Audio-Visual Tracking of Multiple Moving Speakers

Volkan KILIÇ

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

January 2016

© Volkan KILIÇ 2016

Abstract

In this thesis, a novel approach is proposed for multi-speaker tracking by integrating audio and visual data in a particle filtering (PF) framework. This approach is further improved for adaptive estimation of two critical parameters of the PF, namely, the number of particles and noise variance, based on tracking error and the area occupied by the particles in the image. Here, it is assumed that the number of speakers is known and constant during the tracking. To relax this assumption, the random finite set (RFS) theory is used due to its ability in dealing with the problem of tracking a variable number of speakers. However, the computational complexity increases exponentially with the number of speakers, so probability hypothesis density (PHD) filter, which is first order approximation of the RFS, is applied with sequential Monte Carlo (SMC), namely particle filter, implementation since the computational complexity increases linearly with the number of speakers. The SMC-PHD filter in visual tracking uses three types of particles (i.e. surviving, spawned and born particles) to model the state of the speakers and to estimate the number of speakers. We propose to use audio data in the distribution of these particles to improve the visual SMC-PHD filter in terms of estimation accuracy and computational efficiency. The tracking accuracy of the proposed algorithm is further improved by using a modified mean-shift algorithm, and the extra computational complexity introduced by mean-shift is controlled with a sparse sampling technique. For quantitative evaluation, both audio and video sequences are required together with the calibration information of the cameras and microphone arrays (circular arrays). To this end, the AV16.3 dataset is used to demonstrate the performance of the proposed methods in a variety of scenarios such as occlusion and rapid movements of the speakers.

Summary

The problem of detection and tracking of multiple moving speakers in indoor environments using audio-visual (AV) modalities has attracted an increasing amount of attention in the last decade due to its potential applications in e.g. automatic camera steering in video conferencing, individual speaker discrimination in multi-speaker environments, and surveillance and monitoring in security applications. Several challenges are associated with AV tracking including fusion of multiple modalities, estimation of the variable number of speakers and their states, and dealing with various conditions such as occlusion, limited view of cameras, illumination change and room reverberation. This thesis aims to address part of these challenges under the Bayesian framework. This leads to three main contributions summarised as follows. First, a novel approach is proposed for combining audio and video modalities under the particle filter (PF) framework. Audio information such as the direction of arrival (DOA) angles of the audio sources is incorporated into the PF based visual tracking to reshape the typical Gaussian noise distribution of particles in the propagation step and to weight the observation model in the measurement step. The proposed algorithm is further improved to provide adaptive estimation of two critical parameters of the PF: the number of particles and noise variance. In regular implementation, these parameters are determined in the initialization step by rule of thumb and are kept fixed, which makes the tracker inconsistent in practice. With our approach, which is based on tracking error and the area occupied by the particles in the image, the number of particles and noise variance are estimated adaptively during the tracking process.

Next, a more realistic and complex scenario is considered where the number of speakers varies with time. The random finite set (RFS) theory is employed here, due to its ability to deal with a variable number of targets. A particle filter algorithm under the RFS framework is devised for AV tracking. In the RFS approach, the computational cost becomes expensive when the number of speakers increases. To address this problem, the probability hypothesis density (PHD) filter, which is the first order approximation of the RFS, is used together with its sequential Monte Carlo (SMC) implementation. Unlike the single type of particles in generic particle filtering, in the SMC-PHD filter, three different types of particles, namely surviving, spawned and born particles are used, to model the state of the speakers and to jointly estimate the number of speakers with their states. In our proposed AV-SMC-PHD algorithm, audio data is used to determine when to propagate and re-allocate these particles based on their types. Finally, the AV-SMC-PHD algorithm is further improved in its estimation accuracy and computational efficiency. The mean-shift method is integrated in our AV tracking system to shift the particles to a local maximum of the distribution function which moves the estimated position closer to the ground truth position of the speaker. With the integration of the mean-shift method, the tracking error in the proposed AVMS-SMC-PHD algorithm is reduced. However, the computational cost is increased due to the application of the mean-shift method to all the particles. To address this issue, a sparse sampling technique is proposed which samples a small subset, named sparse particles, from the source particles using one dimensional bins based on the KLD-Sampling algorithm. The mean-shift method is applied only to the sparse particles rather than all the particles, which reduces computational cost significantly.

Key words: Audio-visual speaker tracking, particle filter, adaptive particle filter, random finite set, PHD filter, SMC implementation, multi-speaker tracking.

Email: v.kilic@surrey.ac.uk

Özet

Kapalı mekânlarda çoklu hareketli konuşmacıların işitsel-görsel (İG) modaliteleri kullanarak tespiti ve takibi problemi son on yılda, video konferanslarında otomatik kamera yönlendirilmesi, çoklu konuşmacı ortamlarında konuşmacının bireysel olarak ayırt edilmesi, güvenlik uygulamalarında gözetleme ve izleme gibi potansiyel uygulamalarından kaynaklı olarak giderek artan bir ilgi topladı. Çeşitli sorunlar İG takiple bağlantılı olup bunlar, çoklu modalitelerin birleştirilmesi, değişken sayıdaki konuşmacıların sayısının ve durumlarının kestirimi, örtüşme, sınırlı kamera görüntüsü, ışık değişimi ve oda yankılanmaları gibi farklı koşulların üstesinden gelinmesidir. Bu tez, sorunların bir kısmını Bayes teorisi çerçevesinde ele almayı hedeflemiştir. Bu da aşağıda özetlenen üç ana katkı bölümünün oluşmasını sağlamıştır.

İlk olarak, işitsel ve görsel modalitelerin parçacık süzgeci çerçevesi altında birleştirilmesi için özgün bir yaklaşım önerilmiştir. Ses kaynaklarının varış yönü açısı gibi işitsel bilgileri parçacık süzgeci tabanlı görsel takipte, parçacıkların tipik Gaussian gürültü dağılımının yeniden şekillendirilmesinde ve ölçüm adımında gözlem modelinin katsayılarının güncellenmesinde kullanılmıştır. Önerilen algoritma daha da geliştirilerek parçacık süzgecinin iki kritik parametresinin, parçacık sayısı ve gürültü varyansı, uyarlanabilir kestirimi sağlanmıştır. Alışılmış uygulamalarda, bu parametreler başlangıç adımında yaklaşık olarak belirlenir ve sabit tutulur ki bu izleyicinin pratikte istikrarsız olmasına neden olmaktadır. Takip hatası ve parçacıkların resim üzerinde kapladığı alana dayalı olan uyarlanabilir yaklaşımımızla, parçacık sayısı ve gürültü varyansı takip işlemi boyunca uyarlanabilir olarak hesaplanmaktadır.

Sonrasında, daha gerçekçi ve karmaşık bir senaryo olan konuşmacı sayısının zamanla değişmesi incelenmiştir. Rassal sonlu set teorisi değişken sayıdaki hedeflerle baş etme yeteneği dolayısıyla burada uygulanmıştır. İG takip için rassal sonlu set çerçevesi altında parçacık süzgeci tasarlanmıştır. Rassal sonlu set yaklaşımında, konuşmacı sayısı artınca hesaplama maliyeti pahalı olmaktadır. Bu sorunu çözmek için, rassal sonlu setin birinci dereceden yakınsaması olan olasılıksal hipotez yoğunluğu (OHY) süzgeci sıralı Monte Carlo (SMC) uygulamasıyla birlikte kullanılmıştır. Genel parçacık süzgecindeki tek tip parçacık kullanımından farklı olarak, SMC-OHY süzgecinde üç farklı; yani hayatta kalan, çoğalmış ve doğma, parçacıkları konuşmacıların durumlarını modellemek ve konuşmacı sayısını durumlarıyla birlikte hesaplamak için kullanır. Önerdiğimiz algoritmada, İG-SMC-OHY, ses verisi parçacıkların çeşidine göre ne zaman yayılacağında ve parçacıkların yeniden konumlandırılmasında kullanılmıştır.

Son olarak, İG-SMC-OHY süzgecinin kestirim doğruluğu ve hesaplama verimliliği daha da geliştirilmiştir. Ortalama kayma yöntemi, parçacıkları dağılım fonksiyonunun yerel maksimumuna kaydırmak için İG takip sistemimizle bütünleştirilmesiyle kestirilen konum konuşmacının gerçek konumuna yaklaşmıştır. Ortalama kayma yönteminin bütünleştirilmesiyle önerdiğimiz İGOK-SMC-OHY algoritmasının kestirim hatası düşürülmüştür. Ancak, ortalama kaymanın tüm parçacıklara uygulanmasıyla hesaplama maliyeti de artmıştır. Bu soruna çözüm bulmak için, KLD örnekleme algoritmasına dayalı tek boyutlu gözenek kullanarak esas parçacıklardan seyrek parçacık adında daha küçük bir küme örnekleyen seyrek örnekleme tekniği önerilmiştir. Böylelikle, ortalama kayma

yöntemi tüm parçacıklar yerine sadece seçilen seyrek parçacıklara uygulanarak hesaplama maliyeti önemli ölçüde düşürülmüştür.

Anahtar Kelimeler: İşitsel-görsel konuşmacı takibi, parçacık süzgeci, uyarlamalı parçacık süzgeci, rassal sonlu set, OHY süzgeci, SMC uygulaması, çoklu konuşmacı takibi.

Email: v.kilic@surrey.ac.uk

Acknowledgements

I could probably have written a full chapter which is of comparable size to the contribution parts of this thesis to acknowledge those, who, in different ways, have helped me in my PhD life at the University of Surrey.

First and foremost, Dr. Wenwu Wang, my supervisor, whose extremely experienced and very kind way to provide guidance has been helpful in transforming my research work into enjoyable endeavour during the period of my PhD study. His ongoing encouragement has motivated me strongly. I want to thank him very much also for helping me in improving my writing skills. I feel very fortunate and grateful to have Dr. Wang as a mentor and a friend.

I would like to thank Prof. Josef Kittler greatly for his beneficial guidance and constructive comments, from which I gained great insights into my research. Should you meet him, you will see that a great intellectual can also be modest and humble. I also would like to show my great thanks to Dr. Mark Barnard for his valuable ideas and encouragement throughout this study.

I sincerely thank Prof. Müştak Erhan Yalçın for his encouragement to opt for an academic career. His endless motivation helped me to open a new page in my life.

I am grateful to the CVSSP staff, Elizabeth James, Anna Korzeniowska, and Katherine Skinner, for helping me solve my administrative problems. I would like to also thank our FEPS faculty that created this unique environment at the University of Surrey for me and other students to study.

I have had many fun times with my friends at Surrey, particularly office friends, Jing Dong, Zhenhua Feng, Patrik Huber, Guosheng Hu, Elissavet Chotzoglou, Qingju Liu, and have benefited from the time I have spent with them.

I am most grateful to my close friends, lovely couple, Armin Mustafa and Gowhar Syeed Shah who are always sister and brother to me. Their continuous support and great interest let me feel that I am part of their family. I am certain it will continue to be so.

I also thank all my friends from Turkey, Gökhan Öztürk, Seda Durucasu, Behlül Odacılar, Serhat Oral, Ayça Çölgeçen, Fırat Çölgeçen, Meltem Hür Pembe, Okan Pembe and Vildan İçlier who have supported and encouraged me during the years of my PhD study.

Finally, I would like to express my deepest appreciation to my family: Muharrem Kılıç, my father, whose vision and wisdom inspired me in life; Semiye Kılıç, my mother, whose great strength enabled me to try harder at my problems; Hakan Kılıç, my brother, whose trust encouraged me on all the way. I also wish to thank Cengiz Yıldırım, İbrahim Kılıç, Zennure Kılıç, Özlem Kılıç, Özge Kılıç, Buket Kılıç, Irmak Kılıç for their continuous support.

List of Figures

2.1 Steps of the particle filter 21
2.2 The RFS theory in multi-target tracking. 22
2.3 An example for the PHD filter. 23
2.4 Intuitive descriptions of the mean-shift 24
2.5 Schematic diagram of particle filter based tracking. 25
2.6 Physical setup of the $AV16.3$ corpus. 40
2.7 Some frames from the $AV16.3$ dataset. 42

3.1 The SSM method. 50
3.2 Failure in Visual PF after occlusion for single speaker. 53
3.3 Failure in Visual PF after occlusion for multi-speaker. 53
3.4 Schematic diagram of the proposed AV particle filter. 54
3.5 Distribution of particles with DOA line. 55
3.6 The steps of the AV-PF algorithm. 57
3.7 Single speaker occlusion in AV-PF. 58
3.8 Mapping tables. 60
3.9 Occupied area by the particles. 61
3.10 The cumulative beta distribution function for different α and β values. . 63
3.11 Frames of Sequence 11 camera #3 from comparison of the V-PF and AV-PF. 67
3.12 Error plots for Sequence 11 camera #3 68
3.13 Error plots for Sequence 11 camera #1 and Sequence 24 camera #1 . . 69
3.14 Frames of Sequence 24 camera #1 from comparison of the V-PF and AV-PF. 70

3.15 Frames of Sequence 40 camera #1 from comparison of the V-PF and AV-PF. 70

3.16 Frames of Sequence 45 camera #2 from comparison of the V-PF and AV-PF. 72

3.17 Error plots for Sequence 40 camera #1 and Sequence 45 camera #2. . . 72

3.18 Average error comparison between the V-PF and AV-PF algorithms for a variable number of particles N. 74

3.19 The comparison between the AV-A-PF and AV-PF. 76

3.20 The comparison between the AV-A-PF algorithm and KLD-sampling in single speaker tracking. 77

3.21 The comparison between the AV-A-PF algorithm and KLD-sampling in multi-speaker tracking. 78

4.1 Distribution of 50 particles for the visual and audio-visual cases. 85

4.2 Tracking results of the Visual RFS-PF and the proposed AV RFS-PF in Sequence 30 camera #2. 99

4.3 Comparison of the Visual RFS-PF and the proposed AV RFS-PF in estimation of the number of active speakers for Sequence 30 camera #2. 100

4.4 Position estimates of the Visual RFS-PF and proposed AV RFS-PF trackers for Sequence 30 camera #2. 101

4.5 Tracking results of the Visual RFS-PF and proposed AV RFS-PF in Sequence 25 camera #3. 101

4.6 Comparison of the Visual RFS-PF and proposed AV RFS-PF in estimation of the number of active speakers for Sequence 25 camera #3. . . . 102

4.7 Position estimates of the Visual RFS-PF and proposed AV RFS-PF trackers for Sequence 25 camera #3. 102

4.8 Performance comparison of the Visual RFS-PF and proposed AV RFS-PF in terms of the OSPA error. 102

4.9 Computational cost measurement of the V-SMC-PHD and the proposed AV-SMC-PHD filter in Sequence 24 camera #1 and Sequence 45 camera #3. 104

4.10 Performance comparison of the V-SMC-PHD and proposed AV-SMC-PHD for Sequence 24 camera #1. 105

4.11 Comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in estimation of the number of speakers for Sequence 24 camera #1. . . 106

4.12 Position estimates of the V-SMC-PHD and the proposed AV-SMC-PHD filters for Sequence 24 camera #1. 106

4.13 Performance comparison of the V-SMC-PHD and proposed AV-SMC-PHD for Sequence 45 camera #3. 107
4.14 Comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in number of speakers estimation for Sequence 45 camera #3. 107
4.15 Performance comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in terms of the OSPA-T error. 108
4.16 Variation of scale value and OSPA-T error for the proposed AV-SMC-PHD. 109
4.17 The contribution of OSPA components to the final error. 110

5.1 Tracking scheme of the improved tracking system via the integration of the mean-shift, particle and PHD filters. 115
5.2 Basic mean-shift process is illustrated. A single hypothesis is shifted from base to the peak of the distribution function. 117
5.3 Saddle point in multi density distribution. 118
5.4 The mean-shift process for the particles. 120
5.5 The sparse sampling. 124
5.6 Error sensitivity with respect to ζ in the mean-shift process. 127
5.7 Error change with respect to the iteration number in the mean-shift process. 128
5.8 Performance comparisons of the V-SMC-PHD, AV-SMC-PHD and AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3. . . 129
5.9 Mean absolute OSPA-T errors of the V-SMC-PHD, AV-SMC-PHD and AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3. 130
5.10 Average errors of the V-SMC-PHD, AV-SMC-PHD and AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3 131
5.11 The sensitivity of error to parameter τ. 134
5.12 Computational cost comparison between the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD filters. 135
5.13 Performance comparisons of the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3. 136
5.14 Comparison of the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD using mean absolute OSPA-T error. 136
5.15 Performance comparison of the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD with bar plots. 137

A.1 Audio detection and localization results. 147

List of Tables

3.1 Curve fitting coefficients of the mapping table for the proposed AV-A-PF. 62

3.2 Description of sequences used for experiments 66

3.3 Experimental results for V-PF and proposed AV-PF. 71

3.4 Experimental results with TBM, MOT and MAE metrics for V-PF and proposed AV-PF. 74

3.5 Experimental results for KLD-sampling and proposed AV-A-PF. 80

3.6 Experimental results with TBM, MOT and MAE metrics for KLD-sampling and proposed AV-A-PF. 80

4.1 Experimental results for Visual RFS-PF and proposed AV RFS-PF . . . 103

4.2 Experimental results for V-SMC-PHD and the proposed AV-SMC-PHD 110

4.3 Experimental results for V-SMC-PHD, the proposed AV-SMC-PHD, Visual RFS-PF and the proposed AV RFS-PF. 111

5.1 Experimental results for V-SMC-PHD, AV-SMC-PHD and AVMS-SMC-PHD 131

5.2 Experimental results for V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse AVMS-SMC-PHD. 138

5.3 Comparison of tracking results in terms of Wasserstain distance. 139

B.1 Full descriptions of experimental results with TBM and MOT metrics for V-PF and proposed AV-PF. 149

B.2 Full descriptions of experimental results with TBM and MOT metrics for KLD-sampling and proposed AV-A-PF. 150

List of Algorithms

2.1 V-PF tracking algorithm. 27

3.1 Proposed AV-PF algorithm. 57

3.2 Proposed AV-A-PF algorithm. 65

4.1 RFS-PF algorithm for multi-speaker tracking. 92

4.2 Proposed AV-SMC-PHD filtering algorithm. 98

5.1 $\mathbb{MS}$ function for the mean-shift iteration. 119

5.2 Proposed AVMS-SMC-PHD filtering algorithm. 121

5.3 $\mathbb{SS}$ function for sparse sampling. 125

5.4 Proposed sparse-AVMS-SMC-PHD filtering algorithm. 126

Contents

List of Figures **x**

List of Tables **xiii**

List of Algorithms **xv**

1 Introduction **1**

1.1 Speaker Tracking Problem and Motivation 1

1.2 Contributions 5

1.3 Outline of the Thesis 7

2 Literature Survey and Background Knowledge **9**

2.1 Tracking Modalities 9

2.1.1 Visual Cues 9

2.1.2 Audio Cues 11

2.2 Audio-Visual Speaker Tracking 13

2.3 Tracking Algorithms 15

2.3.1 Particle Filtering 25

2.3.2 KLD Sampling 27

2.3.3 Random Finite Set 29

2.3.4 PHD Filtering with SMC Implementation 32

2.3.5 Mean Shift Tracking 34

2.4 Relevant Datasets 39

2.5 Performance Metrics 41

2.6 Chapter Summary 45

3 Audio Assisted Robust Visual Tracking with Adaptive Particle Filtering 47

3.1 Introduction 48

3.2 Audio Detection and Localization 49

3.3 Particle Filter Based Audio Constrained Visual Tracking Algorithm . . 52

3.4 Improved AV Tracking with Adaptive Particle Filter 59

3.5 Experimental Evaluations 64

3.5.1 Visual PF vs. Audio-Visual PF 67

3.5.2 Audio-Visual PF vs. Audio-Visual A-PF 75

3.6 Chapter Summary 79

4 Tracking a Variable Number of Speakers 83

4.1 Introduction 83

4.2 RFS Multiple Speaker Tracking 86

4.2.1 Visual Likelihood 86

4.2.2 Audio Likelihood 88

4.2.3 Particle Filter Implementation of RFS 89

4.3 Particle PHD Filtering 94

4.3.1 Colour Likelihood Model 94

4.3.2 Proposed AV-SMC-PHD Filtering 95

4.4 Evaluation Results and Analysis 97

4.4.1 Visual RFS-PF vs. AV RFS-PF 99

4.4.2 Proposed AV-SMC-PHD 103

4.5 Chapter Summary 111

5 Sparse Sampling for Mean-Shift Based SMC-PHD Filtering 113

5.1 Introduction 114

5.2 Improved AV Tracking with Mean Shift Based SMC-PHD Filtering . . . 116

5.3 Computational Cost Refinement with Sparse Particles 122

5.4 Experimental Results and Analysis 125

5.4.1 AVMS-SMC-PHD 125

5.4.2 Sparse-AVMS-SMC-PHD 132

5.5 Chapter Summary 139

6 Conclusions and Future Research **141**

6.1 Conclusions 141

6.1.1 Audio Assisted Robust Visual Tracking with Adaptive Particle Filtering 142

6.1.2 Tracking a Variable Number of Speakers 142

6.1.3 Sparse Sampling for Mean-Shift Based SMC-PHD Filtering . . . 143

6.2 Future Research 144

A Audio Detection and Localization Results **147**

B Detail Analysis of Performance Metrics **149**

C List of Acronyms, Nomenclature and Glossary **151**

D List of Publications **159**

References **161**

Chapter 1

Introduction

This introductory chapter first presents tracking problems and motivation of this thesis in Section 1.1. Then, Section 1.2 specifies our contributions of the work. Finally, outline of the thesis is given in Section 1.3.

1.1 Speaker Tracking Problem and Motivation

Speaker tracking in enclosed spaces has received much interest in the fields of computer vision and signal processing, driven by applications such as automatic camera steering in video conferencing [78], individual speaker discriminating in multi-speaker environments [127], acoustic beamforming [143], audio-visual speech recognition [103], video indexing and retrieval [95], human-computer interaction [118], and surveillance and monitoring [51] in security applications. However, speaker tracking is a challenging task in real life scenarios as several distinctive issues influence the tracking process such as, estimation of the variable number of speakers and their states, and dealing with various conditions such as occlusion, limited view of cameras, illumination change and room reverberations.

One approach to overcome these challenges is to use multi-modal information, as it provides additional observations about the state of the speaker compared to single-modal tracking. Speaker tracking is a relevant problem to the field of multi-modal

target tracking. The multi-modal information used for tracking can be collected by sensors such as audio, video, thermal vision, laser range finders and RFID [3], [48]. In speaker tracking, audio and video sensors are widely applied compared to others, for their easier installation, cheaper cost, and more data processing tools. Here, we also apply audio and video data in our tracking system.

Early efforts in speaker tracking often use either visual only or audio only data despite the fact that both audio and visual information are readily available in many real world scenarios. The method of video-only tracking [109], [130] is generally reliable and accurate when the targets are in the camera field of view, but limitations are introduced when the targets are occluded by other speakers, when they disappear from the camera field of view, or the appearance of the targets or illumination is changed [56], [66]. Audio tracking [37], [104], [108] is not restricted by these limitations, however, audio data is intermittent over time and may be corrupted by background noise and room reverberations, which may introduce non-negligible tracking errors. In addition, spatial resolution (tracking resolution in the world space) of audio is in general worse than that of video. Using both audio and visual data has the potential to improve the tracking performance in the case that either modality is unavailable or both are corrupted.

Earlier techniques were designed to track one person in a static and controlled environment. However, theoretical and algorithmic advances together with the increasing capability in computer processing have led to the emergence of more sophisticated techniques for tracking in dynamic and less controlled (or natural) environments with multiple speakers [6], [43] and [66]. The type of sensors used to collect the measurements is also evolving from single- to multi-modal.

To use multi-modal information generated by these sensors, several approaches have been proposed which can be categorized into two methods as one is deterministic and data-driven while the other is stochastic and model-driven [125], [152]. Deterministic approaches are often considered as an optimization problem by minimizing an appropriate cost function. The definition of the cost function is an important issue. A common method in the literature is mean-shift [26], [29] method where the cost function is de-

fined in terms of colour similarity measured by Bhattacharyya distance. The stochastic and model-driven approaches use a state-space approach based on the Bayesian framework which is suitable for processing multi-modal information [45]. The particle filter (PF) is one of the popular stochastic approaches, which became widely used in tracking after being proposed by Isard and Blake [55].

To track speaker with multi-modalities, data association methods such as joint probability data association (JPDA) or multiple hypothesis tracking (MHT), are needed to run with PF [115] which results in an increased computational cost. Here, a novel approach is proposed in Chapter 3 for combining audio and video modalities under the PF framework. The key idea is to integrate the audio and visual modalities in the steps of the PF framework, rather than performing any *a priori* data fusion algorithm.

The PF also has some limitations in its computational complexity. The samples (particles) used in PF are weighted in order to approximate the filtering distributions. The quality of the sample based representation rises with the number of particles, N. A key question is: How many particles should be used for a specific estimation problem. In most cases, the choice is made experimentally and users tend to choose N as large as possible to get accurate results, leading to an increased computational cost.

Adaptive particle filtering (A-PF) approaches have therefore been proposed in [24], [42] and [122] to address these problems and to find the optimal N for the PF to use. An early and popular approach, i.e. KLD-Sampling, was proposed by Fox in [42]. This approach aims to bound the error introduced by the sample-based representations of the PF using the Kullback-Leibler divergence between Maximum Likelihood estimates (MLE) of states and the underlying distribution to optimize the number of particles. Specifically, this method is applied to a mobile robot localization problem, where the initial set of particles is generally very large. However, it is not clear how to apply this approach to more general particle filters that provide posterior-based estimates rather than MLE. Moreover, it is assumed that the true posterior is given by a discrete constant piecewise distribution such as a multi-dimensional histogram bins, but the characteristics of robot localization (e.g., binning of the state space) might not be valid in other situations.

Another parameter in a PF that is often fixed is the noise variance (σ^2) which has a critical role in the distribution of particles. This role makes the determination of σ^2 crucial so it should not be chosen randomly or empirically since incorrectly chosen σ^2 may lead to the use of a greater N than actually needed. Therefore, we aim to design an A-PF approach that involves dynamic estimation of σ^2 in order to find the optimal N. In this thesis, we propose an AV-A-PF algorithm in Chapter 3, which is based on our previous AV-PF algorithm. We show the efficiency of our algorithm in single and multi-speaker tracking in comparison with the KLD-Sampling algorithm.

The generic PF applied to multi-speaker AV tracking is often under the assumption that the number of speakers is known and invariant. In a practical tracking environment, however, the speakers to be captured by the audio-visual sensors may appear or disappear in a random manner. As a result, the number of speakers that can be observed from the audio-visual measurements may vary with time. We relax the above assumption and focus on the problem of tracking a variable number of speakers based on the AV data, where the variable number of speakers and their states are jointly estimated in a multi-speaker environment. A Bayesian tracking framework based on the random finite set (RFS) formulation [85], [87], [136] is used to deal with the unknown and variable number of speakers. Our work is based on the PF implementation of the RFS presented in [80] and [151]. Different from [80] and [151], the RFS approach is extended in Chapter 4 to deal with both audio and visual measurements. We show in our experiments that, with this new extension, the proposed AV tracking system is able to track reliably a variable number of speakers in challenging scenarios including occlusion.

One issue with the RFS is that its computational complexity grows exponentially as the number of speakers increases since the complexity order of the RFS is $O(M^{\Xi})$ where M is the number of measurements and Ξ is the number of speakers. To overcome this problem, the RFS is replaced with the PHD filtering [85] approach which is the first order approximation of the RFS with the complexity order $O(M\Xi)$. This framework has recently emerged and it has been found to be promising for multi-speaker tracking [87]. The sequential Monte Carlo (SMC) implementation [135] is introduced to obtain practical solutions of the PHD filter.

Our studies are further extended to improve estimation accuracy and computational efficiency of the tracking system. In that sense, we introduce the mean-shift method to our audio-visual tracking system in order to combine the merits of both stochastic and deterministic tracking approaches in a unified framework using the SMC-PHD filter. The mean-shift iteration is run on the particle set to pull the centre of the particle distribution towards the target centre. This leads to improvement in estimation accuracy as observed in our experiments shown in Chapter 5. However, applying mean-shift method to all the particles increases the computational cost considerably. Therefore, a sparse-sampling technique is proposed which generates a small set of particles, called sparse particles, from original particles set, based on KLD-Sampling algorithm. It is integrated in the proposed algorithm to choose the sparse particles for the mean-shift iteration.

All the methods proposed throughout this thesis are extensively evaluated under different scenarios of multi-speaker tracking using the $AV16.3$ dataset and the results show the advantage of the proposed methods over the existing and baseline methods.

1.2 Contributions

The major contributions of this thesis are summarised as follows:

(1) We propose a new algorithm for joint AV tracking based on the PF. In this algorithm, AV-PF, the direction of arrival angles (DOAs) of the sources, estimated from microphone recordings are used to reshape the distribution of the particles in the propagation step and to weight the observation model in the measurement step of the visual tracker. We show in our experiments that incorporating audio information in this way, not only addresses the occlusion problem, a challenging scenario in visual tracking, but also significantly reduces the number of particles required in visual tracking to robustly model the distribution of the particles and the estimation of the state vector.

This approach is further improved by solving a typical problem associated with the PF. It has been observed that the efficiency and accuracy of the PF usually depend on the number of particles and noise variance. Both of these parameters are normally

specified beforehand and are kept fixed in the regular implementation of the PF. To address these problems, the proposed AV-PF algorithm is extended to adapt both the number of particles and noise variance based on tracking error and the area occupied by the particles in the image.

(2) A more realistic and complex scenario is considered where the number of speakers is unknown and changing with time. The random finite sets framework is taken to represent the multi-target states and measurements and to propagate multi-target posterior for multi-target filtering. We extend the RFS theory for multi-modal data and devise a particle filtering algorithm under the RFS framework for AV tracking. However, the computational complexity increases exponentially as the number of speakers increases in the RFS. This motivates the use of approximations such as the first-order moment method, which leads to the PHD filter, providing a computationally cheaper alternative, but theoretically sound, solution. The particle filter implementation of the PHD filter, the SMC-PHD filter, is developed to handle a variable number of speakers. The SMC-PHD filter employs surviving, spawned and born particles to model the state of the speakers and jointly estimates the variable number of speakers with their states. The born particles play a critical role in the detection of new speakers, which makes it necessary to propagate them in each frame. However, this increases the computational cost of the visual tracker. Here, we propose an algorithm, named AV-SMC-PHD, to use audio data to determine when to propagate the born particles and to re-allocate the surviving and spawned particles.

(3) Our investigations are extended to optimise the proposed AV-SMC-PHD filtering algorithm aiming at improving its estimation accuracy and computational efficiency. To improve the estimation accuracy, the mean-shift method is incorporated into the SMC-PHD filter to shift the particles toward a close local maximum, leading to the AVMS-SMC-PHD filter algorithm which reduces tracking error significantly. However, applying mean-shift method to all the particles increases the computational cost. In that sense, we propose a “sparse particle” concept. The sparse particles are selected from the set of all the particles using one dimensional bins based on the KLD-Sampling algorithm in order to apply the mean-shift method only to the selected particles rather than all the particles which reduces the computation cost significantly. The performance

of all the proposed approaches is extensively studied under different scenarios of multi-speaker tracking. It compares favourably with the existing tracking techniques.

1.3 Outline of the Thesis

The thesis is organised as follows: The background knowledge and a literature review of the related work are introduced in Chapter 2. Audio-visual modalities, datasets and performance metrics used throughout the thesis are described in detail. Chapter 3 presents our proposed audio-visual particle filter (AV-PF) algorithm and our proposed audio-visual adaptive particle filter (AV-A-PF) algorithm. In Chapter 4, tracking of a variable number of speakers is discussed. Both the RFS theory and the PHD filter are examined in multi-speaker tracking. The proposed AV-PF algorithm is combined with the PHD filter to propose the AV-SMC-PHD filtering algorithm. In Chapter 5, mean-shift and sparse sampling are integrated in the proposed AV-SMC-PHD filtering algorithm for further improvement. Chapter 6 concludes the thesis with recommendations for future work. Lists of acronyms and mathematical symbols are also appended, to improve readability of the thesis.

Chapter 2

Literature Survey and Background Knowledge

This chapter firstly introduces the modalities that can be used in speaker tracking in Section 2.1. Section 2.2 and Section 2.3 present a review of the various methods relevant to AV speaker tracking problems in the literature. In addition, the baseline methods, namely particle filtering, KLD-Sampling, random finite set, PHD filtering and mean-shift, used throughout this thesis, are discussed in detail. Section 2.4 reviews the datasets adopted for the evaluation of tracking algorithms. A number of performance metrics proposed in the literature to measure the performance of multi-target tracking methods are discussed in Section 2.5. Finally, closing remarks are given in Section 2.6.

2.1 Tracking Modalities

2.1.1 Visual Cues

Visual tracking of a target in an image sequence is a difficult task in real life scenarios, as the illumination conditions may vary and the target may be occluded by background objects or may perform fast and complicated movements [154]. To overcome these problems, several visual features, i.e., colour, texture, contour and motion [146] are used in existing tracking systems.

A very intuitive approach to target tracking is to use colour features. Colours are commonly used in tracking as they provide an additional source of information to distinguish targets and object of interests. Different approaches have been proposed to track the target with colour. A colour mixture model is employed based on a Gaussian distribution to perform tracking and segmentation in [106]. A tracking algorithm using an adaptive mixture model is then developed in [90]. Using colour information, the target can be easily detected and tracked as long as the representative colours of the target is sufficiently distinct from others or the background. In that sense, the colour models of the target need to be learned either beforehand, in a separate initialization step [66], or for each new initialized track. In addition, the appearance of colours may be influenced by ambient or local changes in lighting [7].

Another intuitive approach in tracking is to use contour-based methods which track the target contour by employing shape matching or contour-evolution techniques [148]. In representing the contours, active models can be used such as snakes, B-splines, and geodesic active contours, or meshes [141]. One general problem in target tracking is occlusion of the target by objects or other targets. To handle occlusion, only the contour of the upper body can be detected [91] and tracked [8] instead of tracking the contour of the whole bodies. This helps to detect a new person since the front and back view of the upper bodies are distinctive for different people.

One commonly used cue is texture which is a measure of the intensity variation of a surface which quantifies properties such as smoothness and regularity [38], [119], [142]. The texture feature is studied with Gabor wavelet in [88]. The Gabor filters can be used as orientation and scale tunable edge and line detectors, and the statistics of these micro-features in a given region are mostly used to characterize the underlying texture information [147]. Recently, local patterns of image have received increasing interest for better detection and recognition. Local patterns that are binarized with an adaptive threshold, provide promising results on various topics, such as face detection and image classification. The Local Binary Patterns (LBP) method was proposed in [97]. It is an efficient texture descriptor defined as a grey-scale invariant texture measure. An advantage of the LBP is its tolerance to illumination changes.

Another cue used in tracking, especially in indoor environments, is motion which is a very strong cue of human presence. This cue can be extracted by using foreground detection. A simple method of foreground detection is to compute the difference frame by frame. Although it has been used in multi-modal tracking systems [99], it fails when the person is stationary since it does not provide any foreground information.

The Scale Invariant Feature Transform (SIFT) [79] is also widely used as a feature in tracking applications. It transforms image data into scale-invariant coordinates relative to local features. The SIFT extracts distinctive invariant features from images and affords reliable matching between different views of an object. The SIFT feature is invariant to translation, rotation, scaling, clutter, lighting and occlusion. In addition, it is robust for matching across a substantial range of affine distortions, change in $3D$ viewpoint, addition of noise, and change in illumination. The features are mostly distinctive which allow a single feature to be correctly matched with high probability against a large database of features from an image database that includes many images. However, the SIFT matching rate and recognition performance can be dramatically decreased in noisy environments with non-rigid targets [58].

Among the visual cues explained above, colour cues have been used throughout this thesis due to their easy implementation and low complexity. Colour information is used by taking the histogram of possible targets at the initialization step as reference images and then using them in detection and tracking of the target. In the literature, the RGB or HSV colour histogram model is commonly used [120]. In our study, HSV is chosen since it is observed to be more robust to illumination variation.

2.1.2 Audio Cues

Despite the fact that a variety of audio information could be used such as sound source localization (SSL) and time delay estimation (TDE), as a proof of concept, the DOA angle is used here which is more feasible in audio processing from the circular microphone array in an indoor environment, employed for collecting the dataset used in our experiments.

The audio source localization methods can be classified into three categories [14],

namely, steered beamforming, superresolution spectral estimation, and time-delay estimation. Beamformer-based source localization provides comparatively low resolution and requires a search over a highly nonlinear surface [104]. In addition, it is computationally expensive which may not have good performance for real-time applications. Super-resolution spectral estimation methods are not well suited for locating a moving speaker as it is assumed that the speaker location is fixed for a number of frames [34]. In fact, the location of a moving speaker can change dramatically from frame to frame. In addition, these methods are not robust to modelling errors caused by room reverberation, and are usually computationally expensive [13], [104]. The time delay of arrival (TDOA) based location estimators estimate the relative time-delay between the wave-front arrivals at microphone positions in order to determine the location of the speaker. It has an advantage, as compared with the other two methods, because of its direct connection with the location of the speaker and its computational efficiency.

The DOA estimation is found similarly to the TDOA estimation problem as the DOA is estimated by first determining the TDOA between the different microphones in the array. The estimation of source locations highly depends on the quality of the DOA measurements and there exist many DOA estimation techniques such as the coherent signal subspace (CSS) [137] and the MUSIC algorithm [113] differing mainly on how to deal with background noise, reverberation and possible movement of the sources [104]. Estimation quality of the DOA is mostly influenced by the following three factors [105]. The first one is the spectral content of the speech segment used to derive the DOA. An utterance is the combination of a succession of high-energy and low-energy segments interspersed with silence parts. Both the low-energy speech and silence parts are sensitive to background noise and are prone to erroneous DOA measurements. Next is the reverberation level of the room which causes the spurious measurements due to reflections from the walls and objects within the enclosure. The third one is the array position with respect to the speakers, and also the number of simultaneous sources present in the receptive field and their relative positioning.

2.2 Audio-Visual Speaker Tracking

Speaker tracking is a key part of multimedia applications which helps us to determine the speaker trajectories and analyse the behaviour of speakers. Speaker tracking can be achieved with the use of audio-only, visual-only or audio-visual information.

Recently, some approaches for speaker tracking based on audio-only information have been proposed [37], [80], [96], [101], [104], [136]. An audio-based fusion scheme was presented in [104] to detect multiple speakers. The location estimates from multiple microphone arrays are fused to determine the same speaker state. Separate KFs are used for all the individual microphone arrays for the location estimation and the probabilistic data association technique is used with an interacting model to deal with motion of the speaker and measurement uncertainty. One problem in [104] is that it cannot handle the tracking problem for a time-varying number of speakers. To track an unknown and time varying number of speakers, Ma *et al.* [80], [136] proposed an approach based on random finite set. The RFS theory and SMC implementation are employed to develop the Bayesian RFS filter which tracks the time-varying number of speakers and their states simultaneously. The random finite set theory is capable of dealing with a time varying number of speakers, but the maximum number of speakers that can be handled is limited as its computational complexity grows exponentially with the number of speakers. Pham *et al.* [101] proposed a cardinalized PHD filter, which is the first-order approximation of the RFS, to alleviate the computational cost caused by the number of speakers. TDOA measurements from microphone pairs are fused by using asynchronous sensor fusion with the CPHD filter to obtain the positions of the speakers. Fallon *et al.* [37] proposed a methodology to track an unknown and time-varying number of speakers based on the concept of an existence grid which detects active regions before tracking. Nguyen *et al.* [96] used a time-frequency method and the PHD filter to localize and track simultaneous speakers. The multiple speakers are located based on the time-frequency method which uses an array of three microphones, then the PHD filter is applied to the localization results as post-processing to deal with miss-detection and clutter.

Speaker tracking with multi-modal information has also received much attention and

many approaches have been developed in the past decade using audio-visual information [6], [8], [20], [45], [46], [52], [54], [118], [127], [128], [133], given the complementary characteristics of each modality. Generally speaking, the differences among these existing works arise from the overall objective like tracking single versus multiple speakers and the specific detection/tracking framework. None of the above works can handle the problems addressed here at the same time under one framework: multiple speakers, simultaneous speakers, unknown and time varying number of speakers.

Beal *et al.* [6] have used graphical models to fuse audio-visual measurements to track a moving speaker in a cluttered, noisy environment. Audio and video observations are modelled jointly by computing their mutual dependencies and the model parameters are learnt using the expectation-maximization algorithm from a sequence of audio-visual data. Talantzis *et al.* [127], [128] proposed to use a hierarchical Kalman filter structure to track people in a three-dimensional space using multiple microphones and cameras. Two separate local Kalman filters are applied for audio and video streams, and then the outputs of these two local filters are fused under one global Kalman filter. Unlike Talantzis *et al.* [127,] Vermaak *et al.* [133] used particle filters to estimate the predictions from audio- and video-based measurements and the fusion of audio-visual information is performed at the feature level. In other words, the independent particle coordinates from the features of both modalities are fused to track the speaker. These works [6], [127], [128], [133] have concentrated on the single speaker case which cannot directly handle the tracking problem for multiple speakers.

Bernardin *et al.* [8] presented two multi-modal systems for the tracking of multiple persons. To detect speech and determine active speaker positions, a joint probabilistic data association filter is used. For visual features, two systems are used where the first is a particle filter using foreground, colour, upper body detection and person region cues from multiple camera images and the other is a blob tracker using only a wide angle overhead view. Then, acoustic and visual tracks are fused using a finite state machine. However, unlike Bernardin *et al.* [8], Perez *et al.* [45], [46] proposed a particle filtering framework which incorporates the audio and visual detections into the particle filtering framework using an observation model. Their probabilistic framework allows one to track multiple people jointly with their speaking activity which is based on a

mixed-state dynamic graphical model defined on a multi-person state-space. Heuer *et al.* [52] also presented a particle filter based multi-modal fusion approach that a single speaker can be identified in the presence of multiple visual observations. To fuse multiple observations and modalities, Gaussian mixtures model was adopted to their approach. Unlike [8], [45], [46] and [52], Shivappa *et al.* [118] did not use particle filtering framework and proposed hidden Markov model based iterating decoding scheme to fuse audio and visual cues for localization and tracking of persons. Although these works [8], [45], [46], [52] and [118] are capable for multiple speaker tracking, they cannot deal with unknown and time-varying number of speakers. Also, their models can detect only single speaking activity rather than simultaneous speaking activity from multiple people.

To the best of our knowledge, the closest works to ours are [20] and [54], both are based on the Bayesian framework and can handle varying number of speakers tracking. The particle filter is employed in [20], and observation likelihoods based on both audio and video measurements are derived to use in the estimation of weights of the particles, and then the number of people is calculated using weights of these particles. In [54], the RFS theory based on multi-Bernoulli approximations is used to integrate audio and visual cues with sequential Monte Carlo implementation. The nature of the random finite set formulation supports their framework to handle varying number of targets tracking. However, they cannot deal with simultaneous speaking activity from multiple speakers. Our work presented in Chapters 4 and 5 substantially differs from existing works in AV multi-speaker tracking with respect to the capabilities for dealing with multiple speakers, simultaneous speakers, unknown and time varying number of speakers.

2.3 Tracking Algorithms

This section presents a brief review of the literature in tracking algorithms relevant to this thesis. This includes visual and audio-visual tracking algorithms from computer vision and multimedia, Bayesian statistical methods and nonlinear filtering approaches from statistics and signal processing.

As we mentioned earlier in Chapter 1, the search paradigms in tracking are categorized

into two methods as one is stochastic and model-driven while the other is deterministic and data-driven [125]. The stochastic and model-driven approaches use a state-space approach based on the Bayesian framework, for example, the Kalman filter (KF) for linear motion and sensor models [4], extensions of KF for the nonlinear models using the first order Taylor expansion including the decentralized Kalman filter (DKF) [124], [128] and extended Kalman filter (EKF) [47], [89], and the PF for nonlinear and non-Gaussian models [3]. In comparison to the KF and EKF approaches, the PF approach is more robust for nonlinear models as it can approach the Bayesian optimal estimate with a sufficiently large number of particles [3]. It has been widely employed for speaker tracking problems [20], [45], [133]. For example, in [45] and [133], the PF is used to fuse object shapes and audio information. In [20], independent audio and video observation models are fused for simultaneous tracking and detection of multiple speakers. One challenge in using PF, however, is to choose an appropriate number of particles. An insufficient number may lead to particle impoverishment (a loss of diversity among the particles) while a larger number (than required) will introduce extra computational burden. Choosing the optimal number of particles is one of the issues that affect the performance of the tracker.

Another approach for tracking is based on finite-set statistics (FISST) theory called the PHD filter [86]. The PHD filter is a first-moment filter, which propagates the first order moment of a dynamic point process. Some applications of the PHD filter with speaker tracking are given in [96] and [136]. The main advantage of the PHD filter over Bayesian (Kalman or PF) approach is that it does not require any *a priori* knowledge of the number of targets, which is actually estimated during the tracking process. However, the PHD filter confines the propagation of the full multi-target posterior to the first order multi-target distribution moment which corresponds to a loss of higher order cardinality information that results in erratic estimates of the number of objects in low signal-to-noise ratio (SNR) scenarios [87]. Propagating the whole multi-target posterior is computationally intractable. The cardinalized PHD (CPHD) filter additionally propagates the cardinality distribution to PHD and leads to better performance over the PHD for the estimation of instantaneous target number [84], [87] and the position of the speakers [101]. The cardinality distribution, nevertheless,

makes the CPHD more computationally demanding than the PHD filter. In addition, the CPHD does not provide explicit models for the spawning of new targets by prior targets.

The mean-shift [29] is a deterministic and data-driven method, which focuses on target localization using representation of the target. Target localization is based on minimizing a cost function. In [29], a spatially smooth similarity function is defined and state estimation problem is reduced to a search of the region of interest of this function. A gradient optimization method is employed, which leads to fast localization. However, it is assumed that the target representation is sufficiently discriminated against the background, which is not always true in real life applications. In addition, despite its efficiency and robustness, the mean-shift method may easily fail in occlusion and the rapid motion cases [10], [72] where the target is out of the region of interest of the search area.

To address these problems in mean-shift tracking, many approaches have been proposed in the literature, which can be categorized into two groups. One group has focused on making improvements [10], [25], [73], [75] to mean-shift tracking itself, such as introducing adaptive estimation of the search area, iteration number and bin number. In the other group, the mean-shift algorithm is integrated with other methods such as particle filter [32], [116], [117], [139] . There are many attempts to use stochastic and deterministic approaches under the same framework. An earlier study is reported in [125] where particle filtering (stochastic and model-driven) and a variational approach (deterministic and data-driven) are fused. The so called "switching search methods" operation is applied to all the particles. Deterministic search (gradient descent) is initialised if the momentum of the particle is smaller than a pre-determined threshold. Otherwise, a stochastic motion model is applied to propagate the particles.

Shan *et al.* [117] combined particle filtering and mean-shift in a different way and proposed the mean-shift embedded particle filter (MSEPF). It was inspired by [125], but the mean-shift was used as a variational method. They aimed to integrate the advantages of the particle filtering and mean-shift method. The MSEPF is capable of tracking the target with a small number of particles since the mean-shift search

concentrates on the particles around local modes (maxima) of the observation. A skin colour model is used and updated every frame to deal with the possible changes due to variable illuminations. As an observation model, colour and motion cues are used. The mean-shift analysis is modified to use a multi-cue observation model and applied to all the particles. Then, resampling (selective resampling) is performed if the effective sample size is too small. Another approach is proposed in [32] where the mean-shift and particle filtering methods are employed independently. At every iteration, the estimated positions of the target obtained by these two methods are compared using Bhattacharyya distance and the best value is chosen as the estimated position of the target. This method is chosen to avoid the algorithm from being trapped by a local maximum, and thus finding the true maximum beyond the local one. Maggio *et al.* [81] proposed a hybrid particle and mean-shift tracker which is almost the same as Shan *et al.* [117]. Alternatively, Maggio *et al.* used the original application of the mean-shift and performed the mean-shift process on all the particles to reach the local maxima. To deal with manoeuvring targets, which have a high speed of movement, an adaptive motion model is used. Chang *et al.* proposed kernel particle filter [17, 18, 19]. After the mean-shift iteration, small perturbations are added to the states of the particles to prevent stopping the gradient ascent too early in the density. Kernel radius is calculated adaptively every iteration in the mean-shift process. The same kernel particle filter is applied to multiple target tracking in [19] using multiple hypotheses. These hypotheses are then evaluated and assigned to possible targets.

