that $\mathbf{z}$ is generated by a target with state $\mathbf{x}$, and $\kappa_k(\cdot)$ denotes the clutter density of RFS Γ_k.

According to both (2.5.11) and (2.5.12), the PHD filter is possible to avoid combinatorial computations and requires lower complexity than Bayesian recursions in (2.5.9) and (2.5.10) operating on $\mathcal{F}(\mathcal{X})$ [103]. However, there is no generally closed-form solutions for the integral in (2.5.11) and (2.5.12). Two implementations of approximating the PHD recursions have been developed by a Gaussian mixture as in the GM-PHD filter [103] or the sequential Monte Carlo method via a set of weighted random particles known as the SMC-PHD filter [44], and thus they will be mainly exploited along with the detection based tracking framework for online multiple human tracking in the next three chapters.

The above methods underpin the Bayesian theory based probabilistic tracking models with a particular interest in RFS filtering. In the next section, datasets and performance measures for the tracking evaluations will be presented.

2.6 Evaluation datasets

To demonstrate the effectiveness and applicability of the proposed filtering methods for real-world tracking scenarios, it is essential to evaluate the performance of these trackers on real video sequences. In this section, several publicly available sequences including training and testing from the MOTChallenge benchmark[2] are explained, including the 2D MOT15, MOT16, and MOT17 datasetS and they will be used to demonstrate the validity of the proposed methods. All the video sequences are suitable for pedestrian tracking. Multi-camera tracking is out of the scope of this thesis.

[2]https://motchallenge.net/

2.6.1 2D MOT 2015 challenge

The 2D MOT15 dataset[3] [3] collects a set of video sequences from other datasets (e.g. PETS2009 [108], TUD [109], ETH [110] and KITTI [111]) and some newly added challenging sequences. This dataset consists of 11 training and 11 testing video sequences captured with different camera motion, viewpoint, and weather conditions. Public object detections are available to be used for fair comparisons. The testing video sequences are used for performance evaluation in comparisons with other available trackers. Table 2.1 details the different characterises (e.g. frame rate, target density, camera motion, and illumination conditions) of the challenging video sequences included in this dataset. Example image frames as shown in Fig. 2.7 and 2.8 provide an overview throughout the sequences.

[3]https://motchallenge.net/data/2D_MOT_2015/

Figure 2.7. Example image frames from the 2D MOT 2015 training dataset [3].

Figure 2.8. Example image frames from the 2D MOT 2015 test dataset [3].

Table 2.1. Details of the sequences presented in the 2D MOT15 Benchmark [3]

Training Sequences									
Name	FPS	Resolution	Length	Tracks	Boxes	Camera	Viewpoint	Conditions	Ref
TUD-Stadtmitte	25	640×480	179 (00:07)	10	1156	static	medium	cloudy	[109]
TUD-Campus	25	640×480	71 (00:03)	8	359	static	medium	cloudy	[109]
PETS09-S2L1	7	768×576	795 (01:54)	19	4476	static	high	cloudy	[108]
ETH-Bahnhof	14	640×480	1000 (01:11)	171	5415	moving	low	cloudy	[110]
ETH-Sunnyday	14	640×480	354 (00:25)	30	1858	moving	low	sunny	[110]
ETH-Pedcross2	14	640×480	840 (01:00)	133	6263	moving	low	sunny	[110]
ADL-Rundle-6	30	1920×1080	525 (00:18)	24	5009	static	low	cloudy	[3]
ADL-Rundle-8	30	1920×1080	654 (00:22)	28	6783	moving	medium	night	[3]
KITTI-13	10	1242×375	340 (00:34)	42	762	moving	medium	sunny	[111]
KITTI-17	10	1242×370	145 (00:15)	9	683	static	medium	sunny	[111]
Venice-2	30	1920×1080	600 (00:20)	26	7141	static	medium	sunny	[3]
Test Sequences									
TUD-Crossing	25	640×480	201 (00:08)	13	1102	static	medium	cloudy	[109]
PETS09-S2L2	7	768×576	436 (01:02)	42	9641	static	high	cloudy	[108]
ETH-Jelmoli	14	640×480	440 (00:31)	45	2537	moving	low	sunny	[110]
ETH-Linthescher	14	640×480	1194 (01:25)	197	8930	moving	low	sunny	[110]
ETH-Crossing	14	640×480	219 (00:16)	26	1003	moving	low	cloudy	[110]
AVG-TownCentre	2.5	1920×1080	450 (03:45)	226	7148	static	high	cloudy	[112]
ADL-Rundle-1	30	1920×1080	500 (00:17)	32	9306	moving	medium	sunny	[3]
ADL-Rundle-3	30	1920×1080	625 (00:21)	44	10166	static	medium	sunny	[3]
KITTI-16	10	1242×370	209 (00:21)	17	1701	static	medium	sunny	[111]
KITTI-19	10	1242×374	1059 (01:46)	62	5343	moving	medium	sunny	[111]
Venice-1	30	1920×1080	450 (00:15)	17	4563	static	medium	sunny	[3]

2.6.2 2D MOT 2016 and 2017 challenges

The 2D MOT16 dataset[4] contains a variety of more challenging video sequences than those included in 2D MOT15. These sequences are recorded by static or moving cameras, and under the complex scenes of illumination changes, varying viewpoints and weather conditions. This dataset [4] consists of 7 training and 7 testing video sequences with the public object detections generated by [2] for fair comparisons. Annotations only for the training sequences in the MOT16 dataset [4] are more accurately obtained via a consistent protocol.

The 2D MOT17 dataset [4] is built on the 2D MOT16 dataset, but with a new and more accurate ground truth. Each sequence included in the dataset is provided with 3 sets of public detections (DPM [2], FRCNN [16], and SDP [113]) for more comprehensive evaluations. The training video sequences with available ground truths are primarily utilized to process the performance analysis, while testing sequences are used to generate quantitative comparisons against existing state-of-the-art tracking methods.

[4]https://motchallenge.net/data/MOT16/

Figure 2.9. Overview of the 2D MOT 2016 training dataset [4]

Figure 2.10. Overview of the 2D MOT 2016 test dataset [4]

Table 2.2. Details of the sequences presented in the 2D MOT16 dataset [4]

Training Sequences								
Name	FPS	Resolution	Length	Tracks	Boxes	Camera	Viewpoint	Conditions	Ref
MOT16-02	30	1920×1080	600 (00:20)	49	17,833	static	medium	cloudy	[4]
MOT16-04	30	1920×1080	1,050 (00:35)	80	47,557	static	high	night	[4]
MOT16-05	14	640×480	837 (01:00)	124	6,818	moving	medium	sunny	[110]
MOT16-09	30	1920×1080	525 (00:18)	25	5,257	static	low	indoor	[4]
MOT16-10	30	1920×1080	654 (00:22)	54	12,318	moving	medium	night	[4]
MOT16-11	30	1920×1080	900 (00:30)	67	9,174	moving	medium	indoor	[4]
MOT16-13	25	1920×1080	750 (00:30)	68	11,450	moving	high	sunny	[4]
Test Sequences									
MOT16-01	30	1920×1080	450 (00:15)	23	6,395	static	medium	cloudy	[4]
MOT16-03	30	1920×1080	1,500 (00:50)	148	104,556	static	high	night	[4]
MOT16-06	14	640×480	1,194 (01:25)	217	11,538	moving	medium	sunny	[110]
MOT16-07	30	1920×1080	500 (00:17)	55	16,322	moving	medium	shadow	[4]
MOT16-08	30	1920×1080	625 (00:21)	63	16,737	static	medium	sunny	[4]
MOT16-12	30	1920×1080	900 (00:30)	94	8,295	moving	medium	indoor	[4]
MOT16-14	25	1920×1080	750 (00:30)	230	18,483	moving	high	sunny	[4]

2.7 Evaluation metrics

This section describes three common evaluation metrics to evaluate the tracking performance of any relevant tracking method. One is termed the optimal subpattern assignment (OSPA) [114] and originates from the signal processing community, so that a direct comparison is viable. The second is to apply CLEAR MOT metrics [115] which are the commonly-used tool from the computer vision community to give a number of individual performance measures for the tracking assessment. And the third provides a set of measures in [95], to evaluate the quality of target trajectories. The second and third measures form the MOTChallenge Benchmark metrics [3], which will be used to assess the benchmark performance of proposed methods, in order to achieve a fair comparison with other state-of-the-art tracking algorithms in the leaderboard.

2.7.1 OSPA metric

The miss-distance has generally played an essential part in the formulation and evaluation of filtering and control algorithms [114]. For single target tracking, the performance measures including the Euclidean errors and mean squared errors are based on the concept of miss-distance [114]. However, those measures may not be suitable for the case in multi-target tracking. A performance metric for evaluating the multi-target tracking was developed in [114], which is intended to capture both cardinality and localization errors. This OSPA metric has been widely used to examine the tracking performance in video-based multiple human tracking [1] [9] [28] [55] [116].

Let $\mathbf{O}_k = \{\mathbf{o}_k^1, ..., \mathbf{o}_k^i, ..., \mathbf{o}_k^m\}$ 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, $\hat{\mathbf{O}}_k = \{\hat{\mathbf{o}}_k^1, ..., \hat{\mathbf{o}}_k^j, ..., \hat{\mathbf{o}}_k^n\}$ 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 [8]. 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 $c > 0$, the metric is defined as,

$$d_{k,p}^{(c)}(\mathbf{O}_k, \hat{\mathbf{O}}_k) := \left(\frac{1}{n} \left(\min_{\pi \in \Pi_n} \sum_{i=1}^{m} d_k^{(c)}(\mathbf{o}_k^i, \hat{\mathbf{o}}_k^{\pi(i)})^p + c^p(n-m) \right) \right)^{\frac{1}{p}} \tag{2.7.1}$$

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 $\{1, 2, ..., n\}$ for any $n \in \mathbb{N} = \{1, 2, ...\}$. The function of $d_{k,p}^{(c)}$ is called as the OSPA metric of order p with cut-off c at time k step. In this thesis, $c = 20$ and $p = 2$ are used in the evaluations as these setting have been commonly used in the literature [5].

2.7.2 CLEAR MOT metrics

As summarized above, the OSPA metric employs the Euclidean distance to establish the correspondence between the ground truth and tracking outputs before computing the metric. While for the CLEAR MOT [115] evaluation, this correspondence can be also obtained by reasoning the bounding boxes in the image plane. Thus, the intersection over union is computed with a typical threshold of 0.5 or 50%, to quantify the closeness between annotated and estimated bounding boxes. The CLEAR MOT metrics [115] mainly entail multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP).

Accuracy

The MOTA as an accuracy score is often considered as the most important metric for summarizing the overall performance, since it is comprised of three following error sources: the total number of false negatives or missed

targets (FN), the total number of false positives (FP), and the total number of identity switches (IDS) [115]. Thus, the MOTA is defined as,

$$MOTA = 1 - \frac{\sum_k (FN_k + FP_k + IDS_k)}{\sum_k GT_k} \tag{2.7.2}$$

where GT_k denotes the actual number of targets present at time step k.

The concept of each source error is further interpreted as follows. False positives can be explained as tracking hypotheses which do not have correspondence with desired targets. For instance, targets which are not humans tracked as humans or there are no desired targets in the tracked bounding boxes. Targets which are missed by any tracking hypotheses are considered as false negatives. An identity switch or mismatched error is likely to occur when targets move close to each other and the tracker mislabels their identities. [115] To achieve a better MOTA score, these three error sources are expected to be as few as possible. This MOTA score is reported as the percentage $(-\infty, 100]$ in all evaluations, the negativity of which is due to situations where the number of errors caused by the tracking method is in excess of the number of all present targets [3].

Precision

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, and it is calculated as,

$$MOTP = \frac{\sum_{k,i} dist_{k,i}}{\sum_k \tau_k} \tag{2.7.3}$$

where $dist_{k,i}$ is the bounding box overlap between a target estimation and its aligned ground truth target i, and τ_k is the number of matches found at time k. MOTP as a localization precision measure is reported with a range from 50% to 100%. This measure is mainly determined by the human

detections and annotations, but less affected by the tracking outputs [3].

2.7.3 Other metrics

The following metrics [95] are specifically designed to evaluate the quality of each target trajectory that is recovered by the tracking method, which are the mostly tracked targets (MT, the ratio of ground truth targets whose trajectories are covered by a tracking result for at least 80%), mostly lost targets (ML, the ratio of ground truth targets whose trajectories are covered by a tracking result for at most 20%), and the total number of times a trajectory is fragmented (Frag) or interrupted. It is desirable to a smaller value of ML and Frag, but a higher MT. In addition, the average number of false alarms per frame (FAF) and the runtime performance in frames per second (Hz) are also included for the tracking evaluations.

2.8 Summary

In this chapter, a large number of tracking algorithms closely related to the subject of this thesis have been reviewed according to two important trends respectively: dealing with the challenging issues and enhancing the tracking components. Firstly, existing tracking methods presented in this chapter were generally classified based on the processing solutions to four representative challenges in multiple human tracking, which are unknown number of targets, noisy detections, missed detections and occlusions. A series of existing algorithms proposed to cope with the challenges as well as their advantages and disadvantages were investigated. Then, the taxonomy of multiple human tracking methods was made based on two primary components in the tracking system: target representation and tracking inference.

This thesis aims to further address the challenges including noisy detections, missed detections and occlusions, as well as strengthening pri-

mary components in the tracking system which are target representation and tracking inference. In order to better represent the targets, the proposed works focused on improving the target appearance model and fusion model using both hand-crafted and CNN features. For the target inference, this thesis has a particular interest in PHD filtering from Bayesian filtering. Bayesian theory based probabilistic models were described in detail to formulate the online multiple human tracking problem. Fundamental solutions such as the Kalman filtering and particle filtering approaches to the multi-target Bayesian filtering recursions have been demonstrated to be restricted to the issues of varying number of targets and non-fixed target state dimensions. The RFS theory with its resulting PHD filter provided a better solution to handle the previous issues, whereas it is more suitable to be used in tracking multiple targets.

In addition to the problem formulation with Bayesian filtering, three benchmark datasets from the MOT Challenge have been presented and the performance evaluation measures as well as the evaluation sequences were given to evaluate the tracking performance. In the next chapter, the SMC-PHD filter based tracking system will be firstly developed with proposed online group-structured dictionary learning to improve the tracking performance.

Chapter 3

MEASUREMENT-DRIVEN SMC-PHD FILTER BASED ONLINE MULTIPLE HUMAN TRACKING USING ONLINE GROUP-STRUCTURED DICTIONARY LEARNING

3.1 Introduction

In video-based human tracking, due to the imperfections in the human detector, it is important for the SMC-PHD filter to accurately select valid measurements to determine the birth intensity of the newborn targets, and also remove the false alarms or clutter in the measurement model.

Conventional PHD filtering methods [46], [44], [26] empirically preset the target birth intensity and the gating threshold for selecting reliable measurements for the PHD update. This may not be effective in dealing with noisy measurements in real world tracking scenarios. 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 [52], [53], [54]. However, these methods may be no longer applicable to achieve effective measurements without the assumptions: 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, recent literature shows structured dictionary learning has been demonstrated to provide better efficiency and robustness in appearance modelling to remove background clutter during tracking [91–94].

Inspired by the aforementioned established methods, this chapter presents an enhanced SMC-PHD filter for multiple human tracking in video. The proposed system mainly exploits two concepts: adaptive gating (AG) based measurement classification and birth intensity estimation driven by online group-structured dictionary learning (Online-GSDL). The AG technique aims to refine the original measurement set, and it is able to adaptively update the gating threshold by considering the spatio-temporal and human size information. This gives the advantages of better discriminating measurements between the survival and new-born targets. In the Online-GSDL step, target appearance features are firstly extracted by the concatenation of histogram of oriented gradients (HOGs) and RGB colour histogram. The hierarchical K-means clustering [77] is used with the combined features to construct a group-structured dictionary. The multi-task group-structured sparsity in the dictionary is achieved by exploiting a collaborative hierarchical Lasso (C-HiLasso) model [117], which is able to strengthen the discriminability of the sparse coefficients at the group level. Then, a maximum voting method based on the sparsity solution is proposed to eliminate the potential false positives induced by noise or clutter from the measurement set. This could

increase the accuracy of birth intensity generalization. Moreover, the SimCO algorithm [118] is utilized along with the proposed structure pattern to implement the dictionary update, which would handle the target appearance variations.

1. A novel adaptive gating strategy is developed in the tracking 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.

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.

This chapter addresses the first and second objectives of the thesis, which matches with the particle PHD filter based multi-target tracking using discriminative group-structured dictionary learning published in [28], and the particle PHD filter based multiple human tracking using online group-structured dictionary learning published in IEEE Access [10]. The rest of this chapter is organized as follows. Section 3.2 details the proposed tracking algorithm as four main parts, including background study of the SMC-PHD filter, adaptive gating based measurement classification, online group-structured dictionary learning for birth intensity estimation and the dictionary update using the SimCO algorithm. Experimental results and comparisons between the proposed approach and other state-of-the-art methods are presented in Section 3.3.

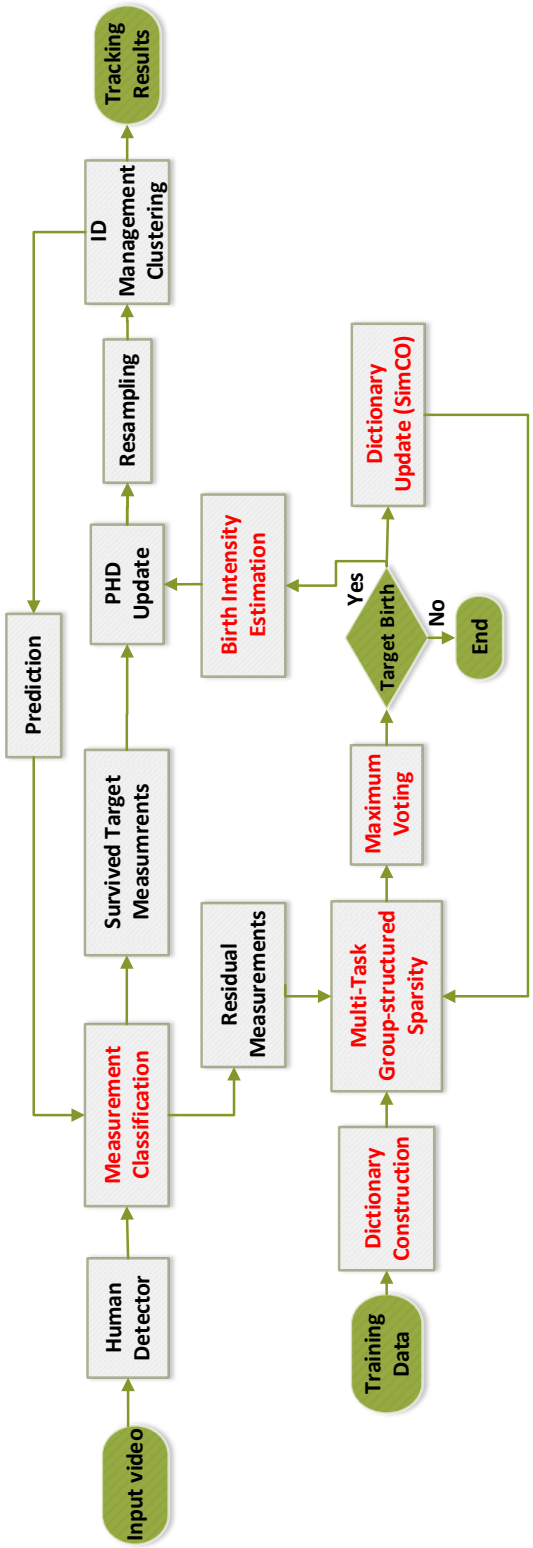


Figure 3.1. Block diagram of the proposed multiple human tracking system: the main contributions are labelled in red.

3.2 The proposed tracking system

3.2.1 Overview of the proposed approach

The overview of the proposed tracking system is shown in Fig. 3.1. Each input video sequence uses a human detector for measurement acquisition. 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 the proposed online group-structured dictionary learning which includes dictionary construction based on training data, multi-task group-structured sparsity, maximum voting technique and dictionary update using the SimCO algorithm. For efficient processing, birth intensity estimation (birth measurements) and dictionary update are performed, only if there exists new-born targets which can be verified by the maximum voting technique. 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.

3.2.2 The measurement-driven 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: $\mathbf{X}_k = \{\mathbf{x}_k^1, ..., \mathbf{x}_k^{M_k}\} \in \mathcal{F}(\mathcal{X})$ and $\mathbf{Z}_k = \{\mathbf{z}_k^1, ..., \mathbf{z}_k^{N_k}\} \in \mathcal{F}(\mathcal{Z})$, 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 [49]. 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 [26] 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 $\nu_{k|k}(\mathbf{x}|\mathbf{Z}_{1:k})$ abbreviated by $\nu_{k|k}(\mathbf{x})$. In this work, the decomposed form of PHD filter [49] is used, since it can distinguish the survival targets $\nu_{k|k}(\mathbf{x}, 0)$ and new-born targets $\nu_{k|k}(\mathbf{x}, 1)$ in both the prediction and update steps. Hence, the PHD prediction equation is given by,

$$\nu_{k|k-1}(\mathbf{x}, 0) = \int e_{k|k-1}(\xi) f_{k|k-1}(\mathbf{x}|\xi) \nu_{k-1|k-1}(\xi) d(\xi) \tag{3.2.1}$$

$$\nu_{k|k-1}(\mathbf{x}, 1) = \gamma_{k|k-1}(\mathbf{x}) \tag{3.2.2}$$

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 [49] can be defined with the available measurements from survival targets $\mathbf{Z}_{k,s}$ and new-born targets $\mathbf{Z}_{k,b}$ as:

$$\begin{aligned} \nu_{k|k}(\mathbf{x}, 0) = & \sum_{\mathbf{z} \in \mathbf{Z}_{k,s}} \frac{\psi_k(\mathbf{x}) \nu_{k|k-1}(\mathbf{x}, 0)}{\kappa_k(\mathbf{z}) + \langle \psi_k(\mathbf{x}), \nu_{k|k-1}(\mathbf{x}, 0) \rangle} \\ & + \nu_{k|k-1}(\mathbf{x}, 0) p_M(\mathbf{x}) \end{aligned} \tag{3.2.3}$$

$$\begin{aligned} \nu_{k|k}(\mathbf{x}, 1) = & \sum_{\mathbf{z} \in \mathbf{Z}_{k,b}} \frac{\psi_k(\mathbf{x}) \nu_{k|k-1}(\mathbf{x}, 1)}{\kappa_k(\mathbf{z}) + \langle \psi_k(\mathbf{x}), \nu_{k|k-1}(\mathbf{x}, 0) \rangle} \\ & + \nu_{k|k-1}(\mathbf{x}, 1) p_M(\mathbf{x}) \end{aligned} \tag{3.2.4}$$

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 this work, the sequential Monte Carlo implementation is used 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,

$$\{\tilde{\omega}^i_{k|k-1}, \tilde{\mathbf{x}}^i_k\}_{i=1}^{i=(M_{k-1}+J_k)\times\mathcal{N}} \tag{3.2.5}$$

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

$$\tilde{\omega}^i_{k|k-1} = \begin{cases} f_{k|k-1}(\tilde{\mathbf{x}}^i_k)\tilde{\omega}^i_{k-1}, i = 1, ..., M_{k-1} \times \mathcal{N}. \\ \\ \frac{\gamma_{k|k-1}(\mathbf{x})}{J_k}, i = M_{k-1} \times \mathcal{N} + 1, ..., (M_{k-1} + J_k) \times \mathcal{N}. \end{cases} \tag{3.2.6}$$

Once the new set of observations is available, the predicted weights $\tilde{\omega}^i_{k|k-1}$ are updated as,

$$\tilde{\omega}^i_k = \left[p_M(\tilde{\mathbf{x}}^i_k) + \sum_{\mathbf{z}\in\mathbf{Z}_k} \frac{\psi_k(\tilde{\mathbf{x}}^i_k)}{\kappa_k(\mathbf{z}) + C_k(\mathbf{z})} \right] \tilde{\omega}^i_{k|k-1} \tag{3.2.7}$$

where

$$C_k(\mathbf{z}) = \sum_{i=1}^{(M_{k-1}+J_k)\times\mathcal{N}} \psi_k(\tilde{\mathbf{x}}^i_k)\tilde{\omega}^i_{k|k-1}. \tag{3.2.8}$$

The likelihood function for each particle is given by,

$$g_k(\mathbf{z}|\tilde{\mathbf{x}}^i_k) = \frac{1}{(2\pi\sigma_g)^{1/2}} \exp\left(- \frac{(\mathbf{z} - \mathbf{H}\tilde{\mathbf{x}}^i_k)^T(\mathbf{z} - \mathbf{H}\tilde{\mathbf{x}}^i_k)}{2\sigma_g^2} \right) \tag{3.2.9}$$

where $\mathbf{H}$ is the observation matrix and σ_g^2 denotes the variance for the likelihood function. The expected number of targets M_k at time k can be estimated by the total mass of the weights from (3.2.7), $M_k = \sum_{i=1}^{(M_{k-1}+J_k)\times\mathcal{N}} \tilde{\omega}^i_k$. Furthermore, a resampling step will be performed with normalized weights $\tilde{\omega}^i_k = \tilde{\omega}^i_k / M_k$ after the update step, aiming to eliminate particles with low

importance weight and avoid the degeneracy problem [15].

The above work underpins the decomposed form of SMC-PHD filter, which has been used extensively in multiple human tracking [49, 52, 53]. 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 [50] can be utilized to extract the current states from all the particles.

3.2.3 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 this end, a novel adaptive gating method is proposed to extract the survival targets from the entire measurement set, and also to discard false positives within the survival measurements set.

In the context of video human tracking, the proposed gating strategy is not only limited to employing the positions and velocities, but also includes an elliptical human shape with height and width. The validation gate threshold T_k is designed to reduce the number of candidate measurements, and it is computed by,

$$T_k = \left[\left(\frac{1}{N_k} \sum_{n=1}^{N_k} \|\bar{\mathbf{s}}_k^n\|_1 \right)^2 + \left(\frac{1}{M_k} \sum_{m=1}^{M_k} \|\mathbf{s}_{k|k-1}^m\|_1 \right)^2 \right]^{\frac{1}{2}} \tag{3.2.10}$$

where $\bar{\mathbf{s}}_k^n = [\bar{w}_k^n, \bar{h}_k^n]^T \in \mathbf{z}_k^n$ offers the size information for the n-th measurement, $\mathbf{s}_{k|k-1}^m = [w_{k|k-1}^m, h_{k|k-1}^m]^T \in \mathbf{x}_{k|k-1}^m$ denotes the width and height of the m-th predicted target state which is 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. For this purpose, the proposed gating technique is developed to update the gating threshold with an adaptive step [40] associated with previous measurements,

$$T_k = (1 - \lambda_k)T_{k-1} + \lambda_k T_k \tag{3.2.11}$$

$$\lambda_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \sum_{j=1}^{N_{k-1}} \exp\left(- \frac{(\mathbf{z}_k^n - \mathbf{z}_{k-1}^j)^T (\mathbf{z}_k^n - \mathbf{z}_{k-1}^j)}{2\sigma_\lambda^2} \right). \tag{3.2.12}$$

The adaptive parameter λ_k is computed by the similarity between two measurement sets $\mathbf{Z}_k$ and $\mathbf{Z}_{k-1}$, 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 $\mathbf{z}_k^n \in \mathbf{Z}_k$, $n = 1, 2, ..., N_k$ as follows,

$$\mathbf{z}_{k,s}^n = \left\{ \mathbf{z}_k^n : \min_n \|\bar{\mathbf{p}}_k^n - \mathbf{p}_{k|k-1}^m\| < T_k \right\} \tag{3.2.13}$$

where the n-th measurement location is $\bar{\mathbf{p}}_k^n = [\bar{p}_{x,k}^n, \bar{p}_{y,k}^n]^T \in \mathbf{z}_k^n$, $\mathbf{p}_{k|k-1}^m = [p_{x,k|k-1}^m, p_{y,k|k-1}^m]^T \in \mathbf{x}_{k|k-1}^m, m = 1, ..., M_k$ denotes the location of the m-th predicted state, and $\|\cdot\|$ denotes the Euclidean distance. Furthermore, duplicate detections that have the same Euclidean distance to a predicted state $\mathbf{x}_{k|k-1}^m$ are discarded. This would 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,

$$\mathbf{Z}_{k,s} = \bigcup_{n=1}^{N_k} \mathbf{z}_{k,s}^n \tag{3.2.14}$$

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

$$\mathbf{Z}_{k,r} = \mathbf{Z}_k \setminus \mathbf{Z}_{k,s}. \tag{3.2.15}$$

The example pseudo-code in Algorithm 1 summarizes the proposed adap-

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. (3.2.10).
Compute the adaptive parameter λ_k with Eq. (3.2.12)
Update T_k with Eq. (3.2.11)
for *each* $\mathbf{x}^m_{k|k-1} \in \mathbf{X}_{k|k-1}$, $m = 1, ..., M_k$ **do**
    for *each* $\mathbf{z}^n_k \in \mathbf{Z}_k$, $n = 1, 2, ..., N_k$ **do**
        Obtain each survival measurement $\mathbf{z}^n_{k,s}$ with Eq. (3.2.13).
    end
end
Compute $\mathbf{Z}_{k,s}$ using Eq. (3.2.14) and $N_{k,s} = |\mathbf{Z}_{k,s}|$.
Compute $\mathbf{Z}_{k,r}$ with Eq. (3.2.15).

tive gating method. It is noteworthy that the adaptive step in [40] is originally developed to reduce the approximation error of delayed measurements, by combining the forward and backward processing (batch method). The proposed approach utilizes this adaptive step as a part of the process to enhance the gating technique for measurement classification. On the other hand, the adaptive parameter λ_k which is computed 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. In this way, the gating region can be better designed, and becomes more robust to parameter changes. Overall, the proposed gating method which fuses temporal, spatial and human size information, can dynamically control the validation region, and thereby improves the measurement refinement. It should be noted that $\mathbf{Z}_{k,r}$ comprising the new-born targets and false detections will be further distinguished via online group-structured dictionary learning in the following sections.

3.2.4 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 and is applied in the training and testing processes. The training phase is processed using measurements 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 HOG with 9 orientation bins [81]. A feature vector $\mathbf{c}_n \in \mathbb{R}^{d_c}$ that consists of transformed coefficients of the RGB colour histogram is employed for characterizing a target image patch, where d_c is the dimensionality of the RGB colour feature. Then, these features form a feature set $\mathbf{F}_c = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n] \in \mathbb{R}^{d_c \times n}$, 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 $\mathbf{F}_h = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n] \in \mathbb{R}^{d_h \times n}$, where d_h is the dimensionality of the HOG features. For further processing, the HOG and RGB colour features are concatenated to a combined feature set, $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n] \in \mathbb{R}^{(d_c+d_h)\times n}$.

Different from imposing data directly to a dictionary, an unsupervised learning method - the hierarchical K-means clustering algorithm [77] is used to learn a dictionary with the group structure information from the combined feature template $\mathbf{F} \in \mathbb{R}^{d \times n}$. This 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 [91]. To be specific, this dictionary $\mathbf{D} \in \mathbb{R}^{d \times n}$ learned with the pre-defined group structure $\mathcal{G} = \{1, ..., q\}$ has q groups, and each group consists of the same l

sub-dictionary atoms. The dictionary can be considered as a column matrix concatenation $\mathbf{D} = [\mathbf{D}_1, ..., \mathbf{D}_g, ..., \mathbf{D}_q]$ 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. 3.2.

In addition, each observed target $\mathbf{z}_k \in \mathbf{Z}_{k,r}$ at time k from the residual measurement set is cropped, and its features are extracted 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. This work intends to exploit similarities between observed signals and the group-structured dictionary in a multi-task approach, which necessitates an observation matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_{N_{k,r}}] \in \mathbb{R}^{d \times N_{k,r}}$, where $N_{k,r}$ denotes the cardinality of $\mathbf{Z}_{k,r}$.

3.2.5 Group-structured dictionary learning for birth intensity estimation

Based on the analysis in Section 3.2.3, the entire measurement set $\mathbf{Z}_k$ has been divided into the set of survival measurements $\mathbf{Z}_{k,s}$ in (3.2.14) and the residual measurement set $\mathbf{Z}_{k,r}$ in (3.2.15). 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 potential false detections $\mathbf{\Gamma}_k$ can be illustrated as,

$$\mathbf{Z}_{k,r} = \mathbf{\Gamma}_k \cup \mathbf{Z}_{k,b} \tag{3.2.16}$$

Hence, it is essential for birth intensity estimation to remove the false alarms from the remaining measurements. To achieve this, online group-structured dictionary learning is proposed to discriminate the new-born targets from false alarms or background clutter, and thus correctly estimate the birth

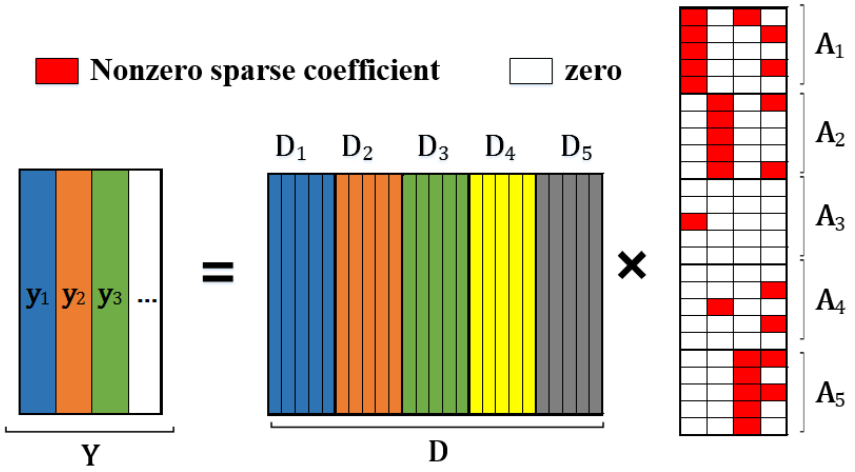


Figure 3.2. Example illustration of multi-task structured sparsity solution induced by the C-HiLasso model. The dictionary $\mathbf{D}$ consists of sub-dictionaries for five different groups, $\mathbf{D}_1, ..., \mathbf{D}_5$, with five atoms in each group. Input signals $\mathbf{Y}$ contain different measurements in the feature space. All input signals within the same class are forced to reveal the group-sparsity structure $\mathbf{A}_1, ..., \mathbf{A}_5$.

intensity. It is known that seeking the sparsity solution $\mathbf{A}$ is NP-hard in Fig. 3.2. Traditionally, the sparse coding solution $\mathbf{a}_i$ for each test target $\mathbf{y}_i$ is performed separately via *Lasso* or *Basis pursuit* [90], because different tasks would choose the dictionary atoms independently. However, the dictionary atoms of the proposed approach have been grouped with the pre-defined structure in the learning process instead of being treated as singletons. This enables multiple test targets to be represented by a few active groups of atoms, in which, a few atoms of each group are selected to be active at a time. For this study, a C-HiLasso model [117] is exploited 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 3.2.4. The sparse coefficient matrix $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_{N_{k,r}}] \in \mathbb{R}^{n \times N_{k,r}}$ can be found by optimizing the following multi-task C-HiLasso model [117],

$$\min_{\mathbf{A} \in \mathbb{R}^{n \times N_{k,r}}} \frac{1}{2} \|\mathbf{Y} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda_2 \sum_{g \in \mathcal{G}} \|\mathbf{A}_g\|_F + \lambda_1 \sum_{j=1}^{N_{k,r}} \|\mathbf{a}_j\|_1 \qquad (3.2.17)$$

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. λ_1 and λ_2 are the first and second regularization parameters respectively. The sparsity pattern is shown in Fig. 3.2, 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. In addition, the nonzero sparse codes for each measurement are gathered within a group g, as depicted in Fig. 3.2, but they would tend to scatter among groups, instead of centralizing in some single group, when candidate targets are outliers and out of the dictionary. [90].

According to the sparsity solution above, a maximum voting method is developed to select the new-born targets from the residual measurements $\mathbf{Z}_{k,r}$. A residual measurement $\mathbf{z}_{k,r}^i$ can be confirmed as a birth measurement $\mathbf{z}_{k,b}^i$ with the following equation,

$$\mathbf{z}_{k,b}^i = \left\{ \mathbf{z}_{k,r}^i : \frac{\max_{g\in\mathcal{G}} \|\mathbf{A}_{g,i}\|_1}{\|\mathbf{A}_{:,i}\|_1} \geq \varepsilon \right\} \tag{3.2.18}$$

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

$$\mathbf{Z}_{k,b} = \bigcup_{i=1}^{N_{k,r}} \mathbf{z}_{k,b}^i. \tag{3.2.19}$$

Then, the false measurements can be removed as follows,

$$\mathbf{Z}_{k,r} \setminus \mathbf{Z}_{k,b} = \emptyset. \tag{3.2.20}$$

Let α_u record each valid index i, where $u = 1, ..., J_k$ $(J_k > 0)$ and J_k is the cardinality of $\mathbf{Z}_{k,b}$, and $\Omega = \{\alpha_1, ..., \alpha_{J_k}\}$ 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,

$$\theta_u = \underset{g \in \mathcal{G}}{\operatorname{argmax}} \|\mathbf{A}_{g,\alpha_u}\|_1, \tag{3.2.21}$$

then each obtained θ_u consists of a set $Q = \{\theta_1, ..., \theta_{J_k}\}$. It is worth noting that 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,

$$\eta_u = \frac{1}{l}\|\mathbf{A}_{\theta_u,\alpha_u}\|_1. \tag{3.2.22}$$

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

$$\gamma_{k|k-1}(\mathbf{x}) = \frac{1}{J_k}\sum_{u=1}^{J_k} \frac{1}{(2\pi\sigma_b)^{1/2}} \exp\Big(- \frac{\eta_u}{2\sigma_b^2}\Big). \tag{3.2.23}$$

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 (3.2.6), and meanwhile both survival measurement set $\mathbf{Z}_{k,s}$ and birth measurement set $\mathbf{Z}_{k,b}$ distinguished by the proposed method will be used to realize the weights update of (3.2.7) (3.2.8) in Section 3.2.2.

3.2.6 Dictionary update with SimCO algorithm

Since the learned dictionary $\mathbf{D}$ is pre-trained, and only contains the target features for a few relevant frames, using such an off-line dictionary may not be able to robustly deal with the target appearance variations [89]. In order to improve the robustness of the learned dictionary, the SimCO algorithm proposed by W. Dai et al. [118] is utilized for the dictionary update, because the key characteristics of this algorithm can update an arbitrary subset of

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 $\mathcal{G} = \{1, ..., q\}$, 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* $\mathbf{z}_{k,r}^{i}, i = 1, 2, ..., N_{k,r}$ **do**
  Obtain each birth measurement $\mathbf{z}_{k,b}^{i}$ with Eq. (3.2.18).
end
Compute the measurement set of new born targets $\mathbf{Z}_{k,b}$ with Eq. (3.2.19).
Remove the false measurements with Eq. (3.2.20).
if $J_k > 0$ **then**
  for *each* $\mathbf{z}_{k,b}^{u}, u = 1, 2, ..., J_k$ **do**
    Compute the voting index θ_u with Eq. (3.2.21).
    Calculate the average of selected sparse codes η_u with Eq. (3.2.22).
  end
  Compute the birth intensity $\gamma_{k|k-1}(\mathbf{x})$ with Eq. (3.2.23).
  Compute the structure patterns Q and Ω.
end

atoms in the dictionary $\mathbf{D}$. In general, the dictionary update problem with the SimCO algorithm can be written as [118],

$$\underset{\mathbf{D} \in \mathbb{R}^{d \times n}}{\operatorname{argmin}} f(\mathbf{D}) = \underset{\mathbf{D} \in \mathbb{R}^{d \times n}}{\operatorname{argmin}} \left(\min_{\mathbf{A} \in \mathbb{R}^{n \times N_{k,r}}} \|\mathbf{Y} - \mathbf{D}\mathbf{A}\|_F^2 \right) \tag{3.2.24}$$

where the dictionary matrix $\mathbf{D}$ contains unit ℓ_2-norm columns, and the sparse coding matrix $\mathbf{A}$ is obtained from (3.2.17). Instead of updating all the atoms of $\mathbf{D}$, the dictionary update can be performed through the structure pattern Q and Ω is determined by the 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 $\mathbf{B} = [\mathbf{D}_{\theta_1}, ..., \mathbf{D}_{\theta_u}, ..., \mathbf{D}_{\theta_{J_k}}] \in \mathbb{R}^{d \times (J_k \times l)}$ 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 $\mathbf{D}$, whereas $\mathbf{B}^c$ is formed by other sub-matrices of $\mathbf{D}$ indexed by Q^c remain unaltered, where Q^c is a set complementary to Q over $\mathcal{G}$. Similarly, define $\mathbf{K} = [\mathbf{A}_{\theta_1}^T, ..., \mathbf{A}_{\theta_u}^T, ..., \mathbf{A}_{\theta_{J_k}}^T]^T \in \mathbb{R}^{(J_k \times l) \times h}$, where $\mathbf{A}_{\theta_u}$ is the sub-matrix of $\mathbf{A}$ containing l rows of the sparse coefficients, 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 [118], the following equation can be defined,

$$\mathbf{Y}_r = \mathbf{Y} - \mathbf{B}^c \mathbf{K}^c \tag{3.2.25}$$

Since $\mathbf{Y} - \mathbf{DA} = \mathbf{Y}_r - \mathbf{BK}$, then the dictionary update problem in (3.2.24) can be rewritten as,

$$\underset{\mathbf{B} \in \mathbb{R}^{d \times (J_k \times l)}}{\operatorname{argmin}} f(\mathbf{B}) = \underset{\mathbf{B} \in \mathbb{R}^{d \times (J_k \times l)}}{\operatorname{argmin}} \left(\min_{\mathbf{K} \in \mathbb{R}^{(J_k \times l) \times h}} \|\mathbf{Y}_r - \mathbf{BK}\|_F^2 \right) \tag{3.2.26}$$

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

$$f(\mathbf{B}) = \min_{(\mathbf{K}_\Omega)_{:,u}} \sum_{u=1}^{J_k} \|(\mathbf{Y}_\Omega)_{:,u} - \mathbf{B}(\mathbf{K}_\Omega)_{:,u}\|_2^2 \tag{3.2.27}$$

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 is applied to provide the search direction $\mathbf{E}$ for updating $\mathbf{B}$, which is defined as follows,

$$\begin{aligned} \mathbf{E} &= -\nabla f(\mathbf{B}) \\ &= -2(\mathbf{Y}_\Omega - \mathbf{BK}_\Omega)\mathbf{K}_\Omega^T \end{aligned} \tag{3.2.28}$$