Wang *et al.* [139] proposed adaptive mean-shift tracking with auxiliary particles. An earlier study was reported in [138]. The mean-shift is used as the main tracker as long as the conditions are met, such as the target remaining in the region of interest, and there are no serious distractions. Auxiliary particles are introduced when sudden motions or distractions are detected by the motion estimator to support the mean-shift tracker. Background/foreground feature selection is applied to minimize the tracking error since the mean-shift may diverge from the target to converge on the background. This study is inspired by [125] where a particle filter is the main tracker. However, Wang used mean-shift as a main tracker and switched trackers to deal with occlusion, sudden movements and distractions. The target model is updated online to continue

tracking even if the appearance of the target is changed because of illumination or view point.

Although the mean-shift and particle filters have been combined in various frameworks in previous studies, none of the frameworks were explicitly designed for a variable number of targets since the structure of both methods was devised for single target tracking scenarios. The PHD filter is a promising method for multi-target tracking which propagates the posterior intensity of the multi-target posterior. However, the PHD filter recursion consists of equations with multi-dimensional integrals that do not have closed-form solutions in general. The particle filter or sequential Monte Carlo (SMC) implementation [135] was introduced to obtain practical solutions of the PHD filter. Therefore, the particle filter and PHD filter are combined under SMC-PHD filtering. Here, we also combine mean-shift with standard SMC-PHD filtering, aiming at improving its computational efficiency and estimation accuracy, in Section 5.2.

Apart from the tracking methods mentioned above, multi-modal usage in speaker tracking poses the problem of associating each measurement with an appropriate target, which is known as data association. Data association methods can be classified into two main categories [11]. The first one is unique-neighbour data association such as multiple hypothesis tracking (MHT) which associates each measurement to one of the existing tracks. The second one is all-neighbours data association, such as joint probabilistic data association (JPDA) which uses all the measurements for updating the entire track estimate. MHT filter has an advantage in maintaining multiple hypotheses of the association between a target state and the measurements in the measurement set. The drawback of MHT is that the number of hypotheses grows exponentially over time [98]. JPDA approximates the posterior target distribution in terms of separate Gaussian distributions for each target [5], [112] which results in an increased computational cost. Data association algorithms with Bayesian methods and the PHD filter in target tracking applications can be found in [16], [57], [64], [104] and [140]. However, some researchers found that classical data association algorithms are computationally expensive, and this led them to fuse multi-modal measurements inside their proposed framework [45], [47], [96], [128], [133]. Here, we also propose to fuse audio and visual information in our framework by using audio information in the steps of the visual

particle filter which is discussed in Chapter 3.

Among the methods explained above, the PF, RFS, PHD filter and mean-shift are the main baseline methods used throughout this thesis. Before introducing the mathematical background of these methods, the main concepts of the methods are presented below.

The PF became widely used in tracking after being proposed by Isard *et al.* [55] due to its ability to deal with nonlinear and non-Gaussian problems. The basic idea of the PF is to represent a posterior density by a set of random particles with associated weights and compute estimates based on these samples and weights [2]. Figure 2.1 is illustrated to give an idea of how the particle filter works. In the first step, ten particles are initialized with equal weights. The particles are weighted based on given measurements in the second step. Here, some particles get small weights while others get larger weights. This is represented by the size of the particles. The weighted particles collectively represent the state distribution. Then, a resampling step selects the particles with large weights to generate a set of new particles with equal weights in the third step. These new particles are distributed to predict the next state in step four. This cycle continues from steps two through four till all the observations are exhausted.

Although there are several extensions of the PF in the literature, the main concept is the same and based on the idea of representing the posterior distribution by a set of particles.

The generic PF is devised for single target tracking. Multi-target tracking is more complex than single target tracking as it is necessary to jointly estimate the number of targets and the state of the targets. One scenario for multi-target tracking is illustrated in Figure 2.2a. There are 5 targets in state space (blue plane) given at the previous time with 8 measurements in observation space (white plane). Here, the number of measurements is larger than the number of targets due to clutter or noise. After the targets are moved to the current time, the number of targets becomes 3 and two targets no longer exist. In multi-target tracking, the variable number of targets and noisy measurements need to be handled for reliable tracking. Therefore, the RFS approach [87] is

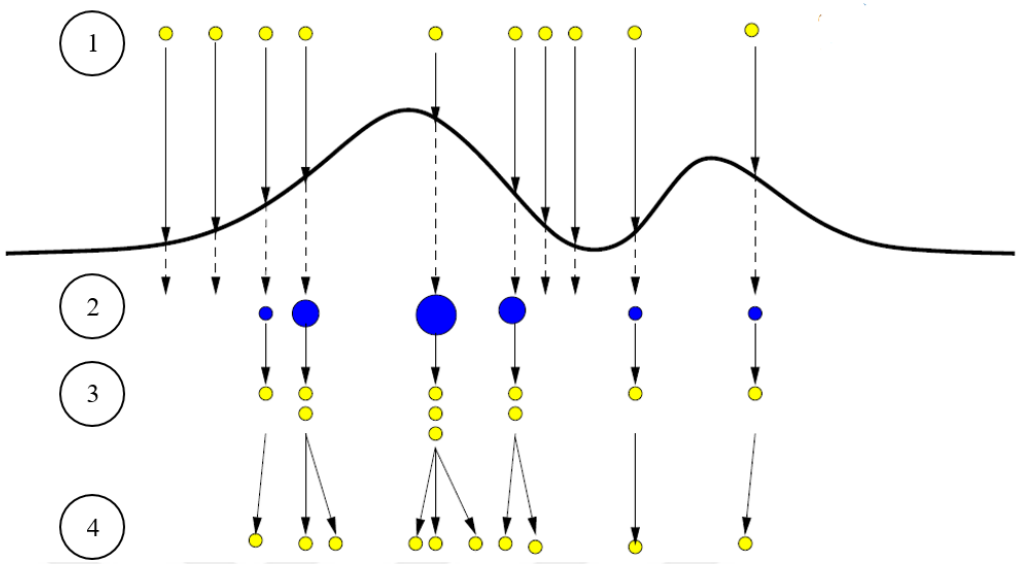
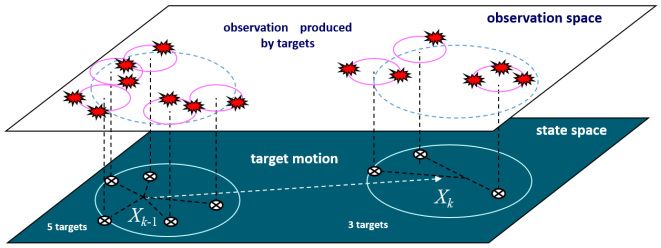


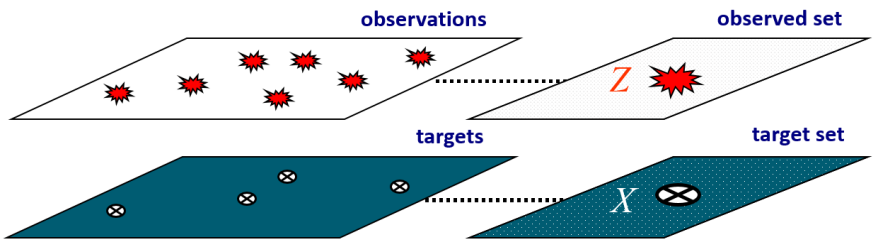
Figure 2.1: Steps of the particle filter. The particles are initialized in the first step with equal weights. Then, they get weights in the second step. A resampling step is performed in the third step. The particles are distributed to predict the next state in the fourth step. Figure is adapted from [121].

an elegant solution for multi-target tracking. The key idea behind the RFS approach is to treat the collection of targets as a set-valued state called the multi-target state and the collection of measurements as a set-valued observation, called multi-observation in order to handle the problem of estimating multiple targets in the presence of clutter and uncertainty by modelling these set-valued entities as random finite sets [134]. The point here is to generalize multi-target tracking from single target tracking. The RFS approach is illustrated in Figure 2.2b where all the targets are collected in one target set and all the measurements are treated as one measurement set. The RFS propagates the full multi-target posterior and the state model of the RFS incorporates individual target dynamics which are target birth, target spawn and target death. Similarly, the observation model of the RFS incorporates the measurement likelihood as target detection uncertainty (miss-detection) and clutter (false alarm). These incorporations are implemented by assigning hypotheses and all possible associations between hypotheses and measurement/targets need to be considered at every time step which gets computationally expensive in the case of a high number of targets and measurements.

The PHD filter is a computationally cheaper alternative to the RFS which is the first-



(a)



(b)

Figure 2.2: The RFS theory in multi-target tracking. One case for multi-target tracking scenario is given in (a), and (b) represents the RFS approach to multi-target tracking. Figures are adapted from [22].

order approximation of the RFS and propagates only the first order moments instead of the full multi-target posterior [135]. The PHD filter function is denoted as the intensity $v(x)$, and the integral of the intensity function on any region of the state space gives the expected number of targets. The peaks of the PHD function indicate the highest local concentration of the expected number of targets, which can be used to provide estimates of individual targets [87]. A simple example [87] of the PHD filter is given in Figure 2.3 which corresponds to Equation (2.1).

$$v(x) = \mathcal{N}_{\sigma^2}(x-a) + \mathcal{N}_{\sigma^2}(x-b) = \frac{1}{\sqrt{2\pi\sigma}}\left[exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) + exp\left(-\frac{(x-b)^2}{2\sigma^2}\right)\right] \tag{2.1}$$

Figure (2.3) is plotted for Equation 2.1 with $\sigma = 1$, $a = 1$, and $b = 4$. The peaks of $v(x)$ occur near the target locations $x = 1$ and $x = 4$.

The integral of $v(x)$ gives the actual number of targets Ξ:

$$\Xi = \int v(x)dx = \int \mathcal{N}_{(\sigma)^2}(x-a)dx + \int \mathcal{N}_{(\sigma)^2}(x-b)dx = 1 + 1 = 2 \tag{2.2}$$

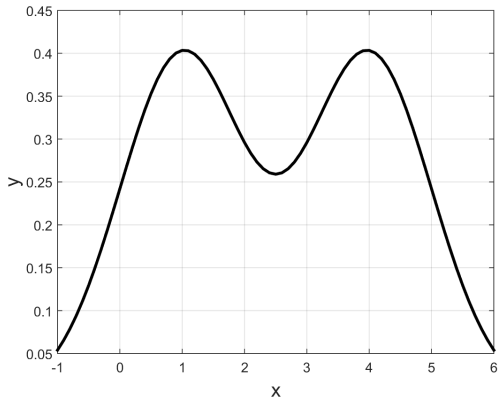


Figure 2.3: An example for the PHD filter.

Contrary to stochastic approaches such as the PF, RFS and PHD filter, the mean-shift is a deterministic method [29] and it can be described as a simple iterative procedure that shifts each data point to the average of data points in its neighbourhood [21]. It has been widely used in clustering [1], mode seeking [149], image segmentation [129] and tracking [27]. One application of the mean-shift method is illustrated in Figure 2.4 where the purpose is to find the densest region of distributed balls. Firstly, an initial point needs to be selected with the region of interest as shown in Figure 2.4a. The blue circle indicates the region of interest centred on the initial point. The centre of the mass is calculated using the balls inside the region of interest in Figure 2.4b. Then, the mean-shift vector is calculated to get the distance and direction for shifting the initial point in Figure 2.4c. The initial point is shifted together with the region of interest to new point in Figure 2.4d. The centre of the mass is calculated again using the balls inside the region of interest and new mass point is obtained in Figure 2.4e. The mean-shift vector is computed to show the direction and distance for shifting and the region of interest is shifted to new point as illustrated in Figures 2.4f and 2.4g, respectively. This iteration continues until the mean-shift method finds the densest point in Figure 2.4h.

To promote a better understanding of the techniques behind the proposed methods in this thesis, the basic mathematical and statistical concepts of the baseline methods are described in the following subsections.

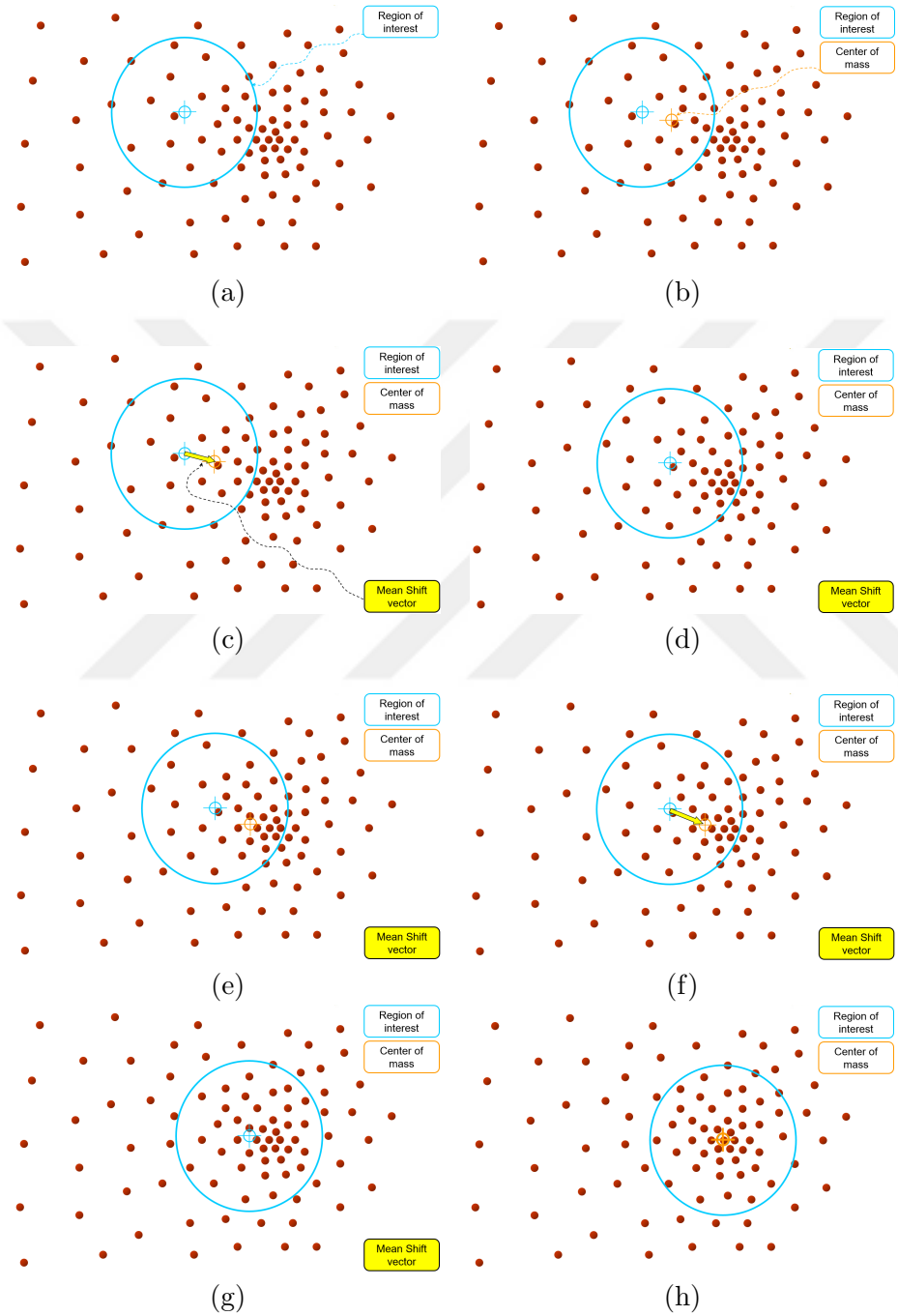


Figure 2.4: Intuitive descriptions of the mean-shift process. Figures are adapted from [131].

2.3.1 Particle Filtering

Here, visual tracking implementation of the PF is described since the AV particle filtering approach proposed in Chapter 3 is built on using audio information in the steps of visual PF.

The PF is an approach for obtaining estimates of the state of a stochastic dynamical system based on observations recursively in time. It is also known as a sequential Monte Carlo (SMC) method based on simulation. It was first introduced by Gordon *et al.* [50]. The PF, which is based on sequential importance sampling and Bayesian theory, is a powerful approach for nonlinear and non-Gaussian problems.

The sampling importance resampling (SIR) is a generalization of the PF framework which can be used in visual tracking to track the position of the face of a speaker (x_1, x_2) in five steps as depicted in Figure 2.5.

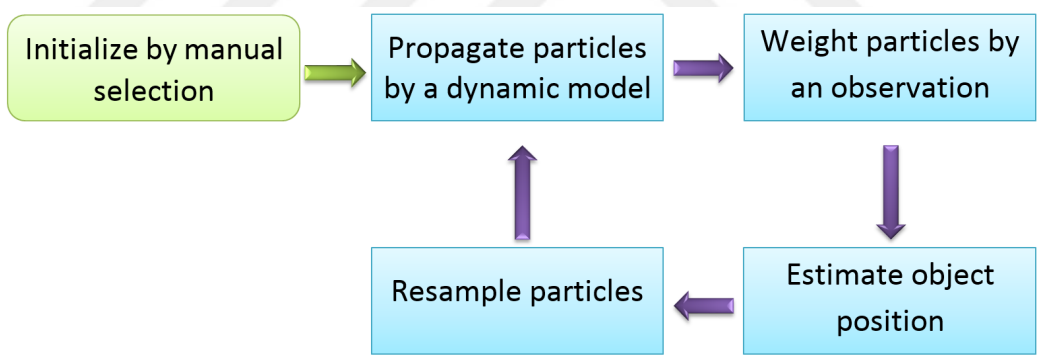


Figure 2.5: Schematic diagram of particle filter based tracking.

The particles are initialized as $\mathbf{x}_0^{(n)} \sim p(\mathbf{x}_0)$, $w_0^{(n)} = \frac{1}{N}$ for $n = 1, ..., N$ in the first step of V-PF. Here, N is the number of particles and $w_0^{(n)}$ are the initial weights of the particles. The state vector is defined as $\mathbf{x} = \begin{bmatrix} x_1 & \dot{x}_1 & x_2 & \dot{x}_2 & s \end{bmatrix}^T$, where x_1 and x_2 are the horizontal and vertical positions of the rectangle centred around the face that we wish to track, $\dot{x}_1$ is the horizontal velocity, $\dot{x}_2$ is the vertical velocity and s is the scale of the rectangle centred around (x_1, x_2). In the second step, particle propagation is employed by a dynamic model [31],

$$\mathbf{x}_k^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)}, \tag{2.3}$$

where $\mathbf{x}_k^{(n)}$ is the state of the nth particle at time frame $k = 1, ..., K$ and $\mathbf{q}_k^{(n)}$ is the zero-mean Gaussian noise with covariance $\mathbf{Q}$, $\mathbf{q}_k^{(n)} \sim \mathcal{N}(0, \mathbf{Q})$ for each particle and $\mathbf{F}$ is the state transition matrix,

$$\mathbf{F} = \begin{bmatrix} 1 & T & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & T & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_s^2 \end{bmatrix},$$

where T is the period between two adjacent frames, σ_s^2 is the variance of the scale and σ^2 is the variance for both the position and the velocity. The third step is the weighting step and the particles are weighted by the observation model [31],

$$w_k^{(n)} = e^{-\lambda (D^{(n)})^2}, \tag{2.4}$$

where λ is the design parameter and $D^{(n)}$ is the Bhattacharyya distance [31]:

$$D^{(n)} = \sqrt{1 - \sum_{u=1}^{U} \sqrt{r_u q_u^{(n)}}}, \tag{2.5}$$

where U is the number of histogram bins, r_u is the histogram of the reference image determined by the user in the initialization step, and $q_u^{(n)}$ is the histogram extracted from the rectangle centred on the position of the nth particle. The RGB or HSV colour model is commonly used in the literature [120] to create a histogram model. In our study, HSV is chosen since it is observed to be more robust to illumination variation. The weights are normalized to ensure that $\sum_{n=1}^{N} w_k^{(n)} = 1$. In the fourth step, the position of the speaker is estimated by [31]:

$$\tilde{\mathbf{x}}_k = \sum_{n=1}^{N} w_k^{(n)} \mathbf{x}_k^{(n)}. \tag{2.6}$$

As a last step, the particles $\mathbf{x}_k^{(n)}$ are resampled to remove the particles with very small weights and duplicate particles with large weights, so a new particle set drawn from $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N}$ is generated. Then it returns to the second step and continues recursively.

To implement the basic resampling method, there are many methods available such as multinomial resampling [2], systematic resampling [65] and residual resampling [77]. Here, multinomial resampling is applied because of its simplicity and being less sensitive to the order in which the particles are ordered during the resampling process [35]. In multinomial resampling, each particle is duplicated by a multinomial distribution with probabilities $w_k^{(n)}$, and then particles in the original set with small variance weights are removed, while those of high weights are duplicated in proportion to these weights.

The pseudo code of V-PF algorithm is given in Algorithm 2.1.

Initialize: N, σ^2, U, T, $\mathbf{F}$, λ, r_u, $\mathbf{x}_0^{(n)}$, $w_0^{(n)}$, k
while $k < K$ **do**
  Propagate particles: $\mathbf{x}_k^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)}$
  Calculate $D^{(n)}$ using Equation (2.5), for $n = 1...N$
  Weighting: $w_k^{(n)} = e^{-\lambda(D^{(n)})^2}$, for $n = 1...N$
  Estimate target position $\tilde{\mathbf{x}}_k = \sum_{n=1}^{N} w_k^{(n)} \mathbf{x}_k^{(n)}$
  Resampling: Generate $\mathbf{x}_k^{(n)}$ from the set $\left\{\mathbf{x}_k^{(n)}, w_k^{(n)}\right\}_{n=1}^{N}$
  $k = k + 1$
end

Algorithm 2.1: V-PF tracking algorithm.

2.3.2 KLD Sampling

The PF uses a weighted set of samples (particles) in order to approximate the filtering distributions and hence the quality of the sample based representation increases with the number of particles. It is, however, not clear how to determine the optimal number of particles to be used for a specific estimation problem. As a rule of thumb, the number of particles is chosen to be as large as possible to get accurate results which leads to an increased computational cost.

A detailed analysis of this trade-off is given by Pitt *et al.* [102] who provided practical guidelines for the estimation of the optimal number of particles in Markov chain Monte Carlo particle filter with the Metropolis Hastings sampler. It is assumed that the standard deviation of the estimated log-likelihood from the PF is around one and inversely proportional to the number of particles. Their results are valid for the Metropolis Hast-

ings sampler, but for other samplers, it is not clear whether the standard deviation of the likelihood from the PF plays the same role in the estimation of the optimal number of particles.

Another potential approach for this problem is variable resolution particle filter (VRPF) [132] which introduces the concept of "abstract particles" in which a particle may represent an individual state or a set of similar states. The VRPF has the advantage that a limited number of particles are sufficient to represent large portions of the state space since a single abstract particle simultaneously tracks multiple similar states. However, this method cannot answer the question of how to determine the optimal number of particles.

Subsequent researchers have therefore proposed adaptive particle filtering (A-PF) approaches in [24], [42] and [122]. The Kullback-Leibler divergence (KLD) sampling algorithm was proposed by Fox [42]. The idea behind this algorithm is to adaptively estimate the number of particles at each step to bound the approximation error introduced by the sample based representation of the PF below a specified threshold. One assumption of KLD-sampling is that a sample based representation of the PF can be used to estimate the posterior by a discrete piecewise constant distribution consisting of a set of multidimensional bins. Subsequent work [122] modified the KLD-sampling criterion to estimate the number of particles and proposed an approach for adaptive propagation of the samples. Recent work [24] uses the innovation error to modify the number of particles being used where a two-fold metric is employed to select the number of particles. The first metric is used to eliminate the particles whose distance to a neighbouring particle is below a predefined threshold, and the second is a basis for setting the threshold on the innovation error to control the birth of particles. These two thresholds should be set prior to running the algorithm, but how they are determined is not mentioned.

The key idea behind the KLD-Sampling algorithm [42] is to estimate the number of particles adaptively to bound the error of the particle filter. To measure the error, the Kullback-Leibler (KL) divergence between the empirical distribution and the true posterior distribution, known as nonparametric maximum likelihood estimate, is used.

KLD-Sampling assumes that the true posterior can be represented by a discrete piecewise constant distribution consisting of a set of multidimensional bins. This assumption allows the use of a chi-square (χ^2) distribution in the likelihood ratio statistic to find a bound for the number of particles, N [42]:

$$N = \frac{1}{2\epsilon}\chi^2_{b-1,1-\delta}, \tag{2.7}$$

where $(1-\delta)$ is the quantile of χ^2 distribution with $b-1$ degrees of freedom, b is the number of bins and ϵ is the upper bound for the error given by the KL-divergence. In order to determine N according to Equation (2.7), a Wilson-Hilferty transformation [42] is applied to compute the quantiles of the chi-square distribution, which yields

$$N = \frac{1}{2\epsilon}\chi^2_{b-1,1-\delta} \doteq \frac{b-1}{\epsilon}\left\{1 - \frac{2}{9(b-1)} + \sqrt{\frac{2}{9(b-1)}}z_{1-\delta}\right\}^3, \tag{2.8}$$

where $z_{1-\delta}$ is the upper $1-\delta$ quantile of the standard normal distribution $\mathcal{N}(0,1)$.

Incorporation of KLD-Sampling into the PF algorithm is achieved by estimating b in the sampling step by incrementally checking for each generated sample whether it falls into an empty bin. The bin size is initially specified depending on the application where the PF is used, and kept constant during implementation. At the beginning of sampling, b goes up with almost every new sample since virtually all bins are empty. After each sample, N is updated by Equation (2.8) required for the current estimate of b. Eventually, more and more bins become non-empty and once N remains unchanged, the update stops.

Detailed information about the KLD-Sampling algorithm can be found in [42].

2.3.3 Random Finite Set

Multi-target tracking is the process of jointly estimating the number of targets and their states from a noisy set of measurements. It is a challenging task in real life environments, as the number of targets may vary due to the appearance and disappearance while the sets of measurements may contain false alarms and measurements of true targets may be missing. The classical Bayesian filter was derived for the case where

the target generates exactly one measurement and there is no clutter. However, this is not the case in a multi-target system. The theory of random finite sets is therefore a natural representation of multi-target states and measurements that allows multi-target filtering by propagation of the multi-target posterior.

This section describes our problem formulation based on the RFS framework for multi-speaker tracking using visual and audio modalities.

In a single speaker tracking system, speaker state is defined as in Section 2.3.1, $\mathbf{x} = \begin{bmatrix} x_1 & \dot{x_1} & x_2 & \dot{x_2} & s \end{bmatrix}^T$, and the constant velocity model given in Equation (2.3) is employed for the evolution of time dependent speaker state [62].

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{q}_k \tag{2.9}$$

Observations are defined as a vector with different dimensions depending on the application. The state and measurement of a single speaker system evolve in time with their dimensions fixed, which is not the case for multi-speaker tracking since the number of speakers and measurements may change [80], [135]. Therefore, the dimensions of the multi-speaker state and measurements also evolve in time.

Since joint detection and tracking of an unknown and time-varying number of speakers is considered, the state to be estimated is no longer a random vector with fixed size. The randomness arises from the number of speakers as well as the positions of the speakers. In our work, such randomness is characterized by using an RFS [80], given by

$$\mathcal{X}_k = \{\mathbf{x}_{1,k}, ..., \mathbf{x}_{\Xi_k,k}\}, \tag{2.10}$$

$$\mathcal{Z}_k = \{\mathbf{z}_{1,k}, ..., \mathbf{z}_{M_k,k}\}, \tag{2.11}$$

where Ξ_k=$|\mathcal{X}_k|$ is the number of speakers, with $|\cdot|$ representing the cardinality of the set. We assume that the maximum number of speakers at each time step is bounded by $\Xi_{\max}$, i.e., $\Xi_k \leq \Xi_{\max}$. Also, $\mathcal{Z}_k$ consists of M_k observations which may be corrupted by noise due to clutter. In a single speaker Bayesian tracking, uncertainty is introduced by modelling $\mathbf{x}_k$ and $\mathbf{z}_k$ as random vectors. Similarly, uncertainty in multi-speaker

tracking is introduced by modelling $\mathcal{X}_k$ and $\mathcal{Z}_k$ as RFSs [80]

$$\mathcal{X}_k = \mathcal{S}_k\left(\mathcal{X}_{k-1}\right) \cup \mathcal{B}_k(\mathcal{X}_{k-1}) \cup \Gamma_k, \tag{2.12}$$

$$\mathcal{Z}_k = \Theta_k(\mathcal{X}_k) \cup \mathcal{C}_k, \tag{2.13}$$

where '$\cup$' denotes union, $\mathcal{S}_k(\mathcal{X}_{k-1})$ denotes the RFS of surviving speakers, $\mathcal{B}_k(\mathcal{X}_{k-1})$ is the RFS of speakers spawned from the previous set of speakers $\mathcal{X}_{k-1}$ and Γ_k is the RFS of the new speakers that appear spontaneously at time k [135]. $\Theta_k(\mathcal{X}_k)$ denotes the RFS of the measurements generated by the speakers $\mathcal{X}_k$ and $\mathcal{C}_k$ is the RFS of clutter or false alarms. We assume that for the birth process at most one speaker is born at a time step and apply the following hypotheses [80], [151]:

$$\mathcal{B}_k(\mathbf{b}_k) = \begin{cases} \emptyset, & \bar{\hbar}_{\text{birth}}, \\ \{\mathbf{b}_k\}, & \hbar_{\text{birth}} \\ \emptyset, & |\mathcal{X}_{k-1}| = \Xi_{\max}, \end{cases} \tag{2.14}$$

where $\hbar_{\text{birth}}$ and $\bar{\hbar}_{\text{birth}}$ are, respectively, the birth and non-birth hypotheses and $\mathbf{b}_k$ is an initial state vector under the birth hypothesis. We denote the probability $\hbar_{\text{birth}}$ by P_{birth}. For the surviving state set $\mathcal{S}_k(\mathcal{X}_{k-1})$, death hypotheses are applied as follows [80], [151]:

$$\mathcal{S}_k(\mathcal{X}_{k-1}) = \begin{cases} \mathcal{S}_k(\mathcal{X}_{k-1}) \setminus \mathbf{x}_{i,k-1}, & \hbar^i_{\text{death}} \\ \bigcup_{i=1}^{|\mathcal{X}_{k-1}|}\{\mathbf{F}\mathbf{x}_{i,k-1} + \mathbf{q}_{i,k}\}, & \bar{\hbar}_{\text{death}} \end{cases} \tag{2.15}$$

where $\hbar^i_{\text{death}}$ and $\bar{\hbar}_{\text{death}}$ are, respectively, the death assumption for the ith speaker and the no-death hypothesis, and '$\setminus$' denotes the set subtraction. Here, we assume that each speaker has the same prior probability of disappearing P_{death}. In the case of the death process, the corresponding state is set as empty, and the other states will evolve according to the dynamic model in Equation (2.9). Besides, the dynamics in the state evolution $\mathcal{X}_k$ are described by the multi-speaker transition density $f_{k|k-1}(\mathcal{X}_k|\mathcal{X}_{k-1})$, while the randomness in the observations is described by the multi-speaker likelihood $g_k(\mathcal{Z}_k|\mathcal{X}_k)$. Then, the RFS formulation can be employed in the optimal multi-speaker Bayesian filter by propagating the multi-speaker posterior density. In this thesis, we use the particle filter implementation of the RFS, which is RFS-PF, and we propose

to integrate the audio information with visual RFS-PF in order to improve the tracker for a varying number of speakers, which is introduced in Chapter 4.

However, it is observed that the RFS approach is computationally intractable since multiple integrals are involved in the recursion of the multi-speaker posterior. To reduce the computational complexity, the PHD filter is proposed which propagates the first-order moment of the multi-speaker posterior instead of the posterior itself [85].

2.3.4 PHD Filtering with SMC Implementation

The first moment approximation of the multi-target posterior, the PHD filter, provides a more computationally efficient alternative, but still theoretically sound, solution. The PHD filter is defined as the intensity $v_{k|k}$ whose integral on any region of the state space gives the expected number of speakers. The local maxima of the PHD function indicate the highest local concentration of the expected number of speakers, which also identify the likely positions of the speakers. The PHD filter has two iterative steps: prediction and update. The prediction step of the PHD is defined as [87]

$$v_{k|k-1}\left(\mathbf{x}_k\right) = \xi_k\left(\mathbf{x}_k\right) + \int \phi_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) v_{k-1|k-1}\left(\mathbf{x}_{k-1}\right) d\mathbf{x}_{k-1}, \tag{2.16}$$

where $\xi_k\left(\mathbf{x}_k\right)$ denotes the intensity function of the new speaker birth RFS Γ_k, and $\phi_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ is the analog of the single-speaker state transition probability [87]

$$\phi_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) = p_{S,k}\left(\mathbf{x}_{k-1}\right) f_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) + \beta_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right), \tag{2.17}$$

where $p_{S,k}\left(\mathbf{x}_{k-1}\right)$ is the survival probability for the speakers still existing and $f_{k|k-1}$ $\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ is the single-speaker state transition density. The intensity function of RFS $\mathcal{B}_k(\mathcal{X}_{k-1})$ is denoted by $\beta_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ for the speaker spawned at time k with previous state $\mathbf{x}_{k-1}$. The PHD update is defined as [87]

$$v_{k|k}\left(\mathbf{x}_k\right) = \left[1 - p_{D,k}\left(\mathbf{x}_k\right)\right] v_{k|k-1}\left(\mathbf{x}_k\right) + \sum_{\mathbf{z}_k \in \mathcal{Z}_k} \frac{p_{D,k}\left(\mathbf{x}_k\right) g_k\left(\mathbf{z}_k|\mathbf{x}_k\right) v_{k|k-1}\left(\mathbf{x}_k\right)}{\kappa_k\left(\mathbf{z}_k\right) + \int p_{D,k}\left(\mathbf{x}_k\right) g_k\left(\mathbf{z}_k|\mathbf{x}_k\right) v_{k|k-1}\left(\mathbf{x}_k\right)}, \tag{2.18}$$

where $p_{D,k}(\mathbf{x}_k)$ is detection probability and $g_k(\mathbf{z}_k|\mathbf{x}_k)$ is the single-speaker likelihood defining the probability that $\mathbf{z}_k$ is generated by a speaker state $\mathbf{x}_k$. The intensity of clutter RFS $\mathcal{C}_k$ is defined as $\kappa_k(\mathbf{z}_k)$ which is $\kappa_k(\mathbf{z}_k) = \psi u(\mathbf{z}_k)$, where ψ is the average number of Poisson clutter points per scan and $u(\mathbf{z}_k)$ is the probability distribution of each clutter point.

With the multiple integrals in the PHD prediction Equation (2.16) and update Equation (2.18) steps, there are no closed-form solutions in general. To obtain a numerical solution for the integrals in PHD recursion, the SMC method has been proposed which approximates the PHD with a set of random samples (particles) [135]. Suppose that at time step $k-1$, the PHD $v_{k-1|k-1}(\mathbf{x}_{k-1})$ is approximated by $\left\{w_{k-1}^{(n)}, \mathbf{x}_{k-1}^{(n)}\right\}_{n=1}^{N_{k-1}}$ of N_{k-1} particles and their corresponding weights as [135]

$$v_{k-1|k-1}(\mathbf{x}_{k-1}) \approx \sum_{n=1}^{N_{k-1}} w_{k-1}^{(n)} \delta\left(\mathbf{x}_{k-1} - \mathbf{x}_{k-1}^{(n)}\right). \tag{2.19}$$

Prediction of the PHD $v_{k|k-1}(\mathbf{x}_k)$ is obtained with particles $\tilde{\mathbf{x}}_k$ and their weights $\tilde{w}_{k|k-1}$, $\left\{\tilde{w}_{k|k-1}^{(n)}, \tilde{\mathbf{x}}_k^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$. Here, N_{k-1} particles of $\tilde{\mathbf{x}}_k$ are first drawn from importance sampling $q_k\left(\tilde{\mathbf{x}}_k^{(n)}|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k\right)$ to propagate the particles from time step $k-1$, then J_k particles of $\tilde{\mathbf{x}}_k$ from the new born importance function $p_k\left(\tilde{\mathbf{x}}_k^{(n)}|\mathcal{Z}_k\right)$ are drawn to model the state of new speakers appearing in the scene [135]

$$v_{k|k-1}(\mathbf{x}_k) \approx \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_{k|k-1}^{(n)} \delta\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_k^{(n)}\right), \tag{2.20}$$

where J_k new particles arise from the birth process. By substituting Equation (2.19) into Equation (2.16) and then applying importance sampling, we get predicted weights $\tilde{w}_{k|k-1}^{(n)}$ [135]:

$$\tilde{w}_{k|k-1}^{(n)} = \begin{cases} \frac{\phi_{k|k-1}\left(\tilde{\mathbf{x}}_k^{(n)}, \mathbf{x}_{k-1}^{(n)}\right) w_{k-1}^{(n)}}{q_k\left(\tilde{\mathbf{x}}_k^{(n)}|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k\right)}, & n = 1, ..., N_{k-1} \\ \\ \frac{\xi_k\left(\tilde{\mathbf{x}}_k^{(n)}\right)}{J_k p_k\left(\tilde{\mathbf{x}}_k^{(n)}|\mathcal{Z}_k\right)}, & n = N_{k-1}+1, ..., N_{k-1}+J_k \end{cases} \tag{2.21}$$

The update step of the PHD recursion is obtained by updating the weight of the pre-

dicted particles when the likelihood $g_k\left(\mathbf{z}_k|\tilde{\mathbf{x}}_k^{(n)}\right)$ is available. Then $v_{k|k-1}\left(\mathbf{x}_k\right)$ is substituted into Equation (2.18) and the predicted weights $\left\{\tilde{w}_{k|k-1}^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$ are updated according to [135]

$$\tilde{w}_k^{(n)} = \left[\left[1 - p_D\left(\tilde{\mathbf{x}}_k^{(n)}\right)\right] + \sum_{\mathbf{z}_k \in \mathcal{Z}_k} \frac{p_D\left(\tilde{\mathbf{x}}_k^{(n)}\right) g_k\left(\mathbf{z}_k|\tilde{\mathbf{x}}_k^{(n)}\right)}{\kappa_k\left(\mathbf{z}_k\right) + C_k\left(\mathbf{z}_k\right)}\right] \tilde{w}_{k|k-1}^{(n)}, \tag{2.22}$$

where

$$C_k\left(\mathbf{z}_k\right) = \sum_{j=1}^{N_{k-1}+J_k} p_D\left(\tilde{\mathbf{x}}_k^{(j)}\right) g_k\left(\mathbf{z}_k|\tilde{\mathbf{x}}_k^{(j)}\right) \tilde{w}_{k|k-1}^{(j)}. \tag{2.23}$$

Note that J_k new particles, sampled for the born speakers at each iteration, are added to the old ones $N_k = N_{k-1} + J_k$ which causes the number of particles to grow over time and makes the PHD filter inefficient. In addition, the low weight particles need to be removed and particles with high weights should be duplicated in order to concentrate the particles on the zones around the targets. To this end, the resampling step is performed after the update step. N_k particles are resampled from $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \tilde{\mathbf{x}}_k^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$ where $\hat{\Xi}_{k|k}$ is the total mass and $\hat{\Xi}_{k|k} = \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_k^{(n)}$. N_k is estimated by $N_k = \eta\hat{\Xi}_{k|k}$ where η is the constant number of particles per speaker. Therefore, the complexity of the SMC-PHD filter increases *linearly* with the number of speakers. After the resampling step, new weights of set $\left\{w_k^{(n)}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$ are normalized to preserve the total mass.

The SMC-PHD filter propagates the surviving, spawned and born particles to model the new and existing speakers. Conventionally, these particles are used every frame which increases the computational complexity. We address this problem by introducing the DOA information which is described in detail in Chapter 4.

2.3.5 Mean Shift Tracking

As mentioned earlier, the search paradigms in tracking can be categorized into two methods as one is stochastic and model-driven while the other is deterministic and data-driven [125]. The PF and PHD filter methods, explained in previous sections are examples of stochastic and model-driven methods. In this section, the mean-shift

algorithm is explained which is perhaps the most popular technique among the deterministic and data-driven methods.

Mean-shift tracking is a robust non-parametric method for climbing density gradients to find the peak of probability distributions [27], [28] and was first adapted for tracking of non-rigid objects in [29] by focusing on a histogram-based target representation and localisation using Bhattacharyya coefficients as similarity measures. Its robustness, ease of implementation and computational efficiency make it a commonly used tracking algorithm [10].

The basic idea behind the mean-shift tracking is to find the target in the next image frame that is most similar to the initialised target (reference model) in the current frame by iteratively searching the next frame with a non-parametric kernel. Similarity is measured as the Bhattacharyya distance between the histogram of the target model and the histogram of the candidate target in the next frame. Before tracking is started, the bin size, kernel function, size of the kernel and maximum iteration number and reference model need to be determined in the initialization step. After the colour histogram of the reference model is found, the probability density function (pdf) of the reference model is calculated as follows [29]

$$r_u = C \sum_{i=1}^{\Delta} \mathfrak{k}\left(\|\mathbf{x}_i^*\|^2\right) \delta[b(\mathbf{x}_i^*) - u], \tag{2.24}$$

where δ is the Kronecker delta function, $\{\mathbf{x}_i^*\}_{i=1,\ldots,\Delta}$ contains the *normalized* pixel locations with Δ number of pixels in the region defined as the target model, $b(\mathbf{x}_i^*)$ assigns one of the bins of the histogram to a given colour at location $\mathbf{x}_i^*$ and $\mathfrak{k}$ is the kernel function that gives higher weights to the pixels at the centre of the target model. The normalization constant C ensures the condition $\sum_{u=1}^{U} r_u = 1$ where u represents the histogram bin index and U is the number of histogram bins. C is derived as [29]

$$C = \frac{1}{\sum_{i=1}^{\Delta} \mathfrak{k}\left(\|\mathbf{x}_i^*\|^2\right)}. \tag{2.25}$$

After calculation of the reference model in the initialization step, the candidate model

at location $\mathbf{y}$ is given by [29]

$$q_u(\mathbf{y}) = C_h \sum_{i=1}^{\Delta_h} \mathfrak{k}\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right) \delta[b(\mathbf{x}_i) - u], \tag{2.26}$$

where h defines the scale of the candidate target, $\{\mathbf{x}_i\}_{i=1,\ldots,\Delta_h}$ contains the *normalized* pixel locations with Δ_h number of pixels in the region defined as the candidate target model and C_h is the normalization constant given by [29]

$$C_h = \frac{1}{\sum_{i=1}^{\Delta_h} \mathfrak{k}\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right)}. \tag{2.27}$$

During the tracking, the pdf of the candidate target is compared with the pdf of the reference model and the target is detected based on the similarity of this comparison. The similarity is measured by the Bhattacharyya distance [29],

$$d(\mathbf{y}) = \sqrt{1 - \rho\left[\mathbf{q}\left(\mathbf{y}\right), \mathbf{r}\right]}, \tag{2.28}$$

where $\mathbf{r} = \{r_u\}_{u=1,\ldots,U}$ $(\sum_{u=1}^{U} r_u = 1)$ is the U-bin reference colour histogram of the target, the centre point $\mathbf{y}$ of the image region in which the colour histogram $\mathbf{q}\left(\mathbf{y}\right) = \{q_u(\mathbf{y})\}_{u=1,\ldots,U}$ $(\sum_{u=1}^{U} q_u = 1)$, and the Bhattacharyya coefficient [29] $\rho\left[\mathbf{q}\left(\mathbf{y}\right), \mathbf{r}\right]$ is given by

$$\rho\left(\mathbf{y}\right) \equiv \rho\left[\mathbf{q}\left(\mathbf{y}\right), \mathbf{r}\right] = \sum_{u=1}^{U} \sqrt{q_u(\mathbf{y}) r_u}. \tag{2.29}$$

The value of ρ is varied between 0 and 1 and a larger ρ means a higher similarity in their pdfs.

To minimize the distance in Equation (2.28), the Bhattacharyya coefficient $\rho\left(\mathbf{y}\right)$ needs to be maximized. The search of the new location of the target in the current frame starts from the location, $\mathbf{y}_0$, which is the location of the target in the previous frame. So, first $\mathbf{q}\left(\mathbf{y}_0\right) = \{q_u(\mathbf{y}_0)\}_{u=1,\ldots,U}$ of the target candidate at location $\mathbf{y}_0$ in the current frame needs to be estimated. The linear approximation of the Bhattacharyya coefficient in Equation (2.29) is obtained by using the Taylor expansion around the values $q_u(\mathbf{y}_0)$

[29],

$$\rho\left[\mathbf{q}\left(\mathbf{y}\right),\mathbf{r}\right] \approx \frac{1}{2}\sum_{u=1}^{U}\sqrt{q_u(\mathbf{y}_0)r_u} + \frac{1}{2}\sum_{u=1}^{U} q_u(\mathbf{y})\sqrt{\frac{r_u}{q_u(\mathbf{y}_0)}}. \tag{2.30}$$

If Equation (2.26) is substituted in Equation (2.30), then [29]

$$\rho\left[\mathbf{q}\left(\mathbf{y}\right),\mathbf{r}\right] \approx \frac{1}{2}\sum_{u=1}^{U}\sqrt{q_u(\mathbf{y}_0)r_u} + \frac{C_h}{2}\sum_{i=1}^{\Delta_h} w_i \mathfrak{k}\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right), \tag{2.31}$$

where

$$w_i = \sum_{u=1}^{U}\sqrt{\frac{r_u}{q_u(\mathbf{y}_0)}}\delta[b(\mathbf{x}_i)-u]. \tag{2.32}$$

Here, to minimize the distance in Equation (2.28), the second term in the right hand side of the Equation (2.31) needs to be maximized as the first term is independent of $\mathbf{y}$. The second term represents the density estimate computed with kernel profile $k(\mathbf{x})$ at $\mathbf{y}$ in the current frame, with the data being weighted by w_i in Equation (2.32). By employing the mean-shift procedure [26], the mode of this density in the local neighbourhood can be found. In this procedure, the kernel is moved from the current location $\mathbf{y}_0$ to the new location $\mathbf{y}_1$ recursively by the relation [29];

$$\mathbf{y}_1 = \frac{\sum_{i=1}^{\Delta_h}\mathbf{x}_i w_i g(\|\frac{\mathbf{y}_0-\mathbf{x}_i}{h}\|)}{\sum_{i=1}^{\Delta_h} w_i g(\|\frac{\mathbf{y}_0-\mathbf{x}_i}{h}\|)}, \tag{2.33}$$

where $g(x) = -\mathfrak{k}'(x)$, under the assumption that the derivative of $\mathfrak{k}(x)$ exists for all $x \in [0,\infty)$, except for a finite set of points.

There are various kernel options such as normal, uniform and Epanechnikov. Here, Epanechnikov kernel is used as it is recommended in [29] since it leads to simplification in Equation (2.33) as explained in the following. The Epanechnikov kernel is defined

as [29] :

$$\mathfrak{k}\left(\mathbf{x}\right) = \begin{cases} (1-\mathbf{x}) & \|\mathbf{x}\| \leq 1 \\ 0 & otherwise \end{cases} \tag{2.34}$$

So, the derivative of the kernel $g(x)$ is [29],

$$g(x) = -\mathfrak{k}'\left(\mathbf{x}\right) = \begin{cases} 1 & \|\mathbf{x}\| \leq 1 \\ 0 & otherwise \end{cases} \tag{2.35}$$

In this case, $g(x)$ is constant and Equation (2.33) reduces to [29],

$$\mathbf{y}_1 = \frac{\sum_{i=1}^{\Delta_h} \mathbf{x}_i w_i g(\|\frac{\mathbf{y}_0 - \mathbf{x}_i}{h}\|)}{\sum_{i=1}^{\Delta_h} w_i g(\|\frac{\mathbf{y}_0 - \mathbf{x}_i}{h}\|)} = \frac{\sum_{i=1}^{\Delta_h} \mathbf{x}_i w_i}{\sum_{i=1}^{\Delta_h} w_i}. \tag{2.36}$$

In order to maximize the Bhattacharyya coefficients, the current search area is shifted iteratively until the similarity of two pdfs is less than a threshold, or when the iteration number has reached to a predefined number. By running the mean-shift tracker on each frame, the target can be tracked over time.