The line search path for this study was created by using the product of

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 B^k ← D^k with the Q structure pattern.
    Find a proper step size δ with the method of golden section
     search [118].
    Compute the search direction, with (3.2.28) and (3.2.29).
    Update B^k → B^{k+1} with (3.2.30).
    Update the dictionary D^{k+1} ← B^{k+1}.
end
```

Grassmann manifolds that was detailed in [118]. Let $\mathbf{e}_j$ be the j-th column of $\mathbf{E}$. the following equation is defined,

$$\bar{\mathbf{e}}_j = \mathbf{e}_j - \mathbf{B}_{:,j}\mathbf{B}_{:,j}^T\mathbf{e}_j, \tag{3.2.29}$$

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

$$\mathbf{B}_{:,j}(\delta) = \begin{cases} \mathbf{B}_{:,j} & \text{if } \|\bar{\mathbf{e}}_j\|_2 = 0, \\ \mathbf{B}_{:,j}\cos(\|\bar{\mathbf{e}}_j\|_2\delta) + \left(\frac{\bar{\mathbf{e}}_j}{\|\mathbf{e}_j\|_2}\right)\sin(\|\bar{\mathbf{e}}_j\|_2\delta) & \\ & \text{if } \|\bar{\mathbf{e}}_j\|_2 \neq 0. \end{cases} \tag{3.2.30}$$

where the step size $\delta \in \mathbb{R}^+$ is properly chosen via the method of golden section search [118]. Besides, the dictionary update stage is summarized in Algorithm 3.

3.3 Experiments

In this section, the datasets used in this experiment are initially introduced, and then the parameter settings are explained. To study the effectiveness of the proposed tracking method, the individual contribution of each proposed

component to the tracking system is validated by conducting experiments on five commonly-used datasets. Also, quantitative comparisons are made between the proposed method and a range of state-of-the-art tracking methods on the MOTChallenge benchmark, as well as including a discussion on the runtime performance.

3.3.1 Datasets

The proposed tracking method is firstly validated on five commonly used video sequences: CAVIAR-EnterExitCrossingPaths1cor (CAVIAR) [119], PETS2009-View001-S2L1 (PETS2009) [108], TUD-Stadtmitte [109], TUD-Campus [109], and TUD-Crossing [109], in order to conduct the analysis of different contribution components.

In addition, the MOTChallenge Benchmark dataset[1] is used to 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. All the video sequences are only suitable for pedestrian tracking. The testing video sequences are used for performance evaluation in comparisons with other recent trackers.

3.3.2 Parameter settings

For this study, 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 $\mathbf{H} = [\mathbf{I}_2, \mathbf{0}_2, \mathbf{0}_2; \mathbf{0}_2, \mathbf{0}_2, \mathbf{I}_2]$ are adopted from the work in [56], 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 [9], 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

[1] https://motchallenge.net/

Table 3.1. 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

(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, which are used for fine-tuning system parameters, and remain fixed for all testing sequences. The system parameters used for all testing video sequences are summarized in Table 3.1.

3.3.3 Effectiveness evaluation of proposed contributions

To achieve a better understanding of the contribution of the individual system component, the OSPA performance measure is used to compare and evaluate the different stage tracking performance of the proposed approach on five commonly used video datasets. In the experiment, the particle PHD filter framework is expressed as SMC-PHD. The proposed method with the adaptive gating and online group-structured dictionary learning are respectively denoted as SMC-PHD-AG and SMC-PHD-Online GSDL.

Table 3.2 shows the average OSPA performance comparison computed on all five sequences using different stage methods. Overall, the highest performance which returns the smallest OSPA error is reported with using both proposed terms regardless of different scenarios. In the meantime the performance can be effectively decreased by removing each individual term.

Table 3.2. Average OSPA (pixel) performance comparison of different system component on five video sequences. The best results are shown in bold.

Dataset	SMC-PHD [44]	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**

By comparing the individual improvement of each proposed contribution, from the third and fourth columns in Table 4.2, it can be observed that performance is relatively more improved by employing SMC-PHD-Online GSDL than in SMC-PHD-AG. This can be explained that the proposed SMC-PHD-AG is fundamentally capable of offering prior information for measurement classification and handling the false detections, whereas the SMC-PHD-Online GSDL method intends to further strengthen the ability of discriminating the targets from the noisy environment, as well as resolving the occlusions.

For the CAVIAR dataset, the combined method achieves the highest improvement of 68.5% over 5 video sequences, where the average OSPA value is reduced from 48.26 to 15.16. This is because the proposed tracker can effectively eliminate a large number of false alarms caused by the raw background subtraction results in the less crowded CAVIAR dataset. In the TUD-Crossing dataset, the tracking accuracy of the proposed method is improved by 47.4% compared with the SMC-PHD method. However, more errors are produced from the proposed method, which is less capable to deal with long-term occlusions without an occlusion reasoning technique.

On the other hand, the effectiveness evaluation can be also visually seen from Fig. 3.4, where each curve demonstrates the change of tracking performance over the entire sequence. From the results, it is clear to see the OSPA performance using SMC-PHD-AG method (green curve) includes instability,

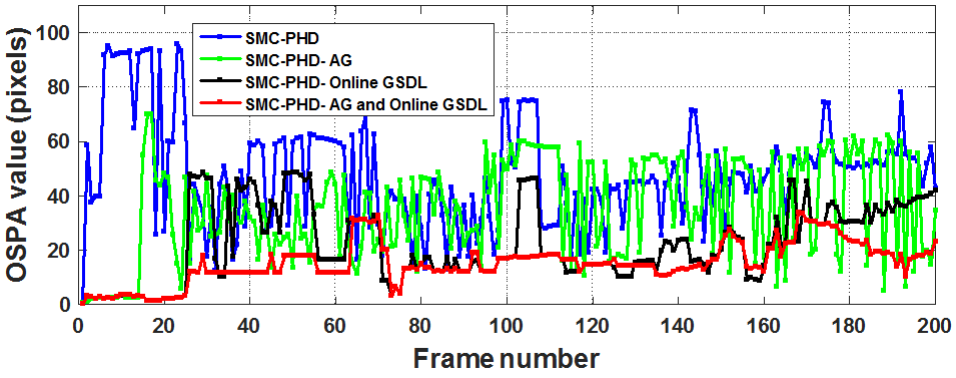
Table 3.3. Average OSPA (pixel) comparison between proposed method and different state-of-the-art methods on five video sequences. The best results are shown in bold

Dataset	SMC-PHD method [44]	MB method [46]	SFM-PHD method [9]	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**

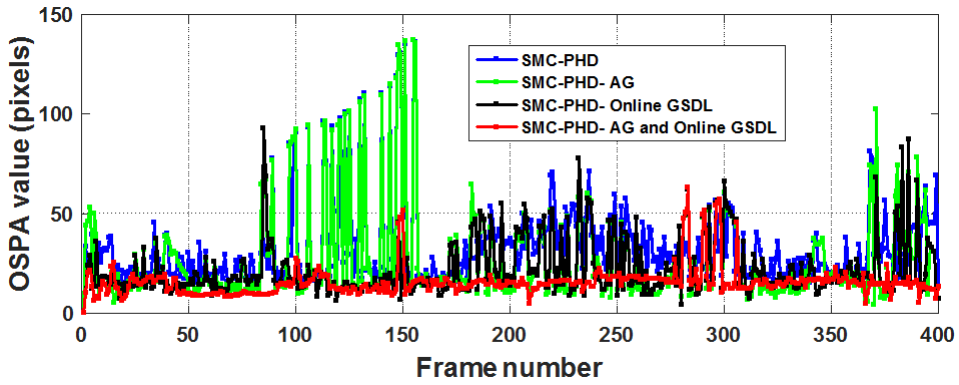
which can be attributed to the inefficiency of dealing with the complex target interactions especially within a crowded scene such as the PETS2009-S2L1 dataset. It can be observed from the black curve in Fig. 3.4, the proposed SMC-PHD-Online GSDL apparently performs better and more robustly than that in SMC-PHD-AG. The improvement results 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, the OSPA value of the combined approach (red curve) is generally shown to be much lower than other baseline methods in most frames, and also the steady performance confirms the robustness of the combined method.

Table 3.3 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 [44], background subtraction based multi-Bernoulli filter (MB) [46], and social force model based particle PHD filter (SFM-PHD) [9]. The comparable results above emphasise the fact that the proposed tracker reports the best OSPA performance over five sequences. Specifically, the proposed tracker outperforms the tracker in [46] with the improvement between approximately 31.8% and 60.2% over five sequences. Moreover, the SFM-PHD method [9] performs worse than the proposed method, which means that the group-structured sparsity of the proposed tracker shows clear advantage over the OCSVM

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

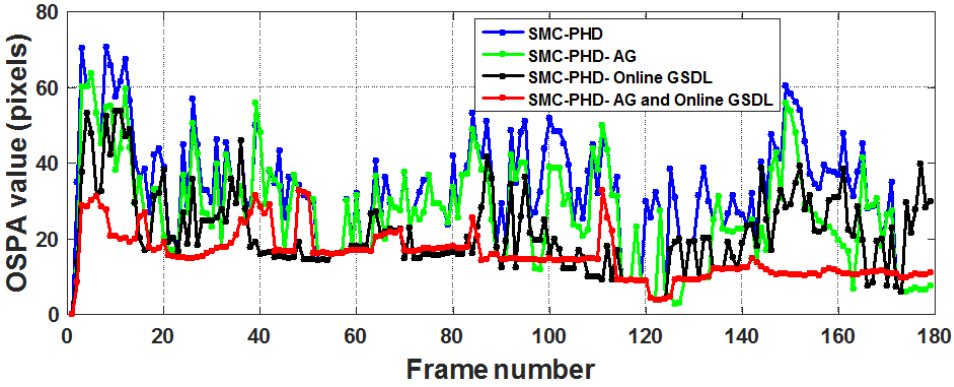


(a) OSPA evaluation on CAVIAR dataset

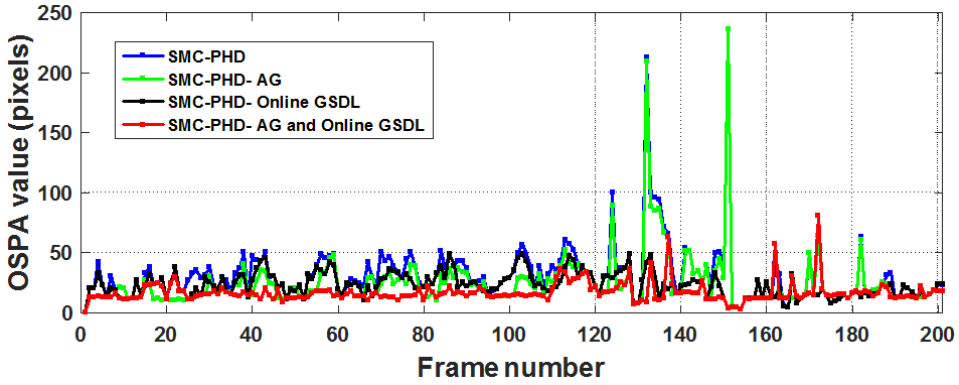


(b) OSPA evaluation on PETS2009-S2L1 dataset

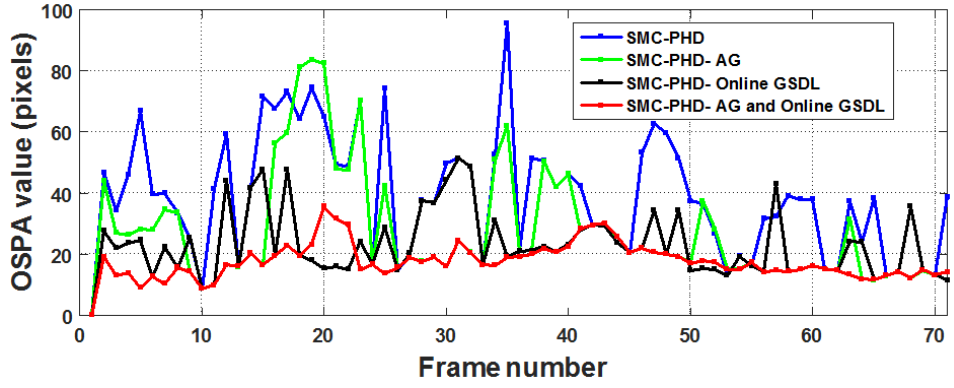
Figure 3.3. OSPA evaluation for different stages of the proposed tracking system on CAVIAR and PETS2009 datasets.



(a) OSPA evaluation on TUD-Stadtmitte dataset



(b) OSPA evaluation on TUD-Crossing dataset



(c) OSPA evaluation on TUD-Campus dataset

Figure 3.4. OSPA evaluation for different stages of the proposed tracking system on TUD datasets.

3.3.4 Evaluations on MOTChallenge

In this section, the proposed method denoted by PHD-GSDL is evaluated on the test set of the 2D MOTChallenge 2015 Benchmark [3] and MOT17 Challenge [4]. It is noteworthy that public detections and the centralized evaluation tool provided by the benchmark website are used to make a fair comparison between methods. Tables 3.4 and 3.5 show the quantitative comparisons with a number of state-of-the-art tracking methods. These include online tracking approaches: MDP [120], SCEA [121], RMOT [63], GMPHD_15 [60], EAMTTPub [29], oICF [122], GM_PHD [27], and GM-PHD_KCF [66]. Offline (batch) tracking approaches are also included as: NOMT [20], QuadMOT [18], JointMC [19], SiameseCNN [85], DCO_X [22], jCC [19], EDMT17 [51], IOU17 [99] and DP_NMS [17].

As illustrated in Table 3.4, 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 3.5 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, the proposed 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.

Table 3.4. Quantitative comparison with other state-of-the-art methods on the 2D MOTChallenge 2015 benchmark with public detections. The proposed method is denoted as 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 [123]	Online	**35.0**	72.6	1.5	11.4%	42.2%	8,455	31,140	**358**	1,267	4.6
TDAM [124]	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 [120]	Online	30.3	71.3	1.7	13.0%	**38.4%**	9,717	32,422	680	1,500	1.1
SCEA [121]	Online	29.1	71.1	**1.0**	8.9%	47.3%	**6,060**	36,912	604	**1,182**	6.8
oICF [122]	Online	27.1	70.0	1.3	6.4%	48.7%	7,594	36,757	454	1,660	1.4
EAMTTPub [29]	Online	22.3	70.8	1.4	5.4%	52.7%	7,924	38,982	833	1,485	12.2
RMOT [63]	Online	18.6	69.6	2.2	5.3%	53.3%	12,473	36,835	684	1,282	7.9
GMPHD_15 [60]	Online	18.5	70.9	1.4	3.9%	55.3%	7,864	41,766	459	1,266	**19.8**
JointMC [19]	Batch	**35.6**	71.9	1.8	**23.2%**	39.3%	10,580	**28,508**	457	969	0.6
QuadMOT [18]	Batch	33.8	**73.4**	1.4	12.9%	**36.9%**	7,898	32,061	703	1,430	3.7
NOMT [20]	Batch	33.7	71.9	1.3	12.2%	44.0%	7,762	32,547	**442**	823	11.5
SiameseCNN [85]	Batch	29.0	71.2	**0.9**	8.5%	48.4%	**5,160**	37,798	639	1,316	**52.8**
DCO_X [22]	Batch	19.6	71.4	1.8	5.1%	54.9%	10,652	38,232	521	**819**	0.3

Table 3.5. Quantitative comparison with other state-of-the-art methods presented in the MOT2017 Challenge benchmark using public detections. The proposed method is denoted as 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 [27]	Online	36.2	76.1	1.3	4.2%	56.6%	23,682	328,526	8,025	11,972	**38.4**
GMPHD_KCF [66]	Online	30.5	74.3	6.1	9.6%	41.8%	107,802	277,542	6,774	**7,833**	3.3
jCC [19]	Offline	**51.2**	75.9	1.5	20.9%	37.0%	25,937	247,822	**1,802**	**2,984**	1.8
EDMT17 [51]	Offline	50.0	**77.3**	1.8	**21.6%**	**36.3%**	32,279	**247,297**	2,264	3,260	0.6
IOU17 [99]	Offline	45.5	76.9	1.1	15.7%	40.5%	19,993	281,643	5,988	7,404	**1522.9**
DP_NMS [17]	Offline	43.7	76.9	**0.6**	12.6%	46.5%	**10,048**	302,728	4,942	5,342	137.7

Figure 3.5. Qualitative performance of the 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.

Fig. 3.5 depicts some selected qualitative tracking results produced by the proposed method on the test video sequences of the 2D MOTChallenge 2015. It can be observed that some pedestrians with similar appearances that are partially or even almost fully occluded are successfully tracked through the proposed tracker. This can be attributed by the online update mechanism with the SimCO algorithm giving the benefits of dealing with the target appearance changes.

In order to further demonstrate the advantages of the proposed RFS-based tracker, the proposed method is compared with other recent PHD filter based methods, EAMTTPub [29], GMPHD_15 [60], GM_PHD [27] and GMPHD_KCF [66]. As compared to [29] and [60] on MOT15 dataset, the MOTA of proposed method is improved by 8.2% and 12.0% respectively. More importantly, the proposed method outperforms the algorithms in [27]

and [66] with large margins on the MOT17 dataset. All above evaluations indicate that the 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 the proposed 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 the proposed 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.

3.3.5 Runtime performance

All the 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. It is found that most of the running time of the proposed 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 3.4 and 3.5, where the runtime of the proposed 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 the proposed tracker runs slower than the methods in [29], [60] and [27] using same the PHD

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

filter framework, it returns a significant improvement regarding the tracking accuracy.

3.4 Summary

In summary, this chapter contributed to improving the particle PHD filter for multiple human tracking by proposing two enhancements: adaptive gating and online group-structured dictionary learning. The adaptive-gating strategy which adaptively updates the gating threshold was firstly developed to refine the measurement set of survival targets thereby strengthening the measurement-driven mechanism. The group-structured dictionary learning was exploited to improve the discriminative power of sparse coding. By taking advantage of the improved sparsity solution, the proposed maximum voting method was demonstrated to better distinguish the birth measurements from noisy measurements. Additionally, the SimCO algorithm with the proposed structure pattern was feasible to efficiently implement the dictionary update stage for handling the target appearance changes.

The OSPA evaluation in Section 3.3.3 showed the contributions of different proposed components. For the CAVIAR dataset, the full proposed method achieves the highest improvement of 68.5% compared with the baseline method, where the average OSPA value is reduced from 48.26 to 15.16. For the TUD-Crossing dataset, the tracking accuracy of the proposed method is improved by 47.4% compared with the baseline method. Compared with different state-of-the-art RFS-based methods, the proposed method achieved lowest OSPA scores across five datasets. Specifically, the proposed tracker outperforms the tracker in [46] with the improvement between approximately 31.8% and 60.2%. Evaluations on the MOTChallenge benchmark were further shown in Section 3.3.4 to confirm the improved tracking performance compared to state-of-the-art methods presented on the leaderboard, where

the proposed method improves the MOTA by 8.2% and 12.0% as compared to [29] and [60] on the MOT15 dataset. The proposed method also achieves the second best online tracker ranked on the leaderboard of MOT17 Challenge, as well as performs well in terms of tracking precision (high MOTP 77.2%), fewer targets lost (low ML 35.6%) and more targets tracked (high MT 17.1%).

Although the proposed tracking method has achieved improved tracking performance, target appearance representation using hand-crafted features may be limited to identify targets in congested environments. The next chapter focuses on improving the measurement selection in the GM-PHD filter, as well as exploiting CNN features with person re-identification to better model the target appearances.

Chapter 4

MEASUREMENT-DRIVEN GM-PHD FILTER WITH ENHANCED DETECTION RELIABILITY FOR ONLINE MULTIPLE HUMAN TRACKING

4.1 Introduction

In the previous chapter, online group-structured dictionary learning has been developed to improve the birth intensity estimation in the filtering process. This chapter continues to exploit accurate measurement selection of detections which is aimed at addressing the noisy detections but in a more efficient way prior to the tracking process. It is known that handling inaccurate detections plays a key element in the online tracking pipeline. Existing work in [29] has attempted to perform a two-stage measurement selection to address inaccurate detections for online tracking. This selection approach

typically comprises with a confidence score sorting and a non-maximal suppression (NMS), which can be efficiently embedded with most of the online tracking methods. The proposed work falls into this two-stage selection approach. Different from [29], the proposed approach focused on improving each stage by introducing two terms: enhanced confidence rescoring and Soft-aggregated non-maximal suppression (Soft-ANMS). For the first stage, an enhanced confidence rescoring strategy is proposed to remove originally high scored false positives as well as preserving more true detections by exploiting the classification power of mask R-CNN [12] from a global-to-local manner. For the second stage, a novel suppression method is devised, namely Soft-ANMS. The idea is to aggregate the sum of intersection over areas (SIOA) in [125] with the intersection over union (IOU) measure to further refine duplicate detections.

On the other hand, previous chapter employed the hand-crafted features for target appearance representation, while deep convolutional neural networks (CNNs) recently have remarkably outperformed heuristic, hand-crafted features in terms of appearance modelling [82]. Therefore, in this chapter, CNN features based on the person re-identification are also studied to model the target appearances, in order to improve the measurement grouping in the GM-PHD filtering processing. In addition, a track management scheme is used after the PHD updating step, to determine the final tracking results. This chapter targets to fulfill the third and fourth objectives of the thesis, which are the enhanced detection reliability for human tracking based video analytics presented in [32], and the enhanced GM-PHD filter using CNN-based weight penalization for multi-target tracking published in [31].

The rest of this chapter is presented as follows: Section 4.2 demonstrates the enhanced detection reliability as two parts: enhanced confidence rescoring and Soft-ANMS. Section 4.3 describes the tracking process using

measurement-driven GM-PHD filter. Section 4.4 explains the experiments, and discusses the results.

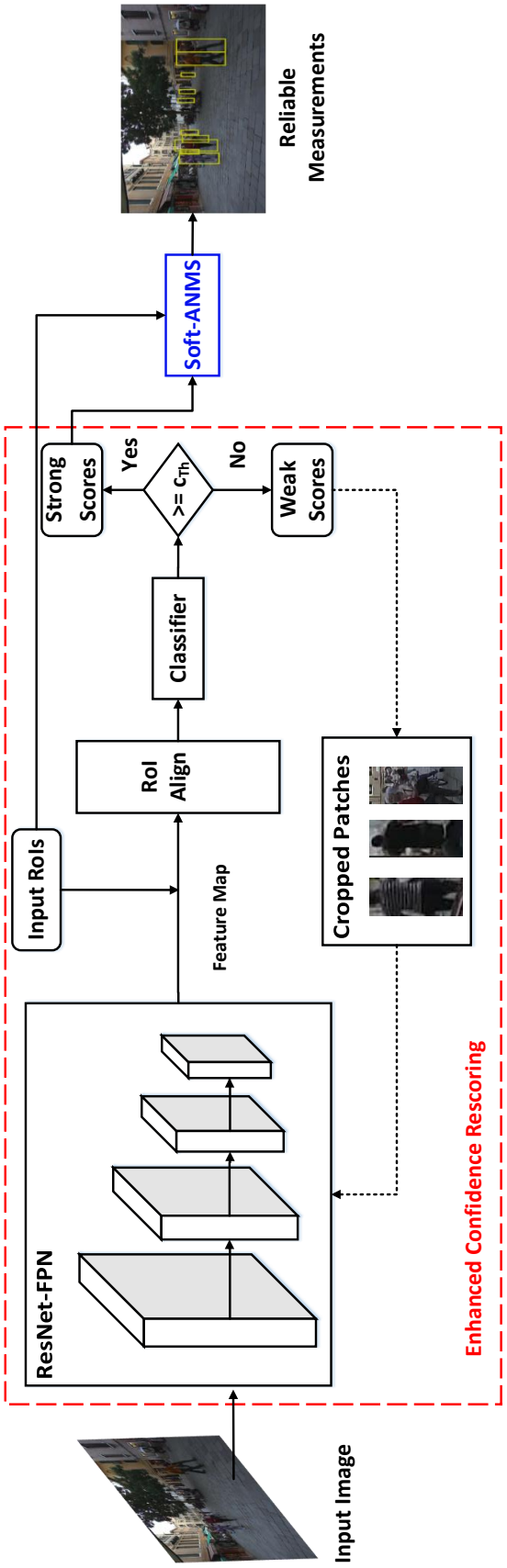


Figure 4.1. Overview of the proposed enhanced detection reliability (EDR) for measurement selection of detections.

4.2 Enhanced detection reliability

4.2.1 Overview of proposed approach

The overview of the proposed enhanced detection reliability for robust measurement selection of detections is shown in Fig. 4.1, which mainly consists of two steps: enhanced confidence rescoring and Soft-ANMS. An input image is firstly fed into ResNet-FPN backbone network [126] to generate multi-level feature maps. Then feature maps associated with the input region of interests (RoIs) are fed into a RoIAlign layer [12] to generate RoI features, which are attached with classifiers to predict the confidence scores. Then a threshold c_{Th} is applied on the obtained scores to categorize the strong and weak RoIs. Cropped patches which are produced by cropping the weak RoIs from the input image, are fed into inference process for the score update. After the enhanced confidence rescoring, only strong RoIs will be moved into Soft-ANMS for the final selection. This Soft-ANMS algorithm is designed as a pruning step to remove the highly scored duplicated detections by aggregating the measures of SIOA with IOU. The detailed implementations on each block of the proposed approach are presented in the following sections.

4.2.2 Enhanced confidence rescoring

Looking through the literature, various online tracking approaches [29,33,59] did not fully exploit the relationship between confidence scores and true detections, but heuristically set up a confidence threshold to select candidates for tracking. This may implicitly create an underlying problem in the beginning of tracking task. To prove this, the misalignment between confidence scores and true positives in the MOT16 Challenge benchmark [4] is analyzed, as shown in Fig. 4.2. It is clear that the number of true positives is not well correlated with the original confidence scores. Large quantities of true positives are aligned with lower scores, which leaves a critical uncertainty in

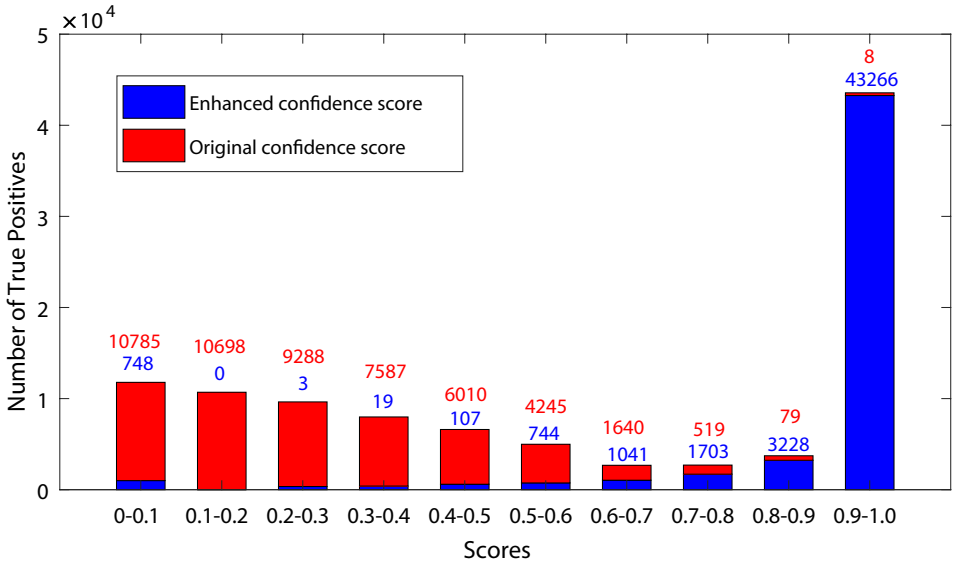


Figure 4.2. Illustration of the number of true positives with different score ranges. The original confidence scores obtained from the benchmark have been normalized to the range of $[0, 1]$ for comparison with the enhanced confidence scores. In this work, a true positive (TP) is not considered if its IoU with the ground truth is less than 0.5.

the selection process. Likewise, this misalignment can also affect the subsequent NMS or Soft-NMS [127] processing. Because these algorithms often rank the detections based on the confidence scores, iteratively suppressing the duplicate detections with a pre-defined threshold.

To tackle this misalignment problem, a global-to-local confidence rescoring strategy is proposed by exploiting the classification power of Mask R-CNN [12], which has achieved convincing performance in a few vision tasks. Fig. 4.1 shows the major steps of the proposed strategy. From a global perspective, an input image $\mathbf{I}$ is firstly fed into Resnet-FPN backbone network [126] for feature extraction. This network outputs multi-level feature maps with both rich semantic and spatial information from an input image. Input RoIs $\mathbf{Z}$ are provided from the original detection results. Each RoI is defined as a measurement vector, $\mathbf{z} = [\bar{p}_x, \bar{p}_y, \bar{\omega}, \bar{h}]^T$, which contains the position and size information. Then an assignment strategy in [16] is adopted to associate each scale-variant RoI with a specified level of feature maps.

Followed by a RoIAlign layer [12], the RoI features can be thus achieved. These features are finally attached with a classifier for confidence rescoring $c = p(g|\mathbf{I}, \mathbf{z})$, where g denotes a true human detection. In this work, these scores are used to determine whether input RoIs are belong to human detections or the background.

Then, a threshold c_{Th} is applied on the obtained scores to categorize the strong and weak RoIs, as shown in Fig. 4.1. For the weak RoIs, they are not performed with an immediate elimination, but moved to another verification. Because it can be conjectured that it is likely to have some true detections are underrated as weak RoIs. To this end, weak RoIs are firstly used to directly crop patches from the input image, and then they are resized and fed again into the network for rescoring. The purpose is to allow the network to more focus on the local region context inside cropped patches, which is specifically helpful for identifying the RoIs with smaller sizes [128]. In this way, weak RoIs can be further processed by exploring the local contextual information. Thereinto, some of them can be elevated to strong RoIs, if their updated scores are greater than the threshold c_{Th}. As a consequence, only the strong RoIs $\mathbf{Z}_R$ are proceeded to the Soft-ANMS for further enhancement. It is worth noting the proposed rescoring method is built upon the inference model of the Mask R-CNN without requiring any heavy training process. Results with enhanced confidence scores in Fig. 4.2 verify the proposed rescoring strategy is able to provide a better correspondence between the number of true positives and confidence scores, which is helpful to supply more reliable measurements for the filtering process.

4.2.3 Soft-ANMS

In spite of achieving a set detections $\mathbf{Z}_R$ with strong scores from previous section, it is likely that there still exist some duplicate detections, which can yield a lower precision. To address this, the second step is to enhance

the detection reliability which is built upon commonly-used NMS algorithm. The NMS has been an essential post processing step to reduce false positives in most state-of-the-art object detectors. Recently, an improved version namely Soft-NMS [127] has been proposed to decay the confidence scores of overlapped detections rather than directly remove them.

However, most of NMS related techniques merely use the IOU in (4.2.1) to measure the target closeness. This could be constraint if overlapped detections have large size variations, as depicted in Fig. 4.3 (a).

$$IOU = \frac{Area(a) \bigcap Area(b)}{Area(a) \bigcup Area(b)} \tag{4.2.1}$$

Motivated by the work in [125], a measure of SIOA was developed to deal with merging targets in the group management.

$$SIOA = 0.5 \times \left(\frac{Area(a) \bigcap Area(b)}{Area(a)} + \frac{Area(a) \bigcap Area(b)}{Area(b)} \right) \tag{4.2.2}$$

Accordingly, the SIOA measure presented in (4.2.2) is aggregated as a complement with the IOU in the Soft-NMS [127] to formulate an improved pruning step namely Soft-ANMS to refine each confidence score c_i, which is expressed,

$$c_i = \begin{cases} c_i(1 - IOU(\mathbf{z}_{max,R}, \mathbf{z}_{i,R})), & \text{if } IOU(\mathbf{z}_{max,R}, \mathbf{z}_{i,R}) > U_T \\ c_i(1 - SIOA(\mathbf{z}_{max,R}, \mathbf{z}_{i,R})), & \text{if } IOU(\mathbf{z}_{max,R}, \mathbf{z}_{i,R}) \leq U_T \\ & \quad \& \; SIOA(\mathbf{z}_{max,R}, \mathbf{z}_{i,R}) > S_T \\ c_i, & \text{otherwise} \end{cases} \tag{4.2.3}$$

where U_T and S_T respectively are the thresholds of IOU and SIOA measures.

$\mathbf{z}_{max,R}$ denotes a detection with the maximum score. The above improved penalty function can mitigate the aforementioned issue in Fig. 4.3 (a) by linearly decreasing the scores if detections have a higher SIOA but lower IOU with $\mathbf{z}_{max,R}$. Fig. 4.3 (b) visually reveals that the advantage of the proposed approach. After performing the Soft-ANMS, finally a reliable measurement set $\mathbf{Z}_{+} = \{\mathbf{z}_R : c \geq \sigma\}$ can be achieved for the tracking process, where σ denotes the threshold for the final selection.

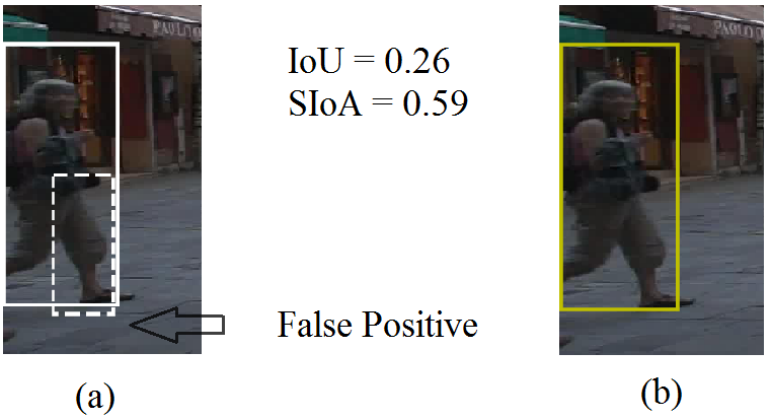


Figure 4.3. Illustration of detection suppression with different overlapping measures. (a) fails to suppress the false positive by only using IOU measure. The threshold of IOU is $U_T = 0.3$. (b) shows the proposed approach effectively eliminates the false positive by leveraging both measures of IOU and SIOA. The threshold of SIOA is $S_T = 0.5$ (better viewed in color version).

4.3 Measurement-driven GM-PHD visual tracker

In this work, a practical GM implementation [103] of the PHD recursion is used to formulate the tracking model. The GM-PHD filter proposed by Vo and Ma [103] introduces a closed-form solution to the PHD recursion. The posterior PHD intensity function can be represented by a sum of weighted Gaussian components that are propagated analytically in time [129]. Using the GM-PHD filtering framework, a set of target states at time k is modelled as: $\mathbf{X}_k = \{\mathbf{x}_k^1, ..., \mathbf{x}_k^{M_k}\}$, and a set of reliable measurement as

$\mathbf{Z}_{k,+} = \{\mathbf{z}_{k,+}^1, ..., \mathbf{z}_{k,+}^{N_k}\}$ where M_k and N_k denote the number of targets and measurements at time k.

4.3.1 Prediction

The motion of each target in the surveillance region from time $k-1$ to time k follows a linear Gaussian dynamical model [103],

$$f_{k|k-1}(\mathbf{x}|\xi) = \mathcal{N}(\mathbf{x}; \mathbf{F}\xi, \mathbf{Q}_{k-1}) \tag{4.3.1}$$

where $\mathbf{F}$ is the state transition matrix which models target propagation, $\mathbf{Q}_{k-1}$ is the process noise covariance matrix, and ξ is the previous state. A posterior intensity ν_{k-1} in a Gaussian mixture form at time $k-1$ is given as,

$$\nu_{k-1}(\mathbf{x}) = \sum_{j=1}^{J_{k-1}} w_{k-1}^j \mathcal{N}(\mathbf{x}; \mathbf{m}_{k-1}^j, \mathbf{P}_{k-1}^j) \tag{4.3.2}$$

where $\mathcal{N}(\cdot; \mathbf{m}, \mathbf{P})$ denotes a Gaussian component with mean $\mathbf{m}$ and covariance matrix $\mathbf{P}$, J_{k-1} is the number of Gaussian components at time $k-1$, and w_{k-1}^j is the corresponding weight of the j-th Gaussian component [6]. Then, a labelling method in [129] is used to manage the target identities, which is to assign a unique label I_{k-1}^j as a hidden identity to individual Gaussian components to achieve an identity set $\mathcal{I}_{k-1} = \{I_{k-1}^1, ..., I_{k-1}^{J_{k-1}}\}$ [6].

For the current time step k, the prediction is performed independently for each individual target survived from the previous time, which is given by [103],

$$\nu_{k|k-1,S}(\mathbf{x}) = e_{k|k-1} \sum_{j=1}^{J_{k-1}} w_{k-1}^j \mathcal{N}(\mathbf{x}; \mathbf{m}_{k|k-1,S}^j, \mathbf{P}_{k|k-1,S}^j) \tag{4.3.3}$$

$$\mathbf{m}_{k|k-1,S}^j = \mathbf{F}\mathbf{m}_{k-1}^j \tag{4.3.4}$$

$$\mathbf{P}_{k|k-1,S}^j = \mathbf{Q}_{k-1} + \mathbf{F}\mathbf{P}_{k-1}^j(\mathbf{F})^T \tag{4.3.5}$$

where $\nu_{k|k-1,S}(\mathbf{x})$ represents the predicted intensity of survival targets, and $e_{k|k-1}$ is the survival probability. The identities of Gaussian components remain unaltered during the prediction, $I^j_{k|k-1,S} = I^j_{k-1},\ j = 1, ..., J_{k-1}$.

4.3.2 Measurement grouping

In the previous chapter, an adaptive gating technique based on spatial constraints was developed to perform the measurement classification, which is intended to improve the efficiency in the updating step. While the performance of this design may degrade when targets move in close proximity, because only relying on target motion is not robust enough to address the target ambiguity in the image plane. For instance, there are two targets moving near each other, one of which is matched with multiple effective measurements while the other suffers from this special type of miss-detection [130].

To avoid misuse of ambiguous measurements in the filtering process, the measurement-driven mechanism in here is not only limited to the spatial information but also extended with target appearance models, which is assisted by CNN features based on person re-identification (ReID). These ReID features are extracted from a deep neural network [131] which is pre-trained on a large-scale person ReID dataset [132]. This CNN model is generally constructed with two convolutional layers, followed by a max pooling layer and six residual layers. The employed network uses the same input size of 128×64 RGB image patch for feature extraction.

Then the Bhattacharyya distance [40] is utilized to calculate the following similarity score in the feature space between the j-th predicted target and n-th measurement at time k,

$$\Lambda^A_k(j,n) = \frac{1}{\sqrt{2\pi\sigma^2_\theta}} \exp\left(- \frac{\{S_k(j,n)\}^2}{2\sigma^2_\theta} \right) \tag{4.3.6}$$

where

$$S_k(j,n) = \sqrt{1 - (\mathbf{f}_k^j)^T \mathbf{d}_k^n}. \tag{4.3.7}$$

$\mathbf{f}_k^j$ and $\mathbf{d}_k^n$ are the feature vectors of the j-th predicted target and the n-th measurement respectively. Features are normalized using L2 normalization. σ_θ^2 denotes the variance of the similarity score. Therefore the overall association cost is jointly constructed by the CNN based appearance models and spatial constraints,

$$\Lambda_k(j,n) = \Lambda_k^S(j,n) \times \Lambda_k^A(j,n). \tag{4.3.8}$$

where $\Lambda_k^S(j,n)$ denotes the spatial cost computed by the Euclidean distances between the position and size information. Then the Hungarian algorithm [133] with the association cost is used to perform the association between the selected measurements $\mathbf{Z}_{t,+}$ and predicted states.

After the association, the measurement set is divided into two groups: $\mathbf{Z}_{k,+} = \mathbf{Z}_{k,\gamma} \cup \mathbf{Z}_{k,A}$, in which, the associated and un-associated measurements $\mathbf{Z}_{k,A}$, $\mathbf{Z}_{k,\gamma}$ are respectively used for target survival and initialization.

4.3.3 Initialization

Standard formulation of PHD filtering methods [103] [44] often presets the target birth model to cover the entire region of interest. In the video tracking context, detections from the object detector are used to handle the target birth and death that are highly random and unpredictable, so as to avoid the need for prior knowledge of the scene information. Therefore, newborn targets at time k will be adaptively estimated by using non-associated measurements $\mathbf{Z}_{k,\gamma}$,

$$\gamma_k(\mathbf{x}) = \sum_{j=1}^{J_{k,\gamma}} w_{k,\gamma}^j \mathcal{N}(\mathbf{x}; \mathbf{m}_{k,\gamma}^j, \mathbf{P}_{k,\gamma}^j) \tag{4.3.9}$$

$$\mathbf{m}_{k,\gamma}^j = \mathbf{H}^{-1} \mathbf{z}_{k,\gamma}^j \tag{4.3.10}$$

$$\mathbf{P}_{k,\gamma}^{j} = \mathbf{H}^{-1}\mathbf{R}_k(\mathbf{H}^{-1})^T \tag{4.3.11}$$

where $\mathbf{H}$ is the observation matrix, and $\mathbf{R}_k$ means the observation noise covariance matrix. In the meantime, new identities $\mathcal{I}_{k,\gamma} = \{I_{k,\gamma}^{1}, ..., I_{k,\gamma}^{J_{k,\gamma}}\}$ are assigned to new-born targets, where $J_{k,\gamma} = |\mathbf{Z}_{k,\gamma}|$ denotes the number of new-born targets [6].

4.3.4 Update

For the update step, both survival and new-born targets are combined to form the predicted PHD intensity $\nu_{k|k-1}(\mathbf{x}) = \nu_{k|k-1,S}(\mathbf{x}) + \gamma_k(\mathbf{x})$, which can be expressed as a Gaussian mixture [103],

$$\nu_{k|k-1}(\mathbf{x}) = \sum_{j=1}^{J_{k|k-1}} w_{k|k-1}^{j}\mathcal{N}(\mathbf{x}; \mathbf{m}_{k|k-1}^{j}, \mathbf{P}_{k|k-1}^{j}) \tag{4.3.12}$$

The predicted identity set is also updated as, $\mathcal{I}_{k|k-1} = \mathcal{I}_{k|k-1,S} \cup \mathcal{I}_{k,\gamma}$. Then the PHD update step [103] at time k can be achieved by associating the predetermined reliable measurements $\mathbf{Z}_{k,+}$,

$$\nu_k(\mathbf{x}) = p_M\nu_{k|k-1}(\mathbf{x}) + \sum_{\mathbf{z}\in\mathbf{Z}_{k,+}} \sum_{j=1}^{J_{k|k-1}} w_k^{j}(\mathbf{z})\mathcal{N}(\mathbf{x}; \mathbf{m}_{k|k}^{j}(\mathbf{z}), \mathbf{P}_{k|k}^{j}) \tag{4.3.13}$$

where

$$w_k^{j}(\mathbf{z}) = \frac{(1-p_M)w_{k|k-1}^{j}q_k^{j}(\mathbf{z})}{\kappa_k(\mathbf{z}) + (1-p_M)\sum_{i=1}^{J_{k|k-1}} w_{k|k-1}^{i}q_k^{i}(\mathbf{z})} \tag{4.3.14}$$

$$q_k^{j}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{H}\mathbf{m}_{k|k-1}^{j}, \mathbf{R}_k + \mathbf{H}\mathbf{P}_{k|k-1}^{j}\mathbf{H}^T) \tag{4.3.15}$$

$$\mathbf{m}_{k|k}^{j}(\mathbf{z}) = \mathbf{m}_{k|k-1}^{j} + \mathbf{K}_k^{j}(\mathbf{z} - \mathbf{H}\mathbf{m}_{k|k-1}^{j}) \tag{4.3.16}$$

$$\mathbf{P}_{k|k}^{j} = [\mathbf{I} - \mathbf{K}_k^{j}\mathbf{H}]\mathbf{P}_{k|k-1}^{j} \tag{4.3.17}$$

$$\mathbf{K}_k^{j} = \mathbf{P}_{k|k-1}^{j}(\mathbf{H})^T(\mathbf{H}\mathbf{P}_{k|k-1}^{j}(\mathbf{H})^T + \mathbf{R}_k)^{-1} \tag{4.3.18}$$

where p_m is the missed detection probability, and κ_k is the clutter density. For each predicted Gaussian term, its identity is assigned to $(1 + |\mathbf{Z}_{k,+}|)$ updated terms, $I_k^j = I_{k|k-1}^j$.

4.3.5 Track management

After the PHD update step, a track management scheme in [129] is adopted in here to correctly extract the confirmed tracks, and discard other tracks that are least reliable. Firstly, targets with the maximum weights can be selected as a collection of possible tracks. Then targets can be confirmed with $w_k^j \geq w_{th}$ and labelled with the same identity as that in prediction, where w_{th} denotes the threshold of a target confirmation. In contrast, the rest of the targets which fail to reach w_{th} are tentatively eliminated after a certain value of T_{miss} frames. The above only summarizes the key steps of the track management after the PHD filtering, while the details have been given in Chapter 3.

4.4 Experiments

In this section, there are two experiments implemented on the widely-used MOT16 Challenge Benchmark [4]. This benchmark provides various pedestrian tracking scenarios with different camera motion and lighting conditions. The first experiment is performed on the entire benchmark training set to justify the effectiveness of the proposed method for measurement selection of detections. Then two video sequences MOT16-02 (static camera) and MOT16-10 (moving camera) are employed to evaluate the proposed method on the tracking task.

4.4.1 Experimental settings

The process noise covariance matrix $\mathbf{Q}_{k-1} = Diag([25, 25, 16, 16, 4, 4])$, and the observation noise covariance matrix is $\mathbf{R}_k = Diag([25, 25])$ [9]. Parameters to implement the GM-PHD filter are empirically set to: $p_M = 0.01$, $e = 0.95$, and $\kappa = 10^{-4}$ [10]. In order to reduce the effect of potentially noisy detections still existing in the measurement model, the birth weights are empirically assigned with a smaller value 0.001 by following the setting as in [134].

For enhanced confidence rescoring, the ResNet-101-FPN backbone of mask R-CNN [12] is retained for feature extraction, but the network heads of bounding-box regression and instance segmentation are removed. The classification branch of the mask R-CNN is merely used during the model inference. The model was pre-trained on the MS-COCO detection dataset [135]. Readers are referred to [12] for further details of the network training. Additionally, RoIs are no longer generated from a regional proposal network [16] which is replaced by an input layer. The public detections provided by the MOT16 benchmark are used in here. The threshold for enhanced confidence scoring is set to $c_{Th} = 0.1$. Parameters for implementing Soft-ANMS are set to $U_T = 0.3$, $S_T = 0.5$, and $\sigma = 0.6$.

4.4.2 Results of measurement selection

In this experiment, the MOT16 trainging set is used to investigate the impact of the proposed EDR module on the measurement selection. In this module, enhanced confidence rescoring and Soft-ANMS are respectively denoted as *ECR* and *S-ANMS*. For performance comparison, a recently developed two-step measurement selection in [29] is included, which consists of confidence score sorting (*CCS*) and non-maximal suppression (*NMS*). The original detection set provided by the benchmark is used as the baseline (RAW) [4]. For this evaluation, the number of true positives (TPs) and false positives

Table 4.1. Selection performance comparison with different approaches on MOT16 training set.

Method	TP(↑)	FP (↓)
RAW [4]	50,859	28,931
CCS [29]	40,074	4,884
CCS-NMS [29]	39,550	2,553
ECR	50,111	13,366
ECR-NMS	48,187	4,416
ECR-S-ANMS	48,433	3,731

(FPs) is report by calculating the IoU ratios between the boxes of detection and ground truth. Specifically, a TP is counted if the IoU ratio is greater than 0.5; ground truth boxes are considered if they are visible and with human-related class labels.

Table 4.1 shows the results on the measurement selection task. The effectiveness of the proposed *ECR* is firstly justified by comparing with *CCS*, since both of which are designed to apply a threshold to rectify the raw detections. It can be observed that *CCS* reduces a larger number of FPs, but it returns a significant deduction in TPs. The proposed *ECR* also produces fewer FPs, and simultaneously retain most of the TPs. This can be explained that the proposed *ECR* improves the correspondence between the scores and TPs. It is noteworthy that the combined method (*ECR-S-ANMS*) reduces 87.1% of FPs and retains 95.2% of TPs, in comparison with the baseline (*RAW*). By comparing the last two methods in Table 4.1, it can be found that the proposed *S-ANMS* outperforms the traditional *NMS* by reducing 685 FPs, as well as recovering 246 TPs. This is because aggregating the scores between SIOA and IOU can provide better robustness in the suppression process.

Furthermore, the trade-off between the precision and recall is analyzed on the MOT16 training set, as shown in Fig. 4.4. The proposed full module (*ECR-S-ANMS*) favorably surpasses others in terms of both precision and recall. On the other hand, Fig. 4.5 also visually displays the advantages of

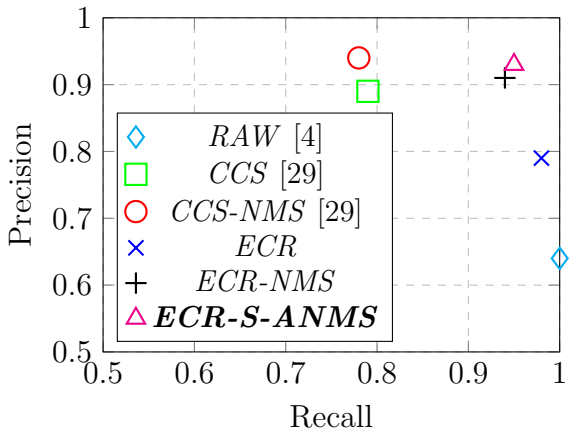


Figure 4.4. Precision and Recall ranking plots on the MOT16 Challenge training set. Methods closer to the upper right corner perform better.

Table 4.2. Tracking Performance Comparison on MOT16-02 sequence. Bolded results indicate the best.

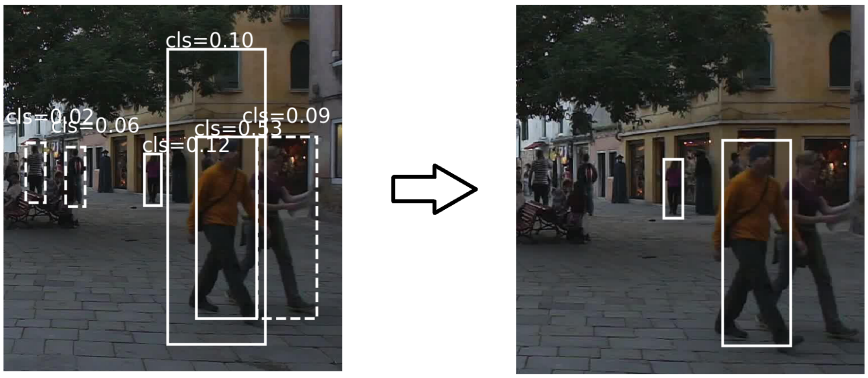
Tracker	MOTA (↑)	FP (↓)	FN (↓)	MT (↑)	ML (↓)
T-RAW [4]	15.8	1,783	13,067	11.1%	55.6%
T-CCS [29]	16.4	474	14,353	7.4%	66.7%
T-CCS-NMS [29]	18.6	**70**	14,367	7.4%	66.7%
T-ECR	18.5	1,351	**13,009**	11.1%	55.6%
T-ECR-NMS	24.6	229	13,070	11.1%	57.4%
T-ECR-S-ANMS	**24.8**	200	13,066	11.1%	57.4%

the proposed method in different steps. These comparisons above verify the proposed EDR algorithm can effectively improve the measurement selection, so as provide a premise for the subsequent tracking process.

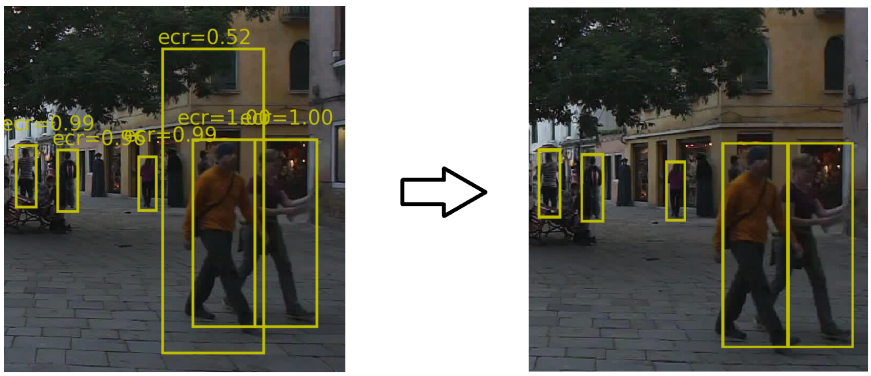
4.4.3 Tracking evaluations

In this section, previously acquired selection results are incorporated with the same measurement-driven GM-PHD tracker for tracking evaluations. The following metrics [115] are specifically used to assess the tracking performance, which are multiple object tracking accuracy (MOTA), false positives (FP), false negatives (FN), mostly tracked targets (MT) and mostly lost targets (ML)

Tables 4.2 and 4.3 detail the tracking results on two scenarios. a short



(a) Selection with original confidence scores and NMS



(b) Selection with enhanced detection reliability

Figure 4.5. Visualization of measurement selection on the MOT16-02 video sequence. Detections with scores smaller than the threshold $c_{Th} = 0.1$ are presented with dashed boxes. (a) shows that the selection original confidence scores improves the false detections but increases the number of missed detections. (b) shows the proposed approach can better preserve the true targets and remove false detections (better viewed in color version).

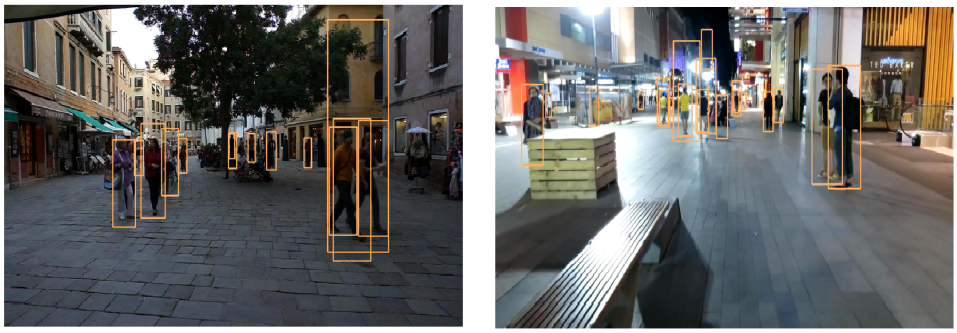
name (T) is added with each selection method to indicate tracking methods. For MOT16-02 sequence, it is shown that tracking with each step of the proposed method can provide better performance in some ways, even if only using single step *T-ECR* can achieve the least number of FN and almost same MOTA compared with traditional two-step measurement selection (*T-CCS-NMS*). The full method (*T-ECR-S-ANMS*) achieves the best MOTA which is the most important metric for the overall evaluation. Comparing with the method *T-ECR-NMS*, the full method performs better in MOTA, FP, and FN. This suggests that proposed *S-ANMS* is more effective than traditional *NMS* for promoting the tracking task.

Table 4.3. Tracking Performance Comparison on MOT16-10 sequence. Bolded results indicate the best.

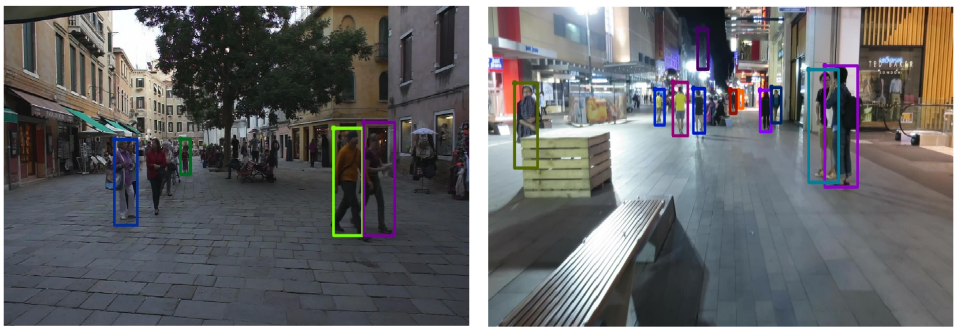
Tracker	MOTA (↑)	FP (↓)	FN (↓)	MT (↑)	ML (↓)
T-RAW [4]	21.6	2,767	**6,749**	14.8%	40.7%
T-CCS [29]	31.7	489	7,826	9.3%	59.3%
T-CCS-NMS [29]	33.4	**256**	7,850	9.3%	59.3%
T-ECR	32.1	1,440	6,785	14.8%	40.7%
T-ECR-NMS	39.5	465	6,859	14.8%	40.7%
T-ECR-S-ANMS	**39.8**	472	6,819	14.8%	40.7%

For MOT16-10 sequence, the proposed full method (*T-ECR-S-ANMS*) also reports the highest MOTA score, and significantly improves the baseline (*T-RAW*) by 18.2%. In addition, Improvements on the MT and ML further demonstrate the proposed method benefits the tracking robustness and consistency. To sum up, the proposed EDR method benefits on the early stage of tracking pipeline by supplying more reliable measurements, and therefore the tracking process can be better realized.

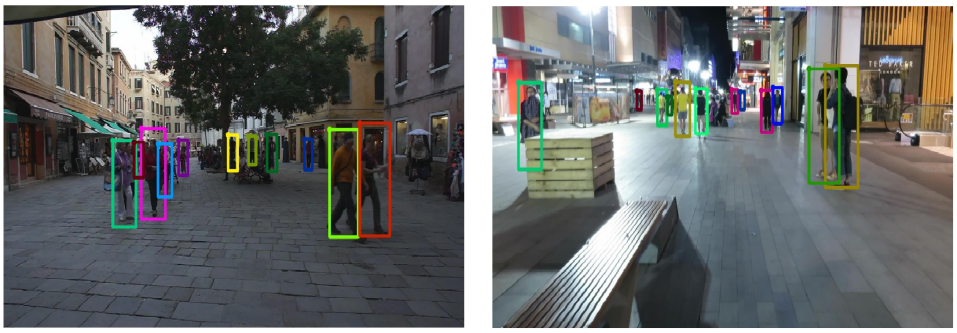
Furthermore, Fig. 4.6 also visually reveals the merits of the proposed method for different tracking scenarios. To be specific, it can be seen from Fig. 4.6 (a), there are several false alarms present in the original detection results. Ideally, the tracker should be able to output fewer false detections and recover more missed detections. As shown in Fig. 4.6 (b), tracking with original measurement selection method is not adversely affected by the false alarms, whereas fails to track a number of targets which have been identified in the detection step. This clearly exposes the weakness of the baseline selection method, which is unable to be well deal with two challenging issues simultaneously. However, Fig. 4.6 (c) shows that the aforementioned problems have been well addressed by tracking with the proposed EDR algorithm. This instance is further to confirm the validity of the proposed method on the tracking task.



(a) Original detection results



(b) Tracking with original confidence scores and NMS



(c) Tracking with enhanced detection reliability

Figure 4.6. Qualitative tracking comparison on the video sequences of MOT16 dataset.

4.5 Summary

In this chapter, a novel enhanced detection reliability approach was proposed to improve the measurement selection for online multiple human tracking. The misalignment between confidence scores and true positives on the MOT16 Challenge benchmark was firstly statistically analyzed. Much uncertainty in the selection process caused by the misalignment issue necessitates the enhancement of the detection reliability. By performing the enhanced confidence rescoring in a global-to-local manner, the correspondence between target candidates and confidence scores has been well improved, which is helpful to supply more reliable measurements for tracking. Then, an improved pruning algorithm Soft-ANMS was developed to eliminate ambiguous detections, and therefore provides better robustness in the suppression process. Moreover, CNN features based target appearance models were employed to help mitigate the target ambiguity in the filtering process.

The obtained results on the selection experiment verifies the effectiveness of the proposed EDR algorithm in the measurement selection, where the proposed method produces fewer FPs, and simultaneously retain most of the TPs . Specifically, the proposed measurement selection method reduces 87.1% of FPs and retains 95.2% of TPs, in comparison with the baseline method. For the tracking evaluations, tracking improvements in FP and MOTA demonstrate the proposed approach is able to reduce false detections and promote the tracking accuracy. The proposed method achieves the best MOTA scores in both evaluated sequences, and in particular outperforms the baseline method by 9.0% and 18.2% respectively.

Overall, this approach can be applied as a stand alone algorithm to any online tracking methods for robust measurement selection. However, missed detections as one of the tracking challenges has not been fully investigated. The next chapter aims to address this issue and improve the tracking per-

formance from a fusion perspective, which integrates two human detectors to perform a multi-level cooperative fusion of GM-PHD filters.

Chapter 5

MULTI-LEVEL COOPERATIVE FUSION OF MEASUREMENT-DRIVEN GM-PHD FILTERS FOR ONLINE MULTIPLE HUMAN TRACKING

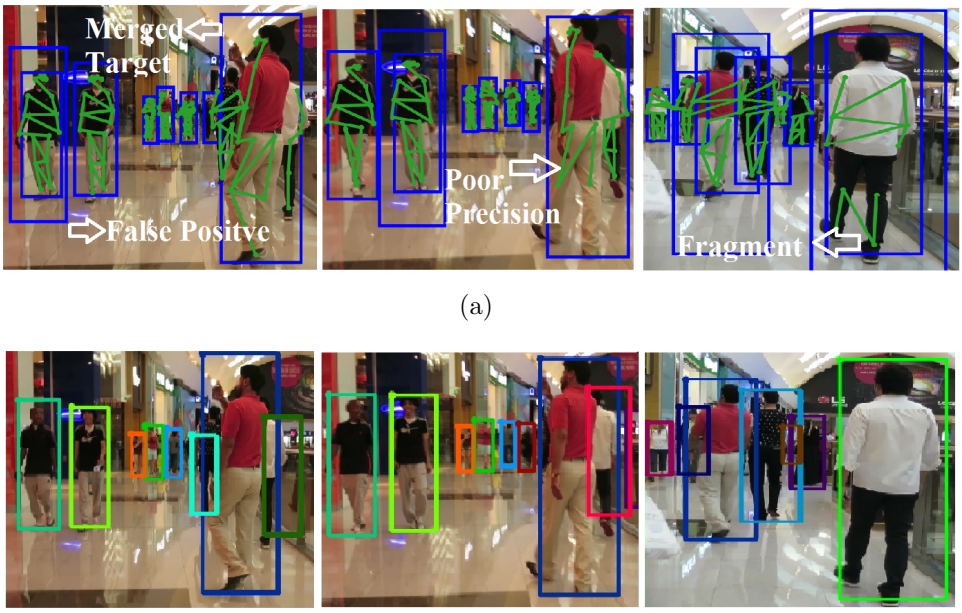
5.1 Introduction

Previous chapters contributed to improving the measurement driven filtering in birth intensity estimation, measurement classification and selection. However, missed detections as a challenging issue presented in this thesis have not been properly addressed. It may be that previously proposed methods solely relying on a single detector are limited to exploring the image context more comprehensively, as depicted in Fig. 5.1 (a). The full-body detector (blue) fails to consider three pedestrians on the right as a merged one when they are in close proximity, and it is prone to false positives. The body-parts

detector (green) improves the false positives and merged targets, whilst it has less promising performance in precision and fragmentation. Therefore, a multi-detector fusion approach can be an effective solution to reinforce the tracking process.