The mean-shift tracking has fast convergence speed and small computational cost. In addition, it is a non-parametric method which provides a general and reliable solution independently from the features representing the targets. However, it is prone to converge to local maxima in cases of clutter or occlusion since the mean-shift process only aims to find the densest (most similar) region based on the starting position within the region of interest. In this sense, the mean-shift trackers tend to fail easily in tracking small and fast moving targets as the region of interest may not cover the targets, so it cannot recover a track after a complete occlusion. In addition, it is derived for single target tracking, so it cannot handle a variable number of targets.

In this study, we integrate mean-shift with our proposed algorithm given in Chapter 4 to address the problems in PHD filtering and benefit the advantages of both deterministic and stochastic approaches. The proposed algorithm is introduced in Chapter 5.

2.4 Relevant Datasets

Several datasets are presented in the literature that combine multiple audio and video sources for tracking.

One of them is the AMI (Augmented Multi-party Interaction) [15] corpus which includes 100 hours of meetings, which were recorded in English using three different rooms. Some of the meetings contain natural conversations, and many others, in particular those using a scenario in which the participants play different roles in a design team are also reasonably natural. Among the natural conversations, the number of speakers varies from three to five. In one type of artificial meeting, four speakers are involved, taking four pre-arranged roles (as industrial designer, interface designer, marketing, and project manager). Other artificial meeting types also appear in the AMI corpus, such as a film club scenario. However, this dataset is not chosen in the evaluation of the proposed algorithms as the speakers are mostly static or with small movements. In addition, calibration information is not available which is required in the proposed algorithms for projecting the coordinates from the $2D$ image into $3D$ space.

The second dataset is CLEAR (CLassification of Events, Activities and Relationships) which is an international effort to evaluate systems that are designed for perceiving identities of people, activities, interactions and relationships in human-human interaction, and related scenarios [93], [123]. This database features recordings of multiple users in realistic small meeting scenarios, captured in a variety of smart rooms equipped with a multitude of audio-visual sensors. It offers five calibrated and synchronized visual streams, four from cameras mounted in the room corners and one panoramic ceiling-mounted camera, as well as synchronized audio streams from a minimum of four microphone arrays on the walls. The speakers are mostly seated around the table and speaking one by one. Therefore, the CLEAR dataset is not chosen. In addition, the audio signals were acquired with linear microphone arrays which cannot be run by our audio-detection algorithm that is designed for a circular microphone array.

Surveillance performance evaluation initiative (SPEVI) [126] is a dataset which contains uni-modal and multi-modal (audio and visual) people detection and tracking. It in-

cludes three sequences for different scenarios recorded by a video camera and two linear microphone arrays. Two of the sequences (motinas_Room160 and motinas_Room105) are recorded in rooms with reverberation and the last sequence (motinas_Chamber) is recorded in a room with reduced reverberation. However, audio signals were recorded with linear microphone arrays and the calibration information is not available. Therefore, this dataset is also not chosen.

One of the most challenging datasets that can be used for testing audio-visual tracking is $AV16.3$ corpus which is developed by IDIAP research institute [70]. The corpus $AV16.3$ has many sequences for different scenarios where subjects are moving and speaking at the same time whilst being recorded by three calibrated video cameras and two circular eight-element microphone arrays. The physical setup of the corpus is illustrated in Figure 2.6.

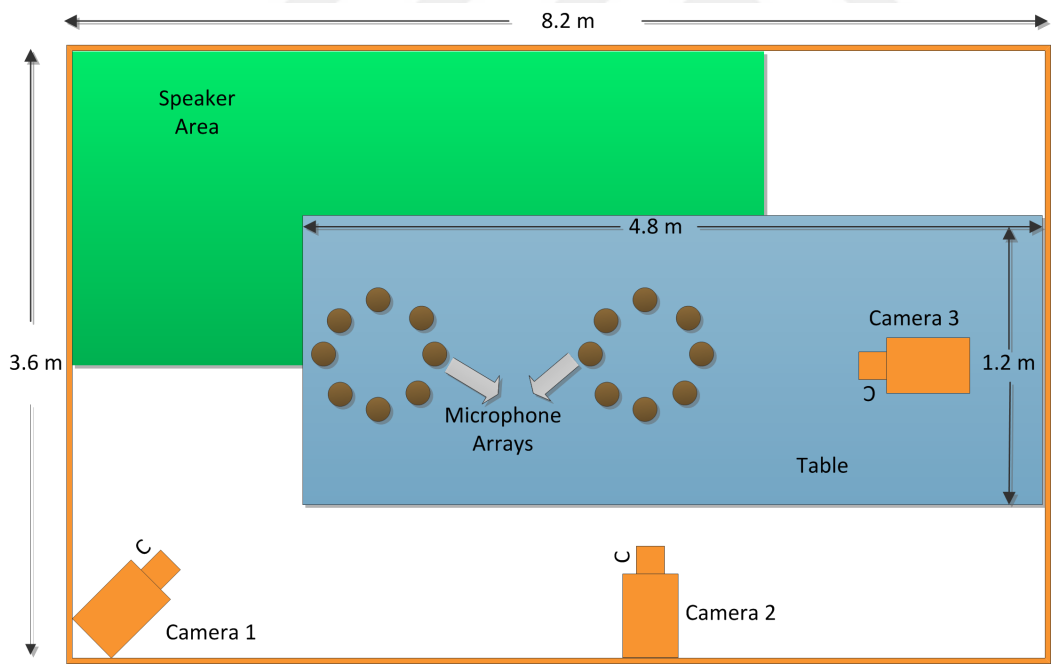


Figure 2.6: Physical setup of the $AV16.3$ corpus. Two circular microphone arrays are on the table and three cameras are located in different positions.

Recordings in the $AV16.3$ include overlapped speech, close and far locations, small and large angular separations, object initialization, variable number of objects, partial and total occlusion and illumination. The audio and video were recorded independently from each other. The audio signals were recorded at 16 kHz and the concurrent video

sequences were recorded at 25 Hz. As we are running our tracking system with both audio and visual modalities, they are then synchronized before being used in our system. Each video frame is a colour image of 288×360 pixels. Some sequences are annotated to get the ground truth speaker position which allows us to measure the accuracy of each tracker and compare the performance of the algorithms. We have also used some sequences from the $AV16.3$ dataset throughout this thesis to evaluate the algorithms as it has all specifications we need such as: circular microphone arrays, calibration information, challenging scenarios like occlusions and moving speakers.

Some frames from $AV16.3$ are shown in Figure 2.7. The frames from (a) to (c) show the view from three different angles. From (d) to (i), the same frames are shown from different sequences. In (d), the random walk of a single speaker is shown. The three-speaker case is also given in (i). Some challenging situations such as rotation of head, two-speaker occlusions, contiguous faces, and three-speaker occlusions are given respectively, in (e), (f), (g) and (h) which make tracking much more difficult than an ordinary case.

2.5 Performance Metrics

Performance evaluations and comparison with existing methods are an important factor in the development of new algorithms and techniques. There are no standard protocols to measure the performance of multi-target tracking methods, so several metrics are proposed in the literature. Four different types of metrics are used in this thesis to evaluate the performance of the proposed and baseline algorithms.

The first one is the mean absolute error (MAE) which is estimated as the Euclidean distance in pixels between the estimated and the ground truth positions, then divided by the number of frames. This metric is chosen because of its simplicity and explicit output for the performance comparison.

The second one is the multiple object tracking (MOT) metric proposed in [9], together with its quantities, MOT precision (MOTP) and MOT accuracy (MOTA). The MOTP measures the precision of the tracking system by comparing it with a threshold value

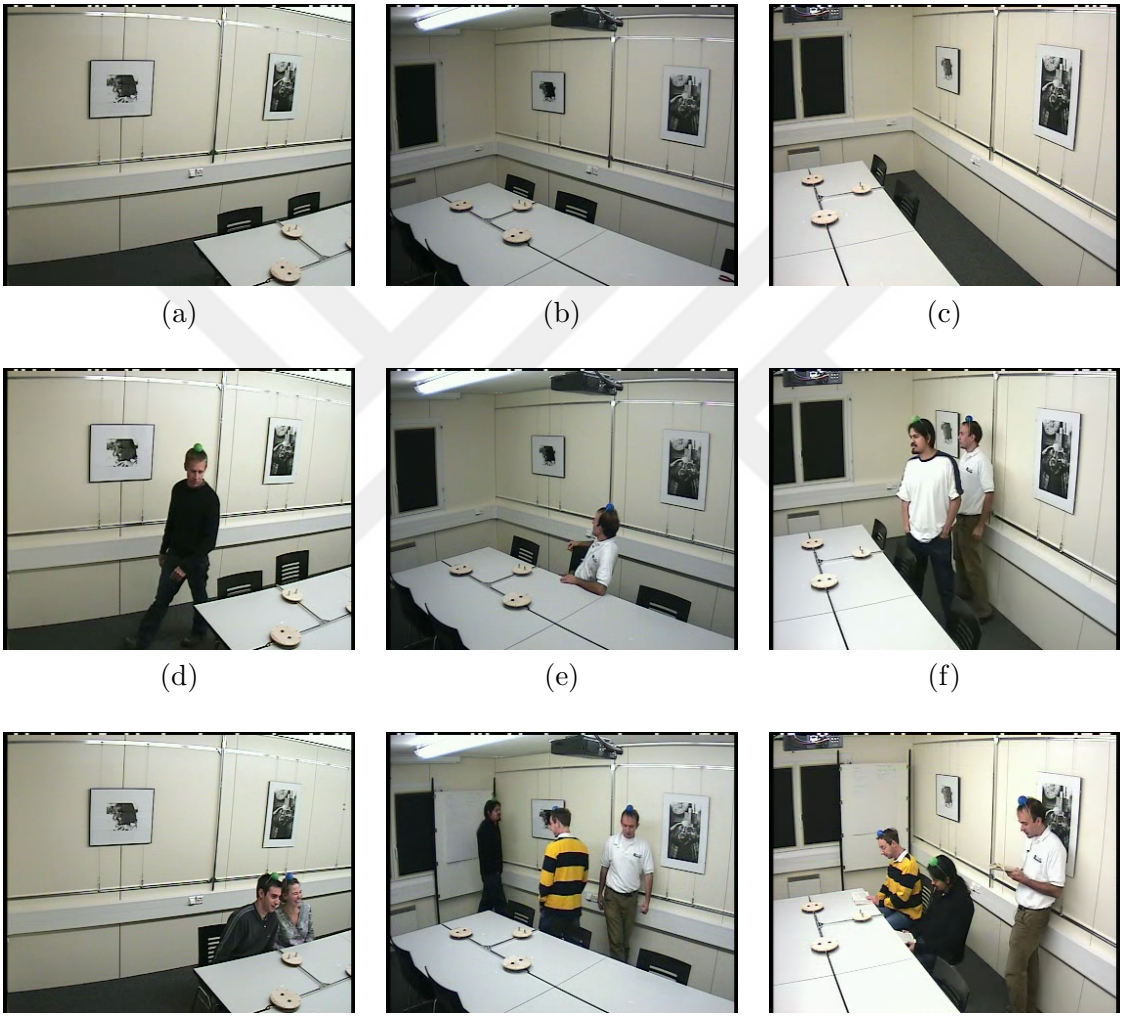


Figure 2.7: Some frames from the $AV16.3$ dataset. The first row shows the views of three different cameras. The second and third rows show the frames from single-speaker, two-speaker and three-speaker sequences.

pre-defined in terms of the Euclidean distance (either in metres [9] or pixels [111]).

$$MOTP = \frac{\sum_{i,k} \bar{d}_k^i}{\sum_k \mathfrak{c}_k}, \tag{2.37}$$

where $\bar{d}_k^i$ is the distance between the ith object and its corresponding hypothesis and $\mathfrak{c}_k$ is the number of matches between the objects and hypotheses for the time frame k.

On the other hand, the MOTA measures the tracking errors, consisting of the false positives (i.e. the case where the error is greater than the threshold value), false negatives (if the speaker is not tracked with the accuracy measured by the threshold) and mismatches (when the speaker identity is switched) [9].

$$MOTA = 1 - \frac{\sum_k (m_k + fp_k + mm_k)}{\sum_k g_k}, \tag{2.38}$$

where m_k, fp_k, mm_k, g_k represent the number of misses (false negatives), of false positives, mismatches, and objects present, respectively, for time frame k.

Another metric is the trajectory-based measures (TBM) proposed in [74] and [144] which measures the performance on the basis of trajectory. According to their definitions, a trajectory can be categorized as mostly tracked (MT) or mostly lost (ML) if, respectively, at least 80% or less than 20% of its ground truth (GT) trajectory is covered by the tracker. Otherwise, it is considered as partially tracked (PT). Additionally, track fragmentation (Frag) is the total number of times that GT is interrupted in tracking result, and identity switches (IDS) measures the total number of times that a tracked trajectory changes its matched GT identity.

As a last performance metric, the well-known OSPA-T (Optimal Subpattern Assignment for Tracks) metric [107] is chosen which is a mathematically consistent metric for the evaluation of multi-speaker tracking systems. The OSPA-T is based on the OSPA metric [114] and extends it for tracking management evaluation. The OSPA employs a penalty value to transfer the cardinality error into the state error and is able to present

the performance on source number estimation as well as speaker position estimation:

$$e_{\mathrm{OSPA}}(\hat{\mathcal{X}}_k, \mathcal{X}_k) = \min_{\pi \in \Pi_{\hat{\Xi}_k, \Xi_k}} \sqrt[a]{\frac{1}{\Xi_k}\left(\sum_{i=1}^{\hat{\Xi}_k} \bar{d}^{(c)}(\hat{\mathbf{x}}_{i,k}, \mathbf{x}_{\pi_i,k})^a + c^a(\Xi_k - \hat{\Xi}_k)\right)}, \tag{2.39}$$

where it is assumed that $\hat{\mathcal{X}}_k = \{\hat{\mathbf{x}}_{1,k}, ..., \hat{\mathbf{x}}_{\hat{\Xi}_k,k}\}$ is an estimation of the ground truth state set $\mathcal{X}_k = \{\mathbf{x}_{1,k}, ..., \mathbf{x}_{\Xi_k,k}\}$ and $\Pi_{\hat{\Xi}_k,\Xi_k}$ is the set of maps $\pi : 1, ..., \hat{\Xi}_k \rightarrow 1, ..., \Xi_k$. Here the state cardinality estimation $\hat{\Xi}_k$ may not be the same as the ground truth Ξ_k. The OSPA error given in Equation (2.39) is for $\hat{\Xi}_k \leq \Xi_k$. If $\Xi_k < \hat{\Xi}_k$, then $e_{\mathrm{OSPA}}(\hat{\mathcal{X}}_k, \mathcal{X}_k) = e_{\mathrm{OSPA}}(\mathcal{X}_k, \hat{\mathcal{X}}_k)$. The function $\bar{d}^{(c)}(\cdot)$ is defined as $\min(c, \bar{d}(\cdot))$ where c is the cut-off value which determines the relative weighting of the penalties assigned to cardinality and localization errors. In addition, a describes the metric order which determines the sensitivity to outliers. The OSPA-T metric differs from other metrics as it considers not only the position estimation of the speaker, but also the estimation of the number of speakers in the evaluation of the tracking results. As OSPA-T measures the error based on these two terms, state (position estimation) and cardinality (number of speaker estimation), it causes ambiguities about how much error is coming from each term to the final error. In addition to x_1 and x_2 variables of the state vector, the scale variable, s, can be considered in the evaluation. However, this will increase the ambiguities in the contributions of the terms to the final error and deteriorate the reliability of the metric. Also, scale information is needed as a ground truth which is not the case in $AV16.3$. As we perform point-based tracking in this study, we prefer to use OSPA-T metric as it is.

As a summary, four metrics were introduced which evaluate the methods from their own perspectives. To see how well the tracker follows its trajectory, the TBM is used to measure its performance. If the tracking error needs to be shown, the MAE or the more advanced option MOT can be used to see how accurately the tracker follows the target. If an unknown and variable number of targets needs to be tracked, then the OSPA-T metric is suitable as it considers both position estimation and the estimated number of targets in the performance evaluation.

2.6 Chapter Summary

In this chapter, a review has been provided on modalities, existing tracking techniques, datasets and performance metrics that have been developed over the past few decades.

As a visual feature, colour information is used for easy implementation and to extract DOA information, the sam-spare-mean method is employed which complied with circular microphone arrays.

After a broad survey of the tracking methods, a technical background of the methods such as, particle filtering, KLD-sampling, random finite set, PHD filter and mean-shift, used throughout this thesis is introduced with their basic mathematical, statistical concepts and definitions, which are required for understanding the mathematics and techniques behind the proposed tracking algorithms.

In order to perform a quantitative evaluation of the proposed algorithms, both audio and video sequences are required, together with the calibration information on the cameras and microphone arrays (circular arrays). Apart from "$AV16.3$", we have also explored the suitability of several other publicly available audio-visual datasets, such as "CLEAR" "AMI" and "SPEVI", and concluded that only the $AV16.3$ dataset is suitable for the evaluation of our proposed methods. It complies with our requirements in terms of having circular microphone arrays with calibration information, mostly talking speakers, and challenging scenarios such as occlusion and rapid movements of the speakers. The other datasets do not fit at least one requirement of this study.

All proposed algorithms are tested and compared with baseline algorithms on the $AV16.3$. Unfortunately, there is no clearly defined protocol in the literature for quantitative evaluations of tracking methods and their comparison with other methods. Therefore, we have used several performance metrics to make comparisons between the proposed and baseline methods.

So far, a comprehensive overview of multi-speaker tracking problems and a number of traditional multi-speaker tracking techniques have been provided. In addition, modalities, datasets and performance metrics have been discussed. Our contributions to multi-speaker tracking problems are discussed in the following chapters.

Chapter 3

Audio Assisted Visual Tracking with Adaptive Particle Filtering

In this chapter, we propose a novel approach for multiple speaker tracking by combining audio and visual data. In our framework, we employ audio data as an aid to PF based visual tracking, by using the direction of arrival angles of the audio sources to reshape the typical Gaussian noise distribution of particles in the propagation step and to weight the observation model in the measurement step. This approach is further improved by solving a typical problem associated with the PF. It has been observed that the efficiency and accuracy of the PF usually depend on the number of particles and noise variance used in the estimation and propagation functions for re-allocating these particles at each iteration. Both of these parameters are specified beforehand and are kept fixed in the regular implementation of the PF which makes the tracker inconsistent in practice. To address these problems, we design an algorithm that adapts both the number of particles and noise variance based on the tracking error and the area occupied by the particles in the image. Experiments on the $AV16.3$ dataset show the advantage of our proposed methods over the baseline PF method and an existing adaptive PF algorithm for tracking occluded speakers with a significantly reduced number of particles.

3.1 Introduction

Speaker tracking may be achieved in a single modality domain through video or audio. Video tracking is generally accurate, but it suffers from a limited field of view, occlusions, and changes in appearance and illumination. On the other hand, audio tracking is not restricted by these limitations, but it is prone to the errors caused by acoustic noise, room reverberations and the intermittency between utterance and silence. As shown in some already published works [46], [52], [118], [127], fusing both audio and visual modalities can provide more robust tracking in comparison to the use of only a single modality, as is the focus here.

Among the approaches presented in Chapter 2, the PF framework has been chosen for tracking multiple speakers in our first study. Compared to other sequential Bayesian estimation techniques, the advantage of PFs lies in their flexibility with respect to the types and numbers of features they support, their robustness in the presence of noise, and the non-parametric fashion in which they represent the belief about the target state, which makes them applicable for highly nonlinear, non-Gaussian estimation problems.

Here, we focus on two challenging problems associated with PF based visual tracking. The first problem stems from the limitations of using the single modality of vision which affects the accuracy and reliability of the tracker because of the limited field of view and occlusion. To address this problem, audio data is used as a second modality to improve the performance of visual tracker.

Researchers have presented fusion strategies for integrating audio localization information with video tracking [6], [46], [118], [127]. These strategies are performed by modifying the observation model [46], using the likelihood function composition of different sensor information [118], state association of two modalities [127], or a graphical model for characterising mutual dependencies of the two modalities [6]. These methods, however, are sensitive to the outliers in audio data, and noisy audio data can easily cause deviation in the estimation of the target position.

Unlike these methods, in this study, we propose integrating audio and visual data in the steps of the PF framework, by weighting the contribution of the audio in order to

minimize the negative effect of outliers and noise coming from the audio data, rather than performing any *a priori* data fusion algorithm. One benefit of the proposed approach, audio-visual PF, is that running a data fusion algorithm is not required which would introduce extra computational cost. To the best of our knowledge, audio information has not been previously fused with visual information in a PF as we do here.

The second problem originates from the PF itself. The PF approximates the posterior density with a set of particles, and this approximation approaches the true posterior as $N \rightarrow \infty$. However, computational cost increases with respect to the number of particles and it is not clear how to determine the optimal number of particles to be used for a specific estimation problem. To address this problem, the proposed AV-PF algorithm is further improved by proposing an adaptive approach to PF, based on the occupied area by the particles in each frame. Our adaptive approach allows us to estimate dynamically not only the number of particles but also the noise variance which makes it different from the adaptive approaches mentioned in Section 2.3.2 with the advantage that adaptive noise variance is used in the estimation of the optimal number of particles. Finally, we demonstrate the results using simulations to compare the performance of the proposed algorithms with [42] and the visual PF.

3.2 Audio Detection and Localization

There are two types of microphone arrays, circular and linear, used commonly in speaker tracking. In our application, circular microphone arrays are chosen as their characteristics are almost invariant with direction in horizontal plane and they cover 360° range while linear arrays cover only 180° [44].

In this study, we use the sam-spare-mean (SSM) method [69] to estimate the DOA information which is incorporated to improve the tracking performance of the visual tracker. The SSM method is a two-step method. The first step consists of a sector based combined detection and localization. The space around a circular microphone array of eight microphones $\{\mathbb{M}_1,..., \mathbb{M}_8\}$ is divided into 18 sectors $\{\mathbb{S}_1,..., \mathbb{S}_{18}\}$ in Figure 3.1a and for each sector an "activeness" measure is evaluated at each time frame [68].

Then, as shown in Figure 3.1b this measure of activeness is compared to a threshold in order to give a binary decision regarding the presence of an active source in that sector. The second step is a point based search conducted in each of the sectors labelled as having at least one active source in Figure 3.1c. The parametric approach [68] is then used for localization, and the location parameters are optimized with respect to a cost function such as SRP-PHAT [33].

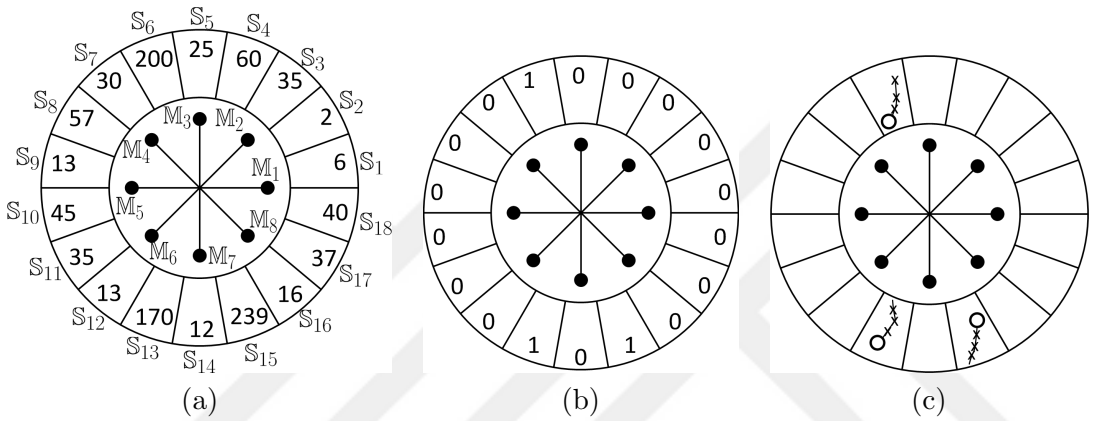


Figure 3.1: Sector based activeness measure is depicted in (a), sector based detection in (b), and points based localization in (c).

Lathoud described DOA estimation process in his PhD study [67] and reported that the average localization precision in degrees on 8 recordings is 91%. Also, he conducted another experiment to see the average false reject ratio for false alarm rate in detection of active speakers. This ratio was reported as 0.3005, 0.4199, 0.0322, 0.0672, 0.0322, the lower is better, respectively for sequences: seq01, seq37, loud01, loud02 and loud03. The high ratios in seq01 and seq37 were caused by many short silences between words and syllables. Lathoud stated that elevation was also estimated during the estimation of DOA, but it was not precise and the dependency of above errors on elevation was not discussed.

More details on the computation of the DOA angle can be found in [67].

The DOA estimates given by the SSM method can be noisy for reverberant audio measurements in some sequences. To mitigate the noise effect, we apply a third order

AR model [49] to improve the estimate of the azimuth:

$$\theta_k = \sum_{i=1}^{3} \varphi_i \theta_{k-i} + \varepsilon_k, \tag{3.1}$$

where θ_k is the DOA (azimuth) angle (in degrees) of the speaker estimated from the audio frame that is synchronized with image frame k, φ_i are the parameters of the model and ε_k is white noise. After applying a third order AR model, the estimate of the azimuth is improved while the small gaps caused by silence of the speaker can be filled by performing an interpolation method. Note that, to estimate the DOA angles, it is not necessary for the microphone array to appear in the field of view of the cameras, as the DOA is estimated from the acoustic recordings acquired by the microphone arrays which have a listening range of 360 degrees no matter whether cameras are present in the room.

Here, we assume that the calibration information of the microphone array, such as its position in $3D$ space, is available. If the calibration information is not available, the positions of the microphone arrays could be estimated via microphone self-calibration [30] or combined microphone and camera calibration [71], which is however beyond the scope of this study.

The robustness and resolution of the DOA data are affected by background noise, reverberations and positioning which include the number of simultaneous sources present with their relative positioning, and the array position with respect to the speakers [105]. The type of microphones can also affect the robustness and resolution of the DOA data as circular microphone arrays have wider range to resolve sound location compared to linear microphone arrays [110], and the fluctuation in circular array data is extremely small and the resolution of azimuth is almost constant [44] which increase the quality of DOA measurements.

3.3 Particle Filter Based Audio Constrained Visual Tracking Algorithm

In this section, we present a new method to enhance the visual tracker described in Section 2.3.1 by introducing audio information.

Although the V-PF algorithm works well in regular conditions, it fails in challenging situations like occlusion. This case is depicted in Figure 3.2 and Figure 3.3 using sequences recorded by the calibrated cameras in $AV16.3$ dataset described in Section 2.4. Figure 3.2 shows an occlusion case where the speaker re-appears in the scene after going out for a while, and another occlusion case is shown in Figure 3.3 where one speaker occludes another. The visual tracker has no visual cues during the occlusion which causes losing the speaker. Even when the speaker becomes visible again after the occlusion, the tracker is unable to detect the speaker as it is depicted in the first row of Figure 3.2. In the second row of Figure 3.2, the particles of the tracker, shown as red spots, are propagated to detect the face of the speaker. Once the tracker loses the speaker, the particles focus on objects similar to the speaker, causing divergence from the speaker.

To address this problem, several methods could be used, such as occlusions map [66] and BraMBLe tracker [56]. Here we present an alternative method by introducing audio information.

The idea behind our approach is to relocate the distributed particles around the DOA line and then re-calculate the weights of the relocated particles according to their distance to the DOA line [61]. The DOA line can be drawn as follows. First, the 3-D position of the speaker's head (A, B_k, C) is determined based on the estimated DOA angle and the following assumptions: (1) A is the distance from the centre of the microphone array to the wall in metres (which is 1.75 metres in our experiments), (2) C is the estimated height of the speaker, typically chosen as 1.80 metres in our experiments. Then B_k is calculated using the standard trigonometric identity as

$$B_k = \tan\left(\theta_k \times \frac{\pi}{180}\right) \cdot A. \tag{3.2}$$

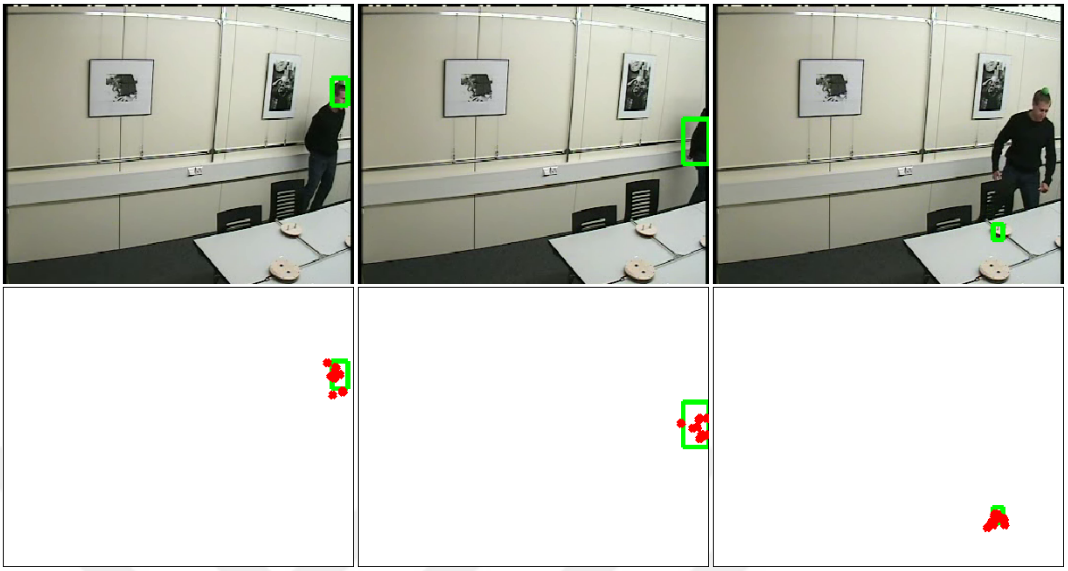


Figure 3.2: Failure in Visual PF after occlusion. The first row shows Sequence 11 camera #1 of the $AV16.3$ dataset where a single speaker disappears for a while and re-enters the scene. The second row shows the propagation of the particles to detect the speaker.

Figure 3.3: Sequence 24 camera #1 of the $AV16.3$ dataset shows multiple speakers occluding each other and the visual tracker failing after occlusion.

The $3D$ coordinate (A, B_k, C) is then projected to the image frame to obtain the $2D$ coordinate (a_k, b_k) using the calibration matrix, formed from the calibration information of the microphone arrays and cameras available in the dataset, e.g. the 3-D coordinates of the centre of the two microphone arrays $(0, 0, 0)$, and the three cameras positions, $(-1.56, 2.02, 1.40)$, $(1.52, -2.25, 1.13)$ and $(-0.25, -3.03, 1.26)$ (unit in meters), for cameras #1, #2, and #3, respectively. The DOA line is drawn from (a_k, b_k) to the $2D$ coordinate of the centre of the microphone array which is estimated only once using the same calibration matrix at the initialization step, since all the cameras are stationary and the positions of the microphone arrays are always constant for all

the camera views.

The V-PF approach described in Section 2.3.1 is enhanced by the DOA information and the proposed AV-PF algorithm is depicted in Figure 3.4. When the particles are

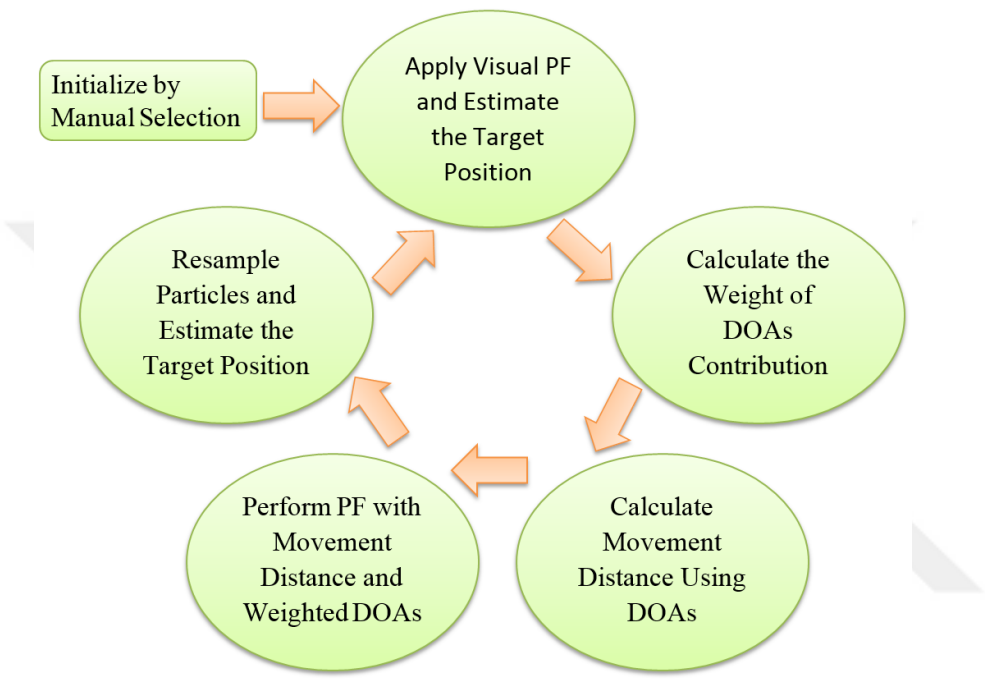
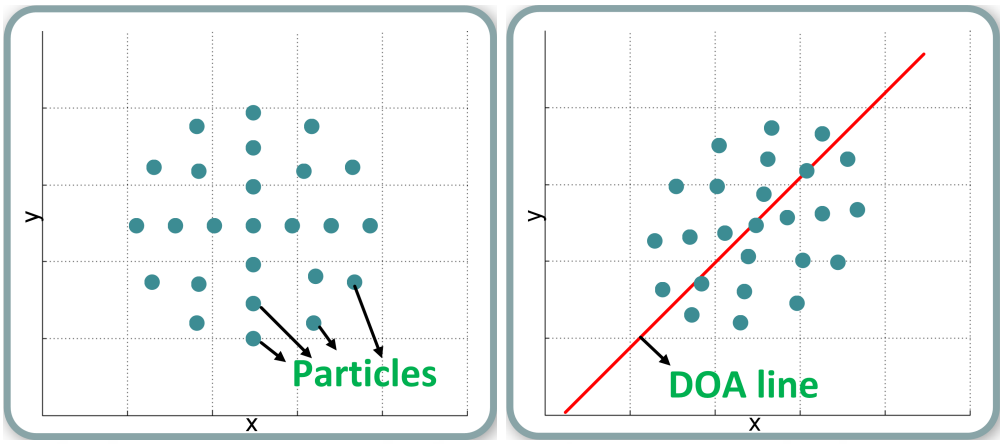


Figure 3.4: Schematic diagram of the proposed AV particle filter.

propagated, we want to concentrate on particles located around the DOA line. Concentrating around the DOA line is likely to increase the possibility of speaker detection by the particles since the DOA indicates the approximate direction of the sound emanating from the speaker. If the location of particles is assumed to be initially distributed in a circular area, then after relocation, it is expected to be elliptical in Figure 3.5b instead of being exactly on the DOA line in order to avoid deviation in the detection in the case of noisy DOAs measurements.

To obtain an elliptical distribution, the moving distance of the particles should be proportional to their initial distances to the DOA line which allows the farthest particle to move more than the closest particle thus maintaining the relative distance to the DOA line. To this end, perpendicular Euclidean distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & \dots & d_k^{(N)} \end{bmatrix}^T$ of the particles to the DOA line are first calculated. These distances are then normalized



(a) (b)

Figure 3.5: Initial distribution of particles in a circular form is given in (a). The particles distribution becomes elliptic after the particles being relocated around DOA line in (b).

to obtain distance coefficients to be used to derive the movement distances $\hat{\mathbf{d}}_k$ as follows:

$$\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_1} \odot \mathbf{d}_k, \tag{3.3}$$

where $\hat{\mathbf{d}}_k = \begin{bmatrix} \hat{d}_k^{(1)} & \dots & \hat{d}_k^{(N)} \end{bmatrix}^T$ and $\odot$ is the element-wise product and $\|.\|_1$ is the ℓ_1 norm. Then $\hat{\mathbf{d}}_k$ is used to guide how much the particles should be moved towards the DOA line. This information is then used to relocate the particle distribution during the propagation step in Equation (3.4).

The noise within the audio measurements can affect the reliability and accuracy of the DOAs. To deal with these effects, the impact of audio to the calculation of particle propagation and importance weighting is controlled by γ_k, which is calculated as the Bhattacharyya distance to measure the similarity between q_u, i.e. the image patch centred on the estimated position, and the reference image patch r_u, by substituting q_u for $q_u^{(n)}$ in Equation (2.5). The dynamic model given in Equation (2.3) is then revised to:

$$\hat{\mathbf{x}}_k^{(n)} = \mathbf{x}_k^{(n)} \oplus \hat{d}_k^{(n)} \mathbf{h}_k \gamma_k, \tag{3.4}$$

where $\oplus$ is the element-wise addition and $\mathbf{h}_k = \begin{bmatrix} cos(\theta_k) & 0 & sin(\theta_k) & 0 & 0 \end{bmatrix}^T$. The movement distance of each particle $\hat{d}_k^{(n)}$ is weighted by γ_k and this is multiplied by $\mathbf{h}_k$

to update only position (x_1, x_2) of the particle state vector $\begin{bmatrix} x_1 & \dot{x}_1 & x_2 & \dot{x}_2 & s \end{bmatrix}^T$ in order to provide the perpendicular movement to the DOA line. Since the positions of the particles are changed, the importance weights are also revised by multiplying them with the inverse of the distance coefficients calculated in the previous step to make sure that the particles that are close to the DOA line in terms of the Euclidean distance still have high importance weights:

$$\hat{w}_k^{(n)} = (e^{-\lambda (D^{(n)})^2}) \frac{\|\mathbf{d}_k\|_1}{d_k^{(n)}}. \tag{3.5}$$

The weights are then normalized to ensure that $\sum_{n=1}^{N} \hat{w}_k^{(n)} = 1$. The fourth and fifth steps of the PF algorithm are performed in the same way as in Algorithm 2.1 in Section 2.3.1. Position estimation follows the weighting step and it is calculated using Equation (2.6) and denoted as $\tilde{\mathbf{x}}_k^{av}$. Before the resampling step, to prevent the tracker to be deceived by noise in audio, γ_k is calculated again with $\tilde{\mathbf{x}}_k^{av}$ and denoted as γ_k^{av}. If γ_k^{av} is smaller than γ_k, the AV tracker results are used in the next step and iteration. Otherwise, audio is assumed to be noisy and the visual-only tracker results are used in the next step and iteration. Then the resampling step is performed to generate the new particles $\mathbf{x}_k^{(n)}$ from the set $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N}$. The pseudo code of the proposed AV-PF algorithm is depicted in Algorithm 3.1.

To further clarify our approach, the steps of the algorithm are given in Figure 3.6. We define the current and next positions of the target speaker in Figure 3.6a and 3.6b. Particles with weights from previous iteration are shown in Figure 3.6c. Resampled and normalized particles are drawn in Figure 3.6d. The particles are propagated in Figure 3.6e to detect the target. After that, the DOA line that indicates where the target is likely to be located is drawn in Figure 3.6f. The particles are moved towards the DOA line in Figure 3.6g. Lastly weights of the particles are estimated in Figure 3.6h.

With our proposed modifications in Equation (3.4) and Equation (3.5), the tracking algorithm can preserve the position of the face even if the visual tracker is lost, due to the use of the DOA as depicted in Figure 3.7. As opposed to the visual tracker in Figure

Initialize: N, σ^2, U, T, $\mathbf{F}$, λ, r_u, $\mathbf{x}_0^{(n)}$, $w_0^{(n)}$, k
while $k < K$ **do**
  Propagate particles: $\mathbf{x}_k^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)}$
  Calculate $D^{(n)}$ using Equation (2.5), for $n = 1...N$
  Calculate weights: $w_k^{(n)} = e^{-\lambda(D^{(n)})^2}$, for $n = 1...N$
  Estimate the target position $\tilde{\mathbf{x}}_k$ using Equation (2.6)
  Calculate γ_k using Equation (2.5)
  Get corresponding DOA angle θ_k
  Calculate distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & \dots & d_k^{(N)} \end{bmatrix}^T$
  Find movement distances: $\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k \odot \mathbf{d}_k}{\|\mathbf{d}_k\|_1}$
  Re-propagate particles: $\hat{\mathbf{x}}_k^{(n)} = \mathbf{x}_k^{(n)} \oplus \hat{d}_k^{(n)} \mathbf{h}_k \gamma_k$
  Re-weighting: $\hat{w}_k^{(n)} = (e^{-\lambda(D^{(n)})^2}) \frac{\|\mathbf{d}_k\|_1}{d_k^{(n)}}$
  Re-estimate target position $\tilde{\mathbf{x}}_k^{av}$ using Equation (2.6)
  Calculate γ_k^{av} using Equation (2.5)
  if $\gamma_k^{av} < \gamma_k$ **then**
    $\mathbf{x}_k^{(n)} = \hat{\mathbf{x}}_k^{(n)}$, $w_k^{(n)} = \hat{w}_k^{(n)}$, $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_k^{av}$
  end
  Resampling: Generate $\mathbf{x}_k^{(n)}$ from the set $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^N$
  $k = k + 1$
end

Algorithm 3.1: Proposed AV-PF algorithm.

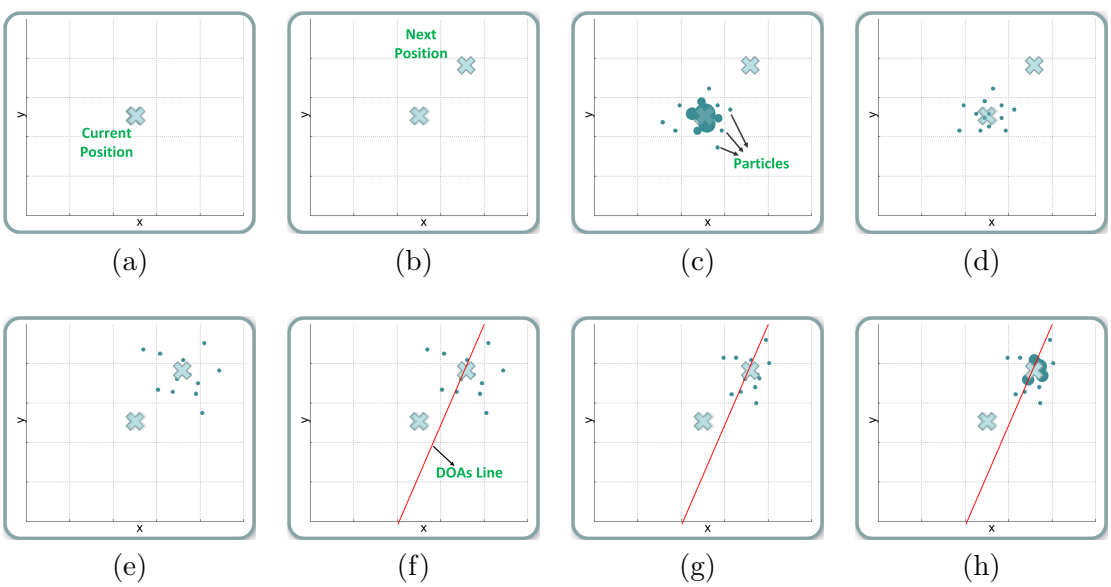


Figure 3.6: The steps of the AV-PF algorithm. Current and next positions of the target speaker are denoted in (a) and (b). Particles with weights from previous iteration are shown in (c). Resampled and normalized particles are drawn in (d). After propagation of the particles in (e), the DOAs line is drawn as shown in (f). The particles are moved towards the DOAs line in (g) and weights of the particles are estimated in (h).

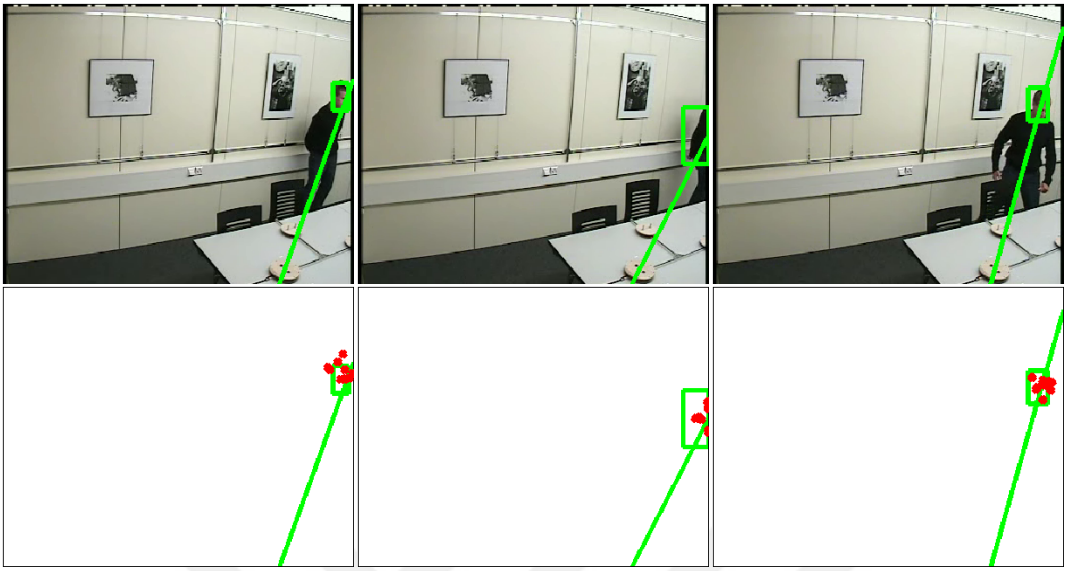


Figure 3.7: The first row shows Sequence 11 camera #1 of the $AV16.3$ dataset where the single speaker disappears for a while and re-enters the scene. After occlusion, the AV tracker continues tracking. The second row shows the distribution of the particles which are relocated by the DOA line.

3.2, the AV tracker continues tracking after the speaker comes back to the camera view in the first row of Figure 3.7. The second row shows how the particles are distributed around the DOA line. Concentrating particles around the DOA line increases the efficiency of the particles in terms of speaker detection since all the particles converge to the potential location of the speaker. This allows us to use a smaller number of particles than required in visual-only PF.

In the $AV16.3$ dataset that we used in our experiments, the speakers are talking continuously most of the time in the video sequence which therefore provides the advantage of using DOA information to improve visual tracking. In the case of missing audio, the DOA is estimated by interpolation, based on those obtained from the previous frames where the DOAs may be available. If the gap of the missing audio is large, the accuracy of such interpolation will be limited. However, by making small changes in the proposed algorithm, such as first checking whether DOA exists or not, the audio-visual tracker can be reduced to visual-only tracker when the DOA information is missing.

3.4 Improved AV Tracking with Adaptive Particle Filter

The PF approach in Section 3.3 uses a fixed number of particles which, as discussed in Section 3.1, has limitations in practice. To address these limitations, we propose a new adaptive approach to estimate the optimal number of particles at each iteration.

Fox [42] proposed an adaptive approach called KLD-sampling where the number of particles is estimated adaptively by bounding the tracking error of the PF as discussed in Section 2.3.2. It uses the KL divergence between the empirical distribution and the true posterior distribution, known as non-parametric maximum likelihood estimate, to measure the error. One assumption in this approach is that the true posterior can be represented by a discrete piecewise constant distribution consisting of a set of multi-dimensional bins. However, there is no certain way to estimate the size of these bins, and incorrect determination may cause deviation in the estimation of N. In addition, it does not mention anything about the second fixed parameter of the PF, i.e. noise variance σ^2 whose selection affects the distribution of the particles, causing the tracker to become potentially incoherent.