Recently, fusion of multiple data sources has been proven to enhance the tracking robustness and reliability, since this approach can provide redundancy in different aspects, and eliminate uncertainties between individual sources [6, 67, 136]. Most existing fusion based tracking methods focus on single level fusion, such as fusing multiple features [78], grouping detections [36,65] or integrating the tracking outputs [67]. Based on the processing level, sequential [26, 66], parallel [27, 70, 71] and hybrid data fusion [137] approaches are possible. Different from the aforementioned fusion approaches, this chapter presents multi-level (feature-level and decision-level) cooperative fusion within the GM-PHD filter framework for online multiple human tracking. As shown in Fig. 5.1 (b) Fusion of both detectors show that false positives are eliminated, missed targets are recovered from the occlusion area, and that tracking precision is improved.

The proposed fusion approach is composed of two key ideas: enhanced identity association and a robust fusion center. In general, this approach integrates two human detectors with different characteristics (full-body and bodyparts), and investigate their complementary benefits for tracking multiple targets. For the enhanced identity association, a novel discriminative correlation matching (DCM) model in the feature-level fusion is proposed, and it is fused with spatio-temporal information to mitigate the ambiguous identity associations in the GM-PHD filter. The DCM model exploits discriminative correlation filters (DCFs) learned from features of multiple convolutional layers as target-specific classifiers, which discriminate the desired target from background and other existing targets. Features are obtained from the outputs of both top and lower convolutional layers, which can en-



(a)

(b)

Figure 5.1. Justification of using multiple detectors on the MOT16-11 video sequence [6].

code the target appearances with better discriminability of the background and targets in the same category.

For the robust fusion center, it is designed with virtual and real zones to perform the fusion at the decision level. The real zone is mainly responsible for performing the cooperative track fusion which applies survival and birth track fusion independently on the tracking estimates of survival and new-born targets. The virtual zone manages to reconstruct the missed targets and remove false detections. The intuition of this design is to enable the fusion process in the tracking system to better exploit the maximum strengths from the two detectors, and also mitigates the sensitivity of missed detection occurring in the original GCI rule [68]. Besides, an identity reassignment mechanism which is similar to [72] is developed to overcome the identity mismatching problem. MOTChallenge Benchmark evaluations are provided to confirm improved performance over other state-of-the-art RFS based tracking methods. The main contributions of this chapter are sum-

marized as follows:

1. A novel DCM model at the feature-level fusion is proposed, which is fused with spatio-temporal information to enhance the ambiguous identity associations in the filtering process.

2. A robust fusion center at the decision-level fusion is proposed to improve the fusion process and tracking consistency.

3. A novel multi-level (feature-level and decision-level) cooperative fusion within the GM-PHD filtering for online multiple human tracking.

This chapter addresses the fourth and fifth objectives of this thesis, which relate to the GM-PHD filter based online multiple human tracking using deep discriminative correlation matching published in [30], the collaborative detector fusion of the data-driven PHD filter for online multiple human tracking published in [6], and the multi-level cooperative fusion of GM-PHD filters for online multiple human tracking published in the IEEE Transactions on Multimedia [33].

The rest of this chapter is presented as follows: Section 5.3 describes the formulation of the multi-target dynamical and measurement model using the RFS framework. Section 5.4 presents the proposed enhanced identity association which performs the feature-level fusion on each detector domain. Section 5.5 explains the tracking process, which is realized by applying the data-driven GM-PHD filters on the association results from both detectors. Then the proposed robust fusion center is introduced to provide a global analysis on the preliminary tracking estimates in Section 5.6. Experimental results and discussions are given in Section 5.7, verifying the effectiveness of the proposed multi-level cooperative fusion with GM-PHD filters.

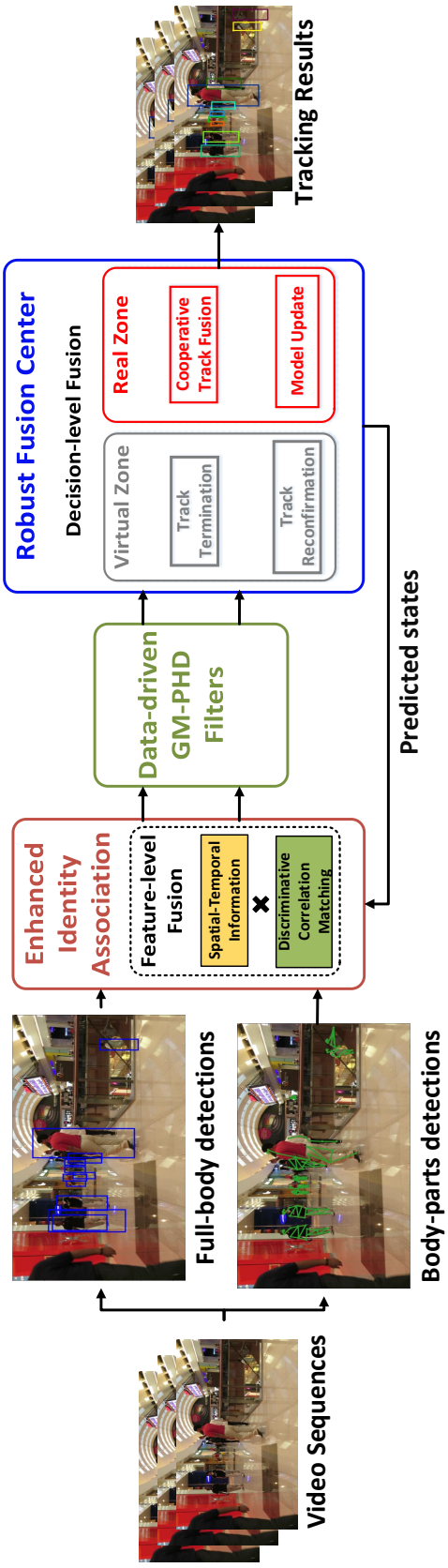


Figure 5.2. Overview of the proposed approach for online multiple human tracking.

5.2 Overview of the proposed tracking approach

The overview of the proposed tracking approach is illustrated in Fig. 5.2. The designed fusion based tracking system consists of feature-level and decision-level fusion and performs in a cooperative way to fulfil the tracking objective. Given an input image, both full-body and body-parts detectors locally achieve valid detections. In the enhanced identity association stage, detections from each domain are locally associated with predicted states by performing the feature-level fusion which exploits spatio-temporal information and discriminative correlation matching. The resulting association outputs from each domain are further processed through the GM-PHD filters to achieve the local tracking estimates. These preliminary tracking results are further moved into the designed robust fusion center for the decision-level fusion. This fusion center consists of real and virtual zones. In the real zone, cooperative track fusion and appearance model updates are performed on the significant tracks, which yields the final tracking results. The virtual zone in the robust fusion center is designed to manage track termination and reconfirmation. The detailed implementations on each block of the proposed approach are presented in the following sections.

5.3 Formulation of Motion Prediction and Measurement Model

5.3.1 Motion Prediction

Let $\nu_{k-1}(\mathbf{x}) = \sum_{j=1}^{J_{k-1}} w_{k-1}^{j} \mathcal{N}(\mathbf{x}; \mathbf{m}_{k-1}^{j}, \mathbf{P}_{k-1}^{j})$ denote PHD intensity at time $k-1$. Following the method in [129], an identity set $\mathcal{I}_{k-1} = \{I_{k-1}^{1}, ..., I_{k-1}^{J_{k-1}}\}$ is initialized to couple with $\nu_{k-1}(\mathbf{x})$. For the current time k, the motion prediction for each potentially survived target in $\nu_{k-1}(\mathbf{x})$ is performed, which

is given by [103],

$$\nu_{k|k-1,S}(\mathbf{x}) = e_{k|k-1} \sum_{j=1}^{J_{k-1}} w_{k-1}^{j} \mathcal{N}(\mathbf{x}; \mathbf{F}\mathbf{m}_{k-1}^{j}, \mathbf{Q}_{k-1} + \mathbf{F}\mathbf{P}_{k-1}^{j}\mathbf{F}^{T}) \quad (5.3.1)$$

where $e_{k|k-1}$ represents the survival probability, J_{k-1} is the number of Gaussian components, and w_{k-1}^{j} is the weight of the j-th Gaussian component. $\mathcal{N}(\cdot; \mathbf{m}, \mathbf{P})$ denotes a Gaussian component with mean $\mathbf{m}$ and covariance $\mathbf{P}$. $\mathbf{F}$ and $\mathbf{Q}_{k-1}$ are the matrices of motion transition and process noise covariance, respectively [6]. The identities of Gaussian components are predicted as, $I_{k|k-1,S}^{j} = I_{k-1}^{j},\ j = 1, ..., J_{k-1}$.

In reality, these predicted Gaussian terms may contain *ghost* targets which are potentially missed at current time k, which increases the uncertainty in the update step. This issue was not practically resolved in the previous chapters. Therefore, in this chapter, it is required and necessary to examine the predictions in the early stage before implementing the filtering process.

5.3.2 Measurement Model

To build a robust measurement model, two sets of measurements (full-body and body-parts detections) are used at each time step k for parallel processing, as illustrated in Fig. 5.2. First of all, a full body detector [2] is applied to form a set of full-body detections $\mathbf{Z}_{k,a}$, where a represents the full-body detector. Furthermore, it is likely that each person can have multiple measurements that are spatially surrounding the target for each time step, which means that each human target consists of a certain number of body parts that share common dynamics or attributes [6,138]. To this end, another body parts detector [139] is adopted to acquire body-parts detections $\mathbf{Z}_{k,b}$, where b demonstrates the body-parts detector. There are ideally 14 body parts which can be obtained per time step k, namely ankles, knees, hips, wrists,

elbows, shoulders with left/right symmetry, and the head top/bottom [6]. A rectangle shape is considered for grouping the body-parts measurements to model the human contour. In this work, the reshaped measurements will be discarded if the number of body-parts is less than 6. For simplicity, $\mathbf{Z}_{k,\theta}$, $\theta \in \{a, b\}$ is denoted concisely as $\mathbf{Z}_k$, in which, the detector index θ on the obtained detections is ignored. The measurement model for each detection under the Gaussian assumption is given by,

$$g_k(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}_k; \mathbf{H}\mathbf{x}_k, \mathbf{R}_k) \tag{5.3.2}$$

where $\mathbf{H}$ denotes the observation matrix, $\mathbf{R}_k$ denotes the observation noise covariance matrix at time k. In fact, noisy measurements can degrade the efficiency of the data-driven PHD filter and birth prediction [30]. Then a pre-processing step is implemented on the original detections before performing the target association. It takes advantage of the confidence score $c_k \in [0, 1]$ which is built in each full-body detection, to remove the noises $\mathbf{\Gamma}_k = \{\mathbf{z}_k^f : c_k < c_{th}\}$ in the original detections $\mathbf{Z}_k$, where c_{th} is the threshold value [30], thus reliable measurements are obtained as, $\mathbf{Z}_k^+ = \mathbf{Z}_k \setminus \mathbf{\Gamma}_k$.

5.4 Enhanced Identity Association

This section presents an enhanced labelling system for the PHD filtering process. Unlike the early association in [29], the proposed approach integrates DCM appearance models with spatio-temporal information to address the target ambiguity in the labelling system.

5.4.1 Spatio-Temporal Information

Spatio-temporal information has been widely used for measuring affinity between detections and targets [29, 30, 129]. It is able to capture geometric relations between the bounding boxes with low computational cost. The

means of Gaussian terms given by (4.3.4) are regarded as predicted states of individual targets for the calculation of association cost. Reliable measurements $\mathbf{Z}_k^+$ are previously obtained in Section 5.3.2. Then, the association cost Λ_k^{ST} is measured by spatio-temporal relation between a reliable detection $\mathbf{z}_k \in \mathbf{Z}_k^+$ and a predicted state $\mathbf{m}_{k|k-1,S}$, which is computed by,

$$\Lambda_k^{ST}(\mathbf{m}_{k|k-1,S}, \mathbf{z}_k) = \exp\left(- \frac{\|\mathbf{H}\mathbf{m}_{k|k-1,S} - \mathbf{z}_k\|}{2\sigma_s^2} \right) \tag{5.4.1}$$

where σ_s^2 represents the variance. These affinity scores Λ_k^{ST} obtained from the spatio-temporal relation construct a cost matrix $\mathbf{\Delta}_k^{ST}$. However, only acquiring the geometric relations may generate ambiguous affinity scores when targets are in close proximity. To this end, target relations on the visual content will be explored in the next section.

5.4.2 Discriminative Correlation Matching

Recently, discriminative correlation filters have been widely used in single object visual tracking for better accuracy and efficiency. The DCF based tracking approaches that learn correlation filters to encode target appearances can achieve high computational efficiency [96] with the use of fast Fourier transforms (FFTs). For visual tracking, DCFs are mainly learned as linear classifiers to discriminate between target and background appearances. In this work, the discriminative power of DCFs is exploited and a discriminative correlation matching scheme is proposed to formulate the appearance model. The major task is to learn the discriminative correlation filters with features of multiple convolutional layers as target-specific classifiers to discriminate the desired target from noisy background and also other intra-class targets. Preliminary ideas have been appeared in the previous conference paper [30]. Fig. 5.3 shows the workflow of the proposed discriminative correlation matching. For each target, two DCFs are learned

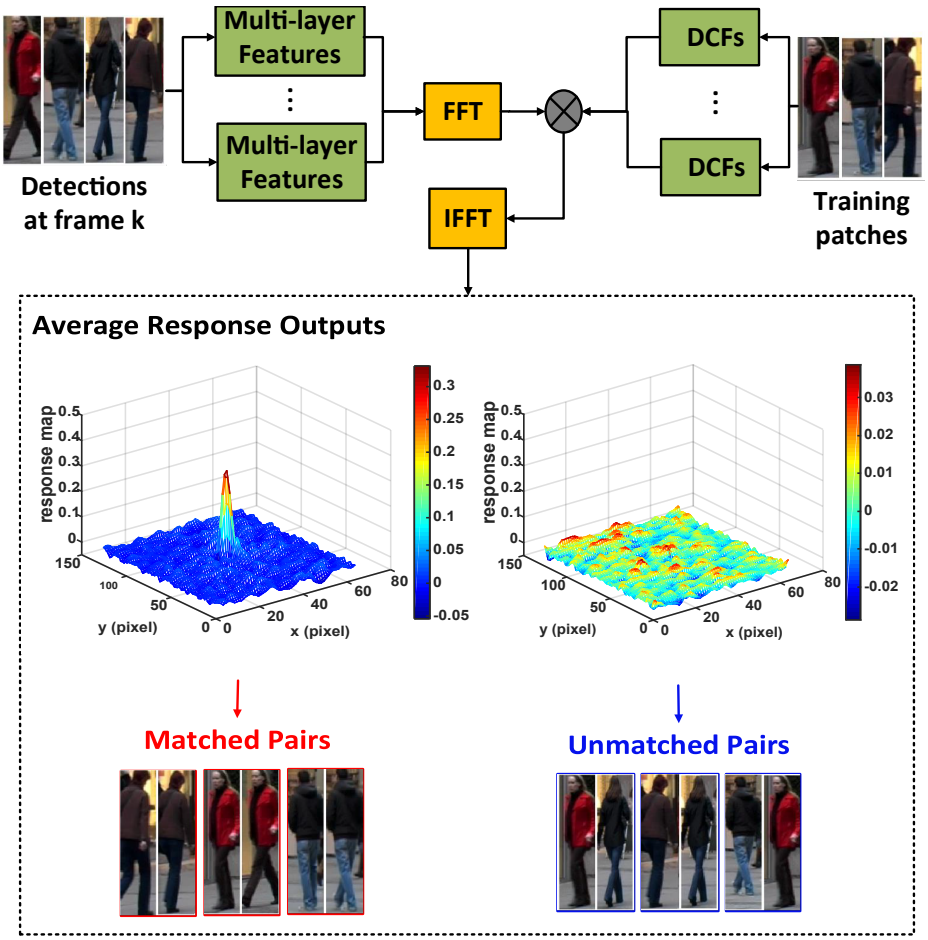


Figure 5.3. An example workflow of the proposed discriminative correlation matching scheme which uses DCF-based target-specific classifiers with the average response outputs for target appearance matching.

from the outputs from both top and lower convolutional layers to encode the target appearances, one (top layer) is used to distinguish between targets and the background, the other (lower layer) is used for handling intra-class targets variations. The affinity scores are measured by computing the average peak-to-sidelobe ratios (PSRs) from the correlation responses.

Training Phase

There are two discriminative correlation filters $\{\mathbf{c}^{(l)}\}$ trained for each predicted target. The intuition of applying two filters together is to help match

between a predicted target and a new detection with the exploitation of both semantic and spatial details of target appearances. Feature maps used for training are from the outputs of both the top and lower convolutional layers, and each of them $\mathbf{f}^{(l)}$ with the size of $A \times B \times D$ is extracted from the output of the l-th convolutional layer, where A, B, and D denote the width, height, and the number of channels, respectively [30]. Training samples for discriminative correlation filters are generated from all circular shifts $\mathbf{f}^{(l)}_{a,b}, (a,b) \in \{0,...,A-1\} \times \{0,...,B-1\}$. Each shifted sample has a desired output $g^{(l)}(a,b) = \exp(-\frac{(a-A/2)^2+(b-B/2)^2}{2\sigma_c^2})$ to form a Gaussian label matrix $\mathbf{g}^{(l)} = \{g^{(l)}(a,b)|(a,b) \in \{0,...,A-1\} \times \{0,...,B-1\}$, where σ_c is the kernel width [30]. The DCF $\mathbf{c}^{(l)}$ with the same size of $\mathbf{f}^{(l)}$ can be learned by minimizing the following loss [30, 96],

$$\underset{\mathbf{c}^{(l)}}{\operatorname{argmin}} \sum_{a,b} \left\| \sum_{d=1}^{D} (\mathbf{c}^{(l)}_{a,b,d})^T \mathbf{f}^{(l)}_{a,b,d} - g^{(l)}(a,b) \right\|^2 + \lambda \|\mathbf{c}^{(l)}\|_2^2 \tag{5.4.2}$$

where λ is the regularization parameter. Following the work in [96], the DCF is trained by performing the fast Fourier transform (FFT) in the frequency domain. Therefore, the solution of (5.4.2) on the d-th ($d \in 1,...,D$) channel can be written as [30, 96],

$$\hat{\mathbf{c}}^{(l)}_d = \frac{\hat{\mathbf{g}}^{(l)} \odot (\hat{\mathbf{f}}^{(l)}_d)^\dagger}{\sum_{d=1}^{D} \hat{\mathbf{f}}^{(l)}_d \odot (\hat{\mathbf{f}}^{(l)}_d)^\dagger + \lambda} \tag{5.4.3}$$

where the hat stands for FFT operator, and the dagger represents complex conjugation operation. The operator $\odot$ defines the Hadamard (element-wise) product.

Correlation Matching

Prior to performing the matching scheme, it is essential for a newly detected target to extract each feature map $\mathbf{y}^{(l)} \in \mathbb{R}^{A \times B \times D}$ by using the same layers in the training phase. The goal of performing correlation matching in a

many-to-many scenario is to find all the correlation responses between each predicted and newly detected target. In practice, each correlation filter $\mathbf{c}^{(l)}$ is correspondingly associated with each feature map $\mathbf{y}^{(l)}$ to compute a correlation response map $\mathbf{r}^{(l)} \in \mathbb{R}^{A \times B}$ at the l-th layer [30, 96],

$$\mathbf{r}^{(l)} = \mathscr{F}^{-1}\left\{ \sum_{d=1}^{D} \hat{\mathbf{c}}_d^{(l)} \odot (\hat{\mathbf{y}}_d^{(l)})^{\dagger} \right\} \tag{5.4.4}$$

where $\mathscr{F}^{-1}\{\cdot\}$ denotes the inverse fast Fourier transform (IFFT). Given the set of response maps $\{\mathbf{r}^{(l)}\}$, PSRs are utilized to obtain the set of scores $\{PSR^{(l)}\}$ from each layer l:

$$PSR^{(l)} = \frac{\max(\mathbf{r}^{(l)}) - \mu^{(l)}}{\sigma_r^{(l)}} \tag{5.4.5}$$

where $\mu^{(l)}$ and $\sigma_r^{(l)}$ denote the mean value and the standard deviation of the sidelobes. Similar works [62] and [66] used the PSRs as a gating technique to confirm the predicted state or detect tracking failures. However, the proposed work focuses on enhancing the association step. Firstly, the average sum of the $\{PSR^{(l)}\}$ is computed to achieve the overall matching score ρ, which results in the matched pairs with $\rho < \varpi$ and unmatched pairs with $\rho \geq \varpi$, where ϖ denotes the matching threshold. Then a generalized *sigmoid* function is utilized to compute affinity scores between each predicted and newly detected target based on the correlation matching results, since this function limits the overall score ρ to a range of $[0, 1]$. Each pairwise affinity score Λ_k^A is calculated based on an overall matching score ρ,

$$\Lambda_k^A = \frac{1}{1 + exp^{-(\alpha \times \rho + \beta)}} \tag{5.4.6}$$

where α and β are the coefficients for the calculation. These affinity scores Λ_k^A obtained from the matching scheme construct a cost matrix $\mathbf{\Delta}_k^A$, which is ultimately fused with the cost of spatio-temporal relation $\mathbf{\Delta}_k^{ST}$ to build

Algorithm 4: Enhanced Identity Association ($k > 1$)

Input : Reliable measurements $\mathbf{Z}_k^+$, predicted states: $\nu_{k|k-1,S}$.
Output: Associated and un-associated measurements $\mathbf{Z}_{k,D}^+$, $\mathbf{Z}_{k,\gamma}^+$. Predicted intensities of detected and potentially missed targets $\nu_{k|k-1,D}$, $\nu_{k|k-1,M}$.

Initialize discriminative correlation filters $\{\mathbf{c}_{k-1}^{(l)}|l = 3, 4\}$ for each $\mathbf{m}_{k|k-1,S} \in \nu_{k|k-1,S}$, using Eq. (5.4.2) and Eq. (5.4.3).
Compute the pairwise spatio-temporal association cost Λ_k^{ST}, using $\mathbf{m}_{k|k-1,S} \in \nu_{k|k-1,S}$, $\mathbf{z}_k \in \mathbf{Z}_k^+$ and Eq. (5.4.1).
Crop out the image patches based on $\mathbf{z}_k \in \mathbf{Z}_k^+$ and extract multi-layer convolutional features $\{\mathbf{y}^{(l)}\}$.
for *each layer* l **do**
    Compute the correlation response map $\mathbf{r}^{(l)}$ using l-th feature map $\mathbf{y}^{(l)}$, $\mathbf{c}_{k-1}^{(l)}$ and Eq. (5.4.4).
    Compute the matching score $PSR^{(l)}$ via Eq. (5.4.5).
end
Compute the overall matching score ρ, and use it to achieve the pairwise appearance association cost Λ_k^A using Eq. (5.4.6).
Compute the total association cost $\mathbf{\Delta}_k$ using Eq. (5.4.7).
Perform the optimal association to obtain $\mathbf{Z}_{k,D}^+$, $\mathbf{Z}_{k,\gamma}^+$, $\nu_{k|k-1,M}$, $\nu_{k|k-1,D}$.

the total association cost as follows:

$$\mathbf{\Delta}_k = \mathbf{\Delta}_k^{ST} \odot \mathbf{\Delta}_k^A \tag{5.4.7}$$

The benefit of this feature-level fusion is that it can compensate for unreliability present in the individual association cost, especially when target ambiguities occur in either motion dynamics or visual content [30]. In practice, each set of valid detections is locally processed in the enhanced identity association. The Hungarian algorithm [133] is used to achieve the optimal association. The overall algorithm of enhanced identity association is given with details in Algorithm 4.

5.5 Measurement-Driven Filtering

The filtering process is applied on each set of association results from both detectors to achieve candidate tracking estimates. In the traditional PHD filter, all the input measurements in $\mathbf{Z}_k$ are used for the update steps. However, it may degrade the updating performance because of the misuse of measurements and false alarms. For this tracking task, taking advantage of association results in the previous section, the reliable measurement set $\mathbf{Z}_k^+$ achieved in Section 5.3.2 can be categorized into,

$$\mathbf{Z}_k^+ = \mathbf{Z}_{k,D}^+ \cup \mathbf{Z}_{k,\gamma}^+ \tag{5.5.1}$$

where $\mathbf{Z}_{k,D}^+$ and $\mathbf{Z}_{k,\gamma}^+$ denote the associated and un-associated measurements, respectively. The clutter measurement set $\mathbf{\Gamma}_k$ which often misleads the tracking process is not included in the update step.

Different from the GM-PHD updating in the previous chapter, this section presents a measurement-dependent updating process, where new-born and survival targets are performed independently with the corresponding measurements. Associated measurements are only considered for the update of survival targets, while un-associated measurements are used for the target initialization.

5.5.1 Target Survival

Owing to the enhanced identity association, the predicted intensity of survival targets can be reformulated as,

$$\nu_{k|k-1,S}(\mathbf{x}) = \nu_{k|k-1,M}(\mathbf{x}) + \nu_{k|k-1,D}(\mathbf{x}) \tag{5.5.2}$$

where $\nu_{k|k-1,M}(\mathbf{x})$ and $\nu_{k|k-1,D}(\mathbf{x})$ are the predicted intensities of potentially missed and detected targets, respectively. In the meanwhile, the identity set

of $\nu_{k|k-1,S}(\mathbf{x})$ is modified as, $\mathcal{I}_{k|k-1,S} = \mathcal{I}_{k|k-1,M} \cup \mathcal{I}_{k|k-1,D}$. For the update of predicted intensity of potentially missed targets $\nu_{k|k-1,M}(\mathbf{x})$, target states and covariance matrices are effectively inherited from the prediction, while the weights are modified with the missing detection probability p_M,

$$\nu_{k,M}(\mathbf{x}) = \sum_{j=1}^{J_{k|k-1,M}} w_{k,M}^{j} \mathcal{N}(\mathbf{x}; \mathbf{m}_{k|k,M}^{j}, \mathbf{P}_{k|k,M}^{j}) \tag{5.5.3}$$

$$w_{k,M}^{j} = P_M w_{k|k-1,M}^{j} \tag{5.5.4}$$

$$\mathbf{m}_{k|k,M}^{j} = \mathbf{m}_{k|k-1,M}^{j}, \mathbf{P}_{k|k,M}^{j} = \mathbf{P}_{k|k-1,M}^{j} \tag{5.5.5}$$

In addition, the identity set of $\nu_{k,M}(\mathbf{x})$ remains as, $\mathcal{I}_{k|k,M} = \mathcal{I}_{k|k-1,M}$. The updated intensities of potentially disappearing targets $\nu_{k,M}(\mathbf{x})$ are finally moved into the virtual zone of the proposed fusion center for target termination or reconfirmation.

On the other hand, the predicted intensity of detected targets $\nu_{k|k-1,D}(\mathbf{x})$ can be updated by employing the associated measurements $\mathbf{Z}_{k,D}^{+}$ [49],

$$\nu_{k,D}(\mathbf{x}) = \sum_{\mathbf{z} \in \mathbf{Z}_{k,D}^{+}} \sum_{j=1}^{J_{k|k-1,D}} w_{k,D}^{j}(\mathbf{z}) \mathcal{N}(\mathbf{x}; \mathbf{m}_{k|k,D}^{j}(\mathbf{z}), \mathbf{P}_{k|k,D}^{j}) \tag{5.5.6}$$

where $\mathbf{m}_{k|k,D}^{j}$, $\mathbf{P}_{k|k,D}^{j}$, and $w_{k|k,D}^{j}$ are obtained by (4.3.14)-(4.3.18). Each predicted Gaussian component increases to $|\mathbf{Z}_{k,D}^{+}|$ updated components labelled with the same identity, i.e., $I_{k|k,D}^{j} = I_{k|k-1,D}^{j}$ [6]. Then, the strategy in [129] is used to select updated components with the maximum weights as a set of possible estimated states $\tilde{\nu}_{k,D}(\mathbf{x})$. These tracking estimates are thus taken as inputs to the real zone of the fusion center for the survival fusion process.

5.5.2 Target Initialization

All selected measurements at the initial time $k-1$ are considered as newborn targets, since there are no targets tracked yet. New-born targets at time k will be initialized from measurements $\mathbf{Z}_{k,\gamma}^{+}$ [33],

$$\gamma_k(\mathbf{x}) = \sum_{j=1}^{J_{k,\gamma}} w_{k,\gamma}^{j} \mathcal{N}(\mathbf{x}; \mathbf{m}_{k,\gamma}^{j}, \mathbf{P}_{k,\gamma}^{j}) \tag{5.5.7}$$

These newborn targets are labelled with new identities $\mathcal{I}_{k,\gamma} = \{I_{k,\gamma}^{1}, ..., I_{k,\gamma}^{J_{k,\gamma}}\}$. where $J_{k,\gamma} = |\mathbf{Z}_{k,\gamma}^{+}|$ denotes the number of new-born targets. According to [49], newborn targets that are adaptively initialized from measurements can be considered to be always detected. Hence, the missed detection probability is always zero ($P_M = 0$) for the update step of newborn targets. Then, the update step for the newly initialized targets is given by,

$$\nu_{k,\gamma}(\mathbf{x}) = \sum_{\mathbf{z} \in \mathbf{Z}_{k,\gamma}^{+}} \sum_{j=1}^{J_{k,\gamma}} w_{k|k,\gamma}^{j}(\mathbf{z}) \mathcal{N}(\mathbf{x}; \mathbf{m}_{k|k,\gamma}^{j}(\mathbf{z}), \mathbf{P}_{k|k,\gamma}^{j}) \tag{5.5.8}$$

where $\mathbf{m}_{k|k,\gamma}^{j}$, $\mathbf{P}_{k|k,\gamma}^{j}$, and $w_{k|k,\gamma}^{j}$ can be achieved by the similar reasoning as in (4.3.14)-(4.3.18). Then, the management scheme in [129] is adopted again to extract the significant components to form $\tilde{\nu}_{k,\gamma}(\mathbf{x})$, which are taken as inputs to the real zone of the fusion center for the birth fusion process.

5.6 Robust fusion center

After parallel processing from measurement-driven GM-PHD filters, preliminary tracking estimates which implicitly include the target bounding boxes from each detector domain are passed to a fusion center, where a global decision is performed. The overview of the proposed robust fusion center is depicted in Fig. 5.4.

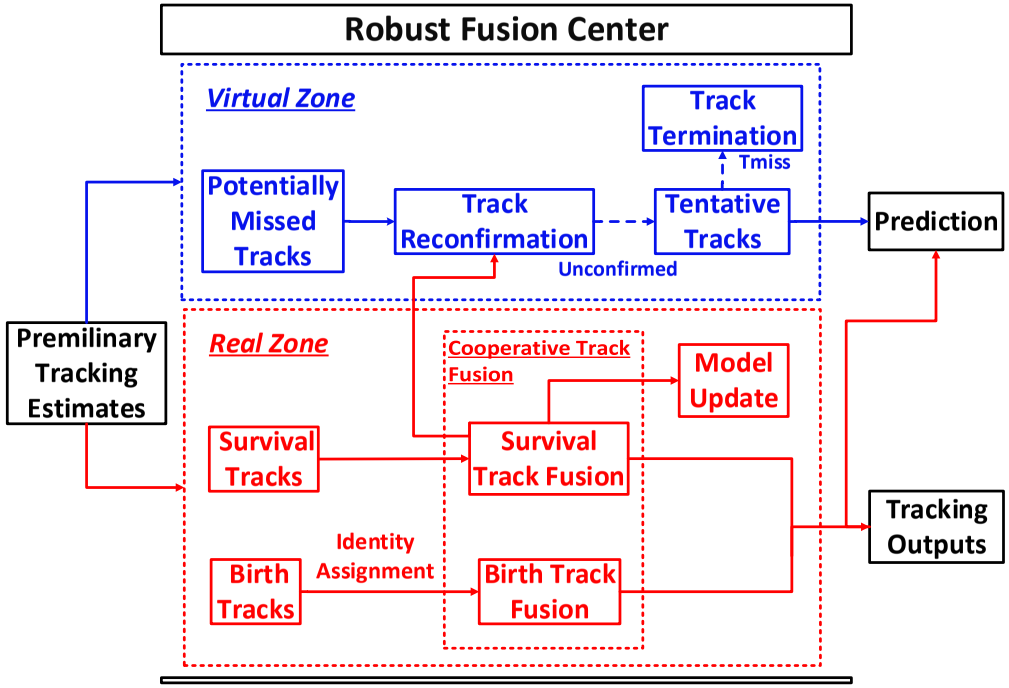


Figure 5.4. Overview of the robust fusion center. Real zone (red): the cooperative track fusion is performed, which applies survival and birth track fusion independently on the survival and birth tracks. An identity reassignment mechanism prior to the birth track fusion process is performed to overcome the identity mismatching issue. Model update is only performed on the fused survival tracks to deal with the appearance variations. Virtual zone (blue): potentially missed tracks require further reconfirmation by communicating the non-fused survival tracks from the real zone. Tracks yet reconfirmed are considered as tentative tracks. Track termination is performed to eliminate the tentative tracks with a threshold T_{miss}. Finally, tentative tracks still remaining in the virtual zone are not added to the final tracks, but are used for prediction in the next time step.

5.6.1 Real zone

In the real zone, the major processes as shown in Fig. 5.4 can be divided into two stages: cooperative track fusion and model update. The proposed cooperative track fusion follows the data-driven scheme, which performs survival and birth track fusion independently on the tracking estimates of survival and new-born targets. The overall cooperative track fusion algorithm is given with details in Algorithm 5. Model update is only performed on the fused survival tracks to deal with the appearance variations.

Cooperative track fusion

In this context, fusing tracks are actually to communicate the GM-PHD intensities, so the GCI rule which has been widely used in multi-sensor fusion with the PHD filter is exploited. The GCI fusion rule was originally proposed by Mahler [68] for fusion of multi-object functions, providing a suboptimal solution to preserve maximal information in the fused posterior from local posteriors. Battistelli et al. [71] employed exponential mixture densities (EMDs), specifically to realize the GCI fusion of GM-PHD intensities. The GCI fusion rule provides an effective solution to fuse two Gaussian mixtures v_a and v_b. It outputs a fused intensity $\nu_\varphi(\mathbf{x})$ with a fusing parameter $0 \leq \varphi \leq 1$:

$$\nu_\varphi(\mathbf{x}) = \frac{\nu_a^\varphi(\mathbf{x})\nu_b^{1-\varphi}(\mathbf{x})}{\int \nu_a^\varphi(\mathbf{x})\nu_b^{1-\varphi}(\mathbf{x})d\mathbf{x}} \tag{5.6.1}$$

Using the exponential Gaussian mixture, the power of a Gaussian mixture model can be formulated by the following approximation [71],

$$\left(\sum_{j=1}^{J} w^j \mathcal{N}(\mathbf{x}; \mathbf{m}^j, \mathbf{P}^j)\right)^\varphi \approx \sum_{j=1}^{J} (w^j)^\varphi \epsilon(\varphi, \mathbf{P}^j) \mathcal{N}(\mathbf{x}; \mathbf{m}^j, \frac{\mathbf{P}^j}{\varphi}) \tag{5.6.2}$$

where

$$\epsilon(\varphi, \mathbf{P}^j) \triangleq \frac{\left[\det\left(2\pi\mathbf{P}^j\varphi^{-1}\right)\right]^{\frac{1}{2}}}{\left[\det\left(2\pi\mathbf{P}^j\right)\right]^{\frac{\varphi}{2}}} \tag{5.6.3}$$

Figure 5.5. Qualitative comparison between the use of the original GCI fusion and the proposed fusion center on the MOT16-09 video sequence. Detection results show that a target located in the middle of the scene is detected by the body-parts detector (green) but missed by the full-body detector (blue). In this case, fusion through the original GCI rule would lose the target even though it has been detected by the body-parts detector. In the proposed fusion center, this target only observed by the body-parts detector would be reconfirmed and preserved in the tracking outputs (better viewed in color version).

Therefore, the GCI fusion for GM-PHD intensities can be equivalent to applying covariance intersection (CI) pairwise to Gaussian components from two intensities [6]. Both human detectors used in this tracking scenario have fully overlapped field of views (FOVs), which results in applying the GCI fusion rule in the current fusion process without requiring transformation.

However, a recent study in [73] has suggested that the original GCI fusion rule is prone to missed detections. Missing targets at one detector can degrade the fusion performance with detected targets from other detectors. To be more specific, suppose a target $\mathbf{x}_k$ at time k is detected by detector a, but missed in detector b, this results in $\nu_a(\mathbf{x}_k) > 0$ and $\nu_b(\mathbf{x}_k) \approx 0$. After applying the GCI fusion in (5.6.1), the fused result can be $\nu_\varphi(\mathbf{x}_k) \approx 0$, which implies that the target is lost even though a larger fusion weight φ is given in $\nu_b(\mathbf{x}_k)$. To overcome the aforementioned issue, a robust fusion center is developed with real and virtual zones, aiming to improve the fusion process and tracking consistency. Qualitative results which demonstrate the

advantages of the proposed fusion algorithm over the original GCI rule are shown in Fig. 5.5.

Survival track fusion For the survival track fusion, survival tracks from both detectors possess the same identity library. It means that a survival target processed by both detectors will be given by the same identity. Therefore, it is feasible to directly apply CI on survival tracks with the same identities.

Given each pair of Gaussian components i and j with the same label from the intensities $\tilde{\nu}^{a}_{k,D}(\mathbf{x})$ and $\tilde{\nu}^{b}_{k,D}(\mathbf{x})$, each fused component with the corresponding weight can be reformulated with the following characteristics [140],

$$\mathbf{m}^{ij}_{ab,k} = \mathbf{P}^{ij}_{ab,k}\left[\varphi\big(\mathbf{P}^{i}_{a,k}\big)^{-1}\mathbf{m}^{i}_{a,k} + \big(1-\varphi\big)\big(\mathbf{P}^{j}_{b,k}\big)^{-1}\mathbf{m}^{j}_{b,k}\right] \tag{5.6.4}$$

$$\mathbf{P}^{ij}_{ab,k} = \left[\varphi\big(\mathbf{P}^{i}_{a,k}\big)^{-1} + \big(1-\varphi\big)\big(\mathbf{P}^{j}_{b,k}\big)^{-1}\right]^{-1} \tag{5.6.5}$$

$$\begin{aligned} w^{ij}_{ab,k} = \ & \big(w^{i}_{a,k}\big)^{\varphi}\big(w^{j}_{b,k}\big)^{1-\varphi}\epsilon\big(\varphi,\mathbf{P}^{i}_{a,k}\big)\epsilon\big(1-\varphi,\mathbf{P}^{j}_{b,k}\big) \\ & \mathcal{N}\left(\mathbf{m}^{i}_{a,k}-\mathbf{m}^{j}_{b,k};0,\frac{\mathbf{P}^{i}_{a,k}}{\varphi}+\frac{\mathbf{P}^{j}_{b,k}}{1-\varphi}\right) \end{aligned} \tag{5.6.6}$$

$$\epsilon\big(\varphi,\mathbf{P}^{i}_{a,k}\big) = \frac{\left[\det\left(2\pi\mathbf{P}^{i}_{a,k}\varphi^{-1}\right)\right]^{\frac{1}{2}}}{\left[\det\left(2\pi\mathbf{P}^{i}_{a,k}\right)\right]^{\frac{\varphi}{2}}} \tag{5.6.7}$$

therefore, these fused components establish the fused intensity of survival tracks $\tilde{\nu}^{ab}_{k,D}(\mathbf{x})$. Conventional fusion approaches [70–72] usually preset the value of φ as 0.5 with the assumption that both sensors have the same sensing abilities in all aspects. However, the aforementioned design of fusion weight may be no longer applicable to real tracking applications, since it is not always feasible for detectors to have the same sensing abilities due to different imaging conditions or camera motions. To this end, an experimental study to determine the parameter φ will be given in the Section 5.7.3. Non-fused survival tracks from both detectors are preserved into final tracking

set.

Algorithm 5: Cooperative Track Fusion ($k > 1$)

Input : Identity sets: $\mathcal{I}^a_{k,\gamma}$ and $\mathcal{I}^b_{k,\gamma}$; $\mathcal{I}^a_{k,D}$ and $\mathcal{I}^b_{k,D}$. Local tracks: $\tilde{\nu}^a_{k,D}$ and $\tilde{\nu}^b_{k,D}$; $\tilde{\nu}^b_{k,\gamma}$ and $\tilde{\nu}^b_{k,\gamma}$.

Output: Final tracking results: ν^{ab}_k and $\mathcal{I}^{ab}_k$.

1 Compute common survival identity sets: $\mathcal{I}^{ab}_{k,D} = \mathcal{I}^a_{k,D} \cap \mathcal{I}^b_{k,D}$.

2 Compute fused survival tracks $\tilde{\nu}^{ab}_{k,D}$: apply $\tilde{\nu}^a_{k,D}$ and $\tilde{\nu}^b_{k,D}$ into (5.6.4)-(5.6.7) based on $\mathcal{I}^{ab}_{k,D}$.

3 Compute independent survival identity sets: $\mathcal{I}^a_{k,D^*} = \mathcal{I}^a_{k,D} \setminus \mathcal{I}^{ab}_{k,D}$; $\mathcal{I}^b_{k,D^*} = \mathcal{I}^b_{k,D} \setminus \mathcal{I}^{ab}_{k,D}$.

4 Obtain independent survival tracks $\tilde{\nu}^a_{k,D^*}$ and $\tilde{\nu}^b_{k,D^*}$ with $\mathcal{I}^a_{k,D^*}$ and $\mathcal{I}^b_{k,D^*}$.

5 Apply the identity reassignment with IOU scores to obtain reassigned identity sets $\tilde{\mathcal{I}}^a_{k,\gamma}$ and $\tilde{\mathcal{I}}^b_{k,\gamma}$.

6 Compute common birth identity sets: $\tilde{\mathcal{I}}^{ab}_{k,\gamma} = \tilde{\mathcal{I}}^a_{k,\gamma} \cap \tilde{\mathcal{I}}^b_{k,\gamma}$.

7 Compute fused birth tracks $\tilde{\nu}^{ab}_{k,\gamma}$: apply $\tilde{\nu}^a_{k,\gamma}$ and $\tilde{\nu}^b_{k,\gamma}$ into (5.6.4)-(5.6.7) based on $\tilde{\mathcal{I}}^{ab}_{k,\gamma}$.

8 Compute independent birth identity sets: $\tilde{\mathcal{I}}^a_{k,\gamma^*} = \tilde{\mathcal{I}}^a_{k,\gamma} \setminus \tilde{\mathcal{I}}^{ab}_{\gamma,k}$; $\tilde{\mathcal{I}}^b_{k,\gamma^*} = \tilde{\mathcal{I}}^b_{k,\gamma} \setminus \tilde{\mathcal{I}}^{ab}_{k,\gamma}$.

9 Obtain independent birth tracks $\tilde{\nu}^a_{k,\gamma^*}$ and $\tilde{\nu}^b_{k,\gamma^*}$ with $\tilde{\mathcal{I}}^a_{k,\gamma^*}$ and $\tilde{\mathcal{I}}^b_{k,\gamma^*}$.

10 Obtain final tracks results:
$\nu^{ab}_k = \tilde{\nu}^{ab}_{k,D} \cup \tilde{\nu}^a_{k,D^*} \cup \tilde{\nu}^b_{k,D^*} \cup \tilde{\nu}^{ab}_{k,\gamma} \cup \tilde{\nu}^a_{k,\gamma^*} \cup \tilde{\nu}^b_{k,\gamma^*}$
$\mathcal{I}^{ab}_k = \mathcal{I}^{ab}_{k,D} \cup \mathcal{I}^a_{k,D^*} \cup \mathcal{I}^b_{k,D^*} \cup \tilde{\mathcal{I}}^{ab}_{k,\gamma} \cup \tilde{\mathcal{I}}^a_{k,\gamma^*} \cup \tilde{\mathcal{I}}^b_{k,\gamma^*}$.

Birth track fusion Different from a survival track labelling system, a newborn target detected by each detector is labelled with two different local identities during the tracking process, which is due to the different detection orders in each detector. This is an issue of inconsistent identity assignment [72] which causes confusions in the fusion process. To remedy this, it is essential to reassign a same identity to matched birth tracks from both detectors. An identity reassignment mechanism is therefore developed before the fusion process, which is designed to calculate pairwise similarity score through intersection-over-union (IOU), i.e. $IOU(a,b) = (Area(a) \bigcap Area(b))/(Area(a) \bigcup Area(b))$, between any two components from

birth intensities of $\tilde{\nu}^{a}_{k,\gamma}(\mathbf{x})$ and $\tilde{\nu}^{b}_{k,\gamma}(\mathbf{x})$ [6]. When the similarity score computed on any two birth components is greater than 0.5, these two birth components will be reassigned with a same identity. Two components with the same reassigned identities from $\tilde{\nu}^{a}_{k,\gamma}(\mathbf{x})$ and $\tilde{\nu}^{b}_{k,\gamma}(\mathbf{x})$, respectively can be fused by the similar reasoning in (5.6.4)-(5.6.7). In addition, non-fused birth tracks are copied into the final tracking results.

Model update

After the fusion process, it is necessary to update the appearance model for the newly achieved tracking estimates in order to handle the appearance variations. To avoid introducing background noise in the model update, since the correlation filters are sensitive to false positives, the proposed strategy is to only update the DCFs with the fused survival tracks $\tilde{\nu}^{ab}_{k,D}$ in the real zone. An update mechanism in [84] is specifically adopted for updating the DCF $\mathbf{c}^{(l)}_{k,d}$ on the l-th layer. Specifically, each DCF $\mathbf{c}^{(l)}_{k,d}$ on the l-th layer at time k is updated as follows,

$$\mathbf{U}^{(l)}_{k,d} = (1-\eta)\mathbf{U}^{(l)}_{k-1,d} + \eta\hat{\mathbf{g}}^{(l)}_{k} \odot (\hat{\mathbf{f}}^{(l)}_{k,d})^{\dagger} \tag{5.6.8}$$

$$\mathbf{V}^{(l)}_{k,d} = (1-\eta)\mathbf{V}^{(l)}_{k-1,d} + \eta\sum_{d=1}^{D}\hat{\mathbf{f}}^{(l)}_{k,d} \odot (\hat{\mathbf{f}}^{(l)}_{k,d})^{\dagger} \tag{5.6.9}$$

where $\mathbf{U}^{(l)}_{k,d}$ and $\mathbf{V}^{(l)}_{k,d}$ are respectively the numerator and denominator of the learned DCF $\mathbf{c}^{(l)}_{k,d}$ in 5.4.3, and η is a learning rate parameter. Besides, each newly-detected track will be initialized with an appearance model using DCFs.

5.6.2 Virtual zone

In this section, the virtual zone is designed to manage further validation and processing on the potentially missed tracks which are strictly excluded from the fusion process as shown in Fig. 5.4. The two major tasks of track

reconfirmation and track termination are included in this zone.

Track reconfirmation

Potentially missed tracks in the virtual zone are required for further reconfirmation by communicating the non-fused survival tracks from the real zone. Typically, a track can be reconfirmed only if its identity can be found in the non-fused survival tracks from each detector domain. Then the reconfirmed tracks are removed from the virtual zone. However, tracks yet reconfirmed are considered as tentative tracks.

Track termination

Tracks which are found from neither detectors are automatically moved into tentative tracks. Tentative tracks missing more than T_{miss} frames are eliminated. Note that tentative tracks remaining in the virtual zone do not contribute to the final tracks, but they are propagated in the next time step.

In the next section, the experiments will be explained, and results with discussions will be given to examine the tracking performance of the proposed method.

5.7 Experiments

This section firstly presents the well-established datasets, then elaborates the detailed setting of the proposed tracker implementation. Next, experiments are made on the validation sequences from the MOT16 benchmark [4] to investigate the impact of the proposed fusion at different levels and the influence of different parameters on the overall tracking performance. Lastly, the proposed method is evaluated on the test set of MOT16 and MOT17 benchmarks [4].

5.7.1 Datasets

Experiments are conducted on the MOTChallenge Benchmark dataset[1] which is the most commonly used for the quantification of multiple human tracking. This benchmark collects various challenging video sequences recorded by static or moving cameras, and under the complex scenes of illumination changes, varying viewpoints and weather conditions. The MOT16 Challenge [4], consists of 7 training and 7 testing fully annotated video sequences, as well as providing public object detections generated by [2] for fair comparisons. MOT17 Challenge [4] is built on the MOT16 Challenge with a new and more accurate ground truth. Each sequence is provided with 3 sets of public detections (DPM, FRCNN, and SDP). The training video sequences with available ground truths are primarily utilized to process the performance analysis, while testing sequences are used to generate quantitative comparisons against existing state-of-the-art tracking methods.

5.7.2 Implementation details

In this work, for the matching scheme, $\sigma_c = 0.1$, set the regularization parameter of (5.4.2) $\lambda = 10^{-4}$, and $\sigma_s^2 = 30$ [30]. In addition, the matching threshold $\varpi = 10$ is experimentally determined in Table 5.3, which results in the settings of two coefficients $\alpha = 0.2$ and $\beta = -2$ in (5.4.6). Likewise, the fusing weight parameter φ is experimentally analyzed in Fig. 5.7. The network in [131] is utilized for the feature generation, where the outputs of the convolutional layers *conv3-3* and *conv4-3* are used as desired features. The extracted features are multiplied with a cosine window to mitigate the boundary effect [30, 96]. Implementation of measurement-driven GM-PHD filtering follows the model settings in the previous chapter. The thresholds of confidence scores c_{th} are set to 0.1 for DPM detections, and set to $-\infty$ for FRCNN and SDP detections. The maximum missing frames T_{miss} is set

[1] https://motchallenge.net/

Table 5.1. Ablation study of the proposed method on MOT16-09 sequence.

Feature	Processer	MOTA (↑)	FP (↓)	FN (↓)	IDS (↓)	MT (↑)	ML (↓)
ST	*FB*	29.4	1603	1981	128	32.0%	16.0%
ST	*BP*	33.1	1461	1943	112	16.0%	8.0%
ST	*FB-BP*	32.8	1839	**1573**	122	**36.0%**	12.0%
ST	*BP-FB*	34.9	1483	1814	125	24.0%	8.0%
ST	*Proposed*	**39.9**	**1112**	1948	**100**	28.0%	**8.0%**
A	*FB*	34.6	1279	2036	124	**28.0%**	16.0%
A	*BP*	39.4	1100	1983	101	12.0%	**8.0%**
A	*FB-BP*	38.5	1208	1923	102	20.0%	16.0%
A	*BP-FB*	41.1	1137	**1847**	111	24.0%	12.0%
A	*Proposed*	**45.5**	**919**	1850	**98**	24.0%	12.0%
ST+A	*FB*	37.4	1136	2031	123	28.0%	16.0%
ST+A	*BP*	41.0	1012	1994	97	16.0%	8.0%
ST+A	*FB-BP*	40.8	1217	1798	95	24.0%	12.0%
ST+A	*BP-FB*	43.0	1108	**1775**	110	20.0%	12.0%
ST+A	*Proposed*	**47.5**	**827**	1838	**93**	**28.0%**	**8.0%**

to 3 in this work, which is experimentally selected in Fig. 5.8.

5.7.3 Performance analysis

In this section, the performance analysis is presented to evaluate the effectiveness of the proposed method, including the ablation study of different proposed components as well as the effects of different parameter settings. For this purpose, experiments are conducted on MOT16-09 and MOT16-11 from the training set of MOT16 Challenge Benchmark [4], as the scene conditions and camera motions are distinct between these two validation sequences.

Ablation study

To investigate the contribution of different components in the proposed method, The ablation study is performed in two aspects. Firstly, the proposed fusion strategy is compared with single detector full-body (FB) or body parts (BP) individually processed by data-driven GM-PHD filter, and also the sequential update PHD fusion [26] with different detector orders, FB before BP (FB-BP) and BP before FB (BP-FB). Next, the impact of dif-

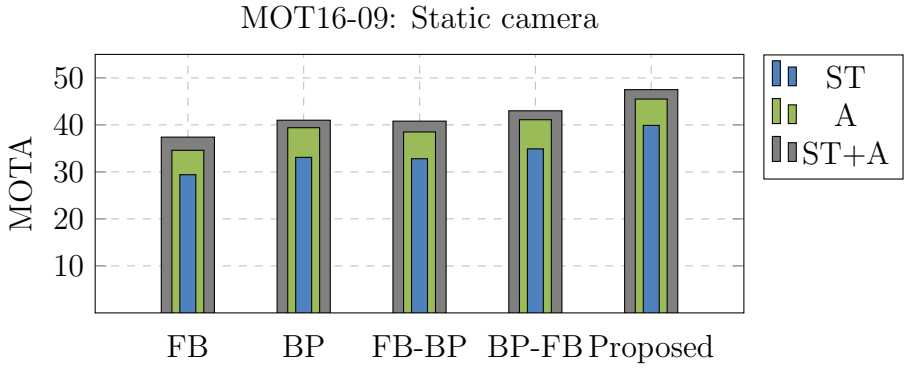
ferent feature models in enhanced identity association is analyzed, including spatio-temporal information (ST) and discriminative correlation matching (A).

Tables 5.1 and 5.2 report the detailed evaluations on the validation sequences. In general, from the results above, the full tracking model achieves noticeably improved performance on almost all presented evaluation metrics. On the one hand, the proposed cooperative track fusion shows the advantage over the sequential fusion approaches (FB-BP & BP-FB) and single detectors (FB & BP), as it improves MOTA and reduces the number of FNs regardless of different feature models. This is because the proposed fusion algorithm exploits well the merits of both human detectors, thereby enabling the tracker to recover the missed detections (low FNs) and provide more reliably consistent tracks. Another finding on these results is that the proposed discriminative correlation matching (A) mainly contributes to reducing the number of FPs, and ID switches. This may be explained by the fact that the proposed appearance model could help the tracking system to establish better mappings between the detections and real targets. In addition, Improvements on the MT and ML further demonstrate the proposed method benefits the tracking robustness and consistency.

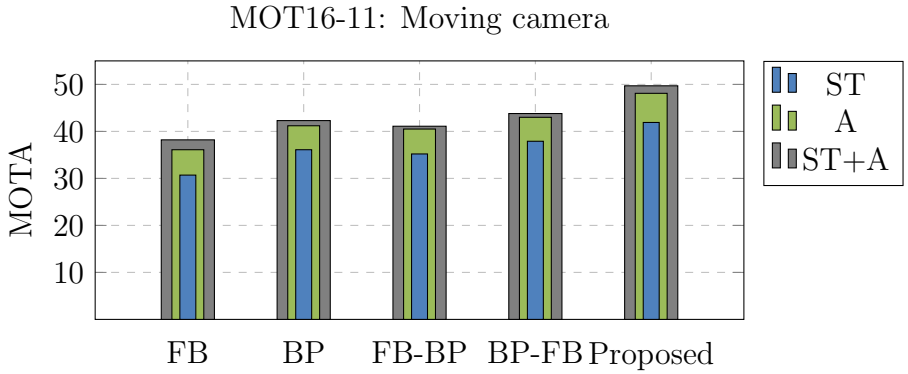
Fig. 5.6 intuitively reveals the advantage of the proposed feature-level fusion. The combined model achieves the best MOTA performance regardless of different fusion solutions, suggesting that fusing features can improve ambiguities which occur in either motion dynamics or visual content. It is shown that the proposed appearance model contributes most to improve the tracking performance. Moreover, spatio-temporal information helps increase the overall accuracy, especially to facilitate localization of the targets with similar appearances. Overall, ablation study results above verify the proposed multi-level fusion is helpful to address target ambiguities, and provide redundancy in each detector domain.

Table 5.2. Ablation study of the proposed method on MOT16-11 sequence.

Feature	Processer	MOTA (↑)	FP (↓)	FN (↓)	IDS (↓)	MT (↑)	ML (↓)
ST	*FB*	30.7	2379	3871	111	21.7%	42.0%
ST	*BP*	36.1	2155	**3607**	102	23.1%	34.7%
ST	*FB-BP*	35.2	2147	3675	122	24.6%	40.5%
ST	*BP-FB*	37.9	1758	3845	95	20.2%	43.4%
ST	*Proposed*	**41.9**	**1625**	3622	**86**	**26.1%**	**31.9%**
A	*FB*	36.1	1808	3953	100	20.3%	42.0%
A	*BP*	41.2	1637	3662	94	20.3%	**34.7%**
A	*FB-BP*	40.5	1775	3574	108	16.0%	44.9%
A	*BP-FB*	43.0	1633	3507	90	20.2%	43.4%
A	*Proposed*	**48.1**	**1359**	**3324**	**80**	**21.7%**	36.2%
ST+A	*FB*	38.2	1776	3793	97	18.9%	43.4%
ST+A	*BP*	42.3	1630	3576	91	23.1%	**34.7%**
ST+A	*FB-BP*	41.1	1713	3596	98	**24.6%**	40.5%
ST+A	*BP-FB*	43.7	1573	3496	91	21.7%	39.1%
ST+A	*Proposed*	**49.7**	**1266**	**3278**	**74**	24.6%	37.7%



(a)



(b)

Figure 5.6. MOTA performance comparison of the proposed method on the validation sequences.

Table 5.3. Comparison of the tracking results with different matching thresholds of appearance model on the validation sequences

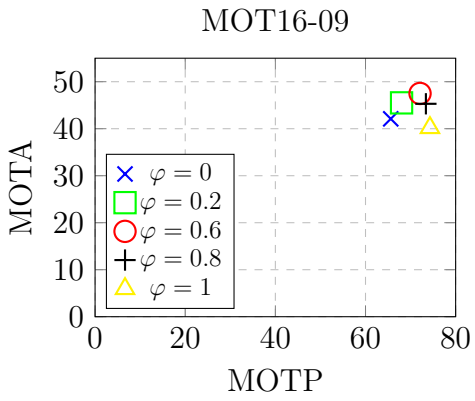
MOT16-09				
ϖ	6	10	14	18
MOTA (↑)	44.3	47.5	46.5	45.9
FP (↓)	1149	827	754	676
FN (↓)	1667	1838	1958	2083
MOT16-11				
ϖ	6	10	14	18
MOTA (↑)	46.0	49.7	48.5	49.1
FP (↓)	1614	1266	1146	1019
FN (↓)	3220	3278	3493	3582

Analysis of parameters

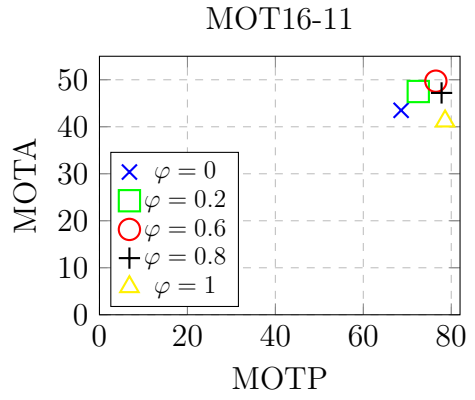
In this section, an experimental study was made on the validation sequences to analyze the influence of different critical parameters on the tracking performance. Firstly, different values of matching threshold ϖ which controls the gate to accept the matched pairs were tested, as illustrated in Table 5.3. Metrics of MOTA, FP, and FN are employed to investigate the relative change on the performance.

For the MOT16-09 sequence, when the value of ϖ is altered from 6 to 18, the number of false positives is largely decreased, whereas the MOTA score is just slightly improved. This is because the higher matching threshold ignores some matched pairs in ambiguous cases. Similar results are also found on the MOT16-11 sequence.

In addition, the fusing parameter φ which determines the relative fusion weight of each detector was analyzed with 5 different settings. The results of different φ are shown in Fig. 5.7. Since both detectors have different detecting abilities, the tracking performance is slightly sensitive to the fusing parameter. As can seen from the extreme cases, the full body detector can provide better precision MOTP but less accuracy MOTA, while the body-parts detector has the opposite impact on the performance. To this end, an appropriate value for φ can be experimentally determined, in order to