In this study, we aim to design a new adaptive approach that addresses the problems in the KLD-sampling algorithm. More specifically we adapt both N and σ^2 dynamically in a simple way which is easily applicable to any implementation. The particles search a rectangular area to detect the face of the speaker before their weights are allocated. The accuracy of the speaker detection partly depends on the size of the area searched. We use this relationship and build our proposed algorithm on the area occupied by the rectangles centred on the positions of the particles [60]. The total area, S, occupied by the rectangles can be defined as:

$$S = f(N, \sigma^2, L), \tag{3.6}$$

where L is the area of each rectangle. The value of S depends on the number of particles, the area of rectangle centred around each particle, and the overlap between the rectangles. The overlap is highly related to the distance between the particles, namely σ^2 which affects the distribution of the particles. One way to formulate the

calculation of S is to analyse the relationships amongst S, N, L and σ^2 using mapping tables. These mapping tables are created by distributing N particles with the variance σ^2 and calculating the area of L pixels occupied by the particles. For each mapping table, N is varied from 5 to 100 with a step size of 5, and σ^2 is varied from 10 to 150 with a step size 10. For each point (for example, $N = 10$, $\sigma^2 = 50$), it is repeated 100 times and the average of the occupied area S is estimated. Therefore, the relationships between S, N and σ^2 are observed in one mapping table for a particular L. Then this process is repeated for ten different Ls, which are found to be adequate based on extensive experimental studies, as illustrated in Figure 3.8.

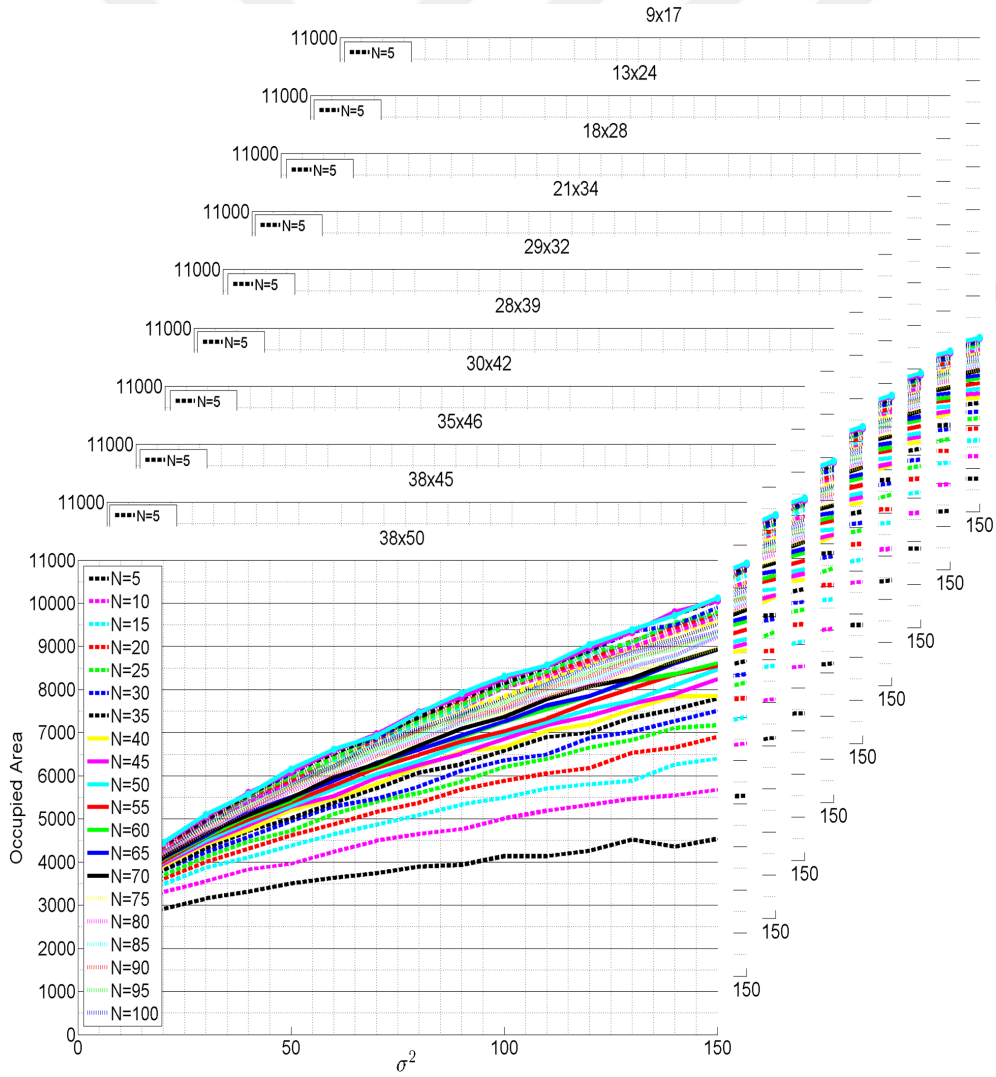


Figure 3.8: 10 mapping tables for different L are created to observe the relation between N, S and σ^2.

An illustration of the occupied area estimation is presented in Figure 3.9. Based on the particle distribution, rectangles are drawn centred on the position of the particles. Since overlaps between rectangles are inevitable, the total occupied area, S, is estimated by counting the number of pixels inside the blue line in Figure 3.9.

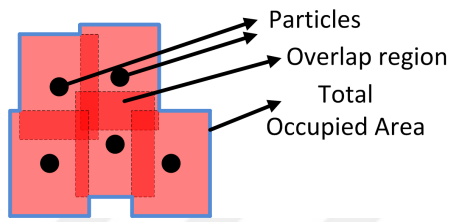


Figure 3.9: The area inside the blue line indicates the total occupied area by five particles.

For adaptive estimation of N, we need to describe 20 different lines in each 10 different mapping tables using a single formula. However, the behaviour of the lines in the mapping tables is nonlinear and this makes the problem intractable. As a solution, a curve fitting process is applied to linearise the nonlinear relation as shown in Figure 3.8. Based on the goodness-of-fit test results, a polynomial model is chosen among the several candidate curve fitting methods. A pth order polynomial model is represented by $p+1$ coefficients. In our mapping tables, the occupied area, S, depends on three variables: σ^2, N and L. Therefore, the number of polynomial coefficients grows with the power of three. Clearly, there is a trade-off between the order of the model and the goodness-of-fit as measured in terms of the sum of squares due to error (SSE). A higher order leads to a lower SSE, but it requires a higher number of polynomial coefficients. As a trade-off, the order of the polynomial model is set to 2. Let us denote $\boldsymbol{\ell} = \begin{bmatrix} L^2 & L & 1 \end{bmatrix}^T$, $\mathbf{n} = \begin{bmatrix} N^2 & N & 1 \end{bmatrix}^T$ and $\mathbf{m} = \begin{bmatrix} (\sigma^2)^2 & \sigma^2 & 1 \end{bmatrix}^T$. These three vectors form a tensor $\mathcal{V} = \boldsymbol{\ell} \otimes \mathbf{n} \otimes \mathbf{m}$ where $\otimes$ is the outer product. Then, the total area, S, can be expressed as:

$$S = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{m=1}^{3} c_{i,j,m} \cdot \nu_{i,j,m}, \tag{3.7}$$

where $\nu_{i,j,m}$ is the element of the tensor $\mathcal{V}$ and $c_{i,j,m}$ is the element of tensor $\mathcal{C}$ containing the coefficients determined by the second order polynomial model fitting. After

rearranging we get:

$$\begin{aligned} S = {} & [(c_{1,1,1}L^2 + c_{2,1,1}L + c_{3,1,1})(\sigma^2)^2 + (c_{1,1,2}L^2 + c_{2,1,2}L + c_{3,1,2})\sigma^2 \\ & + (c_{1,1,3}L^2 + c_{2,1,3}L + c_{3,1,3})]N^2 + [(c_{1,2,1}L^2 + c_{2,2,1}L + c_{3,2,1})(\sigma^2)^2 \\ & + (c_{1,2,2}L^2 + c_{2,2,2}L + c_{3,2,2})\sigma^2 + (c_{1,2,3}L^2 + c_{2,2,3}L + c_{3,2,3})]N \\ & + [(c_{1,3,1}L^2 + c_{2,3,1}L + c_{3,3,1})(\sigma^2)^2 + (c_{1,3,2}L^2 + c_{2,3,2}L + c_{3,3,2})\sigma^2 \\ & + (c_{1,3,3}L^2 + c_{2,3,3}L + c_{3,3,3})]. \end{aligned} \tag{3.8}$$

Equation (3.7) has 27 coefficients calculated by the curve fitting process and they are given in Table 3.1. Then Equation (3.8) can be simplified to:

Table 3.1: Curve fitting coefficients of the mapping table for the proposed AV-A-PF.

$c_{1,1,1}$	-2.14×10^{-12}	$c_{3,2,2}$	3.93×10^{-1}
$c_{2,1,1}$	6.42×10^{-9}	$c_{1,3,2}$	-2.71×10^{-6}
$c_{3,1,1}$	1.73×10^{-6}	$c_{2,3,2}$	1.41×10^{-2}
$c_{1,2,1}$	2.64×10^{-10}	$c_{3,3,2}$	55.02×10^{-1}
$c_{2,2,1}$	-9.5×10^{-7}	$c_{1,1,3}$	7.59×10^{-9}
$c_{3,2,1}$	-2.86×10^{-4}	$c_{2,1,3}$	-4.59×10^{-5}
$c_{1,3,1}$	-3.1×10^{-9}	$c_{3,1,3}$	-1.77×10^{-2}
$c_{2,3,1}$	-9.6×10^{-6}	$c_{1,2,3}$	-1.39×10^{-6}
$c_{3,3,1}$	-2.89×10^{-2}	$c_{2,2,3}$	7.86×10^{-3}
$c_{1,1,2}$	4.52×10^{-10}	$c_{3,2,3}$	29.04×10^{-1}
$c_{2,1,2}$	-1.99×10^{-6}	$c_{1,3,3}$	-7.44×10^{-5}
$c_{3,1,2}$	-2.14×10^{-3}	$c_{2,3,3}$	13.98×10^{-1}
$c_{1,2,2}$	-6.45×10^{-8}	$c_{3,3,3}$	165.18
$c_{2,2,2}$	3.17×10^{-4}		

$$0 = \Upsilon N^2 + \Psi N + \Omega, \tag{3.9}$$

where $\Upsilon = (c_{1,1,1}L^2 + c_{2,1,1}L + c_{3,1,1})(\sigma^2)^2 + (c_{1,1,2}L^2 + c_{2,1,2}L + c_{3,1,2})\sigma^2 + (c_{1,1,3}L^2 + c_{2,1,3}L + c_{3,1,3})$, $\Psi = (c_{1,2,1}L^2 + c_{2,2,1}L + c_{3,2,1})(\sigma^2)^2 + (c_{1,2,2}L^2 + c_{2,2,2}L + c_{3,2,2})\sigma^2 + (c_{1,2,3}L^2 + c_{2,2,3}L + c_{3,2,3})$ and $\Omega = (c_{1,3,1}L^2 + c_{2,3,1}L + c_{3,3,1})(\sigma^2)^2 + (c_{1,3,2}L^2 + c_{2,3,2}L + c_{3,3,2})\sigma^2 + (c_{1,3,3}L^2 + c_{2,3,3}L + c_{3,3,3}) - S$.

From Equation (3.9) N can be readily found as:

$$N = \frac{-\Psi + \sqrt{\Psi^2 - 4\Upsilon\Omega}}{2\Upsilon}. \tag{3.10}$$

Note that L is estimated in every frame after the face of the speaker is detected. In practice, N is implicitly bounded by the choice (or calculation) of σ^2, L and S which usually take a limited range of values. In Equation (3.10), S and σ^2 are unknown parameters that need to be estimated. To this end, we propose an iterative method where the values of S and σ^2 in step k are derived from the initial values confined by $\bar{\gamma}_k$ which is the difference between γ_k and γ_{k-1}. In other words, the calculation of S and σ^2 is linked to the difference of γ in successive frames. We propose to use a statistical model to establish that link. Many distribution functions could be employed. In our case, however, we have several requirements: (1) the input parameter should change between 0 to 1 (to match with the range of γ value); (2) the function may be controlled by at most two parameters (for simplicity); (3) the output of the function should be in the range of 0 to 1 to point out alteration ratio. To meet these requirements, a cumulative beta distribution (CBD) function appears to be the best choice and therefore it is used to model the link between $\bar{\gamma}_k$ and S, as well as σ^2. The CBD function [39] is depicted in Figure 3.10 and given in Equation (3.11).

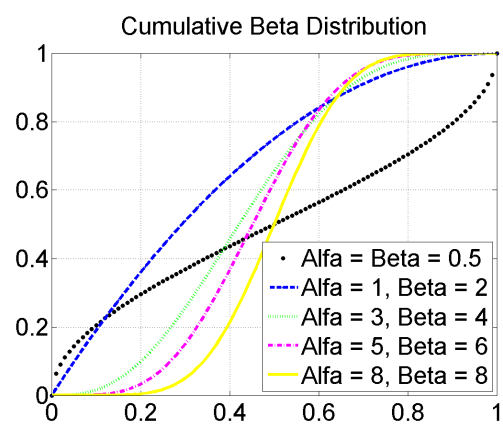


Figure 3.10: The cumulative beta distribution function for different α and β values.

$$I_{\bar{\gamma}}(\alpha,\beta) = \sum_{j=\alpha}^{\alpha+\beta-1} \binom{\alpha+\beta-1}{j} \bar{\gamma}^j (1-\bar{\gamma})^{(\alpha+\beta-1-j)}. \tag{3.11}$$

It needs two control parameters (α and β) and both input and output values change

between 0 to 1. Then, S and σ^2 at time k are defined as:

$$\begin{aligned} S_k &= S_0 \times (\rho_S + sign(\bar{\gamma}_k) \times I_{|\bar{\gamma}_k|}(\alpha_S, \beta_S)), \\ \sigma_k^2 &= \sigma_0^2 \times (\rho_{\sigma^2} + sign(\bar{\gamma}_k) \times I_{|\bar{\gamma}_k|}(\alpha_{\sigma^2}, \beta_{\sigma^2})), \end{aligned} \tag{3.12}$$

where α_S and β_S are the control parameters of the CBD for modelling S_k, and α_{σ^2} and β_{σ^2} are the control parameters of the CBD for modelling σ_k^2. S_0 and σ_0^2 are the initial values of S and σ^2. Absolute value of $\bar{\gamma}_k$ is used in the CBD function, because input values of CBD range between 0 to 1 and $\bar{\gamma}_k$ may be positive or negative depending on the change of γ in successive frames. The output of CBD is multiplied with the sign of $\bar{\gamma}_k$ to make the change of S_k and σ^2 dependent on the change in $\bar{\gamma}_k$.

The proposed AV-A-PF algorithm is an improved version of our proposed AV-PF algorithm explained in Section 3.3. At every iteration, after the comparison of γ_k^{av} with γ_k, S_k and σ_k^2 values are updated using Equation (3.12) in order to find the optimal N_k by Equation (3.10). The last step of the PF algorithm is resampling and since the N_k value has just been changed, this step is also modified for the new N_k. If N_k is decreased, the particles with the smallest weights are removed. The particles with largest weights are duplicated if N_k is increased before the resampling step is performed. The pseudo code of the proposed AV-A-PF algorithm is depicted in Algorithm 3.2.

3.5 Experimental Evaluations

In this section, the proposed and baseline algorithms are evaluated on the $AV16.3$ dataset [70] and the results are presented in plots and tables.

The speakers wear coloured balls in particular sequences which are only used for annotation, and not for tracking in our system. In the experiments with the AV-PF algorithm, the number of particles, N, is selected to be 10. The covariance matrix $\mathbf{Q}$ is a diagonal matrix with $\sigma^2 = 50$, and this is used as the variance for both the position and velocity. For the AV-A-PF algorithm, N and σ^2 are estimated dynamically. T is the period between frames and equals 0.04 seconds and λ in Equation (2.4) is chosen as 150. The number of bins used for Hue-Saturation histogram is 8. The scale factors

Initialize: N_0, σ_0^2, S_0, U, T, $\mathbf{F}$, λ, r_u, $\mathbf{x}_0^{(n)}$, $w_0^{(n)}$, k
while $k < K$ **do**
  // AV Particle Filter - Section 3.3.
  Calculate $\mathbf{x}_k^{(n)}$, $w_k^{(n)}$, $\tilde{\mathbf{x}}_k$ and γ_k using Equations (2.3), (2.4), (2.6) and (2.5), respectively.
  Find movement distances by Equation (3.3)
  Calculate $\hat{\mathbf{x}}_k^{(n)}$ and $\hat{w}_k^{(n)}$ using Equations (3.4) and (3.5), respectively.
  Re-estimate target position: $\tilde{\mathbf{x}}_k^{av} = \sum_{n=1}^{N_k} \hat{w}_k^{(n)} \hat{\mathbf{x}}_k^{(n)}$
  // Adaptive approach modifications - Section 3.4
  Calculate γ_k^{av} using Equation (2.5)
  if $\gamma_k^{av} < \gamma_k$ **then**
    $\mathbf{x}_k^{(n)} = \hat{\mathbf{x}}_k^{(n)}$, $w_k^{(n)} = \hat{w}_k^{(n)}$, $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_k^{av}$, $\gamma_k = \gamma_k^{av}$
  end
  Calculate $\bar{\gamma}_k$ value: $\bar{\gamma}_k = \gamma_k - \gamma_{k-1}$
  Calculate new S value: $S_k = S_0 * (\rho_S + sign(\bar{\gamma}_k) * I_{|\bar{\gamma}_k|}(\alpha_S, \beta_S))$
  Calculate new σ^2 value: $\sigma_k^2 = \sigma_0^2 * (\rho_{\sigma^2} + sign(\bar{\gamma}_k) * I_{|\bar{\gamma}_k|}(\alpha_{\sigma^2}, \beta_{\sigma^2}))$
  Estimate optimal N_k using Equation (3.10)
  Resampling: Generate $\mathbf{x}_k^{(n)}$ from the set $\{\mathbf{x}_k^{(n)}, w_k^{(n)}\}_{n=1}^{N_k}$
  $k = k + 1$
end

Algorithm 3.2: Proposed AV-A-PF algorithm.

ρ_S and ρ_{σ^2} are set to 1. Both α_S and β_S values are chosen as 8, and α_{σ^2} and β_{σ^2} are chosen 0.5 for CBD functions. These α and β values are intuitively chosen based on expected response of the CBD function with respect to the error change (see Figure 3.10). The value of S_0 is taken as 2000 and σ_0^2 is 50 in the simulations. These initial values are found to be appropriate based on cross-validation. In our work, we have used annotated DOAs as *a priori* to avoid mis-correspondence of person-ID after occlusion. Such information may not be available in a practical tracking system, and the person-IDs would have to be modelled and adapted during tracking using methods such as in [6] and [66].

Ten different sequences (3 single speaker, 5 two speakers and 2 three speakers) with three different camera angles from $AV16.3$ corpus have been used to perform the experiments. These sequences are briefly described in Table 3.2 [70].

We have evaluated our proposed algorithms and the baseline algorithms using both MOT and TBM metrics, and the overall average results are given in Tables 3.4 and 3.6. The results in terms of MAE are also given in Tables 3.3 and 3.5 for all the tested

Table 3.2: Description of sequences used for experiments

seq08-1p-0100	A single speaker, mostly moving while speaking and facing the microphone arrays. The speaker is walking back and forth once.
seq11-1p-0100	One moving speaker is doing random motions and talking most of the time.
seq12-1p-0100	One speaker, mostly moving while speaking, facing the microphone arrays and doing random motions.
seq18-2p-0101	Two speakers, both speaking continuously, facing the microphone arrays, slowly getting closer as much as possible to each other then slowly parting twice, sitting and standing.
seq19-2p-0101	Two standing speakers, facing the microphone arrays, slowly getting closer to each other, then further from each other. The speakers are talking most of the time.
seq24-2p-0111	Two moving speakers, both speaking continuously, facing the microphone arrays, walking back and forth, crossing the field of view twice and occluding each other twice.
seq25-2p-0111	Two moving speakers, walking back and forth twice, the second time facing the microphone arrays, each speaker starting from opposite side and occluding the other speaker once. The two speakers are talking most of the time.
seq30-2p-1101	Two moving speakers, walking back and forth once, one behind the other at constant distance, both speaking continuously.
seq40-3p-0111	Three speakers, two seated and one standing, all speaking continuously, facing the microphone arrays, the standing speaker walks back and forth once.
seq45-3p-1111	Three moving speakers, occluding each other many times, all speaking continuously.

algorithms in separate columns.

First, comparative results between V-PF and our proposed AV-PF are given and discussed. Then, the performance of our adaptive algorithm AV-A-PF is compared with our AV-PF algorithm and the baseline KLD-sampling algorithm.

3.5.1 Visual PF vs. Audio-Visual PF

The V-PF and AV-PF algorithms are run in thirty experiments (10 sequences with 3 different camera angles) in order to compare their performance. One of the single speaker experiments is Sequence 11 camera #3 in which the speaker is making random motions as illustrated in Figure 3.11.

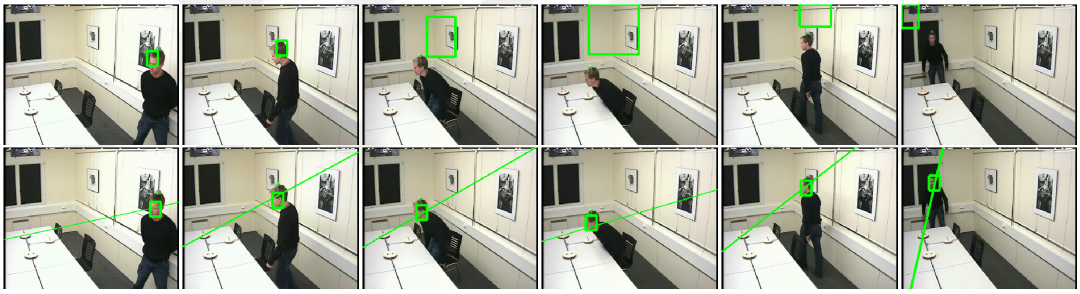


Figure 3.11: Frames are taken from Sequence 11 camera #3. The first row shows the V-PF and the second row shows our proposed AV-PF tracking.

Here, the speaker moves around the table and makes rapid and sudden movements. In the first row, the performance of the V-PF algorithm is shown. At the beginning, the tracker follows the speaker with small errors, but the error increases with sudden movements of the speaker and eventually the tracker fails. In the second row, the results of the proposed AV-PF algorithm are given and here the tracker successfully follows the speaker with the assistance of the audio line. The plot in Figure 3.12a shows the tracking error for this sequence.

Here, the error at frame k is given as the average of the errors from frame 1 to k. This representation is chosen instead of plotting error on corresponding frame k, which would give an oscillating graph since errors may change abruptly in subsequent frames. On the other hand, plotting average error at each frame k gives a smooth graph which

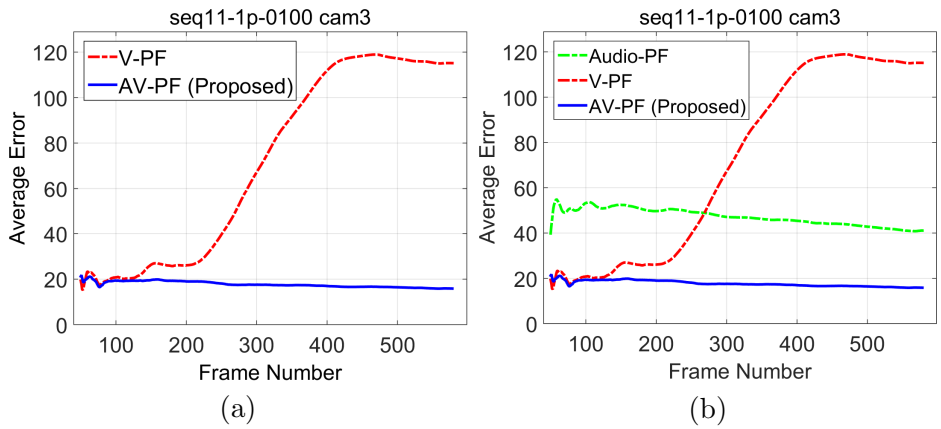


Figure 3.12: In (a) and (b) the performance of algorithms is given for Sequence 11 camera #3.

can be interpreted easily and the overall performance of each tracker can be compared clearly.

To see the contribution of the audio to the proposed AV-PF tracker, the tracker is also run on Sequence 11 camera #3 with only audio information. In the Audio-PF, the particles are weighted based on their Euclidean distances to the position estimated by the DOA information, as explained in Section 3.3, under the assumption that noise on the audio likelihood is Gaussian. Moreover, conversion of the DOA information to the 2D image plane is performed as described in Section 3.3. Firstly, $3D$ coordinates are obtained using the DOA information under the assumption that the height of the speaker and distance of the speaker to the microphone array are known, and then these $3D$ coordinates are projected to the image plane using camera calibration information. The results of the Audio-PF are given together with the V-PF and AV-PF in Figure 3.12b. In the first half of the sequence, the V-PF shows better performance than the Audio-PF. However, the error of the V-PF is increasing with the time and after some point it loses the speaker. On the other hand, the Audio-PF never loses the speaker and it improves the tracking error with the time. The performance of the Audio-PF mainly depends on how accurately the DOA is estimated. As is explained in Section 3.2, the DOA information is extracted using the SSM method [69] and the audio tracker results are given for single and multi-speaker cases in Figure A.1. Even though estimated DOA information has noise and deviation from the ground truth, it is found to be adequate

after post-processing of the DOA data as explained in Section 3.2.

Integration of the audio and visual modalities in the proposed AV-PF outperforms both the V-PF and Audio-PF. In the V-PF, the distribution of the particles may become too diffuse so that the particles spread across the image plane rather than clustering around the speaker. This is prevented in the proposed AV-PF as the particles are concentrated around the DOA line. Also, audio information gives additional information about the state of the speaker in the case of occlusions or when the speaker goes out of the scene for a while. This shows how audio and visual modalities can be complementary to each other in order to highlight their respective strengths and eliminate their weaknesses in tracking.

The experiment for Sequence 11 camera #1 is given in Figure 3.2 and Figure 3.7 shown in Section 3.3 where the speaker comes back after disappearing for a while.

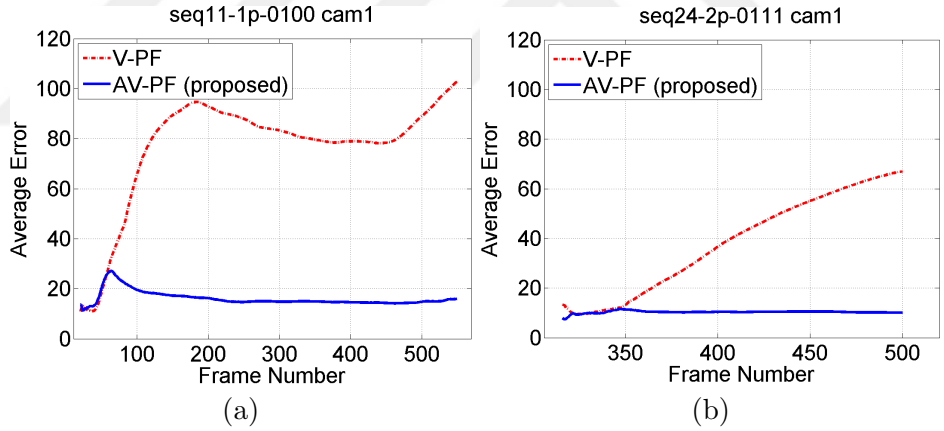


Figure 3.13: In (a) and (b) the performance of algorithms is given for Sequence 11 camera #1 and Sequence 24 camera #1, respectively.

In Figure 3.2, the results of the V-PF approach are given in the first row and as seen from the frames, when the speaker re-appears, the tracker fails to track the face. As opposed to the V-PF, tracking resumes with the reappearance of the speaker in our proposed AV-PF algorithm as shown in Figure 3.7. The plot in Figure 3.13a shows the tracking error for this sequence. After occlusion, the V-PF algorithm lost tracking, but our proposed AV-PF algorithm continued tracking.

Figure 3.14 shows the result for multi-speaker occlusion case, Sequence 24 camera #1 where one speaker is occluded by the other. After the occlusion, our proposed AV-PF

algorithm (in the second row) resumes tracking. The average error of the two speakers for this sequence is shown in Figure 3.13b, and after the 350th frame the V-PF fails, but our proposed AV-PF algorithm continues tracking with small errors. These two

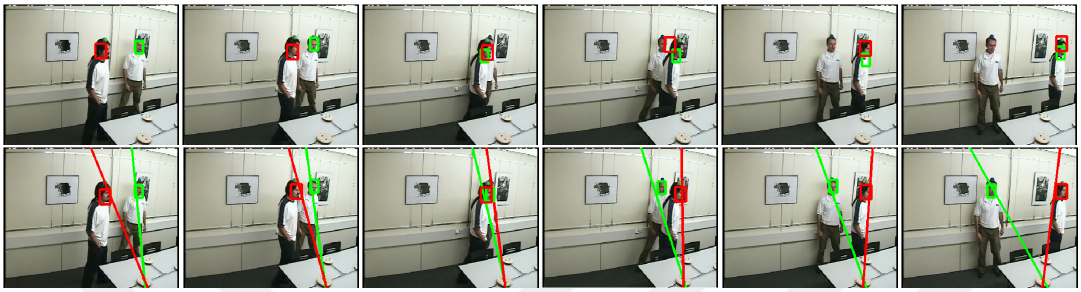


Figure 3.14: Sequence 24 camera #1: Multiple speakers with occlusions. V-PF in the first row cannot track the speaker after occlusion. On the contrary, the proposed AV-PF algorithm keeps tracking.

algorithms are also tested on the case of three speakers with two sequences, Sequence 40 and Sequence 45, respectively. The results for Sequence 40 camera #1 are illustrated in Figure 3.15. Even this sequence is not challenging, V-PF fails to track all three

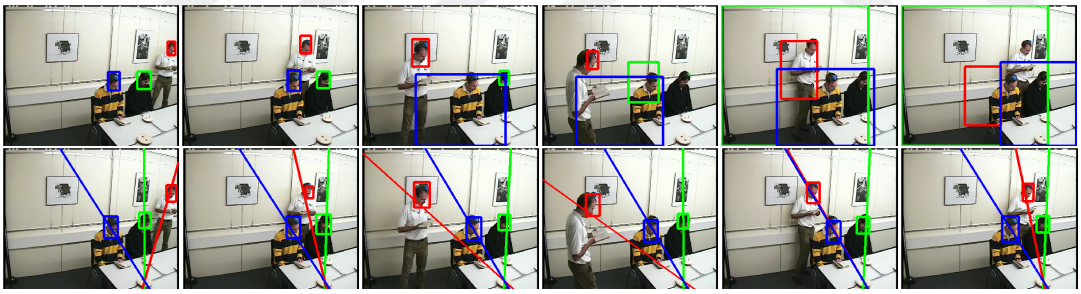


Figure 3.15: Sequence 40 camera #1: Three speakers with occlusions. V-PF performance is shown in the first row, and AV-PF shows better performance in the second row.

speakers. Sequence 45 is the most challenging sequence in this corpus where all the speakers walk and occlude each other many times as shown in Figure 3.16. The V-PF fails as expected. Unlike the V-PF, the proposed AV-PF algorithm successfully tracks the speakers both on Sequence 40 and Sequence 45. The error plots for three speaker experiments are given in Figure 3.17. Both plots show that the proposed AV-PF approach has more stable performance than the V-PF.

Instead of illustrating frames of 30 experiments, all experiments are repeated 10 times and the average results are given in Tables 3.3 and 3.4 with standard error (SE).

Table 3.3: MAE values for V-PF and proposed AV-PF averaged over 10 repetitions and standard errors of the estimated MAE values.

		V-PF		AV-PF	
		MAE	SE-MAE	MAE	SE-MAE
seq08-1p-0100	cam1	109.36	0.76	11.74	0.15
	cam2	120.92	0.81	9.05	0.21
	cam3	83.81	0.9	10.91	0.19
seq11-1p-0100	cam1	81.15	0.71	22.27	0.25
	cam2	117.58	1.34	17.05	0.35
	cam3	130.5	1.22	18.29	0.71
seq12-1p-0100	cam1	145.01	1.07	16.54	1.12
	cam2	170.86	0.77	18.79	1.08
	cam3	157.25	0.87	12.77	0.81
seq18-2p-0101	cam1	115.26	1.24	15.24	0.17
	cam2	110.14	0.62	14.33	0.57
	cam3	115.54	0.71	19.25	0.76
seq19-2p-0101	cam1	62.51	0.5	12.67	0.34
	cam2	61.95	0.44	11.41	0.55
	cam3	56.49	0.67	12.91	0.22
seq24-2p-0111	cam1	73.34	0.72	10.66	0.15
	cam2	51.35	0.35	9.57	0.16
	cam3	44.2	0.21	10.44	0.15
seq25-2p-0111	cam1	33.79	0.82	16.17	0.08
	cam2	20.51	0.36	8.77	0.08
	cam3	33.52	1.24	10.09	0.28
seq30-2p-1101	cam1	33.72	0.23	15.21	0.16
	cam2	19.23	1.12	9.48	0.08
	cam3	26.35	2.52	10.86	0.05
seq40-3p-0111	cam1	75.69	0.33	12.88	0.33
	cam2	72.6	2.38	18.99	0.95
	cam3	78.94	0.48	16.05	0.99
seq45-3p-1111	cam1	73.81	1.58	17.19	0.08
	cam2	41.13	2.21	19.59	0.67
	cam3	71.58	0.34	20.94	1.53
Average		**79.60**	**0.92**	**14.34**	**0.44**

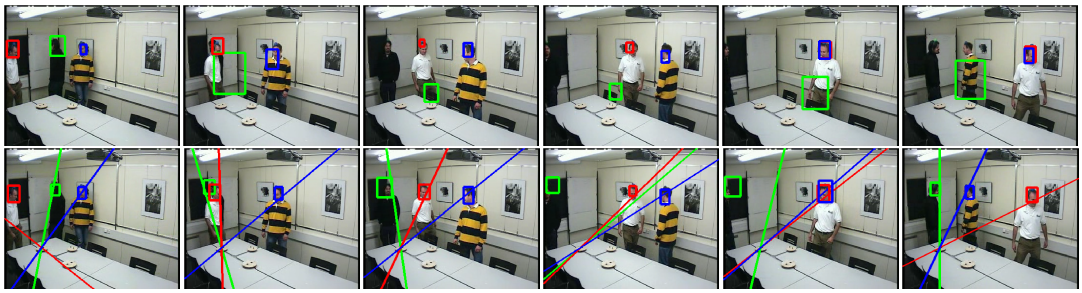
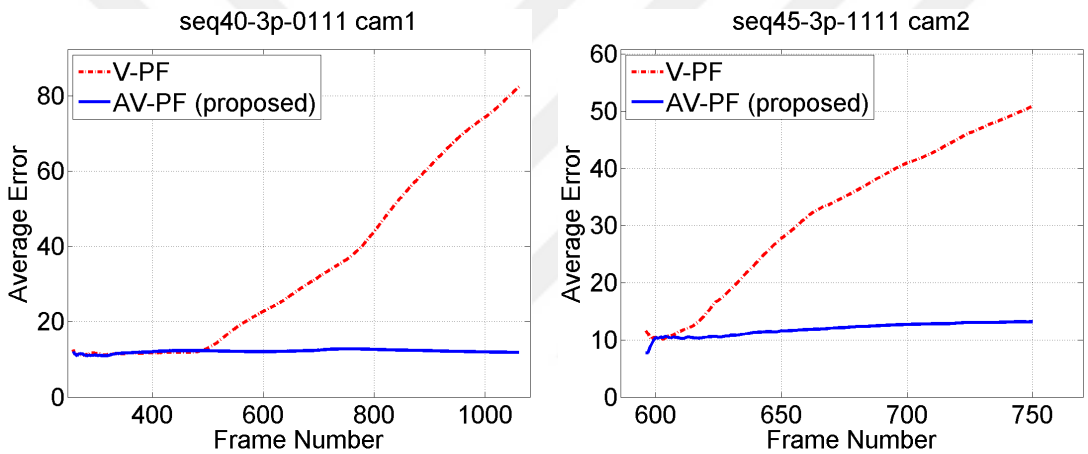


Figure 3.16: Sequence 45 camera #2: Multiple speakers with occlusions. Occlusions occur multiple times. V-PF fails in the first row, but AV-PF continues tracking in the second row.



(a) (b)

Figure 3.17: In (a) and (b) the performance of the V-PF and AV-PF algorithms is given for Sequence 40 camera #1 and Sequence 45 camera #2, respectively.

The V-PF and AV-PF algorithms are compared according to MAE in Table 3.3, and using TBM and MOT metrics in Table 3.4. Since there are no standard protocols to measure the performance of multi-target tracking methods, we have used two commonly used metrics besides the MAE metric, which are widely used for the evaluations of tracking methods. The MAE is firstly used because of its simplicity and it provides explicit output considering only the difference between the estimated position and the ground truth. On the other hand, the MOT metric can be considered an advanced solution as compared to the MAE metric since it measures the number of errors during tracking and evaluates the precision of trackers, in terms of the ability to localize the target with MOTA and MOTP quantities, respectively [9]. The mathematical formulations of the MOTA and MOTP quantities were given in Section 2.5. The

problem with the MOT metric is that it may give a negative accuracy when the number of errors is larger than the number of targets [92]. This case happens especially in visual-only tracking when the tracker fails due to occlusion or rapid movement of the speakers. Also, calculation of MOTP is performed only if the tracker keeps tracking the speakers with the distance less than the threshold value. This may be deceptive if only the MOTP is considered, i.e. without the use of the MOTA. In Table 3.4, MOTP values for V-PF and AV-PF are 14.6 and 12.7, respectively which may give the impression that they have similar performance. However, MOTA values are 20.5% and 90.5% respectively for V-PF and AV-PF which show that the proposed AV-PF outperforms the V-PF.

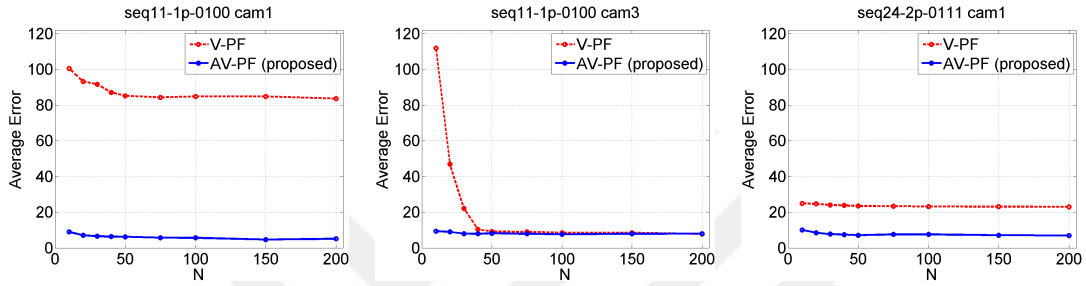
In a simple target tracking scenario, a tracker usually tracks the target only for a certain period of time as the tracker may lose the target in challenging conditions such as occlusion or noisy measurements. This is another aspect for the evaluation of the tracking methods that of how accurately ground trajectory is tracked by the tracker. In that sense, the TBM is a solution as it evaluates the performance on the basis of trajectory. The TBM classifies a trajectory as mostly tracked (MT), partially tracked (PT) and mostly lost. The MT, PT and ML are complementary information, so only MT values are presented in Table 3.4 for simplicity. However, other quantities of the TBM metrics are given in Table B.1 in Appendix B. According to the TBM results given in Table 3.4, the V-PF tracks 43.0% of ground truth trajectories while the proposed AV-PF tracks 92.4%, which demonstrates that the proposed AV-PF outperforms the V-PF based on the TBM metric.

Milan *et al.* [92] conducted a comparative analysis of performance metrics, including the MOT and TBM metrics, and concluded that there is no single objective measure for the quantitative ground-truth evaluation of multi-target tracking algorithms that incorporates all important aspects. Therefore, our tracking results are evaluated with two widely accepted metrics, MOT and TBM, besides MAE metric.

Another experiment has been performed in order to see the effects of particle numbers on the performance of the algorithms. The numbers of particles are selected as: 10, 20, 30, 40, 50, 75, 100, 150 and 200 for all the three sequences. The results for Sequence 11

Table 3.4: Experimental results with TBM, MOT and MAE metrics for V-PF and proposed AV-PF.

Method	**MT(%)**	**MOTA(%)**	**MOTP**	**MAE**
V-PF	43.0	20.5	14.6	79.6
SE	0.2	1.9	0.6	0.9
AV-PF	92.4	90.5	12.7	14.3
SE	0.0	1.5	0.3	0.4



(a) (b) (c)

Figure 3.18: Average error comparison between the V-PF and AV-PF algorithms for a variable number of particles N.

camera #1, Sequence 24 camera #1, and Sequence 11 camera #3 are shown in Figure 3.18a, 3.18b and 3.18c, respectively. In the case of occlusion, the V-PF fails even if it has high numbers of particles as seen in Sequence 11 camera #1 and Sequence 24 camera #1. Sequence 11 camera #3 features a person making a variety of rapid movements, despite the fact that no occlusion is involved. The V-PF has almost the same performance as our proposed AV-PF algorithm when it has a larger number of particles as seen in Figure 3.18b. However, when the number of particles is reduced significantly, e.g. when $N = 10$, the tracking errors increase dramatically in the V-PF while our proposed AV-PF tracking algorithm continues to show excellent performance. We have also compared our proposed AV-PF algorithm with one of the non-PF approaches, i.e. mean-shift tracking [29] which is a non-parametric method and using a single hypothesis for tracking. DOA is also fused in the same way as in our proposed approach by treating this single hypothesis as if it is the only particle of the PF. As an example, here we show the results for Sequence 12 camera #3. Using the AV mean-shift tracking, we got MAE = 33.6, MOTP = 22.6 and MOTA = 60.3%. For our proposed approach, these values are 12.8, 12.3 and 97.9% respectively. It is clear that the PF approach outperforms the mean-shift tracking algorithm.

3.5.2 Audio-Visual PF vs. Audio-Visual A-PF

The results in Figure 3.18 show that the AV tracker is better than the visual-only tracker in handling occlusions even with a small number of particles. Here we demonstrate that we can further reduce the tracking errors by using our proposed adaptive approach, i.e. the AV-A-PF algorithm, as explained in Section 3.4. The AV-A-PF algorithm is also tested on $AV16.3$ and its performance is compared with the baseline algorithm, i.e. KLD-sampling [42]. Since the adaptive approach is based on our proposed AV-PF algorithm, the KLD-sampling is also combined with our proposed AV-PF algorithm in order to make a fair comparison between these two approaches.

To see the advantage of the A-PF, we perform an experiment to compare the proposed AV-A-PF algorithm with the use of a fixed number of particles (AV-PF). Firstly, the AV-A-PF algorithm is run on Sequence 12 camera #1 and we reach an average $\gamma = 0.27$ with an average $N = 15$ in 138.17 seconds. Then, the AV-PF is run with $N = 15$ and γ goes up to 0.35 in 135.87 seconds. The 30% γ difference shows that the AV-A-PF approach is better than the fixed number AV-PF approach with a more accurate tracking result. On the other hand, adaptive estimation of the particle numbers took around extra 2 seconds computational cost. However when we increase N up to 22 to get $\gamma = 0.27$ in the fixed AV-PF, the computational cost became 149.35 seconds. Here, experiments are implemented in Intel core $i7$ 2.2 GHz processor with 8 GB memory under Windows 7 operating system. Adaptive estimation of N and variance adds slight computational cost, but it is reasonable when compared with fixed N usage to reach the same accuracy. It shows that the adaptive approach is beneficial in terms of both accuracy and the computational cost. The plot for this experiment is shown in Figure 3.19. In Figure 3.19a, N is changing with time for AV-A-PF algorithm, while it is fixed for AV-PF and Figure 3.19b shows γ for both approaches.

The KLD-sampling algorithm is also tested on the same sequence and compared with the AV-A-PF algorithm and the results are given in Figure 3.20. The KLD-sampling algorithm needs an average of 68 particles to reach almost the same value, $\gamma = 0.26$. Figure 3.20a and 3.20b show the effect of changing N and γ respectively. σ^2 is set to 50 in the KLD-sampling algorithm. Since σ^2 is adaptive in our proposed approach,

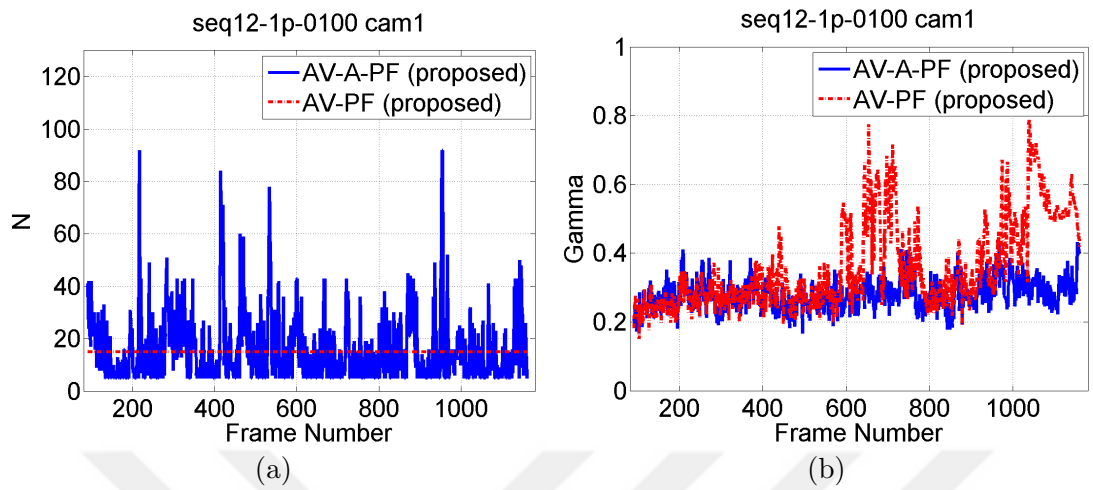


Figure 3.19: The average N for the AV-A-PF and the fixed N for AV-PF is 15 in (a). The average γ for the proposed adaptive and fixed PF is 0.27 and 0.35, respectively in (b).

the average σ^2 is found to be 67.61 as seen in Figure 3.20c. The effect of changing S is shown in Figure 3.20d which is a parameter specific to our proposed approach. In another experiment, we used three multi-speaker sequences with speakers occluding each other. The results of these experiments are shown in Figure 3.21. The AV-A-PF used an average of 27, 12 and 17 particles for Sequence 24 camera #2, Sequence 40 camera #1 and Sequence 45 camera #2, respectively. However, the KLD-sampling used 64, 56 and 63 particles for the same sequences. The difference in γ values is quite small despite the big difference in N. For Sequence 24 camera #2, it is 0.24 for both. For Sequence 40 camera #1, the average γ for the AV-A-PF and KLD-sampling is 0.24 and 0.22, respectively. For Sequence 45 camera #2, the average γ for the AV-A-PF and KLD-sampling is 0.32 and 0.30, respectively.