(a)



(b)

Figure 5.7. Comparisons of MOTA performance with different fusing parameters φ on the validation sequences. Results closer to the upper right corner perform better (better viewed in color version).

manage the trade off between the MOTA and MOTP.

To study the impact of the parameter T_{miss} which controls the number of consecutive missing frames to terminate tentative tracks, a set of pilot tests $T_{miss} = \{1, 3, 5, 7, 9\}$ were made on the validation sequences. The frame rate used in both sequences is 30, which is obtained from the MOT16 Challenge [4]. The results in Fig. 5.8 show that the best setting for the maximum missing frames is $T_{miss} = 3$.

The above analysis demonstrates that the tracking performance of the proposed method is slightly sensitive to the parameter changes in the reasonable range. Parameters with the highest performance are used and remain

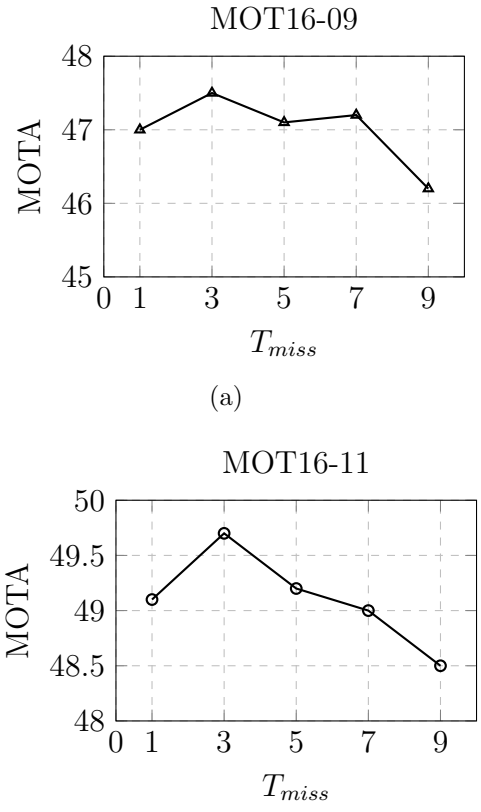


Figure 5.8. Comparisons of MOTA performance with different values of parameter T_{miss} on the validation sequences.

unaltered throughout the benchmark evaluations in the next section.

5.7.4 Benchmark evaluations

The proposed tracking system was evaluated on the test set of the MOT16 and MOT17 Challenge Benchmarks [4]. Quantitative results compared with recent state-of-the-art trackers published on the leaderboard are shown in Tables 5.4 and 5.5. These include online trackers: AMIR [37], DCCRF16 [42], CDA_DDALv2 [59], Deep-align [141], EAMTTPub [29], GM_PHD_N1T [104], GMPHD_HDA [60], MOTDT [39], PHD_GSDL [10] and GMPHD_KCF [66], GM_PHD [27], and also offline trackers: INTERA_MOT [41], FWT [36], MCjoint [19], MHT_DAM [142], EDMT [51], QuadMOT16 [18], and MHT_bLSTM [101]. Evaluation measures with (↑) or (↓) respectively de-

note that higher is better, or lower is better. The MOTA score which is regarded as the most important measure, is employed to rank the trackers. The qualitative tracking results on the MOT17 dataset are also shown in Fig. 5.9.

As can seen from Table 5.4, the proposed method (MTDF) reports the state-of-the-art MOTA and the second best MT compared with online methods, which indicates the proposed method is capable to provide more reliably consistent tracks. Likewise, MTDF achieves the lowest ML and FN, even including offline methods, demonstrating that the proposed method has the advantage to recover missed targets by fusing parts to a cohesive whole. Note that offline methods using future frame information usually achieve more promising performance than online methods.

In terms of the evaluations on the MOT17 Benchmark, Table 5.5 shows that the proposed method (MTDF17) achieves the second best performance in MOTA, best MT among published online methods, and is on a par with state-of-the-art offline methods. Furthermore, MTDF17 records the best ML and FN scores among all listed trackers here. As a trade off, MTDF17 produces more FPs.

Table 5.4. Quantitative comparison between the proposed approach (MTDF) and state-of-the-art approaches on the MOT16 benchmark. The best results of online or offline approaches are shown in bold respectively. (Las submitted on May 15, 2018)

Method	Mode	**MOTA** (↑)	MOTP (↑)	MT (↑)	ML (↓)	FP (↓)	FN (↓)	ID Sw (↓)	Frag (↓)	Hz (↑)
MOTDT [39]	Online	**47.6**	74.8	**15.2%**	38.3 %	9,253	85,431	792	1,858	**20.6**
AMIR [37]	Online	47.2	75.8	14.0%	41.6%	2,681	92,856	774	1,675	1.0
MTDF	Online	45.7	72.6	14.1%	**36.4%**	12,018	**84,970**	1,987	3,377	1.5
DCCRF16 [42]	Online	44.8	75.6	14.1%	42.3%	5,613	94,133	968	1,378	0.1
CDA_DDALv2 [59]	Online	43.9	74.7	10.7%	44.4%	6,450	95,175	676	1,795	0.5
Deep-align [141]	Online	40.8	74.4	13.7%	38.3%	15143	91792	1051	2210	6.5
EAMTT_pub [29]	Online	38.8	75.1	7.9%	49.1%	8,114	102,452	965	1,657	11.8
GM_PHD_N1T [104]	Online	33.3	**76.8**	5.5%	56.0%	**1,750**	116,452	3,499	3,594	9.9
GMPHD_HDA [60]	Online	30.5	75.4	4.6%	59.7%	5,169	120,970	**539**	**731**	13.6
FWT [36]	Offline	**47.8**	75.5	19.1%	**38.2%**	8,886	**85,487**	852	1,534	0.6
MCjoint [19]	Offline	47.1	76.3	**20.4%**	46.9%	6,703	89,368	**370**	**598**	0.6
MHT_DAM [142]	Offline	45.8	76.3	16.2%	43.2%	6,412	91,758	590	781	0.8
INTERA_MOT [41]	Offline	45.4	74.4	18.1%	38.7%	13,407	85,547	600	930	**4.3**
EDMT [51]	Offline	45.3	75.9	17.0%	39.9%	11,122	87,890	639	946	1.8
QuadMOT16 [18]	Offline	44.1	**76.4**	14.6%	44.9%	**6,388**	94,775	745	1,096	1.8

Table 5.5. Quantitative comparison between the proposed approach (MTDF17) and state-of-the-art approaches on the MOT17 benchmark. The best results of online or offline approaches are shown in bold respectively. (Las submitted on May 23, 2018)

Method	Mode	**MOTA** (↑)	MOTP (↑)	MT (↑)	ML (↓)	FP (↓)	FN (↓)	ID Sw (↓)	Frag (↓)	Hz(↑)
MOTDT17 [39]	Online	**50.9**	76.6	17.5%	35.7%	24,069	250,768	**2,474**	5,317	18.3
MTDF17	Online	49.6	75.5	**18.9%**	**33.1%**	37,124	**241,768**	5,567	9,260	1.2
PHD_GSDL17 [10]	Online	48.0	**77.2**	17.1%	35.6%	**23,199**	265,954	3,998	8,886	6.7
EAMTT [29]	Online	42.6	76.0	12.7%	42.7%	30,711	288,474	4,488	**5,720**	1.4
GMPHD_KCF [66]	Online	39.6	74.5	8.8%	43.3%	50,903	284,228	5,811	7,414	3.3
GM_PHD [27]	Online	36.4	74.5	4.1%	57.3%	23,723	330,767	4,607	11,317	**38.4**
FWT [36]	Offline	**51.3**	75.9	21.4%	**35.2%**	24,101	247,921	2648	4279	0.2
jCC [19]	Offline	51.2	75.9	20.9%	37.0%	25,937	247,822	**1,802**	2,984	1.8
MHT_DAM [142]	Offline	50.7	**77.5**	20.8%	36.9%	**22,875**	252,889	2,314	**2,865**	0.9
EDMT17 [51]	Offline	50.0	77.3	**21.6%**	36.3%	32,279	**247,297**	2,264	3,260	0.6
MHT_bLSTM [101]	Offline	47.5	77.5	18.2%	41.7%	25,981	268,042	2,069	3,124	**1.9**

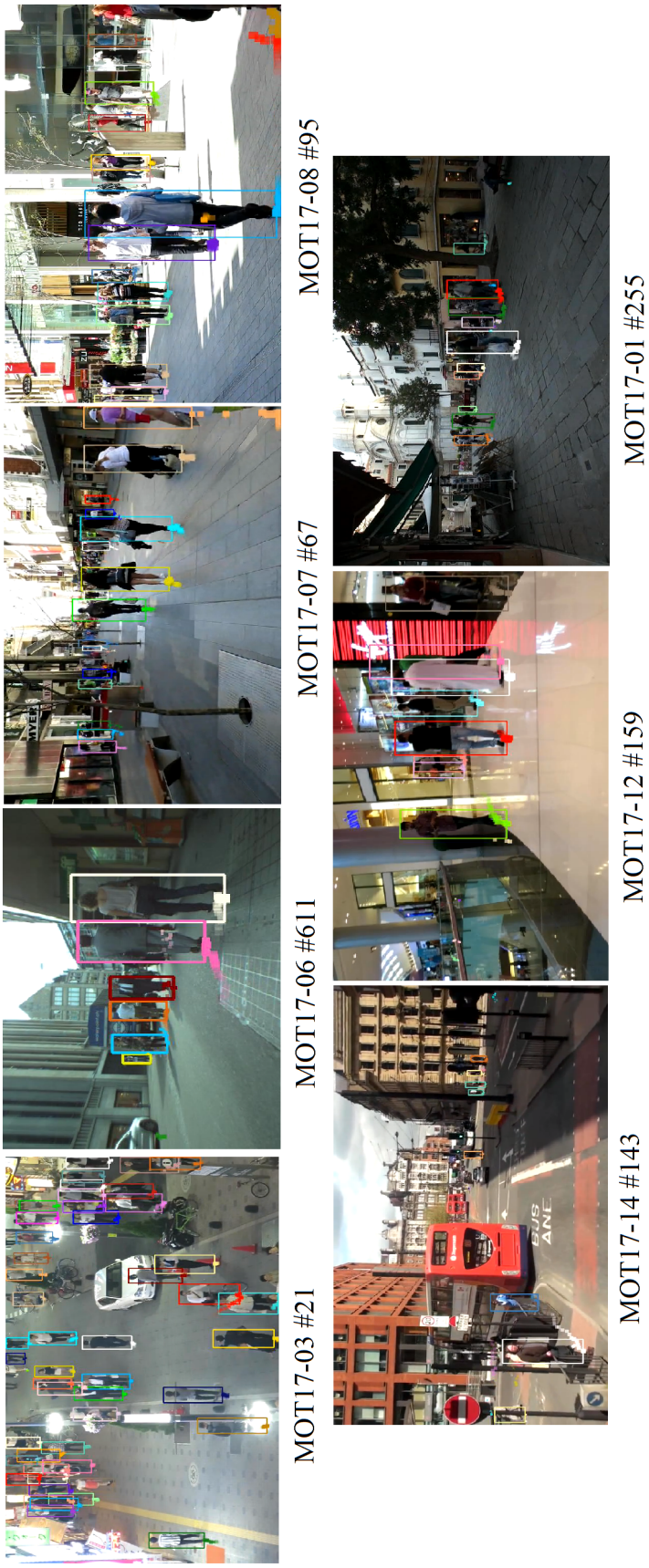


Figure 5.9. Visual tracking results of the proposed tracking system on the test set of MOT17 dataset. Different colors of the bounding boxes represent identities.

5.7.5 Discussions with other MOT methods

In this section, an explicit discussion is made on the benchmark performance between the proposed approach and other MOT methods. Compared with a similar approach [36], the proposed fusion approach achieves better ML and FN, while their fusion approach formulated by a quadratic program performs better in MT and FP due to exploiting the long-term latency. It is worth noting that the proposed method is outperformed by MOTDT [39], which specifically achieves more promising performance in MOTA, FP, ID Sw, and Frag. The effectiveness of this tracker can be attributed to two advantages, one is that a fully CNN based candidate selection is well designed to remove the false positives in the early stage, thereby gaining more reliable detections for the data association. The other is triplet-based person re-identification improves the target appearance model with better discriminativity, so as to reduce the number of ID switches and fragments.

Moreover, the proposed method is specifically compared against other state-of-the-art RFS based methods published on the leaderboard, including EAMTTPub [29], GM_PHD_N1T [104], GMPHD_HDA [60], PHD_GSDL17 [10], GM_PHD [27], and GMPHD_KCF [66]. Overall, the proposed method achieves best tracking performance among all RFS based methods in both benchmarks. In Table 5.5, by comparing with the PHD_GSDL17 [10] which has been presented in Chapter 3, the current approach here effectively reduces a large amount of FN, and improves MOTA by 1.6%, MT by 1.8% and ML by 2.5%. The improvement over EAMTT [29], which also performs early association but only using spatial constraints, verifies the benefits of integrating the target-specific appearance model within the current approach, particularly in establishing much more reliable and stable tracks. Better yet, the proposed method outperforms single-level fusion based approaches in [27] and [66] with large margins on the MOT17 dataset, demonstrating that the proposed multi-level tracking fusion can increase the robustness

and reliability for multiple human tracking task. Evaluations above imply the proposed multi-level fusion approach can considerably strengthen RFS based multiple human tracking.

Table 5.6. Quantitative comparison between the proposed approach (MTDF) and previously published versions on the MOT16 benchmark.

Method	**MOTA** (↑)	MT (↑)	ML (↓)	FP (↓)	FN (↓)	ID Sw (↓)	Frag (↓)	Hz(↑)
MTDF	**45.7**	**14.1%**	**36.4%**	12,018	**84,970**	1,987	3,377	1.5
[6]	39.3	12.5%	40.8%	12,430	93,394	4,934	5,886	**9.7**
[30]	37.7	9.2%	46.5%	**6,515**	105,389	**1608**	**3,372**	2.2

To analyze the advantages of the currently proposed approach over methods given in previous conference papers [6] and [30], experiments were made on the MOT16 benchmark to evaluate their tracking performance. The results in Table 5.6 demonstrate the proposed approach achieves the best scores in terms of MOTA, MT, ML, and FN. The improved performance can be attributed to the major contribution which is the proposed fusion center. It enhances the overall fusion process and contributes to producing lower scores in FN, ML and better MOTA performance. Other improvements ensue from the proposed multi-level cooperative fusion method integrating well the merits from previous works at different stages. By exploiting the complementary benefits of using the two human detectors in [6], more reliable tracks are provided, which yields improved MT score. The work in [30] which improves the identity association helps the tracker to produce better performance in ID Sw and Frag. As a suggestion for future work, FP in the proposed fusion approach can be further improved by exploiting the long-term latency of target trajectories.

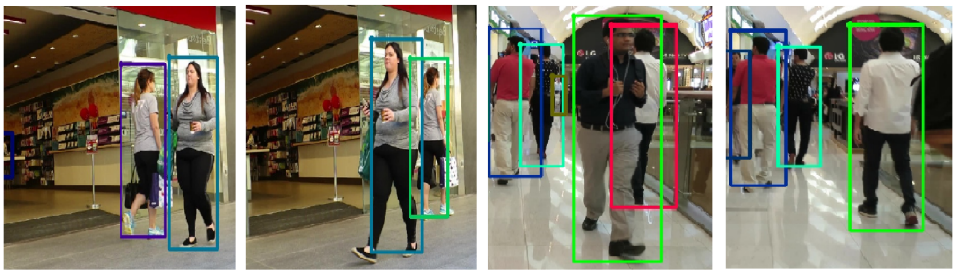
5.7.6 Runtime analysis

In this work, all the experiments were conducted on a Laptop with an Intel i7 3.5GHz CPU, with 32GB of memory and a GeForce GTX 1060 without

parallel speeding. The runtime comparisons with other published methods presented in the MOT Benchmark are summarized in Tables 5.4 and 5.5. The proposed method returns a longer runtime performance compared with the existing RFS based methods, such as EAMTTPub [29], GMPHD_KCF [66] and GM_PHD [27]. In [29], target appearance models which usually consume the most computations were not applied in the association. This increases the tracking speed but generates less promising tracking results. In [66] and [27], single-level based fusion trackers using hand-crafted features for appearance modelling were used, in a sense, which can reduce the computational complexity. To further analyse the computation cost of the proposed multi-level fusion system, runtime comparisons between the journal version and previous conference papers are given in Table 5.6. These results reveal the most consuming component is with the enhanced identity association, particularly in the use of the appearance term. This can be comprehended from the runtime of [30], computing the target appearance models and their update in a many to many scenario clearly slows down the running speed, especially when the environment becomes much congested. However, the collaborative detector fusion approach [6] can run faster by disabling the expensive appearance matching.

5.7.7 Failure cases

This section presents selected tracking failure cases of the proposed approach. In Fig. 5.10 (a), it is clear to see that when a target is occluded for a longer time, it will be labelled with a new identity after occlusion. For instance, the woman on the left is initially labelled with a purple bounding box. When she reappears after occlusion, she is initialized as a new born target with a newly assigned green box. This can be further addressed by maintaining the long-term memory of appearance models for potentially disappeared targets. In addition, it is not difficult to see in Fig. 5.10 (b), two people switch their



(a) MOT16-09 (b) MOT16-11

Figure 5.10. Selected tracking failure cases of the proposed method.

IDs on the right. A man in black shirt is initially labelled with a green bounding box, but his identity (green box) is shifted to the man in white shirt in the next frames. These challenges can be improved by incorporating better deep features in the correlation matching, such as in [143] and [144], so as to better discriminate targets within the occlusion region. Alternatively, occlusion-aware detection [145] can be used in the future work to redetect the targets after long-term occlusion.

5.8 Summary

In summary, this chapter contributed to improving tracking performance by proposing a multi-level cooperative fusion approach within the GM-PHD filter framework. By performing the feature-level fusion between the proposed DCM model and spatio-temporal information, the ambiguities in the identity association were effectively mitigated, especially when targets are in close proximity. For the decision-level fusion, a new fusion strategy was proposed with real and virtual zones to alleviate the issue of missed detections in the original GCI fusion rule, thereby maximizing the complementary benefits using both detectors. As a whole, such a multi-level fusion approach can simultaneously acquire more reliable tracks and recover missed targets under various challenging scenarios.

In Section 5.7.3, performance analysis with an ablation study were made to investigate the contributions from different proposed components. For the feature level fusion, the fused feature model improves the MOTA of spatial-temporal model by 7.8% and the appearance model by 1.6%. For the decision level fusion, The proposed fusion method achieves noticeably improved performance on almost all the presented metrics with particular improvements in MOTA ranging from 4.5% to 10.1%, as compared to baseline methods. Evaluations on the MOTChallenge in Section 5.7.4 were shown to confirm the effectiveness of the proposed approach as well as the improved performance particularly in FN and ML, which means that the issue of missed detections have been effectively alleviated. The proposed fusion method achieves best tracking performance among all RFS based methods in both benchmarks. Typically, comparing with the PHD_GSDL17 [10] which has been presented in Chapter 3, the proposed method reduces a large amount of FNs, and improves MOTA by 1.6%, MT by 1.8% and ML by 2.5%.

On the other hand, selected examples of tracking failures were given and analyzed in 5.7.7. As a suggestion for future work, FP in the proposed fusion approach can be further improved by exploiting the long-term latency of target trajectories. To further improve occlusion issues, techniques using better deep features in [143] [144] as well as occlusion-aware detection [145] will be explored in the future work.

Chapter 6

CONCLUSIONS AND FUTURE WORK

Measurement-driven filtering methods provide several solutions to address the existing issues in the task of online multiple human tracking, particularly to deal with the challenges of unknown number of targets, noisy detections and missed detections. PHD filtering performed together with tracking-by-detection is the fundamental framework which carries through all the studies of multiple human tracking in this thesis. Overall, this thesis has effectively fulfilled the six objectives mentioned in Chapter 1, by investigating three different approaches which are allocated in three contributions to address the challenging problems in multiple human tracking, such as varying number of targets, background clutter and missed detections. The effectiveness of these proposed works was validated by extensive experimental evaluations on real dataset, where in-depth discussions with other state-of-the-art methods on the benchmark leaderboard were made in both quantitative and qualitative ways.

In this final chapter, the contributions of this thesis as well as the limitations are summarized and discussed in Section 6.1. Then the future research directions to further improve multiple human tracking are indicated in Section 6.2.

6.1 Conclusions

In Chapter 3, a new adaptive gating strategy was initially developed in the SMC-PHD filter. The main functionality of this design was to strengthen the measurement-driven mechanism, and then specifically refine the raw measurements by adaptively updating the gating region with the aid of target size and spatial information. The main advantage is that it allows the measurements either from survival targets or new-born targets to be better classified, so that the subsequent PHD updating step can be more efficiently performed. Then in the birth intensity estimation, visual target features including the HOG and color histogram were utilized through a group-structured dictionary learning approach to recover the valid birth measurements. The motivation behind this was to exploit the discriminative power of group structured sparsity to enhance the ability of discriminating the targets from the noisy environment, where the group structure can reinforce the within-class consistency of same targets under different illumination and pose conditions. To deal with the target appearance changes, the SimCO algorithm was efficiently implemented with the designed structure pattern in the dictionary update stage. The OSPA evaluations were made to investigate the effectiveness of individual proposed contributions in this chapter. Compared with different state-of-the-art RFS-based methods, the proposed method achieved the lowest OSPA scores. Evaluations on the MOT15 and MOT17 benchmarks further validated the superiority of the proposed method compared with other state-of-the-art methods, where the proposed method improves the MOTA by 8.2% and 12.0% as compared to [29] and [60] on the MOT15 dataset. The proposed method also achieves the second best online tracker ranked on the leaderboard of MOT17 Challenge, as well as performs well in terms of tracking precision (high MOTP), fewer targets lost (low ML) and more targets tracked (high MT).

Chapter 4 was particularly presented to expose the paramount importance of detection selection to the entire tracking pipeline by performing the analysis in both statistics and visualization, and thus proposed a novel enhanced detection reliability approach. Firstly, by taking advantage of the classification power of mask R-CNN, the original confidence score for each detection was well regained in a global-to-local manner. The benefit of the proposed confidence rescoring strategy is that the correspondence between true target candidates and confidence scores can be better established. Following by a newly developed pruning algorithm Soft-ANMS which was used to further suppress the highly scored duplicate detections, both designs were performed sequentially to supply more reliable measurements for the tracking process. It is noteworthy that the proposed approach can be worked as an independent module for detection dependent tracking systems. Besides, this chapter also utilized the CNN features for target appearance modeling, instead of using hand-crafted features in Chapter 3. Evaluations on the MOT16 benchmark verified the efficacy of the proposed EDR algorithm in the measurement selection, particularly reduces 87.1% of FPs and retains 95.2% of TPs, in comparison with the baseline method. Moreover, the improved tracking performance in FP and MOTA demonstrated the proposed EDR module is able to reduce false detections and promote the tracking accuracy. The proposed method achieves the best MOTA scores in both evaluated sequences, and in particular outperforms the baseline method by 9.0% and 18.2% respectively.

In Chapter 5, the entire tracking process was generally established from a multi-level data fusion perspective. Following the concept of using CNN feature based appearance models above, the first contribution was to further study the CNN features with multiple scales, since they were able to better preserve both spatial and semantic representation of the target appearances compared with features only extracted from the last layer. These features

were then used to learn target-specific classifiers using correlation filters to form the proposed discriminative correlation matching scheme which distinguishes the tracked target between the background and other human targets. This was further combined with spatio-temporal information to accomplish the feature-level fusion with the aim of mitigating the ambiguities in the target association. The second contribution targeted at addressing the issue of missed detection, which was not comprehensively studied in the previous technical contributions. To this end, two human detectors were worked together to gain more complementary information. Then, the proposed robust fusion center was constructed with virtual and real zones to specifically deal with the missed detections at the decision level. By combining the previously mentioned feature-level fusion, such a multi-level fusion approach performed well to remedy the deficiency of the usage of single detectors. The effectiveness of proposed method at different fusion stages was verified on the evaluation datasets. For the benchmark evaluations, the proposed fusion method overall achieves best tracking performance among all RFS based methods in both benchmarks. comparing with the PHD_GSDL17 [10] which has been presented in Chapter 3, the proposed method reduces a large amount of FNs, and improves MOTA by 1.6%, MT by 1.8% and ML by 2.5%. In addition, the performance of FN and ML demonstrate the proposed fusion method is advantageous compared with other state-of-the-art trackers, so the issue of missed detections has been alleviated. However the proposed fusion method unavoidably infused the futile or unreliable measurements into the tracking system, which results in more false positives.

6.2 Future work

Although the proposed methods have achieved promising performance on the research field of multiple human tracking, there are still several areas that

can be further developed to improve this study. This section also aims to provide some potential future research directions for any researchers planning to enter this field.

In this thesis, person re-identification as one of the computer vision tasks has been successfully exploited and transferred to promote the tracking algorithm by quantifying the visual similarities between targets. However, this technique can further push forward the target matching step by incorporating human skeletons and mask related features, since human gaze and silhouettes can be used as additional cues to enhance the data association. It is also possible that other kinds of vision tasks such as scene understanding and gait recognition, can be available for better interpretation of human tracking, since investigating the different scene contexts other than pedestrians can be useful to analyze the human movements.

Handling frequent and long-term occlusions is still a challenging issue within the tracking family, and its quality directly affects almost all the tracking performance measures. Therefore, there are plenty of future works towards addressing this issue. Firstly, the occluded target during the occlusion can be considered a special type of missed detection. Analyzing the association between trajectories in a temporal window could be beneficial to estimate or interpolate the missing positions. Possible solutions in this trend can rely on the deep learning methods such as RNNs and LSTMs which are good at exhibiting temporal dynamic behaviors thereby better investigating the missing nodes of target trajectories. Another issue can be benefited by incorporating a more discriminative CNN features association step, such as in [143] and [144], thereby better reconstruct target trajectories after the occlusions. In addition, occlusion-aware detection [145] as a redetection technique can be also a good option to particularly deal with long-term occlusions.