KLD-sampling is a popular approach in the literature, but one of the limitations of this approach is having only one adaptive parameter, N. Another limitation is that it needs a parameter, the bin size, which also affects the performance of the algorithm. Generally, KLD-sampling shows better performance in the area of robotics in which tracking is done in a vast area with a large number of N (over 1000) [41]. In our adaptive approach, we have used the σ^2 value to find the optimal value for N. The errors can be reduced by adapting σ^2 without changing N. The mapping table also

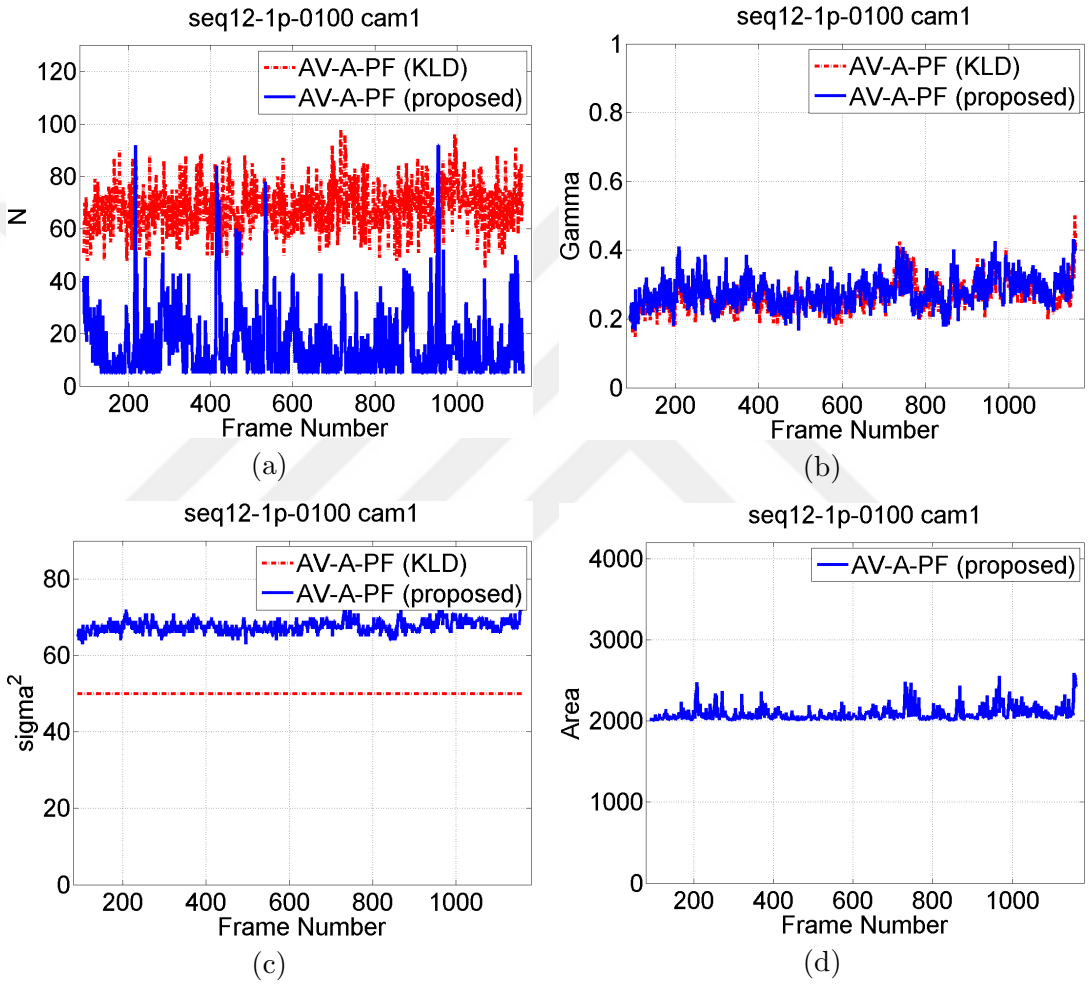


Figure 3.20: The average N for the AV-A-PF algorithm and KLD-sampling is 15 and 68, respectively in (a). The average γ for AV-A-PF algorithm and KLD-sampling is 0.27 and 0.26 respectively in (b). In (c), σ^2 is set 50 in KLD-sampling while σ^2 varies by the time in AV-A-PF algorithm, resulting in average σ^2 is 67.61. In (d), the change of S is given with respect to time frames and the average S is 2083.

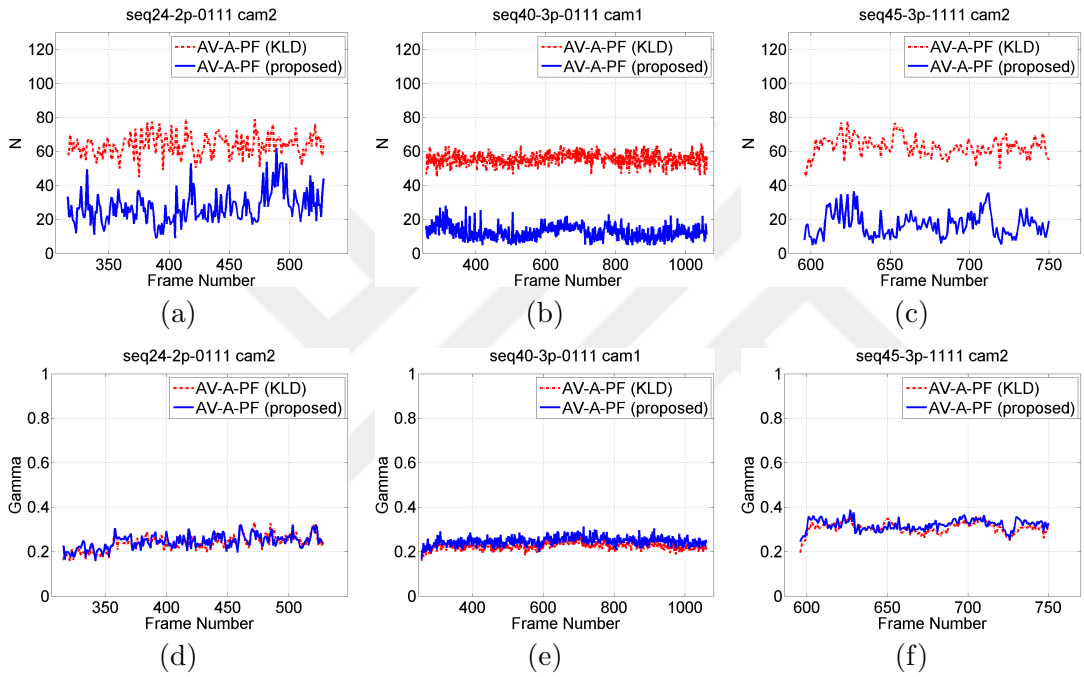


Figure 3.21: Multi-person tracking. In (a), (b) and (c) the average N is given for both sampling algorithms. For Sequence 24 camera #2, the average N for the AV-A-PF and KLD-sampling is 27 and 64, respectively. For Sequence 40 camera #1, the average N for the AV-A-PF and KLD-sampling is 12 and 56, respectively. For Sequence 45 camera #2, the average N for the AV-A-PF and KLD-sampling is 17 and 63, respectively. The average γ is shown in (c), (d) and (e) for both algorithms. For Sequence 24 camera #2, it is 0.24 for both. For Sequence 40 camera #1, the average γ for the AV-A-PF and KLD-sampling is 0.24 and 0.22, respectively. For Sequence 45 camera #2, the average γ for the AV-A-PF and KLD-sampling is 0.32 and 0.30, respectively.

simplifies the calculation of N. These make AV-A-PF algorithm simple and efficient.

We performed KLD-sampling and AV-A-PF algorithm over 30 experiments. The results for all the experiments are given in Tables 3.5 and 3.6. Here also, all experiments are repeated 10 times and the averages of these results are shown in the tables with standard errors.

Overall, our proposed AV-A-PF approach shows almost the same performance as the KLD-sampling despite using a much smaller N, as shown in Tables 3.5 and 3.6. The average of the estimated N by the KLD-sampling method for all sequences is around 53, while our proposed AV-A-PF algorithm gives an estimate of N at around 17. To examine whether the difference in N between these two methods is statistically significant, we have performed one-way ANOVA based F-test [12]. We obtained $F = 559.19$, p-value$= 1.8 \times 10^{-31}$ and the degree of freedom $(1, 58)$. Using the degree of freedom value, the critical value F_{crit} is found to be 4.01 from the F-distribution table given in [12] which is the number that the test statistic must overcome to reject the test. The p-value (or probability value) is the probability of a more extreme result than what we actually achieved when the null hypothesis is true. In other words, the p-value is the certain area under the standard normal distribution curve representing the probability of a particular sample statistic or a more extreme sample statistic occurring if the null hypothesis is true [12]. The F-value is defined as the ratio of the variance of the group means to the mean of the within group variances. The F-test has been carried out at 5% significance level. According to this test, the results are accepted as statistically significant if $F > F_{crit}$ and p-value is less than 0.05 (for a 5% significance level). From the test results, we can observe that the difference in N between the two methods is indeed statistically significant.

3.6 Chapter Summary

We have presented a new audio-visual tracking algorithm in which audio information has been used to modify particle propagation and the weights assigned to the particles. Our proposed algorithm has been tested on both single and multiple speaker sequences and showed significantly improved tracking performance over the V-PF approach for the

Table 3.5: N and MAE values for KLD-sampling and proposed AV-A-PF averaged over 10 repetitions and standard errors of the estimated N and MAE values.

		KLD				AV-A-PF			
		N	SE-N	MAE	SE-MAE	N	SE-N	MAE	SE-MAE
seq08-1p-0100	cam1	67.69	0.06	10.4	0.05	14.22	0.4	10.75	0.13
	cam2	48.58	0.1	7.27	0.02	20.19	0.22	7.33	0.09
	cam3	41.59	0.07	9.33	0.08	16.8	0.3	9.85	0.12
seq11-1p-0100	cam1	50.84	0.18	18.99	0.77	13.65	0.34	14.66	0.34
	cam2	53.55	0.1	15.73	0.1	20.8	0.31	14.01	0.12
	cam3	41.64	0.06	14.85	0.16	17.58	0.48	13.96	0.23
seq12-1p-0100	cam1	67.48	0.04	12.08	0.1	16.47	0.41	12.49	0.16
	cam2	56.48	0.05	9.66	0.07	18.15	0.19	10.81	0.24
	cam3	42.77	0.06	10.32	0.04	12.54	0.12	11.86	0.24
seq18-2p-0101	cam1	47.4	0.04	14.03	0.08	13.41	0.24	14.31	0.13
	cam2	52.67	0.03	11.3	0.08	11.96	0.3	11.66	0.13
	cam3	49.87	0.04	15.37	0.04	12.52	0.17	15.8	0.16
seq19-2p-0101	cam1	52.7	0.08	11.87	0.06	16.95	0.12	11.88	0.11
	cam2	57.65	0.12	9.49	0.06	24.52	0.23	9.62	0.11
	cam3	57.19	0.08	11.96	0.08	21.11	0.32	12.08	0.12
seq24-2p-0111	cam1	53.98	0.08	9.39	0.08	17.08	0.22	9.95	0.16
	cam2	62.6	0.05	8.58	0.04	27.8	0.33	8.85	0.04
	cam3	55.82	0.06	9.7	0.06	17.89	0.25	10.02	0.11
seq25-2p-0111	cam1	43.28	0.1	15.75	0.05	12.25	0.21	14.78	0.03
	cam2	59.52	0.07	7.39	0.06	12.16	0.2	7.7	0.05
	cam3	57.84	0.04	8.67	0.05	15.77	0.32	8.93	0.11
seq30-2p-1101	cam1	60.68	0.04	14.07	0.14	18.42	0.32	13.84	0.13
	cam2	50.22	0.07	8.57	0.02	14.04	0.19	8.85	0.06
	cam3	46.4	0.05	9.92	0.03	11.82	0.12	10.3	0.11
seq40-3p-0111	cam1	55.92	0.05	12.11	0.04	12.71	0.11	12.38	0.09
	cam2	49.19	0.06	11.58	0.15	12.86	0.29	12.04	0.14
	cam3	44.79	0.06	10.85	0.03	14.4	0.19	11.3	0.11
seq45-3p-1111	cam1	52.33	0.06	16.42	0.04	21.92	0.32	16.35	0.09
	cam2	62.59	0.03	14.24	0.63	17.75	0.32	17.22	0.71
	cam3	59.26	0.06	14.52	0.86	27.89	0.19	13.84	0.59
Average		**53.42**	**0.07**	**11.81**	**0.14**	**16.85**	**0.26**	**11.91**	**0.17**

Table 3.6: Experimental results with TBM, MOT and MAE metrics for KLD-sampling and proposed AV-A-PF.

Method	**MT(%)**	**MOTA(%)**	**MOTP**	**MAE**	**N**
KLD	96.7	94.7	11.5	11.8	53.4
SE	0.0	0.6	0.1	0.1	0.1
AV-A-PF	96.2	94.4	11.6	11.9	16.9
SE	0.0	0.5	0.1	0.2	0.3

scenarios where the speaker is either occluded by other speakers or out of the range of the camera view. We demonstrate that by using audio information we can significantly reduce the number of particles, whilst maintaining good tracking performance. This approach has the potential for handling weight degeneracy and particle impoverishment problems due to the significant reduction in the number of particles being used in tracking.

As an enhanced version of our proposed algorithm, we have presented a new adaptive PF algorithm which uses audio and visual information to adapt the number of particles and noise variance dynamically. Our proposed AV-A-PF algorithm has also been tested on both single and multiple speaker sequences and compared with a fixed particle filter and an existing A-PF algorithm. The experiments demonstrate that the proposed algorithm can effectively track moving speakers and it reduces the number of particles without increasing errors.

Despite the fact that our proposed algorithms offer advantages in speaker tracking and the estimation of the optimal number of particles, there are also some constraints and limitations associated with them that we want to point out. The first one is about the audio detection and localization algorithm which assumes that the microphone array is circular. Secondly, the audio information used in tracking is DOA, and as a result, the calibration information is required when projecting the DOA into the 2-D image plane. Third, we assume that the speaker to be tracked is active, from which the DOA information can be obtained. These assumptions or constraints may limit its generalization capability for other scenarios or datasets. However, with some modifications to the proposed algorithm, the proposed method could also be used in these cases. For example, if the audio localisation algorithm used in the proposed tracking system is replaced by a linear microphone array based localisation method together with the microphone calibration information, then the proposed system can also be applied to “CLEAR”, “AMI” or “SPEVI” datasets. If the calibration information of the microphones is not available in the dataset, the proposed system could still be used, provided that the calibration information can be derived by a reliable self-calibration algorithm such as [30] and [71].

In conclusion, we first reduced the number of particles needed for tracking by combining audio information with V-PF, and then we converted AV-PF to AV-A-PF to improve the accuracy. The limitations associated with the proposed algorithms could be interesting directions for future work.

Chapter 4

Tracking a Variable Number of Speakers

In the previous chapter, multi-speaker tracking is studied under the assumption that the number of speakers is known and invariant during the tracking. Here, we have considered a more realistic scenario where the number of speakers is unknown and varies over time. To this end, RFS and PHD filter frameworks, which are capable of dealing with time-varying number of speakers, are investigated for audio-visual multi-speaker tracking.

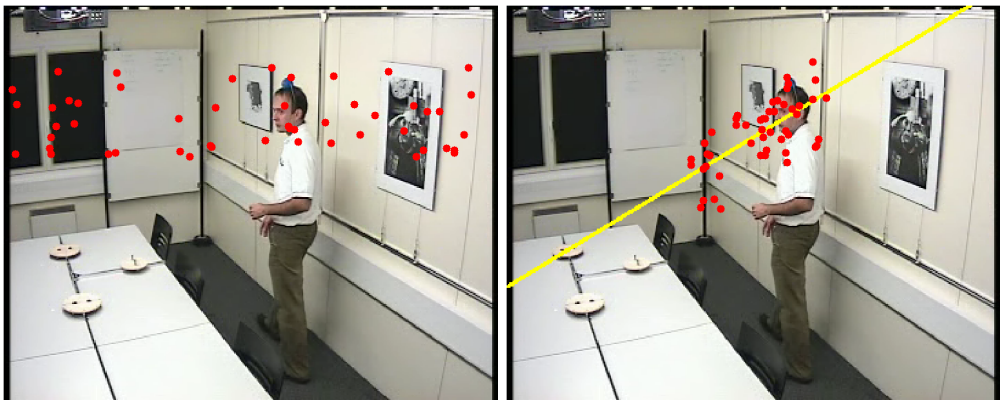
4.1 Introduction

A comprehensive approach for tracking speakers with audio and video data is to use a state-space approach based on a Bayesian framework. The PF [55] is one of the widely employed algorithms which easily approaches the Bayesian optimal estimate with a sufficiently large number of particles [3]. In the previous chapter, the PF is applied to multi-speaker AV tracking under the assumption that the number of speakers is known and constant. In a practical tracking environment, however, the speakers to be captured by the audio-visual sensors may appear or disappear in a random manner. As a result, the number of speakers that can be observed from the audio-visual measurements may vary with time.

In this chapter, we relax the above assumption and focus on the problem of tracking a variable number of speakers based on the AV data, where the number of speakers and

their states are jointly estimated in a multi-speaker environment. A Bayesian tracking framework based on the RFS formulation [85], [87], [136] is used due to its ability to deal with an unknown and variable number of speakers in audio-visual tracking. Compared with existing methods like MHT, RFS presents a more principled way to deal with the birth of new speakers, clutter, missing detections and noise [82]. Our work is based on the PF implementation of the RFS presented in [80] and [151] where speaker tracking is implemented using only audio information. In contrast to [80] and [151], the RFS approach is extended here to deal with both audio and visual measurements. RFS-PF based visual tracker is further improved with fusion of audio and visual measurements which results in reduced tracking error. We show in our experiments that, with this new extension, the proposed AV tracking system is able to track reliably a variable number of speakers in challenging scenarios such as occlusion. However, the computational cost of the RFS Bayesian filter becomes more expensive as the number of speakers grows. To address this problem, it would be appropriate to apply approximations such as the first-order moment method [80], [85]. The PHD filtering [85] approach, as a first order approximation of the RFS, is a framework that has recently emerged and has been found to be promising for multi-speaker tracking. The SMC implementation [135] is introduced to obtain practical solutions of the PHD filter. We investigate and modify the standard SMC-PHD filtering algorithm with the aim of improving its computational efficiency and estimation accuracy under challenging conditions as mentioned above. The SMC-PHD filter uses particles to model the surviving, spawned and born state of the speaker. In the standard implementation of the SMC-PHD filter [135] in visual tracking, the born particles are propagated in every frame to detect the speaker present in the view, which is computationally expensive. In Figure 4.1a, the born particles are distributed to detect the speaker on the restricted region of the frame as this region covers both sides of the scene that new speaker may enter. Here, we propose to use the DOA information obtained from audio for the propagation of the particles. To reduce the computational complexity and improve the estimation accuracy, the propagation of the born particles is decided based on the DOA information and the particles are re-located around the line drawn upon the DOA as illustrated in Figure 4.1b.

To use the PHD filter, a set of random measurements is required to estimate both the



(a) (b)

Figure 4.1: Distribution of 50 particles for the visual case in (a) and the audio-visual case in (b).

position and the number of the targets. Random measurements can be obtained from sensors like GPS in the application of SLAM (Simultaneous localization and mapping) [94] or microphone pairs in audio based speaker tracking [80]. In visual tracking, however, there are no other sensors to generate random measurements except cameras. Therefore, pre-processing of visual data is needed to find measurements. There are two common ways. The first one is to use a detector as in [83] and [153] where background/foreground (B/F) detection algorithm is run on the frame and the centres of the foreground objects are used as the measurement set. Despite being computationally expensive, this method performs well in tracking of moving objects. However, if it is desired to track particular parts like face, it does not work well as the measurements can be any moving objects rather than the face. In that case, face detection algorithms are needed which have a higher computational cost than the B/F detection algorithms. The second method is to use colour histogram templates [145] where the template is created for a single target. In this study, we extend the template for a multi-target scenario. Colour histograms of possible targets are stored in the template and, during the tracking, it is used as the measurement set. This method is computationally cheaper than the first method as the template is created only once at the beginning.

4.2 RFS Multiple Speaker Tracking

The RFS theory is explained in Section 2.3.3. This section describes our contribution to RFS theory for multi-speaker tracking with a variable number of speakers.

We use video and audio modalities which give a visual measurement set $\mathcal{Z}_k^{\mathrm{v}}$ and an audio measurement set $\mathcal{Z}_k^{\mathrm{a}}$ under the assumption that they are independent and follow a normal distribution. $\mathcal{Z}_k^{\mathrm{v}}$ contains the normal distribution of colour histogram measurements, and $\mathcal{Z}_k^{\mathrm{a}}$ contains the DOA angle recorded from microphone arrays which follows a normal distribution.

Since the joint information from both the visual and audio measurements are employed, we can address the complete measurement set at time k as

$$\mathcal{Z}_k = \mathcal{Z}_k^{\mathrm{v}} \cup \mathcal{Z}_k^{\mathrm{a}}, \tag{4.1}$$

where '$\cup$' denotes union. The number of the measurements is the cardinality of the measurement set: $|\mathcal{Z}_k| = M_k^{\mathrm{v}} + M_k^{\mathrm{a}} \triangleq M_k$. The likelihood of the visual and audio measurement sets is explained in the following sections.

4.2.1 Visual Likelihood

The observation from video is the colour histogram q_k extracted from the video frames. We have many colour histogram models of templates $\left\{r_u^{(1)}, r_u^{(2)}, ..., r_u^{(M^{\mathrm{v}})}\right\}$ in multi-speaker tracking which are used as references to compare their similarity with q_k in terms of the Bhattacharyya distance where M^{v} is the number of reference histograms in the template. Recall that in Section 2.3.3 we represented the state vectors of the speakers by a single finite set $\mathcal{X}_k = \{\mathbf{x}_{1,k}, ..., \mathbf{x}_{\Xi_k,k}\}$ where Ξ_k is the number of speakers and $\mathbf{x}_{i,k}$ represents a distinct speaker state vector at time k.

For the measurement model, each $\mathcal{Z}_k^{\mathrm{v}} = \left\{\mathbf{z}_{1,k}^{\mathrm{v}}, ..., \mathbf{z}_{M_k^{\mathrm{v}},k}^{\mathrm{v}}\right\}$ is modelled by [80]

$$\mathcal{Z}_k^{\mathrm{v}} = \left\{ \bigcup_{i=1,...,|\mathcal{X}_k|} \mathcal{D}_k(\mathbf{x}_{i,k}) \right\} \cup \mathcal{C}_k^{\mathrm{v}}, \tag{4.2}$$

where $\mathcal{C}_k^{\mathrm{v}}$ is the finite set of false measurements, and $\mathcal{D}_k(\mathbf{x}_{i,k})$ is the set of colour measurements given by [80]

$$\mathcal{D}_k(\mathbf{x}_{i,k}) = \begin{cases} \emptyset, & \hbar_{\mathrm{miss}} \\ \bigcup_{i=1}^{|\mathcal{X}_k|} D(\mathbf{x}_{i,k}), & \bar{\hbar}_{\mathrm{miss}} \end{cases} \tag{4.3}$$

where $\hbar_{\mathrm{miss}}$ and $\bar{\hbar}_{\mathrm{miss}}$ are, respectively, the miss and detection hypotheses. The hypothesis $\hbar_{\mathrm{miss}}$ happens with probability $P_{\mathrm{miss}}^{\mathrm{v}}$. $D(\mathbf{x}_{i,k})$ is the minimum Bhattacharyya distance between the extracted histogram at $\mathbf{x}_{i,k}$ and the reference histograms [31]:

$$D(\mathbf{x}_{i,k}) = \min_j \left(\sqrt{1 - \sum_{u=1}^{U} \sqrt{q_u(\mathbf{x}_{i,k}) r_u^{(j)}}} \right), \tag{4.4}$$

where U is the number of histogram bins, and $r_u^{(j)}$ is the Hue-Saturation histogram of the reference image from the templates, and $q_u(\mathbf{x}_{i,k})$ is the Hue-Saturation histogram extracted from the rectangle centred on the position of the speaker $\mathbf{x}_{i,k}$.

For the false colour measurements $\mathcal{C}_k^{\mathrm{v}}$, we assume that each $c_k^{\mathrm{v}} \in \mathcal{C}_k^{\mathrm{v}}$ follows a Beta distribution. As for the false colour measurement pdf, it is shown that [80]

$$c^{\mathrm{v}}(\{\mathbf{z}_{1,k}^{\mathrm{v}}, ..., \mathbf{z}_{M_k^{\mathrm{v}},k}^{\mathrm{v}}\}) = P_{|\mathcal{Z}_k^{\mathrm{v}}|}(M_k^{\mathrm{v}}) \left(M_k^{\mathrm{v}}! \prod_{i=1}^{M_k^{\mathrm{v}}} \kappa^{\mathrm{v}}(\mathbf{z}_{i,k}^{\mathrm{v}}) \right), \tag{4.5}$$

where $P_{|\mathcal{Z}_k^{\mathrm{v}}|}(M_k^{\mathrm{v}}) = P[|\mathcal{Z}_k^{\mathrm{v}}| = M_k^{\mathrm{v}}]$ is the probability of the number of false measurements and $\kappa^{\mathrm{v}}(\mathbf{z}^{\mathrm{v}})$ is a Beta distribution function. The number of false measurements $|\mathcal{C}_k^{\mathrm{v}}|$ is assumed to follow a Poisson distribution with an average rate of λ_c^{v}, $P_{|\mathcal{Z}_k^{\mathrm{v}}|}(M_k^{\mathrm{v}}) = e^{-\lambda_c^{\mathrm{v}}} \frac{(\lambda_c^{\mathrm{v}})^{M_k^{\mathrm{v}}}}{M_k^{\mathrm{v}}!}$. Therefore, Equation (4.5) can be expressed as [80]

$$c(\mathcal{Z}_k^{\mathrm{v}}) = e^{-\lambda_c^{\mathrm{v}}} \prod_{\mathbf{z}_k^{\mathrm{v}} \in \mathcal{Z}_k^{\mathrm{v}}} \lambda_c^{\mathrm{v}} \kappa^{\mathrm{v}}(\mathbf{z}_k^{\mathrm{v}}). \tag{4.6}$$

Assuming that noise on the visual likelihood function is Gaussian, then the likelihood function of the measured colour histogram can be written as [31]:

$$g(\mathbf{z}_k^{\mathrm{v}} | \mathbf{x}_{i,k}) \propto \mathcal{N}(z_k^{\mathrm{v}}; 0, \sigma_{\mathrm{v}}^2) = \frac{1}{\sigma_{\mathrm{v}}\sqrt{2\pi}} \exp\left\{ -\frac{\mathcal{D}_k(\mathbf{x}_{i,k})^2}{2\sigma_{\mathrm{v}}^2} \right\}, \tag{4.7}$$

where σ_{v}^2 is the variance of noise.

4.2.2 Audio Likelihood

The previous section described the visual measurement set and the calculation of the likelihood function. This section presents modelling of audio data and the likelihood function for the DOA information.

The DOA information of the speakers is extracted as described in Section 2.1. Then, the DOA measurement model for each $\mathcal{Z}_k^{\mathrm{a}} = \{\mathbf{z}_{1,k}^{\mathrm{a}}, ..., \mathbf{z}_{M_k^{\mathrm{a}},k}^{\mathrm{a}}\}$ takes the form [80]

$$\mathcal{Z}_k^{\mathrm{a}} = \left\{ \bigcup_{i=1,...,|\mathcal{X}_k|} \mathcal{E}_k(\mathbf{x}_{i,k}) \right\} \cup \mathcal{C}_k^{\mathrm{a}}, \tag{4.8}$$

where $\mathcal{C}_k^{\mathrm{a}}$ is the finite set of false DOA measurements and $\mathcal{E}_k(\mathbf{x}_{i,k})$ is the difference between jth DOA angle, $\theta_{j,k}$, and $\psi(\mathbf{x}_{i,k})$. Here, $\psi(\mathbf{x}_{i,k})$ is the ith speaker position in terms of angle with respect to the microphone position [80]

$$\mathcal{E}_k(\mathbf{x}_{i,k}) = \begin{cases} \emptyset, & \hbar_{\mathrm{miss}} \\ \bigcup_{i=1}^{|\mathcal{X}_k|} \left\{ \min_j \left(\psi(\mathbf{x}_{i,k}) - \theta_{j,k} \right) \right\}, & \bar{\hbar}_{\mathrm{miss}} \end{cases} \tag{4.9}$$

where $\hbar_{\mathrm{miss}}$ and $\bar{\hbar}_{\mathrm{miss}}$ are the miss and detection hypotheses, respectively. The hypothesis $\hbar_{\mathrm{miss}}$ happens with probability $P_{\mathrm{miss}}^{\mathrm{a}}$.

We assume that each $c_k^{\mathrm{a}} \in \mathcal{C}_k^{\mathrm{a}}$ follows a uniform distribution for the false measurements. The false DOA measurement pdf can be shown as [80]

$$c^{\mathrm{a}}(\{\mathbf{z}_{1,k}^{\mathrm{a}}, ..., \mathbf{z}_{M_k^{\mathrm{a}},k}^{\mathrm{a}}\}) = P_{|\mathcal{Z}_k^{\mathrm{a}}|}(M_k^{\mathrm{a}}) \left(M_k^{\mathrm{a}}! \prod_{i=1}^{M_k^{\mathrm{a}}} \kappa^{\mathrm{a}}(\mathbf{z}_{i,k}^{\mathrm{a}}) \right), \tag{4.10}$$

where $P_{|\mathcal{Z}_k^{\mathrm{a}}|}(M_k^{\mathrm{a}}) = P[|\mathcal{Z}_k^{\mathrm{a}}| = M_k^{\mathrm{a}}]$ is the probability of the number of false measurements and $\kappa^{\mathrm{a}}(\mathbf{z}^{\mathrm{a}})$ is a uniform density on an interval $[-\theta_{\max}, \theta_{\max}]$. The number of false measurements $|\mathcal{C}_k^{\mathrm{a}}|$ is assumed to follow a Poisson distribution [80] with an average rate

of λ_c^{a}, $P_{|\mathcal{Z}_k^{\mathrm{a}}|}(M_k^{\mathrm{a}})= e^{-\lambda_c^{\mathrm{a}}}\frac{(\lambda_c^{\mathrm{a}})^{M_k^{\mathrm{a}}}}{M_k^{\mathrm{a}}!}$. Equation (4.10) can then be expressed as [80]

$$c(\mathcal{Z}_k^{\mathrm{a}}) = e^{-\lambda_c^{\mathrm{a}}} \prod_{\mathbf{z}_k^{\mathrm{a}} \in \mathcal{Z}_k^{\mathrm{a}}} \lambda_c^{\mathrm{a}} \kappa^{\mathrm{a}}(\mathbf{z}_k^{\mathrm{a}}). \tag{4.11}$$

If noise on the audio likelihood function is also assumed Gaussian, then the likelihood function of the DOA measurements can be written as:

$$g(\mathbf{z}_k^{\mathrm{a}}|\mathbf{x}_{i,k}) \propto \mathcal{N}(z_k^{\mathrm{a}}; 0, \sigma_{\mathrm{a}}^2) = \frac{1}{\sigma_{\mathrm{a}}\sqrt{2\pi}} \exp\left\{-\frac{\mathcal{E}_k(\mathbf{x}_{i,k})^2}{2\sigma_{\mathrm{a}}^2}\right\}, \tag{4.12}$$

where σ_{a}^2 is the variance of noise.

4.2.3 Particle Filter Implementation of RFS

The RFS model described in the previous section can be used in a Bayesian framework to estimate the multi-speaker locations and the number of active speakers.

We can define the pdfs of $\mathcal{X}_k$ and $\mathcal{Z}_k$ for the RFS model. The RFS state transition density is denoted by $f(\mathcal{X}_k|\mathcal{X}_{k-1})$ and the RFS likelihood function is denoted by $g(\mathcal{Z}_k|\mathcal{X}_k)$.

To derive these pdfs, some mathematical concepts are required that are beyond the scope of this thesis. Detailed descriptions of the RFS pdf concepts can be found in [87] and [135]. The derivations of $f(\mathcal{X}_k|\mathcal{X}_{k-1})$ and $g(\mathcal{Z}_k|\mathcal{X}_k)$ based on RFS pdf concepts are given in [80]. Accordingly, the Bayesian recursive estimation of the posterior distribution of the RFS state $p(\mathcal{X}_k|\mathcal{Z}_k)$ can be written as

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}) = \int_{\mathcal{F}} f(\mathcal{X}_k|\mathcal{X}_{k-1}) p(\mathcal{X}_{k-1}|\mathcal{Z}_{1:k-1}) \mu(d\mathcal{X}_{k-1}), \tag{4.13}$$

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}) \propto p(\mathcal{Z}_{1:k}|\mathcal{X}_k) p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}), \tag{4.14}$$

where $p(\mathcal{X}_k|\mathcal{X}_{k-1})$ characterizes the birth, death and survival processes of the state dynamics. The subscript $\mathcal{F}$ is the collection of all finite subsets of the state space, and $\mu(d\mathcal{X}_{k-1})$ denotes an appropriate measure on $\mathcal{F}$ as described in [135]. Considering that the visual measurement set $\mathcal{Z}_k^{\mathrm{v}}$ and the audio measurement set $\mathcal{Z}_k^{\mathrm{a}}$ are independent

[46], [80] the likelihood for the joint measurements can be written as

$$p(\mathcal{Z}_k|\mathcal{X}_k) = p(\mathcal{Z}_k^{\mathrm{v}}, \mathcal{Z}_k^{\mathrm{a}}|\mathcal{X}_k) = p(\mathcal{Z}_k^{\mathrm{v}}|\mathcal{X}_k)p(\mathcal{Z}_k^{\mathrm{a}}|\mathcal{X}_k), \tag{4.15}$$

where $p(\mathcal{Z}_k^{\mathrm{v}}|\mathcal{X}_k)$ and $p(\mathcal{Z}_k^{\mathrm{a}}|\mathcal{X}_k)$ are the likelihood for the visual and audio measurements respectively.

Multiplication of two likelihoods based on independence assumptions is easy and commonly used in tracking applications [46], [52], [76]. It also fits into the RFS theory considering the fact that the joint likelihood is used for the estimation of the weights of the particles which are then used in the estimation of the number of speakers. One drawback of multiplication is that the joint likelihood will be degraded if one of the likelihoods has been corrupted by noisy measurements. In that case, it will cause a wrong estimation of the number of speakers which ends up with speakers being lost. Of course, it depends on the quality of the likelihoods. The RFS theory can tolerate wrong estimation of the number of speakers caused by noisy/corrupted likelihoods as it employs a rounding process (because the summation of the weights generally does not give integer output) after the number of speakers is estimated. Lets assume that there are two speakers in the scene. In a perfect scenario, we get perfect audio and video likelihoods, so the joint likelihood gives accurate results in the calculation of the weights. Assuming the number of speakers will be calculated as 2 with the summation of the weights. Now, suppose that one of the likelihoods (no matter which one) is corrupted by noise and the value of joint likelihood is decreased. In that case, the summation of weights will be less than 2, lets say 1.92. However, after rounding, the estimated number will be $\lceil 1.92 \rfloor = 2$. So it gives correct result even there is a noisy likelihood. However this works until the joint likelihood is decreased to the point where the sum of weights 1.49 as $\lceil 1.49 \rfloor = 1$. So only one speaker will be detected, the second one will be lost. If the quality of one of the modalities (or both of them) is further decreased, then the estimated number of speakers will be 0 when the summation of weights is under 0.50. In this case, both speakers will be lost.

As an alternative to the multiplication of the likelihoods, the sum rule can be employed as it is less sensitive to noise and clutter [145]. However, the two likelihoods cannot

be summed directly as the joint likelihood will cause over-estimation of the number of speakers. In [145], each likelihood is weighted with coefficients estimated based on the spatial uncertainty at each time frame which introduces extra computational cost.

Since the states have a nonlinear relationship with the measurements, closed-form solution for the pdf of the source state is not available. In this chapter, a particle filtering approach is employed to approximate the pdfs. Assume that we have particles $\mathcal{X}_{k-1}^{(n)}$ for $n = 1, \ldots, N$ at the previous time step $k-1$, and the corresponding importance weight $w_{k-1}^{(n)}$. The particles at the current time step k are generated according to [80]

$$\mathcal{X}_k^{(n)} \sim f(\mathcal{X}_k^{(n)} | \mathcal{X}_{k-1}^{(n)}). \tag{4.16}$$

The particles are weighted by [80]

$$w_k^{(n)} = w_{k-1}^{(n)} g(\mathcal{Z}_k | \mathcal{X}_k^{(n)}). \tag{4.17}$$

After resampling, the posterior distribution is thus approximated by [80]

$$p(\mathcal{X}_k^{(n)} | \mathcal{Z}_k) \approx \sum_{n=1}^{N} \tilde{w}_k^{(n)} \delta_{\mathcal{X}_k^{(n)}} (\mathcal{X}_k), \tag{4.18}$$

where $\tilde{w}_k^{(n)}$ is the normalized weight based on $w_k^{(n)}$, and $\delta_{\mathcal{X}}(\mathcal{Y})$ is a set-valued Dirac delta function. For brevity, $\delta_{\mathcal{X}}(\mathcal{Y})$ is defined such that $\delta_{\mathcal{X}}(\mathcal{Y}) = 1$ if $\mathcal{X} \subseteq \mathcal{Y}$ and 0 otherwise.

The proposed tracking algorithm [63], called the RFS-PF algorithm, is presented as Algorithm 4.1. This algorithm describes how to use the RFS-PF for visual or audio-visual tracking. In visual tracking, the colour likelihood function is calculated for $\Xi_{\max} = 2$ (where $\Xi_{\max}$ is the maximum number of speakers) using Equations (4.19), (4.20) and (4.21), respectively for no speaker, one speaker and two speakers [80]. The audio likelihood function is calculated when the DOA information exists, again using Equations (4.19), (4.20) and (4.21), and is fused with the colour likelihood function according to Equation (4.15).

Initialize: N, Ξ_{max}, P_{birth}, P_{death}, P_{miss}, σ^2, U, T, $\mathbf{F}$, λ_c^v, λ_c^a, r, u, k, K
For $n = 1, ..., N$, draw particles $\mathbf{x}_0^{(n)} \sim \mathcal{N}(.|\mathbf{x}_0, \mathbf{Q})$
Set initial weights $\tilde{w}_0^{(n)} = \frac{1}{N}$
while $k < K$ **do**
  Set $\mathcal{X}_k^{(n)} = \emptyset$; Source death, survival and birth:
  Draw a random number $u_d \sim \mathcal{U}[0,1)$; $u_b \sim \mathcal{U}[0,1)$
  if $u_d > P_{death}$ **then**
    Compute $\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{q}_k$
    Set $\mathcal{X}_k^{(n)} = \mathcal{X}_k^{(n)} \cup \{\mathbf{x}_k\}$
  end
  if $u_b > P_{birth}$ *and* $|\mathcal{X}_{k-1}^{(n)}| < \Xi_{max}$ **then**
    Draw an initial state $\mathbf{b}_k$
    Set $\mathcal{X}_k^{(n)} = \mathcal{X}_k^{(n)} \cup \{\mathbf{b}_k\}$
  end
  for $n = 1, ...N$ **do**
    Compute the colour likelihood $g(\mathcal{Z}_k^{\text{v}}|\mathcal{X}_k^{(n)})$ using related Equations (4.19) - (4.21)
    if *DOA exists* **then**
      Compute the audio likelihood $g(\mathcal{Z}_k^{\text{a}}|\mathcal{X}_k^{(n)})$ using related Equations (4.19) - (4.21)
      Compute the likelihood for the joint measurements:
      $g(\mathcal{Z}_k|\mathcal{X}_k^{(n)}) = g(\mathcal{Z}_k^{\text{v}}|\mathcal{X}_k^{(n)})g(\mathcal{Z}_k^{\text{a}}|\mathcal{X}_k^{(n)})$
    else
      // If DOA is not available in time frame k
      $g(\mathcal{Z}_k|\mathcal{X}_k^{(n)}) = g(\mathcal{Z}_k^{\text{v}}|\mathcal{X}_k^{(n)})$
    end
    Compute the importance weight:
      $w_k^{(n)} = \tilde{w}_{k-1}^{(n)} g(\mathcal{Z}_k|\mathcal{X}_k^{(n)})$
  end
  For $n = 1, ..., N$, normalize the weight $\tilde{w}_k^{(n)} = \frac{w_k^{(n)}}{\sum_{n=1}^{N} w_k^{(n)}}$
  Resample the particles
  Cluster with K-means and estimate speaker position
  Detect speaker identity
end

Algorithm 4.1: RFS-PF algorithm for multi-speaker tracking.

$$g(\mathcal{Z}_k|\emptyset) = e^{-\lambda_c}\left(\lambda_c\kappa(z_k)\right)^{|\mathcal{Z}_k|} \tag{4.19}$$

$$g(\mathcal{Z}_k|\{\mathbf{x}_k\}) = g(\mathcal{Z}_k|\emptyset)\left(P_{\text{miss}} + (1-P_{\text{miss}})\sum_{z_k\in\mathcal{Z}_k}\left(\frac{1}{\lambda_c\kappa(z_k)}\right)g(z_k|\mathbf{x}_k)\right) \tag{4.20}$$

$$\begin{aligned} g(\mathcal{Z}_k|\{\mathbf{x}_{1,k},\mathbf{x}_{2,k}\}) = {} & g(\mathcal{Z}_k|\emptyset)\Bigg\{\prod_{i=1,2}\left(P_{\text{miss}} + (1-P_{\text{miss}})\sum_{z_k\in\mathcal{Z}_k}\left(\frac{1}{\lambda_c\kappa(z_k)}\right)g(z_k|\mathbf{x}_{i,k})\right) \\ & -(1-P_{\text{miss}})^2\sum_{z_k\in\mathcal{Z}_k}\left(\frac{1}{\lambda_c\kappa(z_k)}\right)^2 g(z_k|\mathbf{x}_{1,k})g(z_k|\mathbf{x}_{2,k})\Bigg\} \end{aligned} \tag{4.21}$$

The number of speakers at time k is approximated by [80]

$$\bar{\Xi}_k \approx \sum_{n=1}^{N} w_k^{(n)}|\mathcal{X}_k^{(n)}|. \tag{4.22}$$

$\bar{\Xi}_k$ is a floating-point valued number and since the number of speakers should be an integer, a rounding operation is applied $\hat{\Xi}_k = \lceil\bar{\Xi}_k\rfloor$.

After the resampling step, particles need to be clustered to use the centroids of these clusters as final state estimates. In the literature, clustering techniques are categorized into three main classes [36]: agglomerative hierarchical clustering, optimisation methods such as the K-means, and mixture models such as the expectation maximisation (EM). Clark *et al.* [23] conducted a comparative analysis of hierarchical clustering, the K-means and the EM methods. They considered the computational complexity of three approaches with known orders:

- Hierarchical clustering, order $O(N^2 logN)$;
- The K-means, order $O(iN\Xi)$;
- The EM algorithm, order $O(iN\Xi^2)$;

where i is the number of iterations required to converge to a solution, N is the number of particles, Ξ the number of targets. Hierarchical clustering is computationally the most expensive one. The K-means approach is the cheapest in terms of computational cost. However, it is non robust, and statistically biased and inconsistent. On the other hand, the EM algorithm is unbiased and more robust, but computationally intensive

[23]. They used real three-dimensional active sonar data in the comparisons, and the authors reported that the K-means actually outperformed the EM algorithm.

To this end, the K-means algorithm is employed in this study to cluster the particles and the final estimation is obtained from the centroids of these clusters. Here, all state variables (positions, velocities and scale) are used in the clustering algorithm to obtain the position of the detected speaker with other state variables which will be used later to detect the speaker identity. This part can be modified depending on the application such as when only the position of the speaker is needed. Lastly, the identity of the speaker is detected using positions and scale variables of the final state of the speaker. To detect the identity of the speaker, the image patch centred on the estimated position is compared with the reference image from the templates using the Bhattacharyya distance defined in Equation (4.4), and the identity is determined based on the closest distance.

4.3 Particle PHD Filtering

The PHD filter and its particle filter, namely SMC implementation, were discussed in Section 2.3.4. This section first describes the colour likelihood model and then our proposed SMC-PHD filtering approach for audio-visual multiple speaker tracking.

4.3.1 Colour Likelihood Model

The colour information of the state $\mathbf{x}_k$ is represented using a colour histogram. Let the speaker candidate be defined with the rectangle centred around the location (x_k, y_k) on the frame. This rectangle is converted to colour histogram $q_u(\mathbf{x}_k)$ in order to calculate its similarity with the reference speaker models. In multi-speaker tracking, we have many colour models of speakers $\left\{r_u^{(1)}, r_u^{(2)}, ..., r_u^{(M)}\right\}$ where u is the index of histogram bins. The colour similarities between the candidate speaker and the reference models are calculated in terms of the Bhattacharyya distance [31].

$$D^{(m)}(\mathbf{x}_k) = \sqrt{1 - \sum_{u=1}^{U} \sqrt{q_u(\mathbf{x}_k) r_u^{(m)}}}. \tag{4.23}$$

Assuming that noise on the colour likelihood function is Gaussian, then the likelihood function of the each measured colour histogram can be written as [31]:

$$g^{(m)}(\mathbf{z}_k|\mathbf{x}_k) \propto \mathcal{N}(\mathbf{z}_k; 0, \sigma_c^2) = \frac{1}{\sigma_c\sqrt{2\pi}} \exp\left\{-\frac{\{D^{(m)}(\mathbf{x}_k)\}^2}{2\sigma_c^2}\right\}, \tag{4.24}$$

where σ_c^2 is the variance of noise for the colour likelihood.

4.3.2 Proposed AV-SMC-PHD Filtering

The way we introduce the DOA data into the SMC-PHD filter is based on [61] and [62] where the efficiency of the particles is improved with the DOA information under a particle filter framework. The DOA is used to draw a line, named as DOA line, from the centre of the microphone array to a point in the image frame estimated by the projection of DOA to $2D$ image plane. Then, all the particles are re-allocated around the DOA line.

Here, the DOA data is not used in the same way for all the particles as in [61] and [62], instead the contribution of the DOA information is varied by the type of the particles. In the RFS, a multi-speaker state is defined in Equation (2.12) as the union of surviving, spawned and born particles in the SMC-PHD filter. In our proposed algorithm, the born particles are generated only when the detection of a new speaker occurs via audio. The born particles are uniformly distributed around the DOA line as illustrated in Figure 4.1b. The surviving and spawned particles are also concentrated around the DOA line if the DOA data exists at that time. Although for a short silence, the missing DOA data is completed by interpolation, the DOA data will be lost in the case where the speaker stops talking for a long time. In that situation, our proposed algorithm continues tracking without the DOA information.

Reallocating the particles around the DOA line is likely to increase the possibility of speaker detection since the DOA indicates the approximate direction of the sound emanating from the speaker.

The surviving and spawned particles are defined as $\tilde{\mathbf{x}}_{s,k}$ for time k since the DOA information is used for surviving and spawned particles in the same way. In addition,

the born particles are defined as $\tilde{\mathbf{x}}_{b,k}$. In each iteration, the surviving particles from the previous iteration and the particles spawned from them are distributed by a dynamic model given in Equation (2.9). More details about the generation of the surviving, spawned and born particles can be found in [80] and [135]. If the DOA exists in the current frame, the DOA line is drawn [61] and the perpendicular Euclidean distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & ... & d_k^{(N_{k-1})} \end{bmatrix}$ of the particles to the DOA line are calculated. In the case that multiple DOA lines exist, the DOA line closest to the particles is chosen as long as the distance to the DOA line is under a pre-determined threshold value. This is required since one speaker may be silent while the other speaker is still talking, which causes the particles belonging to the silent speaker to converge to the DOA line that belongs to another speaker. Then, the movement distances $\hat{\mathbf{d}}_k$ of the particles are calculated as follows [62]:

$$\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_1} \odot \mathbf{d}_k, \tag{4.25}$$

where $\odot$ is the element-wise product and $\|.\|_1$ is the ℓ_1 norm. $\hat{\mathbf{d}}_k$ is used to relocate the particles $\tilde{\mathbf{x}}_{s,k}$ to around the DOA line:

$$\tilde{\mathbf{x}}_{s,k} = \tilde{\mathbf{x}}_{s,k} \oplus \mathbf{h}_k \hat{\mathbf{d}}_k, \tag{4.26}$$

where $\oplus$ is the element-wise addition and $\mathbf{h}_k = \begin{bmatrix} cos(\theta_k) & 0 & sin(\theta_k) & 0 & 0 \end{bmatrix}^T$. It is multiplied by $\mathbf{h}_k$ to update only the position (x_1, x_2) of the particle state vector $\begin{bmatrix} x_1 & \dot{x}_1 & x_2 & \dot{x}_2 & s \end{bmatrix}^T$ in order to provide the perpendicular movement to the DOA line. The presence of new speakers is checked for using the DOA information. If the number of DOA lines is larger than the number of estimated speakers at time step $k-1$, it means a new speaker may appear in the scene. Here, it is assumed that the new speaker is speaking when he/she first enters the scene. Hence, extra DOA information may be coming from either the new speaker or due to clutter. Then, born particles are propagated to check whether there is a new speaker. The main purpose of this assumption is to reduce the computational cost caused by propagating new born particles every time frame.

As the new speaker is checked when the number of DOA lines is larger than the number of estimated speakers in time step $k-1$, our algorithm does not propagate the born

particles in the case that the number of DOA lines does not exceed the previous number of speakers, meanwhile a new speaker enters the scene. Actually, this is not the case for the sequences of $AV16.3$ as the speakers are mostly speaking. However, this can be easily addressed with a small modification of the algorithm by comparing the current number of DOA with previous ones. After the assignment of speakers to the DOA lines, based on the distance between previous estimated positions of the speakers to the DOA lines, is completed, new born particles can be propagated for the rest of DOA lines.

J_k born particles $\tilde{\mathbf{x}}_{b,k}$ are generated and distributed uniformly around the new DOA line. The born particles are distributed even in occlusion case until occluded speaker is detected.

All the particles are combined under $\tilde{\mathbf{x}}_k$ and the prediction step is employed to calculate the weights of particles $\tilde{w}_{k|k-1}$. After the estimation of colour likelihood using Equation (4.24), the update step is performed to calculate $\tilde{w}_k$. The total mass, which gives the number of estimated speakers, is calculated by summing the weights of the particles. Then, the resampling step is performed. N_k particles are resampled from $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \tilde{\mathbf{x}}_k^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$ to get $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$ where $\hat{\Xi}_{k|k}$ is the total mass and $\hat{\Xi}_{k|k} = \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_k^{(n)}$. N_k is estimated by $N_k = \eta\hat{\Xi}_{k|k}$ where η is the constant number of particles per speaker. After the particles are resampled, the positions of the estimated speakers are estimated using the K-means clustering algorithm. Lastly, the identity of the speakers is detected by measuring the similarity between the colour histogram of the estimated speakers and that of the reference speakers. The pseudo code of the proposed AV-SMC-PHD filtering algorithm [59] is depicted in Algorithm 4.2.

4.4 Evaluation Results and Analysis

In this section, our proposed AV trackers based on both RFS and PHD filters are evaluated in Section 4.4.1 and 4.4.2, respectively. Experiments are performed on $AV16.3$ dataset and the OSPA-T metric, described in Chapter 2, is used for performance comparison.

Initialize: η, σ^2, U, T, $\mathbf{F}$, λ, r, u, p_M, p_D, σ_c^2, σ_s^2, p_S, k, K, N_0, $\mathbf{x}_0$
while $k < K$ **do**
  For $n = 1, ..., N_{k-1}$ sample $\tilde{\mathbf{x}}_k \sim q_k\left(\cdot|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k\right)$, where $\tilde{\mathbf{x}}_k \in \mathbb{R}^{5\times N_{k-1}}$
  Propagate surviving and spawned particles: $\tilde{\mathbf{x}}_{s,k} = \mathbf{F}\tilde{\mathbf{x}}_k + \mathbf{q}_k$
  Get the corresponding DOA angle θ_k
  if *DOA exists* **then**
    // For surviving and spawned particles
    Calculate distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & ... & d_k^{(N_{k-1})} \end{bmatrix}^T$
    Calculate movement distances $\hat{\mathbf{d}}_k$ using Equation (4.25)
    Concentrate $\tilde{\mathbf{x}}_{s,k}$ around the DOA line : $\tilde{\mathbf{x}}_{s,k} = \tilde{\mathbf{x}}_{s,k} \oplus \mathbf{h}_k\hat{\mathbf{d}}_k$
    if *new speaker* **then**
      // For born particles
      For $n = N_{k-1}+1, ..., N_{k-1}+J_k$ sample $\tilde{\mathbf{x}}_{b,k} \sim p_k\left(\cdot|\mathcal{Z}_k\right)$ uniformly around the DOA line
    end
  end
  Combine all the particles:
  $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_{s,k} \cup \tilde{\mathbf{x}}_{b,k}$
  Prediction: For $n = 1, ..., N_{k-1}+J_k$ calculate $\tilde{w}_{k|k-1}^{(n)}$ using Equation (2.21)
  Estimate colour likelihood using Equation (4.24)
  Update: For $n = 1, ..., N_{k-1}+J_k$ calculate $\tilde{w}_k^{(n)}$ using Equation (2.22)
  Calculate the total mass $\hat{\Xi}_{k|k} = \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_k^{(n)}$
  Resampling: Resample $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \tilde{\mathbf{x}}_k^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$ to get $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$
  where $N_k = \eta\hat{\Xi}_{k|k}$
  Multiply the weights by $\hat{\Xi}_{k|k}$ to get $\left\{\tilde{w}_k^{(n)}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$
  Cluster the particles and get the positions of the speakers
  $k = k + 1$
end

Algorithm 4.2: Proposed AV-SMC-PHD filtering algorithm.

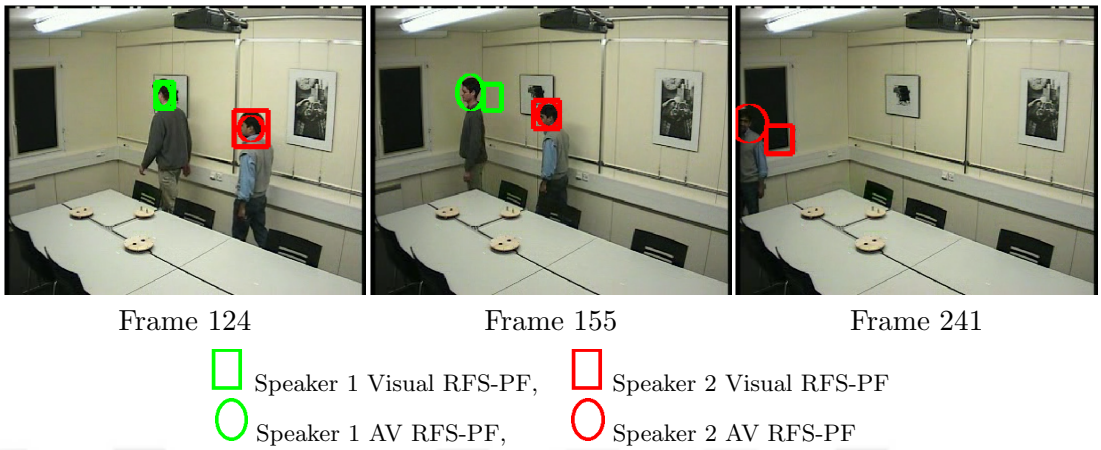


Figure 4.2: Tracking results of the Visual RFS-PF and the proposed AV RFS-PF in Sequence 30 camera #2.

4.4.1 Visual RFS-PF vs. AV RFS-PF

In this study, we used two multi-speaker sequences. The first one is Sequence 30 (camera #2) where two moving speakers walk back and forth once, one behind the other at a constant distance and both speak continuously. Sequence 25 (camera #3) is the second sequence where two moving speakers walk back and forth twice, each speaker starts walking from the opposite side of the room and occludes each other while talking most of the time.

In Sequence 30 and 25, the number of speakers changes between 0 to 2. Sequence 25 is more challenging than Sequence 30 as two speakers occlude each other. Therefore, with these two sequences, we are able to evaluate the proposed AV RFS-PF algorithm on the following two challenging tracking scenarios: a variable number of speakers and speaker occlusion. The parameters for the RFS-PF are set as: $P_{\text{birth}} = 0.2$, $P_{\text{death}} = 0.01$, $P^{\text{v}}_{\text{miss}} = 0.02$, $P^{\text{a}}_{\text{death}} = 0.25$, λ^{v}_{c} and λ^{a}_{c} are set to 0.01, $\theta_{\text{max}} = \pi/2$ and $N = 50$. For the OSPA metric, the cut-off parameter $c = 65$ and order parameter $a = 2$. All these parameters are optimised by a cross-validation in extensive experimental studies.

Figure 4.2 shows some frames with Visual RFS-PF and AV RFS-PF results. To distinguish the trackers and speakers, the visual tracker results are drawn with rectangles, while the AV tracker results are drawn with ellipses. In addition, green and red colours are used to distinguish Speaker 1 from Speaker 2, respectively.

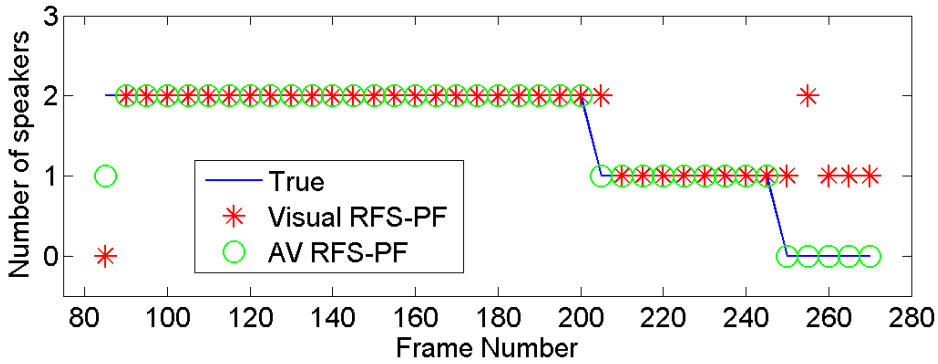


Figure 4.3: Comparison of the Visual RFS-PF and the proposed AV RFS-PF in estimation of the number of active speakers for Sequence 30 camera #2.

At the beginning, both the visual and the AV tracker track the speakers well, but when the speakers go to the corner of the room, the visual tracker starts to drift away because the illumination effects make it difficult for the particles to detect the speaker and tracker starts to deviate from the ground truth position. This deviation affects the distribution of the survival particles on each following frame and finally the tracker loses the speaker. On the contrary, the audio information prevents the deviation of the particles as they are always concentrated around the DOA line in the AV tracker.

Figure 4.3 shows the estimation of the number of active speakers. Here, the number of active speakers is changing from 2 to 0 and the AV tracker shows better performance than the visual tracker. For clearer presentation, down-sampling is performed to draw the plots. The position estimates of the trackers are given in Figure 4.4 where GT is the abbreviation for ground truth positions. It can be observed that the visual tracker starts to deviate from the ground truth trajectory in the last few frames due to the reasons explained above.

The tracking results of the proposed algorithm for Sequence 25 are demonstrated in Figure 4.5. Here, the two speakers occlude each other and after occlusion, the AV tracker is able to follow the second speaker before the visual tracker does. The number of active speakers estimated for Sequence 25 is given in Figure 4.6. It can be observed that the performance of the visual tracker is not as good as the AV tracker.

The position estimates for x- and y- trajectories are given in Figure 4.7a and 4.7b, respectively. The AV tracker trajectories follow the ground truth trajectories closer

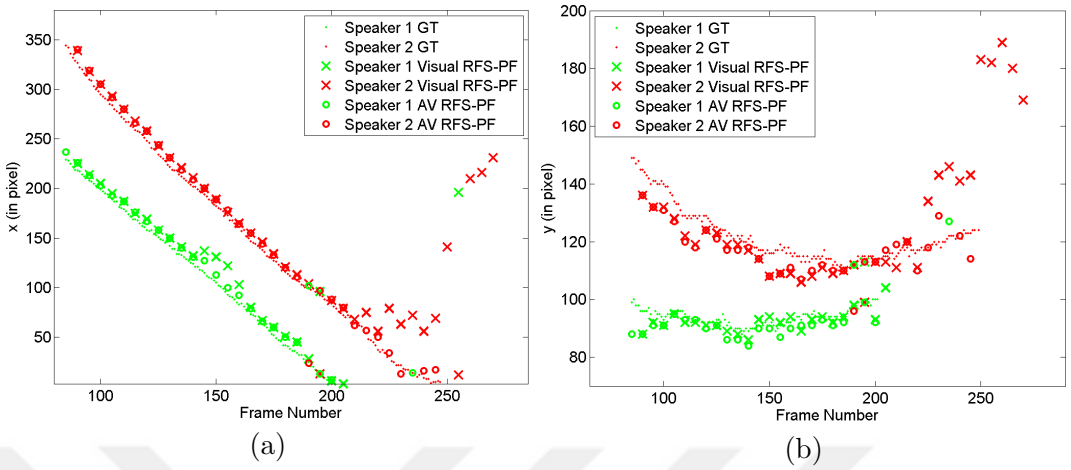


Figure 4.4: Position estimates of the Visual RFS-PF and proposed AV RFS-PF trackers for Sequence 30 camera #2.

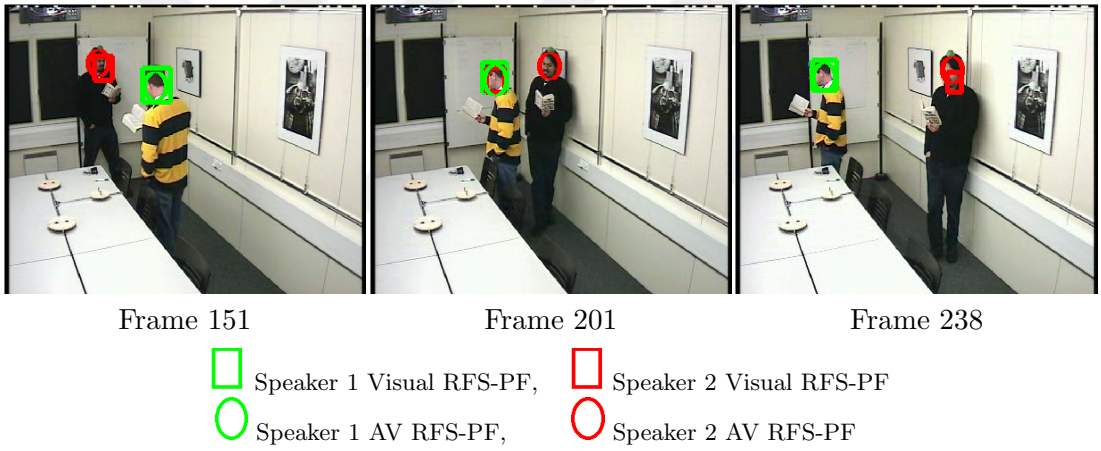


Figure 4.5: Tracking results of the Visual RFS-PF and proposed AV RFS-PF in Sequence 25 camera #3.

than the visual tracker.

To see the performance difference between the trackers, the OSPA errors are plotted for Sequence 30 and Sequence 25 in Figure 4.8a and 4.8b, respectively. To get more reliable results, the experiments are repeated 10 times and the average error is plotted. It is clearly seen that adding the audio information to the visual tracker leads to an increase in performance.

Then, these trackers are also tested on other sequences and the results are given in Table 4.1 with standard errors of experiments results. Here, only two speaker sequences are chosen to test the trackers as the computational complexity increases gradually with the

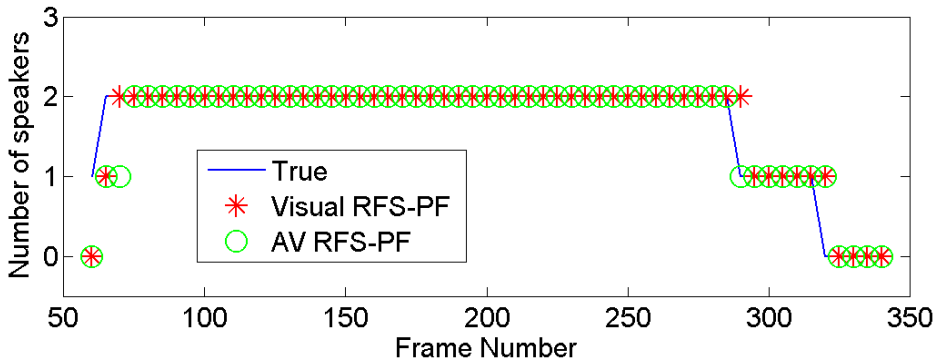
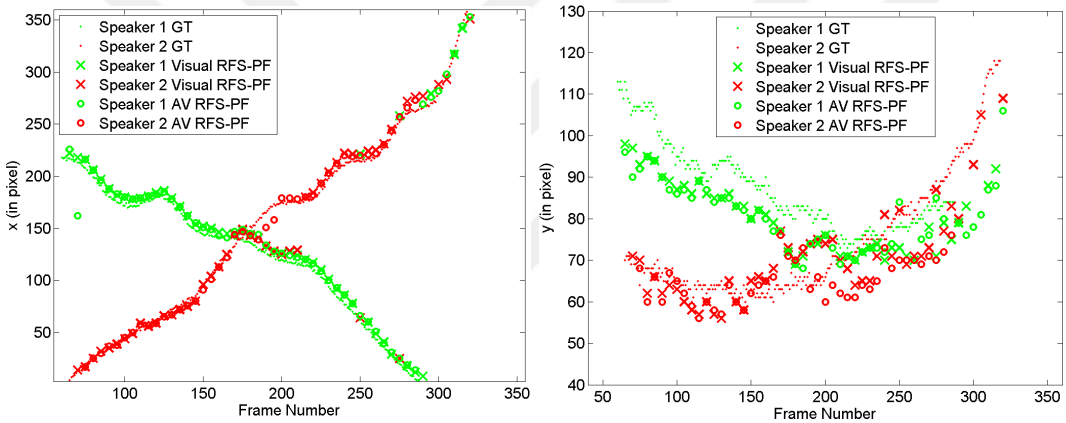
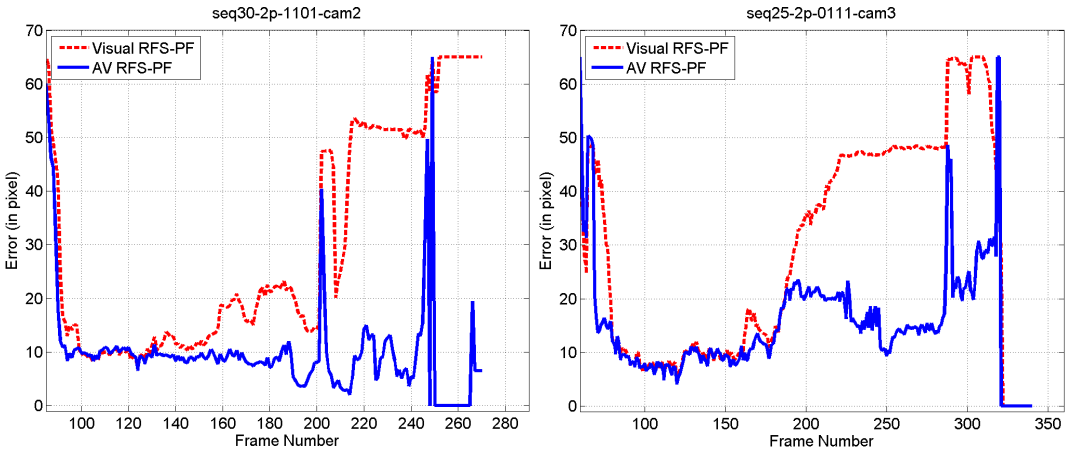


Figure 4.6: Comparison of the Visual RFS-PF and proposed AV RFS-PF in estimation of the number of active speakers for Sequence 25 camera #3.



(a) (b)

Figure 4.7: Position estimates of the Visual RFS-PF and proposed AV RFS-PF trackers for Sequence 25 camera #3.



(a) (b)

Figure 4.8: Performance comparison of the Visual RFS-PF and proposed AV RFS-PF in terms of the OSPA error.

Table 4.1: OSPA-T values for Visual RFS-PF and proposed AV RFS-PF averaged over 10 repetitions and standard errors of the estimated OSPA-T values.

		Visual RFS-PF		AV RFS-PF	
		OSPA-T	SE-OSPA-T	OSPA-T	SE-OSPA-T
seq24-2p-0111	cam1	26.15	0.02	16.84	0.11
	cam2	24.88	0.09	19.02	0.13
	cam3	23.43	0.09	17.66	0.05
seq25-2p-0111	cam1	24.29	0.15	18.37	0.03
	cam2	25.02	0.05	16.92	0.09
	cam3	26.86	0.14	20.58	0.10
seq30-2p-1101	cam1	33.39	0.44	24.7	0.28
	cam2	23.82	0.23	18.41	0.23
	cam3	34.13	0.4	23.19	0.19
Average		**26.89**	**0.18**	**19.52**	**0.13**

number of speakers. In the $AV16.3$ dataset, five different sequences are available and only three of them have challenging conditions such as multiple occlusions and moving speakers. The tracking algorithms are tested on these challenging sequences which are Sequences 24, 25 and 30. The average error is found to be 26.89 for Visual RFS and 19.52 for AV RFS-PF. With the contribution of the audio modality, the tracking error is reduced by 27% in AV RFS-PF. We apply the ANOVA based F-test to see whether the results are significant or not. According to the test, $F = 20.28$ and p-value$= 0.0004$ with the degree of freedom $(1, 16)$. Using the F-distribution table given in [12], the value of $F_{crit} = 4.49$. Based on the ANOVA test, we can say that the improvements shown by the results given in Table 4.1 are significant as $F > F_{crit}$ and the p-value is smaller than 0.05 (for a 5% significance level).

4.4.2 Proposed AV-SMC-PHD

For the evaluation of the proposed AV-SMC-PHD, we used twelve multi-speaker sequences. As it is not feasible to plot the results of all these sequences, only the results of two sequences are illustrated by the plots. The first one is Sequence 24 (camera #1) where two moving speakers are walking back and forth, crossing the field of view twice and occluding each other. Sequence 45 (camera #3) is the second sequence where three moving speakers are occluding each other many times. In these two sequences, the speakers are speaking continuously and the number of speakers is changing up to

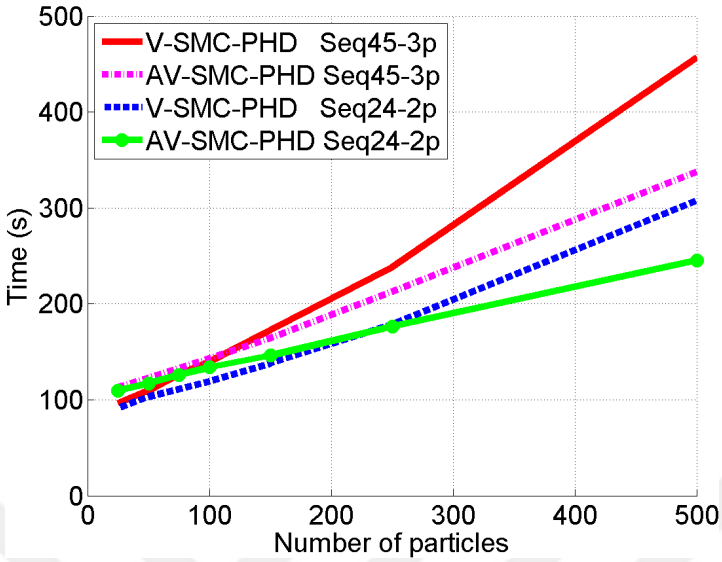


Figure 4.9: Computational cost measurement of the V-SMC-PHD and the proposed AV-SMC-PHD filter in Sequence 24 camera #1 and Sequence 45 camera #3.

three.

The parameters for the SMC-PHD are set as: $p_D = 0.98$, $p_S = 0.99$, $\lambda = 0.26$ and $\sigma_c = 0.1$. The uniform density u is $(360 \times 280)^{-1}$ and the number of particles per speaker is $\eta = 50$. In this case, the cut-off parameter $c = 65$, the OSPA-T metric order parameter $a = 2$. These parameters are also set based on extensive experimental studies.

To test the computational efficiency of the proposed and baseline algorithms, we ran experiments on Sequences 24 and 45. The number of particles per speaker was varied from 25 to 500 on Intel core $i7$ 2.2 GHz processor with 8 GB memory under the Windows 7 operating system. Experiments were repeated 10 times and the average time costs are given in Figure 4.9.

The computational cost increases with the number of particles, and it is observed that the cost of the V-SMC-PHD filter is higher than that of the AV-SMC-PHD filter, by comparing the red line with the pink dash line (for the three-speaker case), and the blue dash line with the green line (for the two-speaker case). The time required for processing Sequence 45 is expected to be higher than for Sequence 24 since the maximum number of speakers is three in Sequence 45 while it is two in Sequence 24. It is true that using

audio information brings some computational cost to our proposed tracking system. However, Figure 4.9 shows this cost is negligible and makes the proposed algorithm computationally more efficient than the baseline algorithm as the born particles are propagated only when necessary.

The following experiments aim to gauge the estimation accuracy of the algorithms. Some frames from Sequence 24 are shown in Figure 4.10. The first row shows the results of V-SMC-PHD filter, and the second row for AV-SMC-PHD filter. In the first

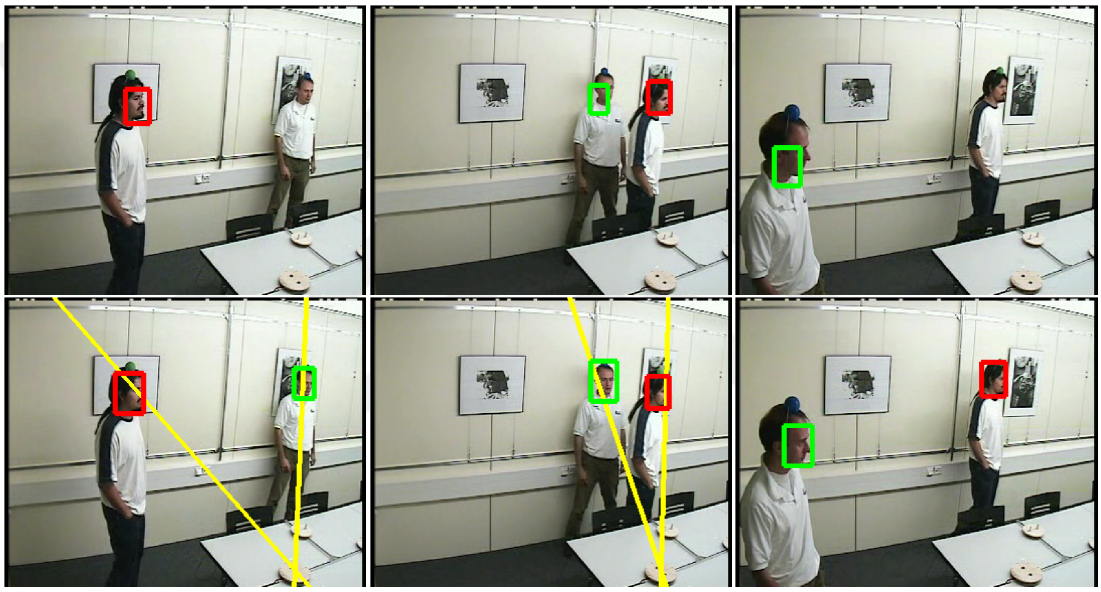


Figure 4.10: Sequence 24 camera #1: Two speakers with occlusions. The first row shows the results of the V-SMC-PHD filter and the second row shows tracking results of our proposed AV-SMC-PHD filter.

column, only one speaker is detected by the V-SMC-PHD filter while both speakers are detected by our proposed AV-SMC-PHD filter. After occlusion in the second column, our proposed AV-SMC-PHD filter tracks the speakers more accurately. There is no DOA information in the third column, but still our proposed AV-SMC-PHD filter manages to track both speakers while the V-SMC-PHD filter lost one speaker. It is because the survival particles are always dense until the DOA information is lost and more particles survive for the next frame in the AV-SMC-PHD filter. Even when the DOA information no longer exists after some points, the AV-SMC-PHD filter has still an advantage on the number of surviving particles compared to the V-SMC-PHD filter. Figure 4.11 shows the estimation of the number of speakers.

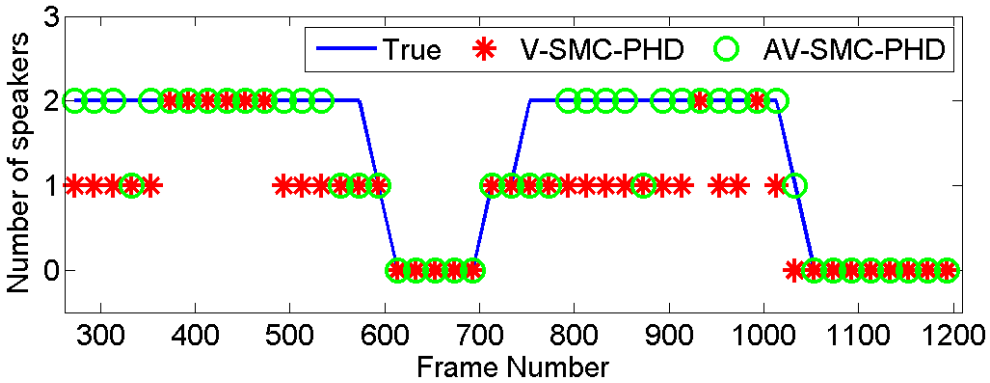
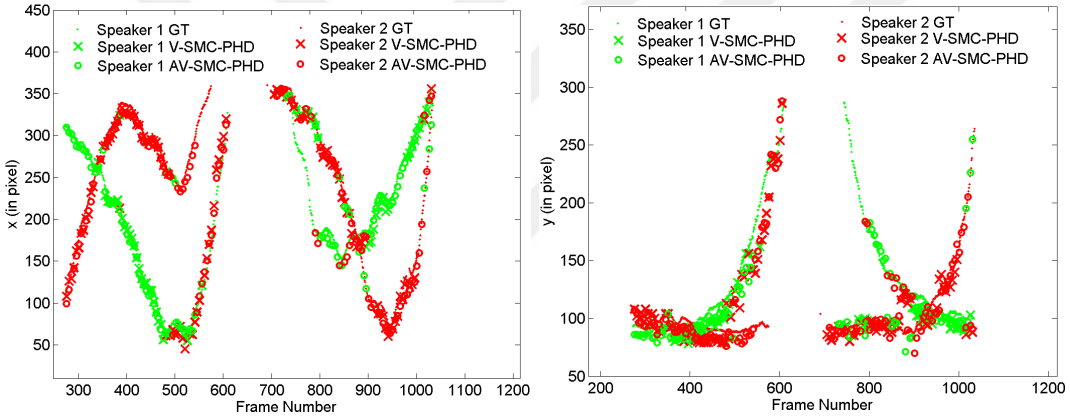


Figure 4.11: Comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in estimation of the number of speakers for Sequence 24 camera #1.



(a) (b)

Figure 4.12: Position estimates of the V-SMC-PHD and the proposed AV-SMC-PHD filters for Sequence 24 camera #1.

Here, the number of active speakers is changing from 2 to 0 and our proposed AV-SMC-PHD filter shows better performance than the V-SMC-PHD filter. For visualization, down-sampling is performed to the plots as in Section 4.4.1. The position estimates of the filters are given in Figure 4.12. The trajectories obtained by the proposed AV-SMC-PHD filter follow the ground truth trajectories closer than the V-SMC-PHD filter which shows that the position estimation of the AV-SMC-PHD filter is more accurate than that of the V-SMC-PHD filter.

The same experiments were conducted on Sequence 45 and some chosen frames are given in Figure 4.13. Here, the three speakers occlude each other many times and the AV-SMC-PHD filter is able to detect and follow all speakers even after the occlusions.

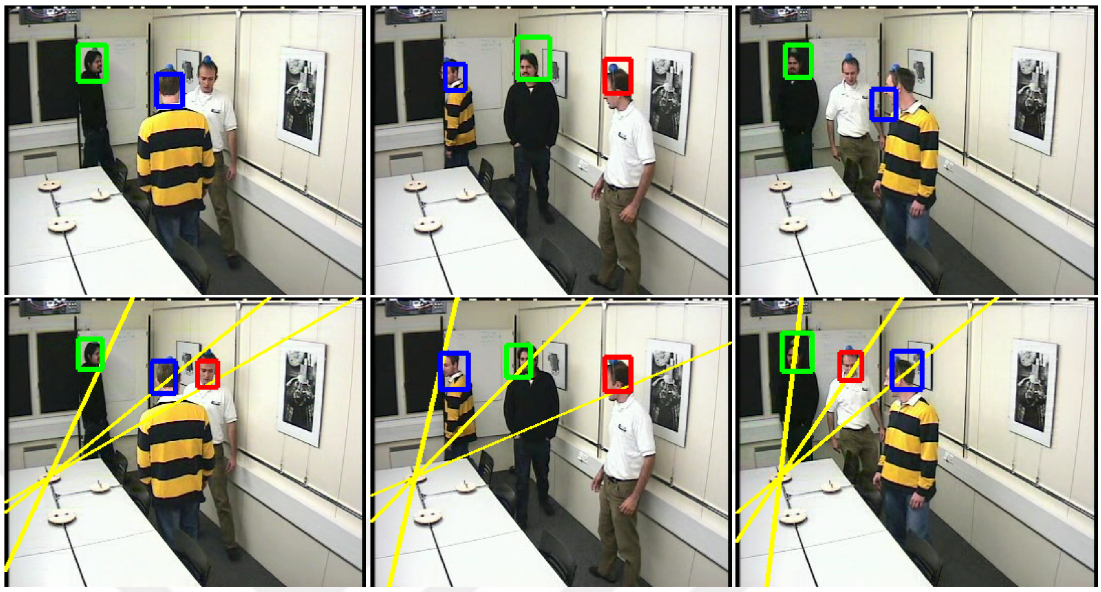


Figure 4.13: Sequence 45 camera #3: Three speakers with occlusions. V-SMC-PHD performance is shown in the first row, and the proposed AV-SMC-PHD shows better performance in the second.

The number of speakers estimated for Sequence 45 is given in Figure 4.14. It can be observed that the performance of the V-SMC-PHD filter is not as good as the AV-SMC-PHD filter.

To see the performance difference between the filters, the OSPA-T errors are plotted for Sequence 24 and Sequence 45 in Figure 4.15a and 4.15b, respectively. To get more reliable results, the experiments are repeated 10 times and the average error is plotted. In Figure 4.15a, the average OSPA-T error is 27.12 for V-SMC-PHD and 17.71 for AV-

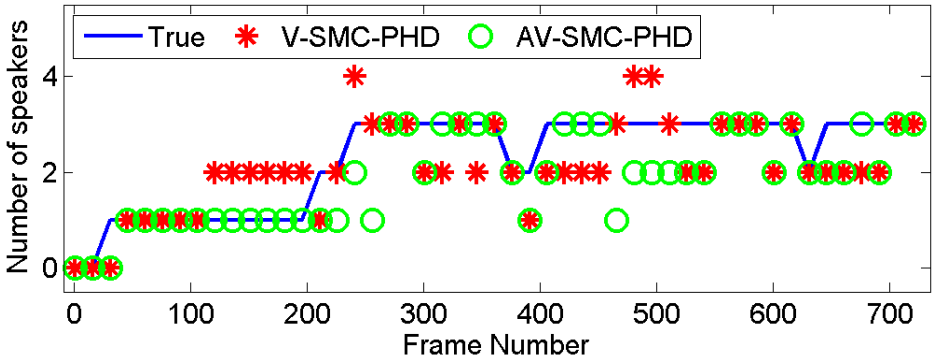


Figure 4.14: Comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in number of speakers estimation for Sequence 45 camera #3.

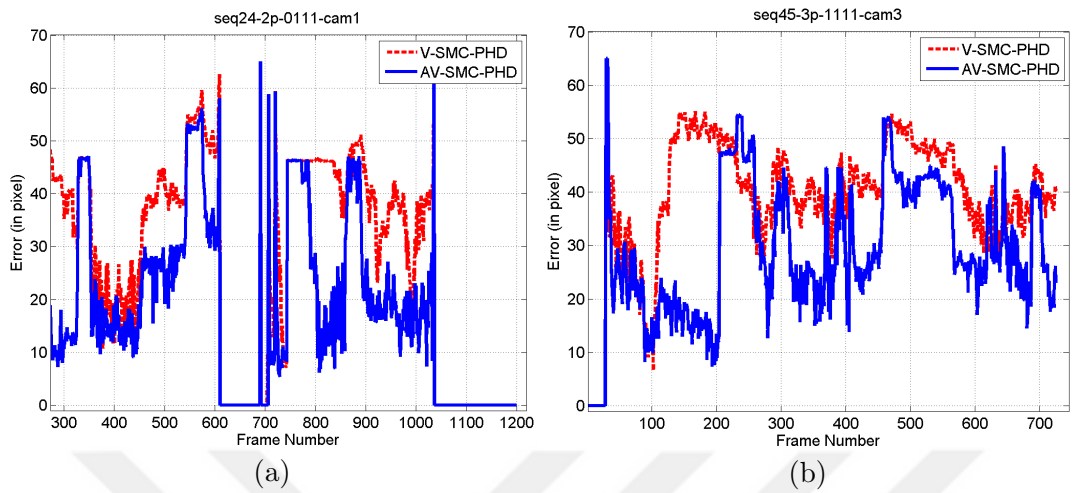


Figure 4.15: Performance comparison of the V-SMC-PHD and the proposed AV-SMC-PHD filters in terms of the OSPA-T error.

SMC-PHD. It means that AV-SMC-PHD offers a 35% improvements over V-SMC-PHD for Sequence 24. The average OSPA-T errors for Sequence 45 in Figure 4.15b are 39.09 and 28.43 for V-SMC-PHD and AV-SMC-PHD, respectively. Again, the AV-SMC-PHD filter performs better with a 27% improvement.

Other experiments are conducted to see the how changes in scale of the rectangle are related with the OSPA error and the contribution of OSPA components, position and cardinality, to the final results and the plots are given in Figures 4.16 and 4.17.

The OSPA-T error is plotted together with the change in the scale parameter in Figure 4.16 where the result of the proposed AV-SMC-PHD on Sequence 24 camera #1 is used. To be able to add another component to the OSPA metric, it is required to have ground truth information. As ground truth information of the scale parameter is not available, the scale parameter is plotted together with the OSPA-T error to see the reaction of scale parameter to change in OSPA-T error. Before coming to frame numbers 400 and 600, the scale value changes suddenly with abrupt change in the error. This is observed also around frame 900. However, this relation may not always be linear, as the scale of the rectangle centred on the face may change with the distance of the speaker to the camera. In other words, the scale of the rectangle can increase independently from the error when the speaker comes close to the camera. This is one reason why scale does not change even error drops before frame number 1000.

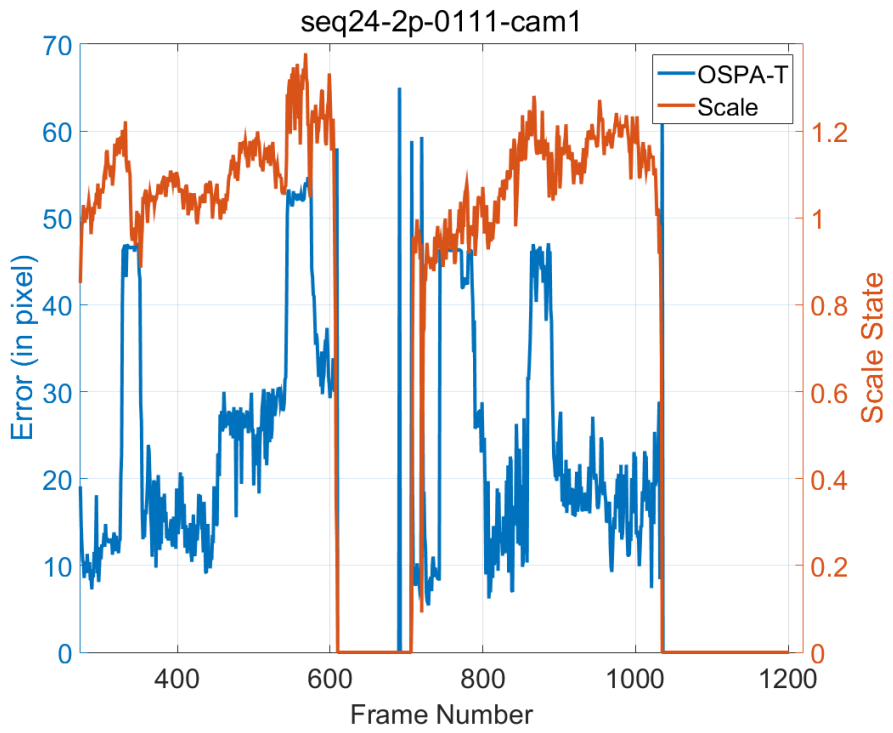


Figure 4.16: Variation of scale value and OSPA-T error for the proposed AV-SMC-PHD.

The OSPA-T error is split into its components, position and cardinality (number of speakers estimation) errors to see the contribution of its components to the final estimation error and the plots are illustrated in Figures 4.17a and 4.17b. The position error is given in Figure 4.17a where the proposed AV-SMC-PHD outperforms the V-SMC-PHD in terms of the position estimation. The cardinality error is given in Figure 4.17b. Here also, the proposed algorithm shows better performance than the baseline method. When the two plots are compared, it is seen that cardinality error is higher than the position error and it makes the cardinality error more dominant in the final error. The reason for this is the selected parameter value penalizing the cardinality error. For other applications, this value can be modified in order to balance the contribution of the position and cardinality errors to the final estimation.

The experiments are extended using other sequences of $AV16.3$ and all the experiments with standard errors are given in Table 4.2. The average error for V-SMC-PHD is 32.01, and 22.75 for AV-SMC-PHD. This result shows that with proposed approach a 29% reduction in tracking error has been achieved.

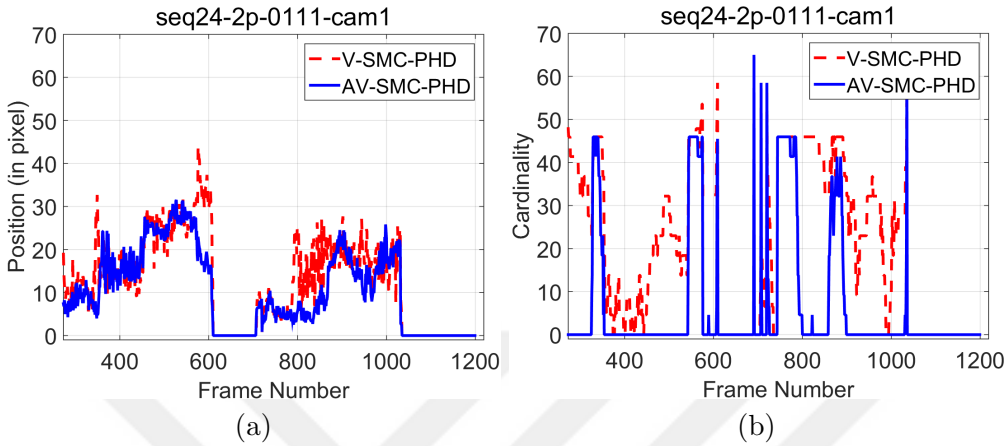


Figure 4.17: The contribution of OSPA components to the final error. Position error is shown in (a) and cardinality error is shown in (b).

Table 4.2: OSPA-T values for V-SMC-PHD and the proposed AV-SMC-PHD averaged over 10 repetitions and standard errors of the estimated OSPA-T values.

		V-SMC-PHD		AV-SMC-PHD	
		OSPA-T	SE-OSPA-T	OSPA-T	SE-OSPA-T
seq24-2p-0111	cam1	27.12	0.17	17.71	0.07
	cam2	25.91	0.23	19.83	0.28
	cam3	24.32	0.16	18.94	0.05
seq25-2p-0111	cam1	25.84	0.09	19.13	0.06
	cam2	25.66	0.3	18.47	0.11
	cam3	29.99	0.31	21.61	0.07
seq30-2p-1101	cam1	35.6	0.23	25.22	0.23
	cam2	24.97	0.1	19.37	0.05
	cam3	37.64	0.36	25.31	0.07
seq45-3p-1111	cam1	48.68	0.09	29.46	0.08
	cam2	39.24	0.19	29.47	0.04
	cam3	39.09	0.15	28.43	0.02
Average		**32.01**	**0.2**	**22.75**	**0.09**

Table 4.3: Experimental results for V-SMC-PHD, the proposed AV-SMC-PHD, Visual RFS-PF and the proposed AV RFS-PF.

		V-SMC-PHD	AV-SMC-PHD	Visual RFS-PF	AV RFS-PF
seq24-2p-0111	cam1	27.12	17.71	26.15	16.84
	cam2	25.91	19.83	24.88	19.02
	cam3	24.32	18.94	23.43	17.66
seq25-2p-0111	cam1	25.84	19.13	24.29	18.37
	cam2	25.66	18.47	25.02	16.92
	cam3	29.99	21.61	26.86	20.58
seq30-2p-1101	cam1	35.60	25.22	33.39	24.70
	cam2	24.97	19.37	23.82	18.41
	cam3	37.64	25.31	34.13	23.19
Average		**28.56**	**20.62**	**26.89**	**19.52**

It is clearly seen that adding the audio information to the visual tracker leads to an increase in performance. To see how significant the results are, the ANOVA based F-test is also applied here and it is found that $F = 12.51$ and p-value$= 0.0019$ with the degree of freedom $(1, 22)$. From the F-distribution table given in [12], the corresponding F_{crit} value for $(1, 22)$ is 4.30. It shows that the results are significant as $F > F_{crit}$ and p-value is smaller than 0.05 (for a 5% significance level).

To make comparison between the SMC-PHD algorithm discussed in this section and the RFS-PF which is discussed in the previous section, the results of these two algorithms for two-speaker cases are given in Table 4.3. As it is seen from the average results for both visual and audio-visual cases, RFS-PF performs slightly better than SMC-PHD. This is an expected outcome since the PHD filter is the first-order approximation of the RFS and it does not propagate the full multi-speaker posterior distribution like RFS. However, the PHD filter has an advantage on the computational cost side which makes it attractive for use in multi-speaker tracking, because computational complexity increases exponentially with the number of speakers in the RFS while it increases linearly in the PHD filter.

4.5 Chapter Summary

In this chapter, tracking of a variable number of speakers is studied under the RFS theory and PHD filtering. Firstly, we have proposed a random finite set approach for

tracking a variable number of speakers in a smart room environment using audio-visual measurements. The proposed RFS-PF algorithm has been evaluated on two-speaker sequences from the $AV16.3$ dataset. In that sense, three sequences with their three camera angles were selected in order to perform quantitative analysis of the proposed algorithm. With our proposed approach, the tracking error in baseline method is reduced 27%. The experimental results demonstrate that the proposed technique can reliably estimate both the number of speakers in the tracking environment and the positions of the speakers for up to three speakers within a challenging tracking scenario involving occlusions. However, as the computational complexity grows exponentially with the number of speakers, a first order approximation, PHD filter, has been considered for further enhancement. We then have proposed the audio-visual SMC-PHD approach for tracking a variable number of speakers. The efficient redistribution of the born particles based on DOA information reduces both the computational complexity and the estimation error. The proposed AV-SMC-PHD algorithm has been evaluated on twelve sequences from the $AV16.3$ dataset including three-speaker sequences. The experimental results show that with proposed approach AV-SMC-PHD, 29% reduction in tracking error compared to V-SMC-PHD has been achieved. It proofs the advantage of proposed method over baseline algorithms in terms of estimating both the number of speakers and the positions of the speakers with significant improvement in challenging tracking scenarios involving occlusion.

Chapter 5

Sparse Sampling for Mean-Shift based SMC-PHD Filtering

In the previous chapter, our proposed AV-PF, explained in Section 3.3, is adapted to PHD filter in order to track a variable number of speakers. In this chapter, we investigate and modify the proposed AV-SMC-PHD filtering algorithm aiming to improve its estimation accuracy and computational efficiency.

In particle filtering, particles are weighted based on their probability distributions which results in some particles being pruned after the resampling step because of their small weights. One way to improve the efficiency of the particles is to shift them to a local maximum of the distribution function which leads to an increase in the weights of the particles and allows more particles to survive for the next iteration. In addition, shifted particles will be closer to the speaker position, compared to their original positions which improves the estimation accuracy. In that sense, the mean-shift is a useful tool as it is searching iteratively by climbing density gradients to find the peak of the probability distribution. We first incorporate the mean-shift process into the particles of the SMC-PHD filter to shift the particles towards a closest local maximum in order to improve the estimation accuracy. In the second part of this chapter, we concentrate on reducing the computational cost which leads us to use sparse sampling, resulting in a significant reduction in processing time with a small sacrifice in terms of estimation

accuracy.

We evaluate the proposed algorithms with the same multi-speaker sequences used in the previous chapter and show their advantages over the corresponding baseline methods.

5.1 Introduction

Particle filtering has emerged as a useful tool for tracking problems. It implements recursive Bayesian filter to approximate the required posterior density by a weighted particle set. The success of particle filtering lies on its ability to maintain a good approximation to the posterior which leads to accurate estimation of the target position. A potential weakness of the particle filter, however, is that the particle set can become too diffuse which causes the particles to spread across the image plane rather than clustering around the target. This causes poor approximation of the posterior as the particles tend to migrate towards local maxima in the evaluation function with the consequence of homing on clutter and losing track of the true target. To address this issue, we propose to apply the mean-shift method to the particle set to pull the centre of the particle distribution towards the target centre which results in better approximation of the posterior as well as in improvement in estimation accuracy of the target position.

The conventional PF has originally been developed for single target tracking and it is a challenging task to track multiple targets with the PF when the number of the targets is unknown and time-varying. The proposed AV tracker in Section 3.3, was integrated with the PHD filter in Section 4.3. So, the proposed AV-SMC-PHD method gained the ability to track a variable number of targets. In this chapter, by applying the mean-shift process to all the particles of the AV-SMC-PHD filter, we present the AVMS-SMC-PHD method for robust and more efficient speaker tracking. The tracking scheme of the new algorithm is illustrated in Figure 5.1.