It should be noted that performing an accurate prediction step in the

Bayesian tracking approaches can also achieve desirable tracking performance, even without the help of visual understanding. This requires an effective human motion forecasting mechanism which can well predict target next movement based on the previous trajectories. Understanding human movements in a data-driven pattern could be a feasible solution, which favors learning from the past to predict the next occurring event. In this way, more realistic human trajectories can be generated, and commonly used linear or non-linear assumptions on the human motion can be avoided.

Interestingly, some of existing challenging issues in multiple human tracking such as noisy and missed detections, are intrinsically due to the poor quality in the detection stages. In fact, most of contributed works in this thesis encompassed the solutions to aforementioned problems which are actually not originated from tracking part. For this purpose, further advancements on the human detector could be a direct approach to benefit the tracing performance, and it can also avoid the unnecessary processing in the tracking part. Therefore, it is expected that a simple but effective real-time tracking system for the next generation of video surveillance would be achieved by combining a nearly perfect detector with a simplified filtering process.

References

[1] Z. Fu, P. Feng, S. M. Naqvi, and J. A. Chambers, "Robust Particle PHD Filter with Sparse Representation for Multi-Target Tracking," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2016, pp. 281–285.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

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

[4] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv:1603.00831 [cs.CV]*, pp. 1–13, 2016.

[5] P. Feng, "Enhanced particle PHD filtering for multiple human tracking," Ph.D. dissertation, Newcastle University, UK, 2016.

[6] Z. Fu, S. M. Naqvi, and J. A. Chambers, "Collaborative detector fusion of data-driven PHD filter for online multiple human tracking," in *Proc. of the International Conference on Information Fusion (FUSION)*, 2018, pp. 1976–1981.

[7] C. S. Regazzoni, A. Cavallaro, Y. Wu, J. Konrad, and A. Hampapur, "Video analytics for surveillance: Theory and practice [from the guest editors]," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 16–17, 2010.

[8] 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 Transactions on Signal Processing*, vol. 64, no. 5, pp. 1320–1335, 2016.

[9] 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 Transactions on Multimedia*, vol. 19, no. 4, pp. 725–739, 2017.

[10] Z. Fu, P. Feng, F. Angelini, J. Chambers, and S. M. Naqvi, "Particle PHD Filter based Multiple Human Tracking using Online Group-Structured Dictionary Learning," *IEEE Access*, vol. 6, pp. 14 764–14 778, 2018.

[11] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] Ata-UR-Rehman, S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-target tracking by using particle filtering and a social force model," in *Proc. of the International Conference on Information Fusion (FUSION)*, 2014, pp. 1–6.

[14] F. Angelini, Z. Fu, S. A. Velastin, J. A. Chambers, and S. M. Naqvi, "3d-hog embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes," in *Proc. of the IEEE International Con-*

ference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4219–4223.

[15] E. Maggio and A. Cavallaro, *Video Tracking-Theory and Practice*. John Wiley and Sons, 2011.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[17] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1201–1208.

[18] J. Son, M. Baek, M. Cho, and B. Han, "Multi-Object Tracking with Quadruplet Convolutional Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5620–5629.

[19] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, and B. Schiele, "A multi-cut Formulation for Joint Segmentation and Tracking of Multiple Objects," *arXiv:1607.06317 [cs.CV]*, pp. 1–18, 2016.

[20] W. Choi, "Near-Online Multi-target Tracking with Aggregated Local Flow Descriptor," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3029–3037.

[21] 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 *European Conference on Computer Vision Workshops (ECCVW)*, 2016, pp. 68–83.

[22] A. Milan, K. Schindler, and S. Roth, "Multi-target tracking by discrete-continuous energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2054–2068, 2016.

[23] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter-Particle Filters for Tracking Applications.* Artech House, 2004.

[24] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASMEJournal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[25] M. S. Arulampalam, S. Maskell, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Tranctions on Signal Processing*, vol. 50, no. 2, pp. 174–187, 2002.

[26] R. P. S. Mahler, "Multitarget Bayes Filtering via First-Order Multitarget Moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, 2003.

[27] V. Eiselein, D. Arp, M. Ptzold, and T. Sikora, "Real-Time Multi-human Tracking Using a Probability Hypothesis Density Filter and Multiple Detectors," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 325–330.

[28] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4376–4380.

[29] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," in *European Conference on Computer Vision Workshops (ECCVW)*, 2016, pp. 84–99.

[30] Z. Fu, F. Angelini, S. M. Naqvi, and J. A. Chambers, "GM-PHD Filter Based Online Multiple Human Tracking Using Deep Discriminative Correlation Matching," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4299–4303.

[31] Z. Fu, S. M. Naqvi, and J. A. Chambers, "Enhanced GM-PHD Filter Using CNN-Based Weight Penalization for Multi-Target Tracking," in *Sensor Signal Processing for Defence Conference (SSPD)*, 2017, pp. 1–5.

[32] Z. Fu, X. Lai, and S. M. Naqvi, "Enhanced detection reliability for human tracking based video analytics," in *submitted to International Conference on Information Fusion (FUSION)*, 2019, pp. 1–7.

[33] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[34] W. Luo, J. Xing, A. Milan, V. Monga, and T. Tran, "Multiple object tracking: A literature review," *arXiv:1409.7618v4 [cs.CV]*, pp. 1–18, 2017.

[35] L. Fan, Z. Wang, B. Cail, C. Tao, Z. Zhang, Y. Wang, S. Li, F. Huang, S. Fu, and F. Zhang, "A survey on multiple object tracking algorithm," in *2016 IEEE International Conference on Information and Automation (ICIA)*, 2016, pp. 1855–1862.

[36] R. Henschel, L. Leal-Taixe, D. Cremers, and B. Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018, pp. 1428–1437.

[37] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 300–311.

[38] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person Tracking by Multicut and Deep Matching," in *European Conference on Computer Vision Workshops (ECCVW)*, 2016.

[39] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[40] P. Feng, W. Wang, S. M. Naqvi, and J. Chambers, "Adaptive Retrodiction Particle PHD Filter for Multiple Human Tracking," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1592–1596, 2016.

[41] L. Lan, X. Wang, S. Zhang, D. Tao, W. Gao, and T. S. Huang, "Interacting Tracklets for Multi-Object Tracking," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4585–4597, 2018.

[42] H. Zhou, W. Ouyang, J. Cheng, X. Wang, and H. Li, "Deep Continuous Conditional Random Fields with Asymmetric Inter-object Constraints for Online Multi-object Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2018.

[43] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House, 1999.

[44] B.-N. Vo, S. Singh, and D. Arnaud, "Sequential Monte Carlo methods for multitarget filtering with random finite sets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 4, pp. 1224–1245, 2005.

[45] A. Dore, M. Soto, and C. S. Regazzoni, "Bayesian tracking for video analytics," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 46–55, 2010.

[46] R. Hoseinnezhad, B. N. Vo, and B. T. Vo, "Visual Tracking in Background Subtracted Image Sequences via Multi-Bernoulli Filtering," *IEEE Transactions on Signal Processing*, vol. 61, no. 2, pp. 392–397, 2013.

[47] B. Vo, B. Vo, and D. Phung, "Labeled random finite sets and the

bayes multi-target tracking filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6554–6567, 2014.

[48] D. Y. Kim, B.-N. Vo, B.-T. Vo, and M. Jeon, "A labeled random finite set online multi-object tracker for video data," *Pattern Recognition*, vol. 90, pp. 377 – 389, 2019.

[49] B. Ristic, D. Clark, B. N. Vo, and B. T. Vo, "Adaptive Target Birth Intensity for PHD and CPHD Filters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 2, pp. 1656–1668, 2012.

[50] E. Maggio, M. Taj, and A. Cavallaro, "Efficient Multitarget Visual Tracking Using Random Finite sets," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1016–1027, 2008.

[51] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, "Enhancing Detection Model for Multiple Hypothesis Tracking," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 2143–2152.

[52] Y. D. Wang, J. K. Wu, A. A. Kassim, and W. Huang, "Data-Driven Probability Hypothesis Density Filter for Visual Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1085–1095, 2008.

[53] Y. Zheng, Z. Shi, R. Lu, S. Hong, and X. Shen, "An Efficient Data-Driven Particle PHD Filter for Multitarget Tracking," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2318–2326, 2013.

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

[55] J. Wu, K. Li, Q. Zhang, W. An, Y. Jiang, X. Ping, and P. Chen, "Iter-

ative RANSAC based adaptive birth intensity estimation in GM-PHD filter for multi-target tracking," *Signal Processing*, vol. 131, pp. 412–421, 2017.

[56] X. Zhou, Y. Li, B. He, and T. Bai, "GM-PHD-Based Multi-Target Visual Tracking Using Entropy Distribution and Game Theory," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1064–1076, 2014.

[57] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.

[58] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online Multi-object Tracking Using CNN-Based Single Object Tracker with Spatial-Temporal Attention Mechanism," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4846–4855.

[59] S. H. Bae and K. J. Yoon, "Confidence-Based Data Association and Discriminative Deep Appearance Learning for Robust Online Multi-Object Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 595–610, 2018.

[60] Y. Song and M. Jeon, "Online multiple object tracking with the hierarchically adopted GM-PHD filter using motion and appearance," in *IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 2016, pp. 1–4.

[61] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *arXiv:1703.07402 [cs.CV]*, 2017, pp. 1–5.

[62] S. H. Park, K. Lee, and K. J. Yoon, "Robust online multiple object tracking based on the confidence-based relative motion network and correlation

filter," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3484–3488.

[63] J. H. Yoon, M. H. Yang, J. Lim, and K. J. Yoon, "Bayesian Multi-object Tracking Using Motion Context from Multiple Objects," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015, pp. 33–40.

[64] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *31st AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017, pp. 4225–4232.

[65] A. J. Ma, P. C. Yuen, and S. Saria, "Deformable Distributed Multiple Detector Fusion for Multi-Person Tracking," *arXiv:1512.05990 [cs.CV]*, pp. 1–9, 2015.

[66] 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 *IEEE International Workshop on Traffic and Street Surveillance for Safety and Security (AVSS)*, 2017, pp. 1–5.

[67] O. Khalid, J. C. SanMiguel, and A. Cavallaro, "Multi-Tracker Partition Fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 7, pp. 1527–1539, 2017.

[68] R. P. S. Mahler, "Optimal/robust distributed data fusion: a unified approach," *Proc. SPIE Signal Processing, Sensor Fusion, and Target Recognition IX*, vol. 4052, pp. 128–138, 2000.

[69] D. Clark, S. Julier, R. P. S. Mahler, and B. Ristic, "Robust multi-object sensor fusion with unknown correlations," in *Sensor Signal Processing for Defence Conference (SSPD)*, 2010, pp. 1–5.

[70] M. Uney, D. E. Clark, and S. J. Julier, "Distributed Fusion of PHD Filters Via Exponential Mixture Densities," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 3, pp. 521–531, 2013.

[71] G. Battistelli, L. Chisci, C. Fantacci, A. Farina, and A. Graziano, "Consensus CPHD Filter for Distributed Multitarget Tracking," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 3, pp. 508–520, 2013.

[72] B. Wang, W. Yi, S. Li, M. R. Morelande, L. Kong, and X. Yang, "Distributed Multi-target Tracking via Generalized Multi-Bernoulli Random Finite Sets," in *International Conference on Information Fusion (FUSION)*, 2015, pp. 253–261.

[73] W. Yi, M. Jiang, S. Li, and B. Wang, "Distributed sensor fusion for RFS density with consideration of limited sensing ability," in *International Conference on Information Fusion (Fusion)*, 2017, pp. 1–6.

[74] M. Khazaei and M. Jamzad, "Multiple Human Tracking using PHD Filter in Distributed Camera Network," in *International Conference on Computer and Knowledge Engineering (ICCKE)*, 2014, pp. 569–574.

[75] Q. Liu, W. Wang, T. de Campos, P. J. B. Jackson, and A. Hilton, "Multiple Speaker Tracking in Spatial Audio via PHD Filtering and Depth-Audio Fusion," *IEEE Transactions on Multimedia*, pp. 1–1, 2017.

[76] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2015.

[77] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust Multi-Speaker Tracking via Dictionary Learning and Identity Modelling," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 864–880, 2014.

[78] X. Zhou, H. Yu, H. Liu, and Y. Li, "Tracking Multiple Video Targets with an Improved GM-PHD Tracker," *Sensors*, vol. 15, no. 12, pp. 30 240–30 260, 2015.

[79] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, p. 42824286, 1995.

[80] K. Nummiaro, E. Koller-Meier, and L. V. Gool, "Object tracking with an adaptive color-based particle filter," in *Proceedings of the 24th DAGM Symposium on Pattern Recognition*, 2002, pp. 353–360.

[81] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[82] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (ANIPS)*, 2012, pp. 1097–1105.

[83] P. Feng, W. Wang, S. M. Naqvi, and J. A. Chambers, "A robust PHD filter with deep learning updating for multiple human tracking," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 1227–1231.

[84] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Convolutional Features for Correlation Filter Based Visual Tracking," in *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015, pp. 621–629.

[85] L. Leal-Taix, C. Canton-Ferrer, and K. Schindler, "Learning by Tracking: Siamese CNN for Robust Target Association," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 418–425.

[86] X. Xu and B. Li, "Head tracking using particle filter with intensity gradient and color histogram," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, pp. 888–891.

[87] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[88] X. Mei and H. Ling, "Robust Visual Tracking and Vehicle Classification via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2259–2272, 2011.

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

[90] W. Z. Lu, C. Bai, K. Kpalma, and J. Ronsin, "Multi-Object Tracking using Sparse Representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 2312–2316.

[91] Y. Suo, M. Dao, U. Srinivas, V. Monga, and T. Tran, "Structured dictionary learning for classification," *arXiv:1406.1943 [cs.CV]*, pp. 1–14, 2014.

[92] Y. Xu, Y. Sun, Y. Quan, and B. Zheng, "Discriminative structured dictionary learning with hierarchical group sparsity," *Computer Vision and Image Understanding*, vol. 136, pp. 59–68, 2015.

[93] T. Zhang, S. Liu, C. Xu, S. Yan, B. Ghanem, N. Ahuja, and M. H. Yang, "Structural Sparse Tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 150–158.

[94] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust Visual Tracking via Structured Multi-Task Sparse learning," *International Journal of Computer Vision*, vol. 101, no. 2, pp. 367–383, 2013.

[95] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybrid boosted multi-target tracker for crowded scene," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, p. 29532960.

[96] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Hierarchical Convolutional Features for Visual Tracking," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3074–3082.

[97] Y. Yang and G. A. Bilodeau, "Multiple object tracking with kernelized correlation filters in urban mixed traffic," in *14th Conference on Computer and Robot Vision (CRV)*, 2017, pp. 209–216.

[98] H. Wu and W. Li, "Robust online multi-object tracking based on kcf trackers and reassignment," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2016, pp. 124–128.

[99] E. Bochinski, V. Eiselein, and T. Sikora, "High-Speed Tracking-by-Detection Without Using Image Information," in *IEEE International Workshop on Traffic and Street Surveillance for Safety and Security (AVSS)*, 2017, pp. 1–6.

[100] J. Shen, Z. Liang, J. Liu, H. Sun, L. Shao, and D. Tao, "Multiobject tracking by submodular optimization," *IEEE Transactions on Cybernetics*, pp. 1–12, 2018.

[101] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear LSTM," in *European Conference on Computer Vision (ECCV)*, 2018.

[102] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," *Radar and Signal Processing, IEE Proceedings F*, vol. 140, no. 2, pp. 107–113, 1993.

[103] B.-N. Vo and W. K. Ma, "The Gaussian Mixture Probability Hypothesis Density Filter," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4091–4104, 2006.

[104] N. L. Baisa and A. Wallace, "Development of a N-type GM-PHD Filter for Multiple Target, Multiple Type Visual Tracking," *arXiv:1706.00672 [cs.CV]*, pp. 1–17, 2017.

[105] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice Hall PTR, 1993, vol. I.

[106] I. Goodman and R. P. S. Mahler, *Mathematics of Data Fusion.* Kluwer Academic Publishers, 1997.

[107] R. P. S. Mahler, "A theoretical foundation for the Stein-Winter probability hypothesis density (PHD) multitarget tracking approach," in *Proc. MSS Nat'l Symp. on Sensor and Data Fusion*, vol. 1, 2000.

[108] 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

[109] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[110] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[111] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[112] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR 2011*, 2011, pp. 3457–3464.

[113] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2129–2137.

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

[115] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[116] N. Chenouard, I. Bloch, and J. C. Olivo-Marin, "Multiple hypothesis tracking for cluttered biological image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2736–3750, 2013.

[117] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-HiLasso: A Collaborative Hierarchical Sparse Modeling framework," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4183–4198, 2011.

[118] W. Dai, T. Xu, and W. Wang, "Simultaneous Codeword Optimization (SimCO) for Dictionary Update and Learning," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6340–6353, 2012.

[119] R. Fisher. (2003) Caviar test case scenarios. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[120] Y. Xiang, A. Alahi, and S. Savarese, "Learning to Track: Online Multi-

object Tracking by Decision Making," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4705–4713.

[121] J. H. Yoon, C. R. Lee, M. H. Yang, and K. J. Yoon, "Online Multi-object Tracking via Structural Constraint Event Aggregation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1392–1400.

[122] H. Kieritz, S. Becker, W. Hbner, and M. Arens, "Online multi-person tracking using integral channel features," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2016, pp. 122–130.

[123] M. Yang, Y. Wu, and Y. Jia, "A Hybrid Data Association Framework for Robust Online Multi-Object Tracking," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5667–5679, 2017.

[124] M. Yang and Y. Jia, "Temporal dynamic appearance modeling for online multi-person tracking," *Computer Vision and Image Understanding*, vol. 153, pp. 16 – 28, 2016.

[125] Y. Song, Y. Yoon, K. Yoon, and M. Jeon, "Online and real-time tracking with the GM-PHD filter using group management and relative motion analysis," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.

[126] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[127] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS – improving object detection with one line of code," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5561–5569.

[128] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting rcnn: On awakening the classification power of faster rcnn," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 453–468.

[129] K. Panta, D. E. Clark, and B. N. Vo, "Data Association and Track Management for the Gaussian Mixture Probability Hypothesis Density Filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 45, no. 3, pp. 1003–1016, 2009.

[130] Y. Wang, H. Meng, Y. Liu, and X. Wang, "Collaborative Penalized Gaussian Mixture PHD Tracker for Close Target Tracking," *Signal Processing*, vol. 102, pp. 1 – 15, 2014.

[131] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 748–756.

[132] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 868–884.

[133] H. W. Kuhn, *The Hungarian method for the assignment problem.* Naval Research Logistics Quarterly, 1955.

[134] M. Vasic and A. Martinoli, "A Collaborative Sensor Fusion Algorithm for Multi-object Tracking Using a Gaussian Mixture Probability Hypothesis Density Filter," in *IEEE International Conference on Intelligent Transportation Systems*, 2015, pp. 491–498.

[135] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, 2014, pp. 740–755.

[136] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68 – 80, 2017.

[137] T. A. Biresaw, A. Cavallaro, and C. S. Regazzoni, "Tracker-Level Fusion for Robust Bayesian Visual Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 776–789, 2015.

[138] K. Granstrom, M. Baum, and S. Reuter, "Extended object tracking: Introduction, overview and applications," *Journal of Advances in Information Fusion*, vol. 12, no. 2, pp. 139 – 174, 2016.

[139] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 34–50.

[140] M. Vasic, D. Mansolino, and A. Martinoli, "A system implementation and evaluation of a cooperative fusion and tracking algorithm based on a gaussian mixture phd filter," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4172–4179.

[141] Q. Zhou, B. Zhong, Y. Zhang, J. Li, and Y. Fu, "Deep alignment network based multi-person tracking with occlusion and motion reasoning," *IEEE Transactions on Multimedia*, pp. 1–13, 2018.

[142] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple Hypothesis Tracking Revisited," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4696–4704.

[143] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2018.

[144] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

[145] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, "Occlusion-aware real-time object tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 763–771, 2017.