The proposed algorithm is designed by integrating mean-shift, particle and PHD filters to bring their respective strengths in tracking and eliminating their weaknesses by complementing one another as indicated by the arrows in Figure 5.1. Each arrow shows their respective contributions to overall tracking process.

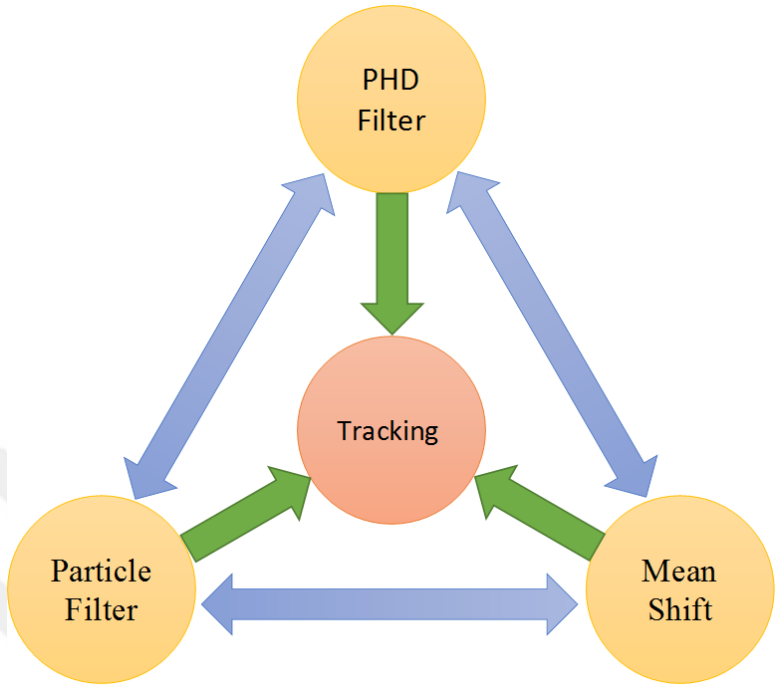


Figure 5.1: Tracking scheme of the improved tracking system via the integration of the mean-shift, particle and PHD filters.

The mean-shift, a typical and popular variational method, is a robust non-parametric method for climbing density gradients to find the peak of the probability distributions using only a single hypothesis [25], [117] . This approach is computationally efficient, but it may converge to saddle points in the case of multi-modal distribution [17, 18, 19]. Another problem with mean-shift tracking is that it may fail to track small and fast moving targets and is unable to recover a track after partial or total occlusions [32], [81], [150]. These problems are easily covered by the PF due to its ability to recover from lost tracks [81], and it uses multiple particles which can help mean-shift to detect the target even if some of the particles fall in local maxima or saddle points. Another problem with the mean-shift is about adaptation to the size or scale of the target. Although the scale is a crucial parameter for the mean-shift algorithm, there is currently no sound mechanism for choosing this scale within the tracking framework. However, this problem is solved with the SMC-PHD filter since the scale is one of the states of the target. The conventional PF and mean-shift tracking are derived for single target tracking. Although multiple trackers can be run in parallel for multi-target tracking under the assumption that the number of targets is known and constant during the

tracking. However, this is not feasible for an unknown and variable number of targets as there is no mechanism to estimate the number of trackers in parallel. The weakness of these two methods is addressed by the PHD filter which estimates the number of targets by propagating the first order multi-target posterior. However, the PHD recursion involves multiple integrals, and does not have closed-form solutions for implementation. At that point, sequential Monte Carlo (the particle filter) provides a numerical solution to the PHD filter.

By combining these three methods, the AVMS-SMC-PHD algorithm is conceived here. It facilitates robust tracking for a variable number of targets as presented in Section 5.2. Evaluations on $AV16.3$ sequences show the advantage of the proposed AVMS-SMC-PHD algorithm over the baseline methods with significant improvement in estimation accuracy.

After improving the estimation accuracy, we concentrate on optimizing the computational cost. The mean-shift process moves the particle towards the target location leading to error reduction, but repeating this process for all the particles increases computational cost. To overcome this problem, we propose a technique based on "sparse sampling" leading to a new concept: "the sparse particle". The traditional way of using sparsity in tracking is to represent the target appearance or features with sparsity. Unlike the traditional way, sparse particles are obtained with a sparse sampling strategy, which has not been done before according to the best of our knowledge. The proposed algorithm is described in Section 5.3 and tested on $AV16.3$ with the same sequences as those used for evaluating the AVMS-SMC-PHD filter, resulting in a significant reduction of the computational cost with a small sacrifice of estimation accuracy.

5.2 Improved AV Tracking with Mean Shift Based SMC-PHD Filtering

In this section, we present our proposed AVMS-SMC-PHD filter which is the integration of the mean-shift tracking with the proposed AV-SMC-PHD filter explained in the previous chapter.

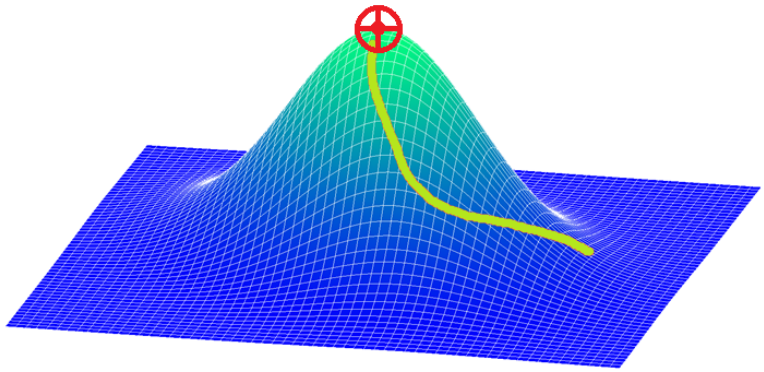


Figure 5.2: Basic mean-shift process is illustrated. A single hypothesis is shifted from base to the peak of the distribution function.

The mean-shift tracking gives a deterministic output at each step which may cause failure in location estimation of the targets due to background clutter or occlusion. A combination of the SMC-PHD filtering with the mean shift algorithm incorporates a mechanism for dealing with uncertainty within the deterministic method. The statistical property of the stochastic approach can improve the robustness of the mean-shift which leads to improvement in the overall estimation accuracy of the tracker.

The mean-shift iteration moves the particles towards the gradient ascent direction in the likelihood until they converge to their neighbouring local peaks as illustrated in Figure 5.2.

Here, the movement of a single particle is shown from the base to the peak of the distribution function. The particle is moved by iterative calculation of the mean-shift vector which points to the direction of the movement. This process is repeated until the particle is converged to the peak points.

During the mean-shift iteration, moving toward the gradient ascent may stop too early because of a local plateau or saddle point as illustrated in Figure 5.3.

Using multiple particles helps to deal with this problem, because the tracker will detect the speaker with the remaining particles even though some of the particles may fall in the saddle points. However, to prevent any of the particles from falling in the saddle points during the mean-shift iteration, a small perturbation is added to the shifted particles at each iteration.

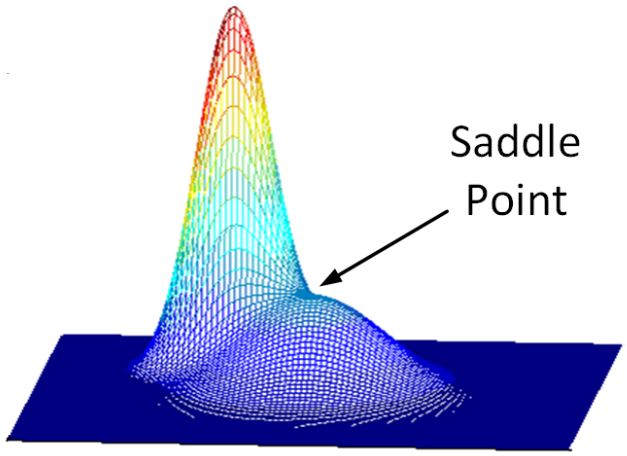


Figure 5.3: Saddle point in multi density distribution.

As mentioned before, the mean-shift is designed for single target tracking. Integration of the mean-shift to the proposed AV-SMC-PHD filter is not straightforward due to its multi-target tracking framework. To use the mean-shift method in multi-target tracking, we have modified the basic mean-shift [29] process which is described in Section 2.3.5. We denote the mean-shift process with function $\mathbb{MS}$. Here, firstly an initialization step is applied before running the mean-shift iteration over the nth particle $\tilde{\mathbf{x}}_k^{(n)}$ at time frame k. In this step, horizontal and vertical positions of $\tilde{\mathbf{x}}_k^{(n)}$ are assigned to $\mathbf{y}_0 = [\tilde{\mathbf{x}}_k^{(n)}(1), \tilde{\mathbf{x}}_k^{(n)}(3)]$ since $\tilde{\mathbf{x}} = \begin{bmatrix} x_1 & \dot{x}_1 & x_2 & \dot{x}_2 & s \end{bmatrix}^T$. The candidate speaker model, $q_u(\mathbf{y}_0)$, with the centre $\mathbf{y}_0$ is compared with all the reference models from the reference template and the closest reference model $\mathbf{r} = \{r_u\}_{u=1,\ldots,U}$ is found. Then, the selected reference model, $\mathbf{r}$, is used in the mean-shift iteration to move the particle towards the speaker in the following steps. Before running the mean-shift iteration for the particle $\tilde{\mathbf{x}}_k^{(n)}$, the Bhattacharyya coefficient $\rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right]$ is calculated. The mean-shift iteration is running in a loop and this loop is controlled with two parameters; iteration flag and iteration number. As long as these two parameters meet the conditions, the mean-shift iteration will progress. Otherwise the loop will be broken and the same process will be repeated for the next particle with iteration number $NumIter$ set to 0 and iteration flag $ContIter$ set $true$. The first step in the loop is to derive the weights according to Equation (2.32). Then, the next location of the particle $\mathbf{y}_1$ is calculated by Equation (2.36). There is a possibility that the new location $\mathbf{y}_1$ may move the particle in the wrong direction. To avoid this problem, the Bhattacharyya coefficient of the

$\mathbf{y}_1$, $\rho\left[\mathbf{q}\left(\mathbf{y}_1\right), \mathbf{r}\right]$ is calculated and compared with $\rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right]$. If $\rho\left[\mathbf{q}\left(\mathbf{y}_1\right), \mathbf{r}\right]$ is smaller than $\rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right]$, it means that $\mathbf{y}_1$ is not a good estimate and the loop will be broken. If $\rho\left[\mathbf{q}\left(\mathbf{y}_1\right), \mathbf{r}\right]$ is larger than $\rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right]$, we need to check the amount of shifting. If $\|\mathbf{y}_1 - \mathbf{y}_0\|$ is larger than the threshold value ζ, $\mathbf{y}_1$ will be set to $\mathbf{y}_0$ and used for the next iteration as long as $NumIter$ is smaller than $MaxIter$. If $\|\mathbf{y}_1 - \mathbf{y}_0\|$ is smaller than ζ, iteration will be broken again. This process will be repeated for all the particles. The pseudo code of the $\mathbb{MS}$ is presented in Algorithm 5.1.

```
Given: x̃_k, N_{k-1}, J_k, MaxIter, U, ζ
for n = 1, ...N_{k-1} + J_k do
    Assign position coordinates of the particle to y_0 = [x̃_k^(n)(1), x̃_k^(n)(3)]
    Find the closest reference model r_u for x̃_k^n by comparing the candidate speaker
    model q_u(y_0) and the reference models.
    Evaluate: ρ[q(y_0), r] = Σ_{u=1}^{U} √(q_u(y_0) r_u)
    Set iteration number: NumIter = 0;
    Set iteration flag: ContIter = true;
    while ContIter == True or NumIter < MaxIter do
        Derive the weights according to Equation (2.32)
        Find the next location y_1 by Equation (2.36)
        Compute: ρ[q(y_1), r] = Σ_{u=1}^{U} √(q_u(y_1) r_u)
        // Continue mean-shift iteration as long as the Bhattacharyya coefficient goes
        up. Otherwise, stop iteration
        if ρ[q(y_1), r] > ρ[q(y_0), r] then
            // If position change exceeds the threshold value ζ, then continue
            mean-shift iteration. Otherwise, stop iteration
            if ‖y_1 − y_0‖ > ζ then
                y_0 = y_1;
                ContIter = true;
            else
                ContIter = false;
            end
        else
            y_1 = y_0
            ContIter = false;
        end
        NumIter = NumIter + 1;
    end
    [x̃_k^(n)(1), x̃_k^(n)(3)] = y_1;
end
```

Algorithm 5.1: $\mathbb{MS}$ function for the mean-shift iteration.

To ensure that $\mathbb{MS}$ is applied to all the particles, this step is performed after the audio contribution is taken into account and all types of particles, born, spawned and

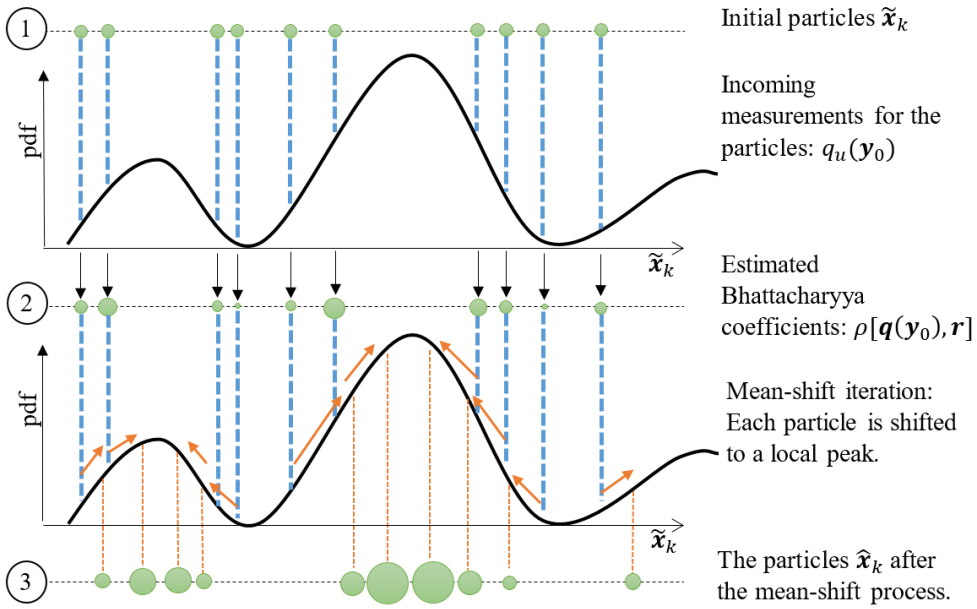


Figure 5.4: The mean-shift process for the particles.

surviving, are combined as $\tilde{\mathbf{x}}_k$. Applying $\mathbb{MS}$ after the DOA check ensures that all the particles are shifted even for the case where the DOA does not exist.

To summarise the main purpose of the mean-shift iteration, the $\mathbb{MS}$ process is illustrated in Figure 5.4. Here, 10 particles $\tilde{\mathbf{x}}_k^{(n)}$, $n = 1, ..., 10$, are given in the first stage. These particles have different Bhattacharyya coefficients $\rho\left[\mathbf{q}\left(\mathbf{y}_0\right), \mathbf{r}\right]$ in the second stage when their measurements are compared with the reference speaker model. After this step, we run the mean-shift iteration to shift the particles to the local maxima of the measurement function in the third stage. With this proposed step, the particles $\hat{\mathbf{x}}_k$ that exhibit higher values of the Bhattacharyya coefficient are likely to be much closer to the speaker position.

In the end, the shifted particles maintain a fair representation of the modes of the distribution, and provide good local characterization of the likelihood which allows the use of a fewer particles to maintain the multi-mode distribution [116]. After the $\mathbb{MS}$, the weight prediction step is performed and the remaining steps are followed as in the AV-SMC-PHD filter. The pseudo code of the proposed AVMS-SMC-PHD filtering algorithm is depicted in Algorithm 5.2.

Initialize: η, σ^2, U, T, $\mathbf{F}$, λ, r, p_D, u, c, a, σ_c^2, p_S, $MaxIter$, ζ, k, K, N_0, $\mathbf{x}_0$
while $k < K$ **do**
  For $n = 1, ..., N_{k-1}$ sample $\tilde{\mathbf{x}}_k \sim q_k\left(\cdot|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k\right)$, where $\tilde{\mathbf{x}}_k \in \mathbb{R}^{5\times N_{k-1}}$
  Propagate surviving and spawned particles: $\tilde{\mathbf{x}}_{s,k} = \mathbf{F}\tilde{\mathbf{x}}_k + \mathbf{q}_k$
  Get the corresponding DOA angle θ_k
  if *DOA exists* **then**
    // For surviving and spawned particles
    Calculate distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & ... & d_k^{(N_{k-1})} \end{bmatrix}^T$
    Calculate movement distances $\hat{\mathbf{d}}_k$ using Equation (4.25)
    Concentrate $\tilde{\mathbf{x}}_{s,k}$ around the DOA line : $\tilde{\mathbf{x}}_{s,k} = \tilde{\mathbf{x}}_{s,k} \oplus \mathbf{h}_k\hat{\mathbf{d}}_k$
    if *new speaker* **then**
      // For born particles
      For $n = N_{k-1}+1, ..., N_{k-1}+J_k$ sample $\tilde{\mathbf{x}}_{b,k} \sim p_k\left(\cdot|\mathcal{Z}_k\right)$ uniformly around the DOA line
    end
  end
  Combine all the particles: $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_{s,k} \cup \tilde{\mathbf{x}}_{b,k}$
  Apply the mean-shift iteration: Each particle will converge to a nearby local mode of the observation distribution. $\hat{\mathbf{x}}_k = \mathbb{MS}\left(\tilde{\mathbf{x}}_k\right)$
  Prediction: For $n = 1, ..., N_{k-1}+J_k$ calculate $\tilde{w}_{k|k-1}^{(n)}$ using Equation (2.21)
  Estimate colour likelihood using Equation (4.24)
  Update: For $n = 1, ..., N_{k-1}+J_k$ calculate $\tilde{w}_k^{(n)}$ using Equation (2.22)
  Calculate the total mass $\hat{\Xi}_{k|k} = \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_k^{(n)}$
  Resampling: Resample $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \hat{\mathbf{x}}_k^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$ to get $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$ where $N_k = \eta\hat{\Xi}_{k|k}$
  Multiply the weights by $\hat{\Xi}_{k|k}$ to get $\left\{\tilde{w}_k^{(n)}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$
  Cluster the particles and get the positions of the speakers
  $k = k + 1$
end

Algorithm 5.2: Proposed AVMS-SMC-PHD filtering algorithm.

5.3 Computational Cost Refinement with Sparse Particles

In the previous section, the mean-shift is integrated into the AV-SMC-PHD tracking algorithm by applying the mean-shift iteration to all the particles. This leads to a reduction in the required number of particles and in estimation error. However, the reduction in particle numbers does not guarantee a lower computational cost, because applying the mean-shift iteration to all the particles is time consuming [150]. To address this issue, a sparse sampling method is proposed to reduce the computational complexity. Sparse sampling is performed on the particle set in order to choose a subset of particles on which the mean-shift iteration is performed instead of the full set particles. This reduces the computational cost significantly with a small sacrifice in estimation accuracy.

Our sparse sampling method is inspired by KLD-Sampling [40], [41], [42] which is proposed for adaptive estimation of the number of particles to run in particle filter based tracking. Detailed information has been given in Chapter 2. Unlike the KLD-Sampling, here one dimensional bins, $\mathbb{B}$ with τ subintervals are created on the interval $[0, 1]$. Each subinterval is denoted by $\mathbb{B}_i$ with range of $[\{1-(i-1)/\tau\},\ \{1-i/\tau\}]$. The critical question here is how to determine the value of τ as the bin number of $\mathbb{B}$ affects the estimation of sparse sampling. Based on experimental studies, which will be detailed in Section 5.4.2, we conclude that τ can be set to a constant number of particles per speaker, η, as a practical choice.

As a demonstration of the proposed sampling algorithm, in Figure 5.5a, a set of 10 particles is given and $\mathbb{B}$ for this set is illustrated in Figure 5.5b. We can consider η as 10 which means τ can be set to 10 as well. So, each bin is divided into $1/10 = 0.1$ subintervals and these subintervals are sorted in a descending order as it is shown on the left side of Figure 5.5b. The top subinterval has a range between 0.9 to 1 while the bottom subinterval has a range between 0 to 0.1.

Initially, each bin is set to $\mathbb{B}_{i=1,\ldots,\tau} = 0$, as they are all empty at the beginning. For each particle, the Bhattacharyya coefficient, ρ is estimated using Equation (2.29). Their

values vary between $[0, 1]$ where 1 is the best similarity matching between reference and candidate histograms, and 0 is the worst. After ρ is estimated, its corresponding range $\mathbb{B}_i$ is found and updated from 0 to 1. Then, the number of particles, $\mathfrak{N}_i$, in $\mathbb{B}_i$ is increased by one. At the end, the Bhattacharyya coefficients of all the particles are estimated and $\mathbb{B}$ becomes as shown in Figure 5.5c. This representation resembles sparse categorization of the particles, that is why the proposed sampling algorithm is called "sparse sampling". The sparse particles are chosen from these subintervals where all the particles are stored based on the Bhattacharyya coefficient. Here, the number of total true bins, b, is 6 as $b = \sum_{i=1}^{\tau} \mathbb{B}_i$, because only 6 subintervals are updated from 0 to 1 in Figure 5.5c.

To estimate the number of particles, $\mathscr{N}$, based on the bin number, Equation (2.8) is employed and $\mathscr{N}$ is estimated as 5. It means only 5 particles will be chosen from the right side of Figure 5.5c and this selection starts from top of $\mathbb{B}$ as the particles are sorted according to their Bhattacharyya coefficients in a descending order. In Figure 5.5d, $\mathbb{B}$ is updated by removing all the particles except those in the first t top subintervals where t is the upper bound for $\sum_{i=1}^{t} \mathfrak{N}_i = \mathscr{N}$. Here, t is 5 since only first 5 subintervals contain $\mathscr{N} = 5$ particles. These five particles can be used as the sparse particles. However, we take one more step in the selection of the sparse particles. It is not necessary to have more than one particle from same interval as the mean-shift process will run afterwards and move the particles from same interval to the same local maxima. So, one particle from each interval with a non-zero number of particles is chosen and added to the sparse particles $\bar{\mathbf{x}}_k$ as illustrated in Figure 5.5e.

We denote the sparse sampling with function $\mathbb{SS}$, and it is incorporated into the AVMS-SMC-PHD filter before the $\mathbb{MS}$ operation. After incorporating the DOA contribution, all the particles are combined as $\tilde{\mathbf{x}}_k$ and then $\mathbb{SS}$ is performed;

$$\bar{\mathbf{x}}_k = \mathbb{SS}\left(\tilde{\mathbf{x}}_k\right) \tag{5.1}$$

to obtain the sparse particles $\bar{\mathbf{x}}_k$. The pseudo code of the $\mathbb{SS}$ is presented in Algorithm 5.3.

In the sequel, the $\mathbb{MS}$ operation is run just for $\bar{\mathbf{x}}_k$ which has fewer particles than $\tilde{\mathbf{x}}_k$.

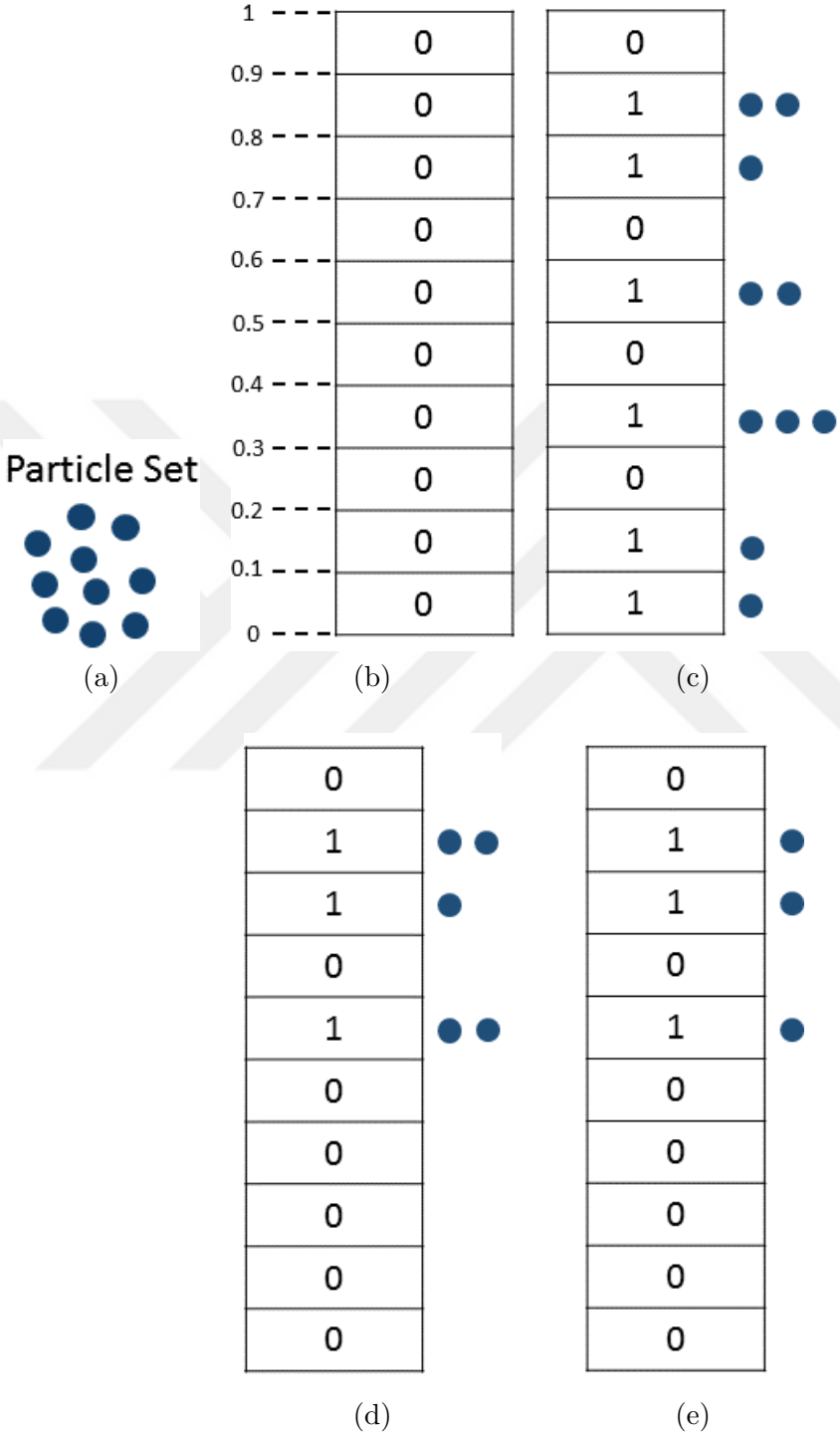


Figure 5.5: A set of 10 particles is given in (a). Initial $\mathbb{B}$ is illustrated in (b), and the distribution of the particles on $\mathbb{B}$ according to Bhattacharyya coefficients is shown in (c). The updated $\mathbb{B}$ after the number of particle estimation with KLD sampling is given in (d). Sparse particle selection is shown in (e).

Given: $\tilde{\mathbf{x}}_k$, N_{k-1}, J_k, τ, ϵ, $z_{1-\delta}$
Create $\mathbb{B}$ with τ subintervals.
for $j = 1, ...N_{k-1} + J_k$ **do**
  Calculate ρ_j using Equation (2.29)
  Find i where $\rho_j \in [\{1-(i-1)/\tau\}, \{1-i/\tau\}]$
  Set $\mathbb{B}_i = 1$
  Increase the particle counter for $\mathbb{B}_i$, $\mathfrak{N}_i = \mathfrak{N}_i + 1$
end
Estimate the total bin number. $b = \sum_{i=1}^{\tau} \mathbb{B}_i$
Calculate $\mathscr{N}$ using Equation (2.8)
Choose the sparse particles $\bar{\mathbf{x}}_k$ by taking one particle from the subintervals $\mathbb{B}_{i=1:t}$ where t is the upper bound for $\sum_{i=1}^{t} \mathfrak{N}_i = \mathscr{N}$

Algorithm 5.3: $\mathbb{SS}$ function for sparse sampling.

With the sparse sampling, the number of particles used in the mean-shift iteration is reduced which leads to a significant reduction in the computational cost. The pseudo code of the proposed sparse-AVMS-SMC-PHD filtering algorithm is depicted in Algorithm 5.4.

5.4 Experimental Results and Analysis

This section contains two parts: evaluations of the proposed AVMS-SMC-PHD and the proposed sparse-AVMS-SMC-PHD methods. Both algorithms are tested on four sequences; Sequence 24, Sequence 25, Sequence 30 and Sequence 45, with all three different camera angles from $AV16.3$ for two and three speaker cases. The OSPA-T metric [107] is used for performance evaluation. Each experiment is repeated 10 times and the average results are presented in plots and tables.

5.4.1 AVMS-SMC-PHD

In this subsection, we show the simulation results for the proposed AVMS-SMC-PHD algorithm and compare them with the results given in the previous chapter.

For the proposed AVMS-SMC-PHD filter, the parameters for the SMC-PHD part are the same as given in Section 4.4.2, namely: $p_D = 0.98$, $p_S = 0.99$, $\lambda = 0.26$, the number of histogram bins, $U = 16$ and $\sigma_c = 0.1$. The uniform density u is $(360 \times 280)^{-1}$ and

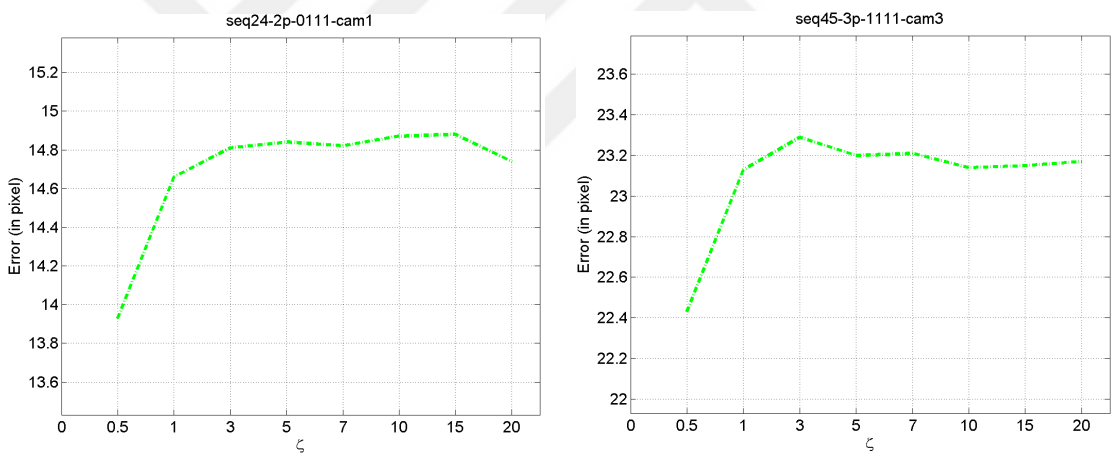
Initialize: η, σ^2, U, T, $\mathbf{F}$, λ, r, p_D, u, c, a, σ_c^2, p_S, $MaxIter$, ζ, ϵ, $z_{1-\delta}$, k, K, N_0, $\mathbf{x}_0$
while $k < K$ **do**
  For $n = 1, ..., N_{k-1}$ sample $\tilde{\mathbf{x}}_k \sim q_k\left(\cdot|\mathbf{x}_{k-1}^{(n)}, \mathcal{Z}_k\right)$, where $\tilde{\mathbf{x}}_k \in \mathbb{R}^{5\times N_{k-1}}$
  Propagate surviving and spawned particles: $\tilde{\mathbf{x}}_{s,k} = \mathbf{F}\tilde{\mathbf{x}}_k + \mathbf{q}_k$
  Get the corresponding DOA angle θ_k
  if *DOA exists* **then**
    // For surviving and spawned particles
    Calculate distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & ... & d_k^{(N_{k-1})} \end{bmatrix}^T$
    Calculate movement distances $\hat{\mathbf{d}}_k$ using Equation (4.25)
    Concentrate $\tilde{\mathbf{x}}_{s,k}$ around the DOA line : $\tilde{\mathbf{x}}_{s,k} = \tilde{\mathbf{x}}_{s,k} \oplus \mathbf{h}_k\hat{\mathbf{d}}_k$
    if *new speaker* **then**
      // For born particles
      For $n = N_{k-1}+1, ..., N_{k-1}+J_k$ sample $\tilde{\mathbf{x}}_{b,k} \sim p_k\left(\cdot|\mathcal{Z}_k\right)$ uniformly around DOA line
    end
  end
  Combine all the particles: $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_{s,k} \cup \tilde{\mathbf{x}}_{b,k}$
  Apply sparse sampling: $\bar{\mathbf{x}}_k = \mathbb{SS}\left(\tilde{\mathbf{x}}_k\right)$
  Apply mean-shift iteration: Each particle will converge to a nearby local mode of the observation distribution. $\hat{\mathbf{x}}_k = \mathbb{MS}\left(\bar{\mathbf{x}}_k\right)$
  Prediction: For $n = 1, ..., N_{k-1}+J_k$ calculate $\tilde{w}_{k|k-1}^{(n)}$ using Equation (2.21)
  Estimate colour likelihood using Equation (4.24)
  Update: For $n = 1, ..., N_{k-1}+J_k$ calculate $\tilde{w}_k^{(n)}$ using Equation (2.22)
  Calculate the total mass $\hat{\Xi}_{k|k} = \sum_{n=1}^{N_{k-1}+J_k} \tilde{w}_k^{(n)}$
  Resampling: Resample $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \hat{\mathbf{x}}_k^{(n)}\right\}_{n=1}^{N_{k-1}+J_k}$ to get $\left\{\tilde{w}_k^{(n)}/\hat{\Xi}_{k|k}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$ where $N_k = \eta\hat{\Xi}_{k|k}$
  Multiply the weights by $\hat{\Xi}_{k|k}$ to get $\left\{\tilde{w}_k^{(n)}, \mathbf{x}_k^{(n)}\right\}_{n=1}^{N_k}$
  Cluster the particles and get the positions of the speakers
  $k = k + 1$
end

Algorithm 5.4: Proposed sparse-AVMS-SMC-PHD filtering algorithm.

the number of particles per speaker is $\eta = 50$. The cut-off parameter $c = 65$, the OSPA-T metric order parameter $a = 2$. T and $\mathbf{F}$ are the parameters for the motion model defined in Section 3.5.

In the mean-shift method, the threshold for shifting distance is $\zeta = 0.5$ pixel and the number of maximum iterations to converge to local maxima is set to 6. These parameters are set based on extensive experimental studies. Note that changing these parameters slightly will not lead to a significant difference in tracking results as shown next.

First, we change the value of ζ from 0.5 to 20 in order to see its effect on tracking error. We use Sequence 24 camera #1 for the two-speaker case and Sequence 45 camera #3 for the three-speaker case as illustrated in Figures 5.6a and 5.6b, respectively.



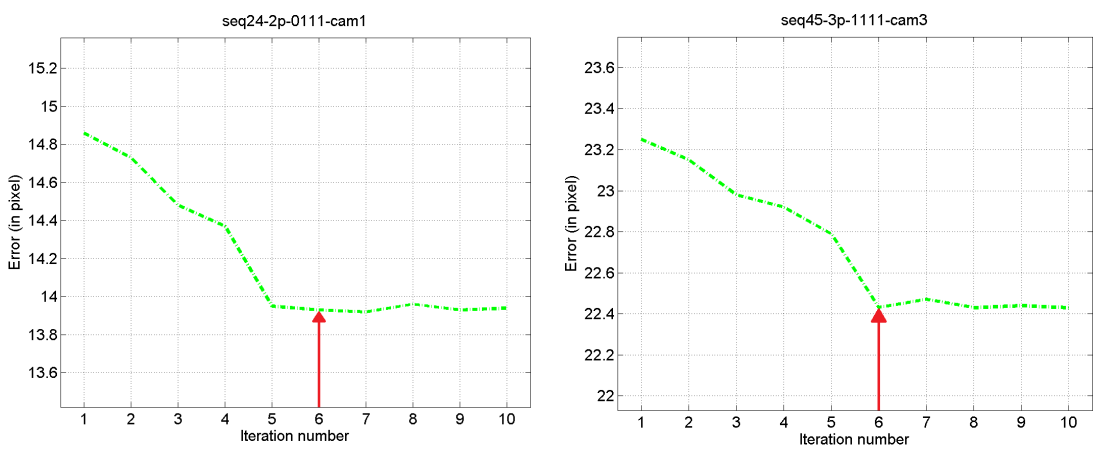
(a) (b)

Figure 5.6: Error sensitivity with respect to ζ in the mean-shift process.

One of the control parameters of the mean-shift iteration loop explained in Section 5.2 is ζ. The value of ζ is critical as bigger values may cause the algorithm to exit the loop early and increase the tracking error. As long as the iteration number does not reach the maximum iteration number, if the shifting distance from initial position to current position $\|\mathbf{y}_1 - \mathbf{y}_0\|$ is larger than ζ, then the mean-shift loop may continue for the next iteration. However, if ζ is not set to an adequate value, the loop will be broken even if there is a reasonable shifting distance. This is evident in Figures 5.6a and 5.6b. The tracking error increases with ζ values in both experiments. However, it is not linear and it tends to an asymptotic convergence for higher values of ζ. There are two

reasons for that. The first is the effect of the second control parameter over the loop which is the maximum iteration number. The loop may be broken since the iteration number reaches the maximum iteration value. The second reason may be the value of ζ. The most common observation [29] in the mean-shift iteration is that the shifting distance is getting smaller at every iteration. After some iterations, it may reduce dramatically and cause loop breaking, because the shifting distance cannot exceed ζ. To get maximum benefit from the mean-shift iteration, we observe that ζ should be as small as possible. It is 0.5 in our application as we check the distance in pixels. This value should be modified for applications where the tracking error is measured with other units like metres or inches.

Another critical parameter for the mean-shift iteration is the maximum number of iterations. We change the maximum iteration number from one to ten and measure the OSPA-T error by running the proposed AVMS-SMC-PHD algorithm in Sequence 24 camera #1 and Sequence 45 camera #3 as illustrated in Figures 5.7a and 5.7b, respectively. As it is seen from Figure 5.7a, error is decreasing by the iteration number.



(a) (b)

Figure 5.7: Error change with respect to the iteration number in the mean-shift process.

In the mean-shift process, the particles are shifted to local maximum step by step, where each step is counted as an iteration. At every iteration, each particle is shifted until it reaches the local maximum or the movement distance becomes less than ζ. That is why the tracking error is reduced until the maximum iteration number reaches 5. After 5 iterations, the tracking error does not change much, because the mean-shift process

stops when the movement distance is less than ζ, although the iteration number does not reach the maximum iteration number. The same observation can be made in Figure 5.7b as the tracking error is again decreasing with the iteration number. However, here it stabilizes after the iteration number reaches 6. Based on these two experimental studies, the maximum iteration number is set to 6 as a practical choice. In principle, setting the maximum iteration number higher can be considered to reduce the tracking error. However, this causes additional computational cost. Optimizing the maximum iteration number is a trade off between the tracking error and computational cost.

After setting the parameters, the proposed AVMS-SMC-PHD algorithm was tested on $AV16.3$. As preliminary results, error comparisons for Sequence 24 camera #1 and Sequence 45 camera #3 are given in Figures 5.8a and 5.8b, respectively. These sequences are chosen as they are the most challenging in terms of movement speed and the number of occlusions.

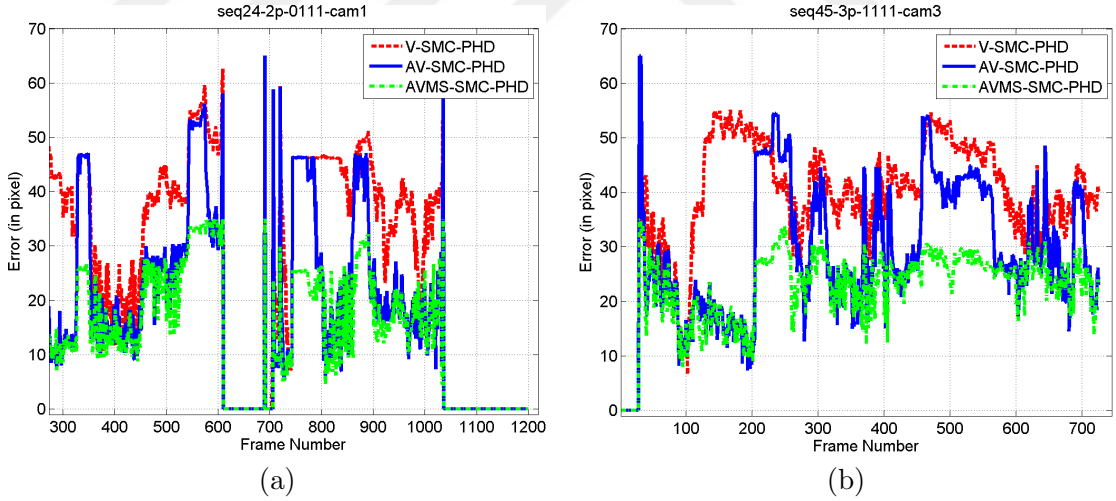
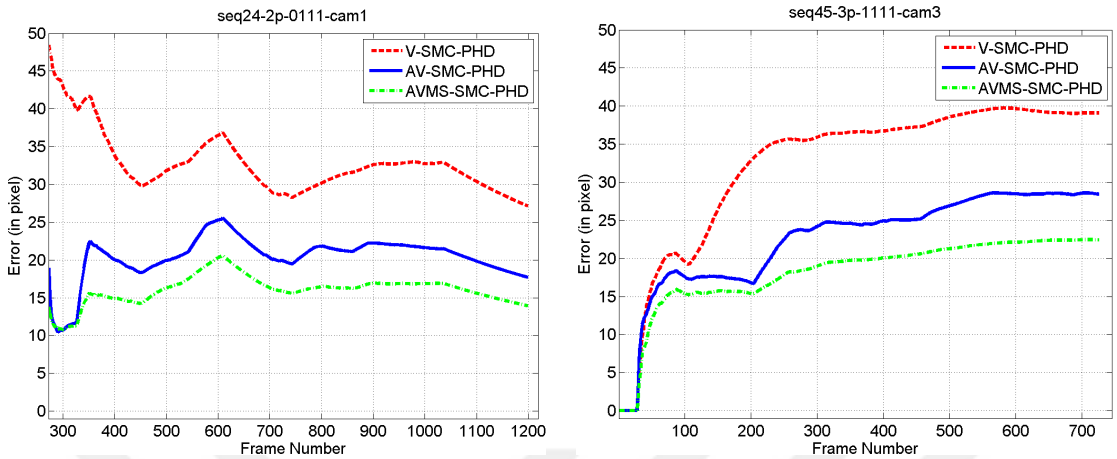


Figure 5.8: Performance comparisons of the V-SMC-PHD with proposed algorithms AV-SMC-PHD and AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3 are given in (a) and (b), respectively.

Our proposed AVMS-SMC-PHD algorithm is depicted by the green dashed line. The contribution of the mean-shift method to the AV-SMC-PHD algorithm for reducing the estimation accuracy is obvious in Figures 5.8a and 5.8b. For most of the tracking, the AVMS-SMC-PHD filter outperforms the V-SMC-PHD and AV-SMC-PHD filters in terms of estimation accuracy. For clarity, mean absolute OSPA-T errors are also



(a) (b)

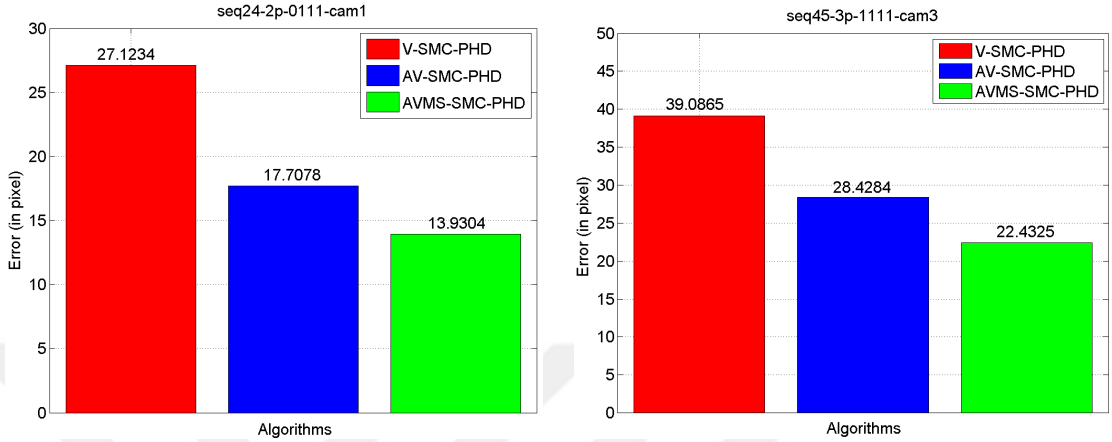
Figure 5.9: Mean absolute OSPA-T errors of the V-SMC-PHD and the proposed algorithms AV-SMC-PHD and AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3 are presented in (a) and (b), respectively.

plotted in Figures 5.9a and 5.9b.

The average errors achieved on Sequence 24 camera #1 are 27.12, 17.71, 13.93, respectively for the V-SMC-PHD, AV-SMC-PHD and AVMS-SMC-PHD algorithms. The AVMS-SMC-PHD algorithm outperforms AV-SMC-PHD by 21% and V-SMC-PHD by 35%. For Sequence 45 camera #3, the average errors are 39.09, 28.43, 22.43, respectively for the V-SMC-PHD, AV-SMC-PHD and AVMS-SMC-PHD algorithms. Here, AVMS-SMC-PHD shows better performance than V-SMC-PHD and AV-SMC-PHD by 43% and 21%, respectively.

The average errors of both sequences are depicted in Figures 5.10a and 5.10b. In both two-speaker (Sequence 24) and three-speaker (Sequence 45) cases, the AVMS-SMC-PHD performs best followed by the AV-SMC-PHD and the V-SMC-PHD. However, all the three algorithms perform better on Sequence 24 rather than Sequence 45. This result is expected as the three-speaker sequence is more complex in terms of the movement of the speakers and the number of occlusions which cause an increase in the estimation error.

Then, the number of sequences was increased for consistency with the previous chapter to evaluate the AVMS-SMC-PHD algorithm. The overall average results with standard errors are given in Table 5.1.



(a) (b)

Figure 5.10: Average errors of the V-SMC-PHD and the proposed algorithms AV-SMC-PHD and AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3 are given in (a) and (b), respectively.

Table 5.1: OSPA-T values for V-SMC-PHD, AV-SMC-PHD and AVMS-SMC-PHD averaged over 10 repetitions and standard errors of the estimated OSPA-T values.

		V-SMC-PHD		AV-SMC-PHD		AVMS-SMC-PHD	
		OSPA-T	SE-OSPA-T	OSPA-T	SE-OSPA-T	OSPA-T	SE-OSPA-T
seq24	cam1	27.12	0.17	17.71	0.07	13.93	0.04
	cam2	25.91	0.23	19.83	0.28	14.97	0.06
	cam3	24.32	0.16	18.94	0.05	14.12	0.06
seq25	cam1	25.84	0.09	19.13	0.06	15.72	0.05
	cam2	25.66	0.3	18.47	0.11	13.93	0.03
	cam3	29.99	0.31	21.61	0.07	17.07	0.04
seq30	cam1	35.6	0.23	25.22	0.23	16.65	0.03
	cam2	24.97	0.1	19.37	0.05	14.86	0.03
	cam3	37.64	0.36	25.31	0.07	19.29	0.05
seq45	cam1	48.68	0.09	29.46	0.08	22.95	0.05
	cam2	39.24	0.19	29.47	0.04	21.47	0.05
	cam3	39.09	0.15	28.43	0.02	22.43	0.03
Average		**32.01**	**0.2**	**22.75**	**0.09**	**17.28**	**0.04**

In the previous chapter, it is shown that the AV-SMC-PHD outperforms the V-SMC-PHD algorithm by 29%. Table 5.1 shows that AVMS-SMC-PHD improves the estimation accuracy by 24% and 46% compared to the AV-SMC-PHD and V-SMC-PHD algorithms, respectively.

In the AVMS-SMC-PHD algorithm, the mean-shift method is adopted to move the particles to the local maxima of the distribution function which improves the accuracy of the speaker detection. As it can be seen from Table 5.1, the tracking error of the AVMS-SMC-PHD algorithm is lower than that of the AV-SMC-PHD and V-SMC-PHD algorithms. This is because the shifted particles are closer to the ground truth position of the speaker than the particles used in the conventional way.

The ANOVA based F-test is also applied to the results of Table 5.1. First, the results of the AVMS-SMC-PHD algorithm are compared with those of the V-SMC-PHD algorithm. We found that $F = 35.59$ and p-value$= 5.28 \times 10^{-6}$ with the degree of freedom $(1, 22)$. Then, same process is repeated for the AVMS-SMC-PHD and AV-SMC-PHD algorithms, and we found that $F = 11.12$ and p-value$= 3 \times 10^{-3}$ with the degree of freedom $(1, 22)$. For both comparisons, the degree of freedom is same and the corresponding F_{crit} value for $(1, 22)$ is obtained from the F-distribution table given in [12] as 4.30. As a result, the results of the AVMS-SMC-PHD algorithm are significantly better in both comparisons as $F > F_{crit}$ and p-value is smaller than 0.05 (for a 5% significance level).

5.4.2 Sparse-AVMS-SMC-PHD

The evaluation results of the sparse-AVMS-SMC-PHD and a comparison with the previously proposed algorithms are given in this subsection. The same sequences from $AV16.3$, as in the above subsection, are used to test the proposed sparse-AVMS-SMC-PHD algorithm. All the parameters of AVMS-SMC-PHD, given in the previous section, are also used in the sparse-AVMS-SMC-PHD algorithm.

A key parameter in the proposed sparse-AVMS-SMC-PHD algorithm is the bin dimension τ which may cause either performance failure, or increase the computation cost, depending on its size. To get a practical guidance for the selection of τ, simulations

were conducted on Sequence 24 camera #1 and Sequence 45 camera #3 as illustrated in Figure 5.11.

In these experiments, four different values for the constant number of particles per speaker, η, are chosen and OSPA-T errors are measured under different τ. In Figure 5.11a, η is set to 25 and τ is varied from 10 to 100. The tracking error starts to decrease with the increase in τ, as τ affects the total bin number and consequently the estimated sparse particles. Larger τ allows the selection of more sparse particles, but the objective here is to find the optimal number of sparse particles. Fewer sparse particles cause poor performance, but more sparse particles do not contribute much to tracking performance. That is why the tracking error becomes stable just before τ reaches 25. Another experiment with η equal 50 is given in Figure 5.11b. The tracker error decreases until τ reaches 50, then after small oscillations, it becomes constant. Better illustration is evident in Figure 5.11c where $\eta = 75$ and the tracking error is reduced until τ increases to 75. The error becomes stable after that. A similar behaviour to Figure 5.11a is also observed in Figure 5.11d where the tracking error is stabilized just before τ reaches the same values as η. The optimal dimension of the bins is not known in reality. In practice, it seems reasonable to set τ as η based on the experiments given in Figure 5.11. In general, setting τ to a larger value leads to a better performance. However, this also results in extra computation cost as each subinterval needs to be managed during the selection of sparse sampling. In other words, the tuning of τ is a trade off between the tracking accuracy and the computational cost.

As the motivation for using sparse sampling is to reduce the computational cost of the AVMS-SMC-PHD filter, we measure the computational cost of the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD filters by running all the filters on Sequence 24 camera #1 as illustrated in Figure 5.12. Here, the experiments are run on Intel core i7 2.2 GHz processor with 8 GB memory under the Windows 7 operating system.

The mean-shift integration to the AV-SMC-PHD filter causes a dramatic increase in computational cost. However, using sparse particles in the mean-shift iteration reduces the computational cost significantly by approximately 10 times as sparse sampling

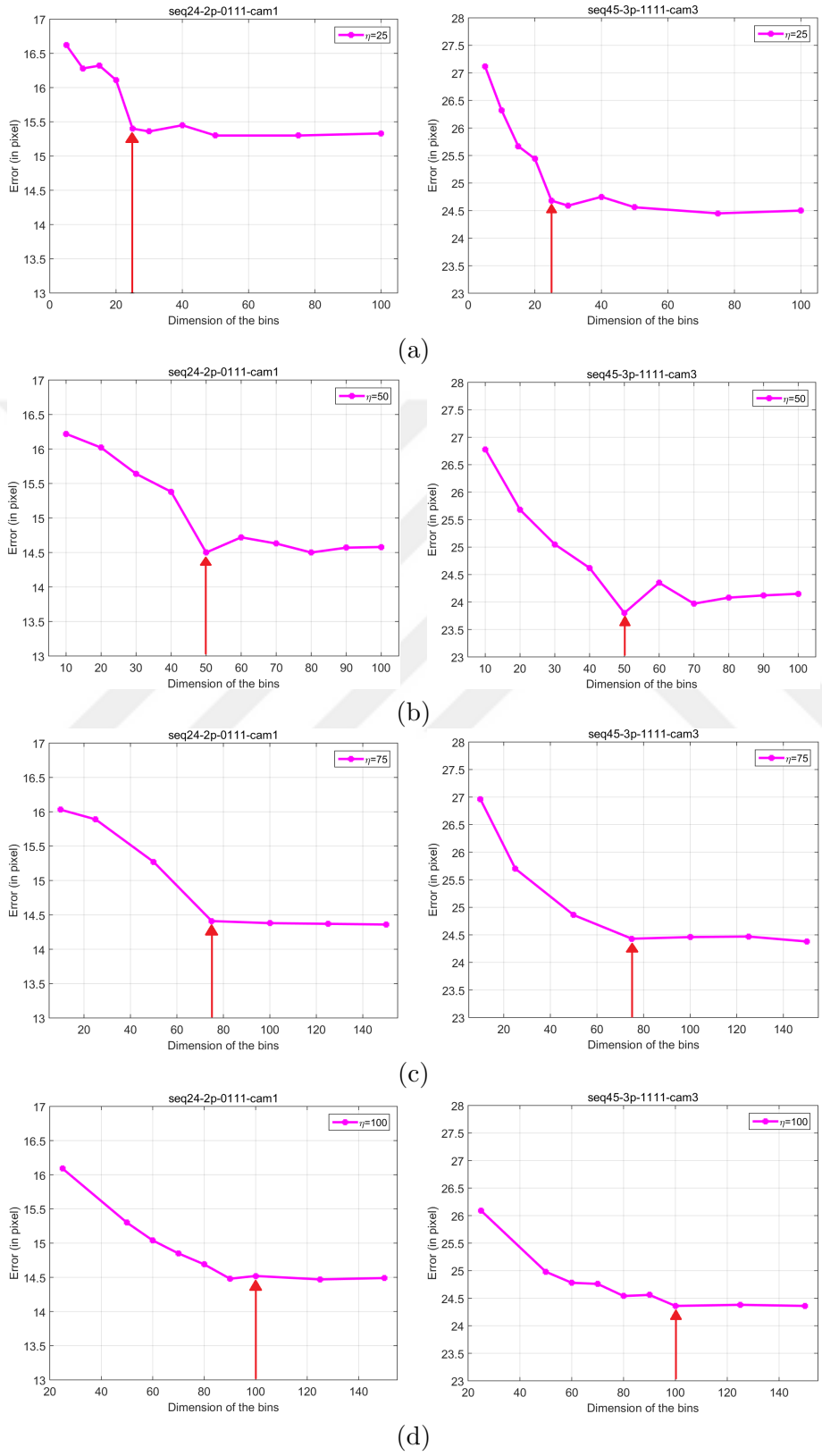


Figure 5.11: The sensitivity of error to parameter τ is analysed using Sequence 24 camera #1 and Sequence 45 camera #3. In (a), a constant number of particles per speaker, $\eta = 25$, and τ is changed from 10 to 100. The same process is repeated for $\eta = 50$, $\eta = 75$ and $\eta = 100$ and the results are presented in (b), (c) and (d), respectively.

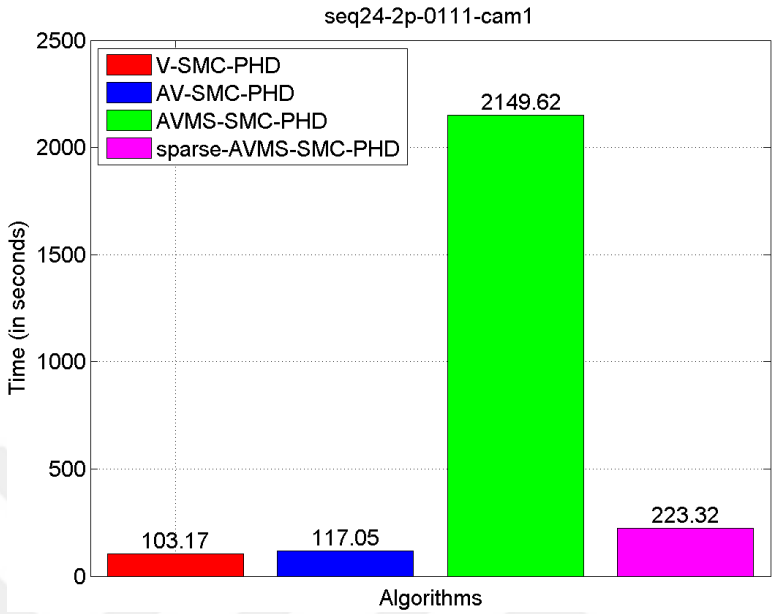


Figure 5.12: Computational cost comparison between the V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD filters.

generates a small subset from the source particles.

To see the effect of this low computational cost on the estimation accuracy, the proposed sparse-AVMS-SMC-PHD algorithm is run on Sequence 24 camera #1 and Sequence 45 camera #3, and to make an easy comparison with the previous proposed algorithms, the results are plotted all together in Figure 5.13.

The estimation accuracy is measured by the OSPA-T metric as before. Figures 5.13a shows the error variation as a function of the frame number for Sequence 24 camera #1, and Figures 5.13b for Sequence 45 camera #3. The pink dashed line is used to represent the sparse-AVMS-SMC-PHD filter. In both sequences, the errors for sparse-AVMS-SMC-PHD and AVMS-SMC-PHD filters seem similar. To see the difference in errors more clearly, the mean absolute error at each time step is depicted in Figures 5.14a and 5.14b for the algorithms.

Figures 5.14a and 5.14b show that sparse-AVMS-SMC-PHD and AVMS-SMC-PHD filters outperform the AV-SMC-PHD and V-SMC-PHD filters. When the comparison is made between the sparse-AVMS-SMC-PHD and AVMS-SMC-PHD filters, it is seen that the AVMS-SMC-PHD filter is slightly better than the sparse-AVMS-SMC-PHD

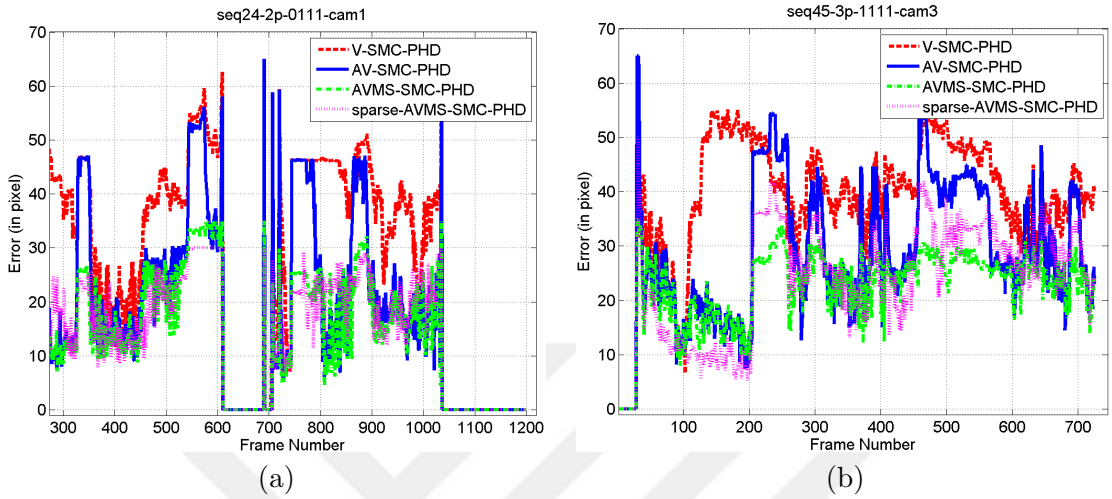


Figure 5.13: Performance comparisons of the V-SMC-PHD with proposed algorithms AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD for Sequence 24 camera #1 and Sequence 45 camera #3 are given in (a) and (b), respectively.

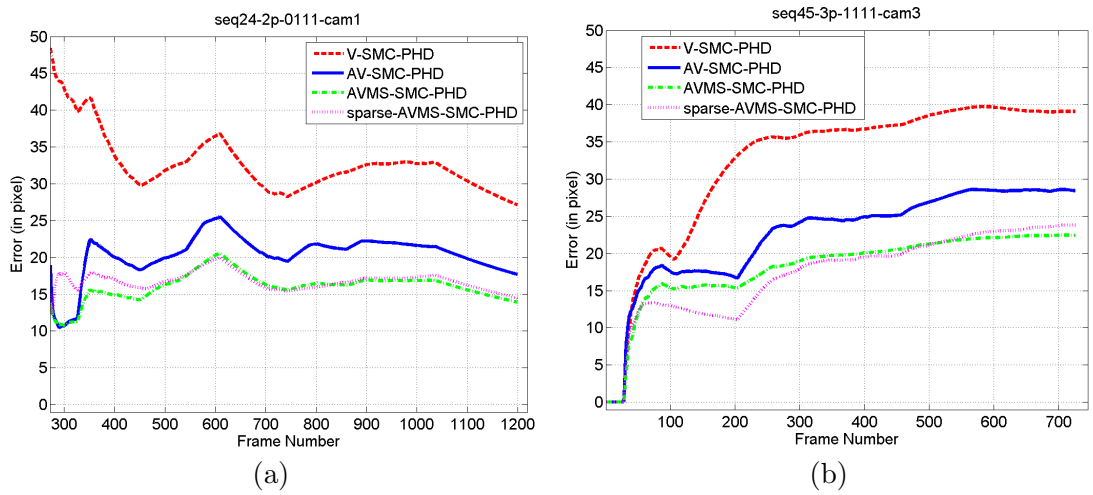


Figure 5.14: Comparison of the V-SMC-PHD and the proposed algorithms AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD using mean absolute OSPA-T error.

filter. Again, the error in Sequence 45 camera #3 is higher than for Sequence 24 camera #1 for the reasons explained above.

To make statistical analysis, the numerical values of the four algorithms are shown with a bar plot in Figures 5.15a and 5.15b. According to these plots, the AVMS-SMC-PHD filter shows only 4% and 6% better than the sparse-AVMS-SMC-PHD filter in Sequence 24 camera #1 and Sequence 45 camera #3, respectively.

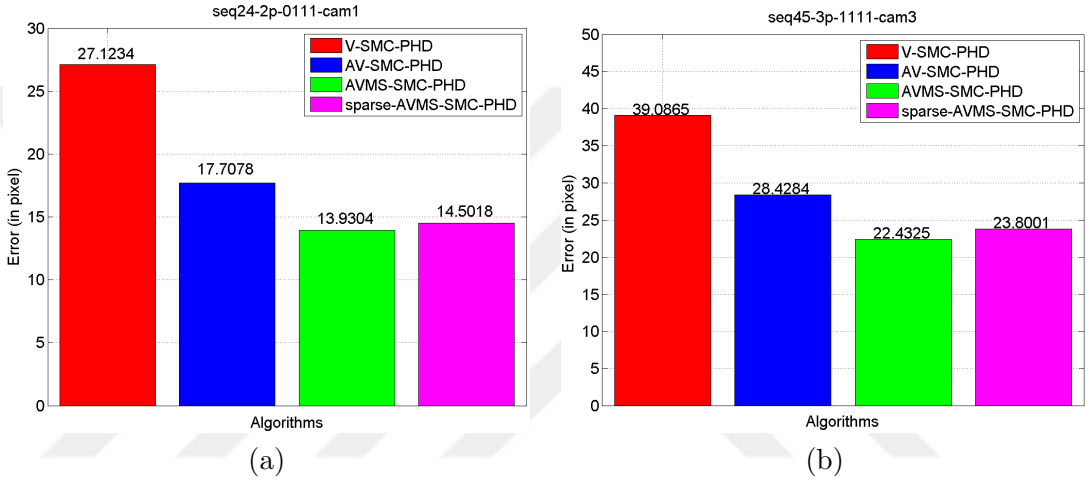


Figure 5.15: Performance comparison of the V-SMC-PHD with proposed algorithms AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD with bar plots.

The sparse-AVMS-SMC-PHD algorithm is also run over the remaining sequences and the results are given in Table 5.2 with standard errors.

Taking the average over all the experiments, sparse-AVMS-SMC-PHD outperforms AV-SMC-PHD and V-SMC-PHD by 19% and 42%. Compared to AVMS-SMC-PHD, the estimation accuracy is decreased by 7%. However, it is a reasonable sacrifice, given a ten-fold reduction in the computational cost as shown in Figure 5.12. To understand the reason behind the cost reduction in sparse-AVMS-SMC-PHD, we calculate the total number of particles used in each frame of the whole sequence for both AVMS-SMC-PHD and sparse-AVMS-SMC-PHD algorithms. The total number of particles in the AVMS-SMC-PHD algorithm is 61853 while it is 6301 in the sparse-AVMS-SMC-PHD algorithm. With the proposed sparse sampling, the number of particles is reduced to almost 10% (from 61853 to 6301) which leads to a ten-fold reduction in computational costs (from 2149.62 to 223.32 seconds) as shown in Figure 5.12.

Table 5.2: OSPA-T values for V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse AVMS-SMC-PHD averaged over 10 repetitions and standard errors of the estimated OSPA-T values.

		V SMC-PHD		AV SMC-PHD		AVMS SMC-PHD		sparse_AVMS SMC-PHD	
		OSPA-T	SE OSPA-T	OSPA-T	SE OSPA-T	OSPA-T	SE OSPA-T	OSPA-T	SE OSPA-T
seq24	cam1	27.12	0.17	17.71	0.07	13.93	0.04	14.5	0.04
	cam2	25.91	0.23	19.83	0.28	14.97	0.06	15.35	0.03
	cam3	24.32	0.16	18.94	0.05	14.12	0.06	15.72	0.03
seq25	cam1	25.84	0.09	19.13	0.06	15.72	0.05	17.17	0.02
	cam2	25.66	0.3	18.47	0.11	13.93	0.03	15.39	0.03
	cam3	29.99	0.31	21.61	0.07	17.07	0.04	17.62	0.07
seq30	cam1	35.6	0.23	25.22	0.23	16.65	0.03	19.27	0.04
	cam2	24.97	0.1	19.37	0.05	14.86	0.03	16.16	0.02
	cam3	37.64	0.36	25.31	0.07	19.29	0.05	19.67	0.03
seq45	cam1	48.68	0.09	29.46	0.08	22.95	0.05	23.4	0.03
	cam2	39.24	0.19	29.47	0.04	21.47	0.05	23.16	0.03
	cam3	39.09	0.15	28.43	0.02	22.43	0.03	23.8	0.02
Average		**32.01**	**0.2**	**22.75**	**0.09**	**17.28**	**0.04**	**18.43**	**0.03**

Table 5.3: Comparison of tracking results in terms of Wasserstain distance.

		Mean Wasserstein Distance (in pixel)				
		V SMC-PHD	AV SMC-PHD	AVMS SMC-PHD	sparse_AVMS SMC-PHD	Tracking algorithm [100]
seq24	cam1	42.13	21.37	18.87	19.97	7.2
	cam2	37.99	13.61	10.96	11.31	4.8

In order to test how significant the results of Table 5.2 are, the ANOVA based F-test is applied to the results of the sparse-AVMS-SMC-PHD algorithm. The first test is between the V-SMC-PHD and sparse-AVMS-SMC-PHD algorithms and we found that $F = 30.23$ and p-value$= 1.59 \times 10^{-5}$ with the degree of freedom $(1, 22)$. The second test is for the AV-SMC-PHD and sparse-AVMS-SMC-PHD algorithms and we found that $F = 6.92$ and p-value$= 1.53 \times 10^{-2}$ with the degree of freedom $(1, 22)$. Both comparisons have same degree of freedom which corresponds to 4.30 for F_{crit} value according to the F-distribution table given in [12]. Hence, the results of the sparse-AVMS-SMC-PHD algorithm are significant in both comparisons as $F > F_{crit}$ and p-value is smaller than 0.05 (for a 5% significance level).

Our tracking results are also compared with the results of Pham *et al.* [100]. His study focussed on visual object tracking using PHD filter and colour measurements. Experimental results were reported in terms of Wasserstein distance [53] for Sequence 24 cameras #1 and #2. So, our results are evaluated in terms of the Wasserstein distance and given in Table 5.3. The V-SMC-PHD, AV-SMC-PHD, AVMS-SMC-PHD and sparse-AVMS-SMC-PHD show similar performance as it is observed in Table 5.2. Even though AVMS-SMC-PHD shows the best performance amongst the four algorithms in both sequences, it is not close enough to [100]. The reason behind this gap will be interesting direction for future work.

5.5 Chapter Summary

In this chapter, the mean-shift method is introduced to improve the estimation accuracy of the AV-SMC-PHD filter proposed in the previous chapter. The mean-shift is an iterative method for climbing density gradients to find the peak of the probabil-

ity distribution. It enhances the AV-SMC-PHD filter by converging the particles to their neighbouring local peaks which leads to better estimation accuracy compared to conventional methods. We have built a tracking system that integrates the respective strengths of mean-shift, particle and PHD filters to robustly track multiple speakers. The proposed AVMS-SMC-PHD filter exhibited improved performance in evaluation tests on two-speaker and three-speaker sequences as compared to the AV-SMC-PHD filter. Further experiments on the $AV16.3$ dataset confirmed the advantages of the proposed AVMS-SMC-PHD algorithm in terms of estimation accuracy.

The mean-shift tracking has fast convergence speed and small computational cost, but applying the mean-shift method to all the particles is time consuming. To reduce the computational cost, we have proposed the use of sparse particles. In sparse sampling, we obtain a small subset from source particles and the mean-shift iteration is run only on this subset. The computational cost is significantly reduced using the sparse particles with a small sacrifice in estimation accuracy. The proposed algorithms have been tested on the $AV16.3$ dataset for two and three-speaker scenarios, where the number of speakers varies over time. The experiments demonstrate that the proposed algorithms can effectively track a variable number of moving speakers.

Chapter 6

Conclusions and Future Research

The problem of tracking a variable number of speakers in a room environment has been investigated in this thesis. This chapter presents a summary of the key contributions of this thesis. Furthermore, based on the discussion on the limitation of our work, potential directions for future work are also suggested.

6.1 Conclusions

In this thesis, the main technical challenges related to multi-modal multi-speaker tracking have been addressed in the context of tracking a variable number of speakers with audio and visual modalities in indoor environments. We have presented the following three key findings in Chapters 3, 4 and 5 respectively: (1) the association of audio information under an adaptive approach improves the performance of the visual tracker, (2) the use of audio information in particle propagation in the SMC-PHD filter enhances the system performance and computational cost, (3) combining the merits of both deterministic and stochastic approaches offers improved tracking performance in terms of estimation accuracy. In the following subsections these contributions will be elaborated in more detail.

6.1.1 Audio Assisted Robust Visual Tracking with Adaptive Particle Filtering

Multi-modal data fusion has been regarded as an effective method to improve the tracking performance in the case of challenging situations like occlusions (by the limited field of view of cameras or by other speakers). However, data fusion algorithms usually suffer from noise corrupting the sensor measurements which cause non-negligible detection errors. Here, we have proposed a novel approach to combine audio and visual modalities. Audio data is employed as an aid to PF based visual tracking. DOA angles of the audio sources are used to relocate the particles in the propagation step and recalculate the weights of the particles in the measurement step. Audio is fused to the V-PF through modifications of the PF steps. That makes the tracker less sensitive to outliers and noise in audio data.

Our approach is further improved by performing adaptive estimation of two critical parameters of the PF: the number of particles and noise variance. The PF uses a set of particles to approximate the posterior density and better approximation can be obtained with a higher number of particles which nevertheless causes an increased computational cost. We proposed an adaptive approach to the PF, based on the area occupied by the particles in each frame.

Our work differs substantially from previous works on AV multiple speaker tracking with respect to audio integration into the PF framework, and adaptive estimation of the particle number and variance of Gaussian noise.

6.1.2 Tracking a Variable Number of Speakers

The generic particle filter can be applied to multi-target tracking under the assumption that the number of target is known and constant. For the case of a time-varying number of speakers, advanced method needs to be used like RFS or its first order approximation PHD filter.

First, we focused on the RFS and extended it use with multi-modal data, and developed a particle filter algorithm under the RFS framework for AV tracking. The experiments

showed that, with this new extension, the proposed AV tracking system is able to track reliably a variable number of speakers in challenging scenarios such as occlusion. However, it was found that the computational cost grows exponentially with the number of speakers. To reduce the computational cost, the PHD filter is also explored.

To handle nonlinear/non-Gaussian speaker tracking problems, SMC implementation of the PHD filter was developed. The SMC-PHD filter propagates three types of particles, surviving, spawned and born particles, to model the state of the speakers and jointly estimates the variable number of speakers with their states. In classical SMC-PHD filter, all types of particles are propagated every frame to detect the speaker presence in the view, which is computationally expensive. To further reduce the computational complexity and improve the estimation accuracy, DOA information is employed differently depending on the particle type. For born particles, it is used to decide when to propagate them, and for surviving and spawned particles, it is used to re-locate them around the line drawn by the DOA.

6.1.3 Sparse Sampling for Mean-Shift Based SMC-PHD Filtering

Our third contribution considers the merits of both stochastic and deterministic tracking approaches in a unified framework for improving the performance of the tracking system in terms of estimation accuracy. The mean-shift, a typical and popular deterministic method, is a robust non-parametric method for climbing density gradients to find the peak of probability distributions using only a single hypothesis. On the other hand, particle filters estimate posterior by a set of random particles and their associated weights. These particles and their weights are then used to estimate the target position approximately. To approach an exact estimation, rather than approximate estimation, the mean-shift method is integrated with the proposed AV-SMC-PHD filter. The mean-shift method moves all the particles close to the local maximum which leads to a significant improvement in estimation accuracy as shifted particles are closer to speaker compared to their original positions.

We also found that the computational cost increases by applying the mean-shift method to all the particles. We developed a new sampling technique to reduce the number of

particles used with the mean-shift method. Our proposed sampling method is named sparse sampling as particles are selected based on a sparse representation of the particles. In consequence, our mean-shift method is run only for the sparse particles rather than all the particles which leads to a significant reduction in processing time with a small sacrifice in estimation accuracy.

6.2 Future Research

There are some important possible extensions to the work discussed in this thesis based on the problems and limitations of the proposed methods. In this section, we highlight the limitations and the drawbacks of the techniques that were developed, and also propose some directions for further improvements. An outline of the possible directions for future research is sketched below.

- The audio detection and localization algorithm used in this thesis assumes that the microphone array is circular. This constrains the potential datasets that can be used to evaluate the performance of the proposed algorithms as some datasets may not feature circular microphone arrays. By extending the audio detection algorithm for other microphone array types, the proposed approaches can be used in different environments.

- The audio information used in tracking is the DOA which makes it necessary to know the calibration information required to project the DOA into the 2-D image plane. As we work in a smart room, it is assumed that calibration information is available. For the environments like outdoor application or a room with no available calibration information, self-calibration algorithms would have to be integrated to the tracking system.

- In the visual part of the tracker, the reference model, which is the colour histogram of the target, would benefit from updating or online learning to enable tracking even in extreme pose changes. More cues such as texture and edge information can be incorporated to improve the tracking accuracy.

- Visual tracking may suffer from limited field of view of cameras. Single camera tracking can be replaced with multi-camera tracking so as to track multiple objects robustly and to mitigate the challenges such as limited field of view, occlusion and target association.

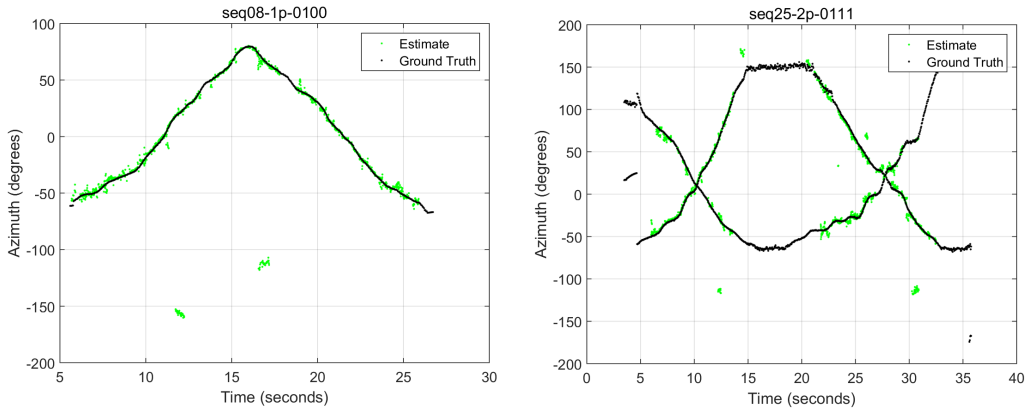
- Processing multi-modal information requires high computational power which is a challenging problem for online or real-time applications. Powerful programming and computing platforms like CUDA (Compute Unified Device Architecture), which is a parallel computing platform implemented by the graphical processing units, can be considered to provide the necessary computational power in order to make the tracking system real-time.

- Finally, more information about the indoor environment can be modelled like the positions of chairs, tables, entrances, windows, noise sources and so forth, which may greatly improve the quality of detection and localization as speaker can be searched in the limited area after some parts of the frame are isolated with prior information.

Appendix A

Audio Detection and Localization Results

As explained in Section 3.2, the SSM method was used to extract the DOA information. At the below, the DOA results are given for single and multi-speaker cases.



(a) (b)

Figure A.1: Audio detection and localization results using the SSM method for single and multi-speaker cases. The single-speaker case is depicted in (a), and (b) shows two-speaker case.

Appendix B

Detail Analysis of Performance Metrics

As mentioned in Chapter 3, three performance metrics, MAE, MOT and TBM, were used in the evaluation of the proposed and baseline algorithms. For simplicity, main terms of these metrics were given in Chapter 3. Here, algorithms are analysed with all terms of the performance metrics.

Table B.1 presents experimental results for V-PF and proposed AV-PF. The TBM evaluates the performance on the basis of trajectory. The first nine measures in Table B.1 belong to the TBM which include GT, MT, PT, ML, Frag and IDS. Definitions of these terms were explained in Section 2.5. For the quantitative evaluations of the proposed algorithm, 10 different sequences with 3 camera angles were used where the total number of speakers in the sequences is 57. So it means we have 57 GT trajectories. 43.0% of the trajectories are mostly tracked (MT) in V-PF while it is 92.4% in proposed AV-PF. Also, 48.6% of the trajectories are partly tracked (PT) in V-PF while it is 7.4% in proposed AV-PF. The MT, PT and ML complement each other to 100%, and so,

Table B.1: Full descriptions of experimental results with TBM and MOT metrics for V-PF and proposed AV-PF.

Method	**GT**	**MT**	**PT**	**ML**	**Frag**	**IDS**	**MOTA**	**MOTP**	**MAE**
V-PF	57	43.0%	48.6%	8.4%	337	159	20.5%	14.6	79.6
SE	-	0.2	0.3	0.1	2	1	1.9	0.6	0.9
AV-PF	57	92.4%	7.6%	0.0%	304	70	90.5%	12.7	14.3
SE	-	0.0	0.0	0.0	1	1	1.5	0.3	0.4

Table B.2: Full descriptions of experimental results with TBM and MOT metrics for KLD-sampling and proposed AV-A-PF.

Method	**GT**	**MT**	**PT**	**ML**	**Frag**	**IDS**	**MOTA**	**MOTP**	**MAE**	**N**
KLD	57	96.7%	3.3%	0.0%	185	39	94.7%	11.5	11.8	53.4
SE	-	0.0	0.0	0.0	0	0	0.6	0.1	0.1	0.1
AV-A-PF	57	96.2%	3.8%	0.0%	234	44	94.4%	11.6	11.9	16.9
SE	-	0.0	0.0	0.0	1	0	0.5	0.1	0.2	0.3

ML (mostly lost) seems redundant in the table, but it is given for clear presentation. The track fragmentation (Frag), number of GT interruptions in tracking result, is 337 and 304 respectively for V-PF and AV-PF. The identity switches (IDS) in AV-PF is more than half of the V-PF. So, based on the TBM evaluations, proposed AV-PF outperforms the V-PF. The MOTA and MOTB are the quantities for the MOT metric. The MOTP measures the precision of the algorithms which is 14.6 and 12.7 for V-PF and AV-PF, respectively. The MOTA measures the accuracy of the tracker which is 20.5% and 90.5% for V-PF and AV-PF, respectively. The standard error (SE) of the metric results are also given at below of the each rows. In terms of SE, there are not much difference between V-PF and AV-PF.

Experimental results with TBM, MOT and MAE metrics for KLD-sampling and proposed AV-A-PF are given in Table B.2. Last column of the table gives the number of particles N, since these algorithms are designed for adaptive estimation of number of particles. When KLD-sampling and proposed AV-A-PF algorithms are compared with three different metrics, it is seen that they show similar performance. However, the point here is to estimate optimum number of particles to use in tracking. If estimation of N is compared for two algorithms, proposed algorithm uses fewer particles than KLD-sampling without sacrificing tracking performance.

Appendix C

List of Acronyms, Nomenclature and Glossary

Acronyms

AMI	Augmented Multi-party Interaction
ANOVA	Analysis of Variance
A-PF	Adaptive Particle Filtering
AR	Auto-regressive
AV	Audio-Visual
AV-PF	Audio-Visual Particle Filtering
AV-A-PF	Audio-Visual Adaptive Particle Filtering
AV-SMC-PHD	Audio-Visual Sequential Monte Carlo Probability Hypothesis Density
AVMS-SMC-PHD	Audio-Visual Mean-shift Sequential Monte Carlo Probability Hypothesis Density
B/F	Background/Foreground
CBD	Cumulative Beta Distribution
CLEAR	Classification of Events, Activities and Relationships
CPHD	Cardinalized Probability Hypothesis Density
CSS	Coherent Signal Subspace
CUDA	Compute Unified Device Architecture
DKF	Decentralized Kalman Filter

DOA	Direction Of Arrival
EKF	Extended Kalman Filter
FISST	Finite-Set Statistics
GCC	Generalized Cross Correlation
GPS	Global Positioning System
GT	Ground Truth
HSV	Hue-Saturation-Value
IDS	Identity Switches
JPDA	Joint Probabilistic Data Association
KF	Kalman Filter
KL	Kullback-Leibler
KLD	Kullback-Leibler divergence
MAE	Mean Absolute Error
MHT	Multiple Hypothesis Tracking
ML	Mostly Lost
MLE	Maximum Likelihood Estimates
MOT	Multiple Object Tracking
MOTA	Multiple Object Tracking Accuracy
MOTP	Multiple Object Tracking Precision
MSEPF	Mean-Shift Embedded Particle Filter
MT	Mostly Tracked
MUSIC	Multiple Signal Classification
OSPA	Optimal Subpattern Assignment
OSPA-T	Optimal Subpattern Assignment for Tracks
PF	Particle Filter
PT	Partially Tracked
PHD	Probability Hypothesis Density
RFS	Random Finite Set
RGB	Red-Green-Blue
SE	Standard Error
SIFT	Scale-invariant Feature Transform

SIR	Sampling Importance Resampling
SLAM	Simultaneous Localization and Mapping
SMC	Sequential Monte Carlo
SMC-PHD	Sequential Monte Carlo Probability Hypothesis Density
SNR	Signal-to-Noise Ratio
SPEVI	Surveillance Performance Evaluation Initiative
SRP-PHAT	Steered Response Power with the Phase Transform
SSM	Sam-Spare-Mean
SSL	Sound Source Localization
TBM	Trajectory-Based Measures
TDE	Time Delay Estimation
TDOA	Time Difference of Arrival
V-PF	Visual Particle Filtering
VRPF	Variable Resolution Particle Filter

Nomenclature

Symbols

$(1-\delta)$	The quantile of χ^2 distribution
b	The number of bins in KLD-Sampling
$\beta_{k\|k-1}(\mathbf{x}_k\|\mathbf{x}_{k-1})$	The intensity function of RFS $\mathcal{B}_k(\mathcal{X}_{k-1})$ at time frame k
$\mathcal{B}_k(\mathcal{X}_{k-1})$	The RFS of speakers spawned from the previous set of speakers $\mathcal{X}_{k-1}$ at time frame k
$\mathbb{B}$	One dimensional bins for sparse sampling
$\mathfrak{c}_k$	The number of matches at time frame k
$\mathcal{C}_k$	The RFS of clutter or false alarms at time frame k
C	The normalization constant
γ_k	Audio contribution weighting at time frame k
Γ_k	The RFS of new speakers that appear spontaneously at time frame k
δ	The Kronecker delta function
$\mathfrak{d}_k^i$	The distance between the object and corresponding match at time frame k
$\hat{\mathbf{d}}_k$	The movement distances at time frame k
$\mathbf{F}$	State transition matrix
ϵ	The upper bound for the error
ε_k	White noise at time frame k
ζ	Threshold for shifting distance
$\mathcal{Z}_k$	Observations set at time frame k
$\mathcal{Z}_k^{\mathrm{v}}$	Visual measurement set at time frame k
$\mathcal{Z}_k^{\mathrm{a}}$	Audio measurement set at time frame k
η	The constant number of particles per speaker
fp_k	Number of false positives at time frame k
$f_{k\|k-1}(\mathbf{x}_k\|\mathbf{x}_{k-1})$	The single-speaker state transition density at time frame k
g_k	Number of objects present at time frame k
$g_k(\mathbf{z}_k\|\mathbf{x}_k)$	The single-speaker likelihood at time frame k

h	The size of the kernel
$\hbar_{\text{birth}}$	The birth hypotheses
$\bar{\hbar}_{\text{birth}}$	The non-birth hypotheses
θ_k	The DOA (azimuth) angle at time frame k
$\Theta_k(\mathcal{X}_k)$	The RFS of the measurements generated by the speakers $\mathcal{X}_k$ at time frame k
J_k	The particles from the new born importance function at time frame k
$k = 1, ..., K$	The image frame index
K	The total number of image frames
$\mathfrak{k}$	The kernel function
λ	The design parameter for calculation of weight of the particles
$\kappa_k(\mathbf{z}_k)$	The intensity of clutter RFS $\mathcal{C}_k$ at time frame k
L	The area of rectangle used for speaker search by the particle
m_k	Number of misses at time frame k
mm_k	Number of mismatches at time frame k
M_k	Number of observations
$\mathbb{M}_1$	Number of microphones
$\mathbb{MS}$	Mean-shift iteration function
$n = 1, ..., N$	The particle index
N	The number of particles
N_{k-1}	Number of particles drawn from importance sampling at time frame $k-1$
$\mathfrak{N}_i$	The number of particles in $\mathbb{B}_i$
$p(.)$	Probability distribution
$p_{D,k}(\mathbf{x}_k)$	The detection probability at time frame k
$p_{S,k}(\mathbf{x}_{k-1})$	The survival probability for the speakers at time frame k
P_{birth}	The probability of $\hbar_{\text{birth}}$
$\rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]$	Bhattacharyya coefficient
$q_u^{(n)}$	The Hue-Saturation histogram extracted from the rectangle centred on the position of the nth particle

$\mathbf{q}_k$	The zero-mean Gaussian noise with covariance $\mathbf{Q}$ at time frame k
r_u	The Hue-Saturation histogram of the reference image
s	The scale of the rectangle centred around (x_1, x_2)
$\mathcal{S}_k(\mathcal{X}_{k-1})$	The RFS of surviving speakers at time frame k
S	The total area occupied on the image by the particles
$\mathbb{S}_1$	Sector number
σ^2	Noise variance
σ_s^2	The variance of the scale
σ_{a}^2	The variance of noise for audio likelihood
$\mathbb{SS}$	Sparse sampling function
T	The period between two adjacent frames
τ	The number of subintervals in $\mathbb{B}$
$\phi_{k\|k-1}(\mathbf{x}_k\|\mathbf{x}_{k-1})$	The analog of the single-speaker state transition probability at time frame k
φ_i	The parameters of the model
U	The number of histogram bins
$v_{k\|k}$	The intensity at time frame k
$\mathbf{x}$	State vectors
x_1	Horizontal positions of the speaker face
x_2	Vertical position of the speaker face
$\dot{x}_1$	The horizontal velocity
$\dot{x}_2$	The vertical velocity
$\{\mathbf{x}_i\}_{i=1,\ldots,r}$	The *normalized* pixel locations in the region
$\tilde{\mathbf{x}}_k^{av}$	Position estimation based on audio-visual modalities at time frame k
$\hat{\mathbf{x}}_{\hat{\Xi}_k,k}$	Estimation of the ground truth state set at time frame k
$\xi_k(\mathbf{x}_k)$	The intensity function of the new speaker birth RFS Γ_k at time frame k
Ξ_k	The number of speakers at time frame k
$\hat{\Xi}_{k\|k}$	The total mass at time frame k

$\mathcal{X}_k$	Multi-speaker state set at time frame k
$u(\mathbf{z}_k)$	The probability distribution of each clutter point at time frame k
ψ	The average number of Poisson clutter points per scan
$\psi_{(\mathbf{x}_i,k)}$	The speaker position in terms of angle with respect to the microphone at time frame k
y	Location of the target in mean-shift process
w	Weights of the particles

Glossary

Analysis of variance (ANOVA): A statistical method used to test a hypothesis concerning the means of three or more populations

Bayesian framework: A framework that allows you to compute the revised probability of an event that occurred before another event when the events are dependent

Data: Measurements or observations for a variable

Dataset: A collection of data values

Degrees of freedom: The number of values that are free to vary after a sample statistic has been computed; used when a distribution consists of a family of curves

F test: A statistical test used to compare two variances or three or more means

Histogram: A graph that displays the data by using vertical bars of various heights to represent the frequencies of a distribution

Monte Carlo method: A simulation technique using random numbers

Multi-modal: A data set with three or more modes

Null hypothesis: A statistical hypothesis that states that there is no difference between a parameter and a specific value or that there is no difference between two parameters

One-way ANOVA: A study used to test for differences among means for a single independent variable when there are three or more groups

Outlier: An extreme value in a data set

Parameter: A characteristic or measure obtained by using all the data values

Quantiles: Values that separate the data set into approximately equal groups

Random sample: A sample obtained by using random or chance methods

Random variable: A variable whose values are determined by chance

Variable: A characteristic or attribute that can assume different values

Variance: The average of the squares of the distance that each value is from the mean

Appendix D

List of Publications

Journal Articles

- V. Kilic, M. Barnard, W. Wang, A. Hilton and J. Kittler, Mean-Shift and Sparse Sampling Based SMC-PHD Filtering for Audio Informed Visual Speaker Tracking, *IEEE Transactions on Multimedia, submitted.*

- 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, pages 186-200, February 2015.

Conference Papers

- V. Kilic, M. Barnard, W. Wang, A. Hilton and J. Kittler, Audio Informed Visual Speaker Tracking with SMC-PHD Filter, In *Proceedings of International Conference on Multimedia and Expo* (ICME 2015). IEEE, July 2015.

- V. Kilic, X. Zhong, M. Barnard, W. Wang, and J. Kittler, Audio-Visual Tracking of a Variable Number of Speakers with a Random Finite Set Approach, In *Proceedings of the 17th International Conference on Information Fusion* (FUSION 2014), pages 1-7. IEEE, July 2014.

- V. Kilic, M. Barnard, W. Wang, and J. Kittler, Adaptive particle filtering approach to audio-visual tracking, In *Proceedings of the 21st European Signal Processing Conference* (EUSIPCO 2013), pages 1-5. IEEE, September 2013.
- V. Kilic, M. Barnard, W. Wang, and J. Kittler, Audio constrained particle filter based visual tracking, In *Proceedings of International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2013), pages 3627-3631. IEEE, May 2013.

Bibliography

[1] S. Anand, S. Mittal, O. Tuzel, and P. Meer. Semi-supervised kernel mean shift clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1201–1215, 2014.

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Feb. 2002.

[3] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, 2010.

[4] Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association*. Waltham, MA, USA: Academic Press, 1998.

[5] Y. Bar-Shalom and X.-R. Li. *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, CT, USA: YBS Publishing, 1995.

[6] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[7] K. Bernardin. *Multimodal Probabilistic Person Tracking and Identification in Smart Spaces*. PhD thesis, Karlsruhe University, 2009.

[8] K. Bernardin, T. Gehrig, and R. Stiefelhagen. Multi-level particle filter fusion of features and cues for audio-visual person tracking. In *Multimodal Technologies for Perception of Humans*, pages 70–81. Springer, 2008.

[9] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[10] C. Beyan and A. Temizel. Adaptive mean-shift for automated multi object tracking. *IET Computer Vision*, 6(1):1–12, 2012.

[11] S. Blackrnan and A. House. *Design and Analysis of Modern Tracking Systems*. Boston, MA, USA: Artech House, 1999.

[12] A. G. Bluman. *Elementary Statistics*. New York, NY, USA: McGraw Hill, 2013.

[13] M. S. Brandstein. *A framework for speech source localization using sensor arrays*. PhD thesis, Brown University, Providence, RI, USA, 1995.

[14] M. S. Brandstein and H. F. Silverman. A practical methodology for speech source localization with microphone arrays. *Computer Speech and Language*, 11(2):91–126, 1997.

[15] J. Carletta, S. Ashby, S. Bourban, et al. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39. Springer, 2006.

[16] R. Chakravorty and S. Challa. Multitarget tracking algorithm-joint IPDA and Gaussian mixture PHD filter. In *Proceedings of the 12th International Conference on Information Fusion*, pages 316–323, 2009.

[17] C. Chang and R. Ansari. Kernel particle filter: Iterative sampling for efficient visual tracking. In *Proceedings of International Conference on Image Processing*, volume 3, pages 977–980. IEEE, 2003.

[18] C. Chang and R. Ansari. Kernel particle filter for visual tracking. *IEEE Signal Processing Letters*, 12(3):242–245, Mar. 2005.

[19] C. Chang, R. Ansari, and A. Khokhar. Multiple object tracking with kernel particle filter. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, volume 1, pages 566–573. IEEE, June 2005.

[20] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 881–884, 2004.

[21] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[22] D. Clark and B.-N. Vo. The random set filtering website. Internet: http://randomsets.eps.hw.ac.uk/tutorial.html [Dec. 01, 2015].

[23] D. E. Clark, J. Bell, Y. Saint-Pern, and Y. Petillot. Phd filter multi-target tracking in 3d sonar. In *Oceans 2005-Europe*, volume 1, pages 265–270. IEEE, 2005.

[24] P. Closas and C. Fernandez-Prades. Particle filtering with adaptive number of particles. In *Proceedings of IEEE Aerospace Conference*, pages 1–7, Mar. 2011.

[25] R. T. Collins. Mean-shift blob tracking through scale space. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 234–241, 2003.

[26] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[27] D. Comaniciu and V. Ramesh. Mean shift and optimal prediction for efficient object tracking. In *Proceedings of International Conference on Image Processing*, volume 3, pages 70–73. IEEE, 2000.

[28] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, volume 2, pages 142–149. IEEE, 2000.

[29] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, May 2003.

[30] M. Crocco, A. D. Bue, M. Bustreo, and V. Murino. A closed form solution to the microphone position self-calibration problem. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2597–2600, 2012.

[31] J. Czyz, B. Ristic, and B. Macq. A color-based particle filter for joint detection and tracking of multiple objects. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 217–220, Mar. 2005.

[32] K. Deguchi, O. Kawanaka, and T. Okatani. Object tracking by the mean-shift of regional color distribution combined with the particle-filter algorithms. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 506–509. IEEE, 2004.

[33] J. H. DiBiase. *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. PhD thesis, Brown University, Providence, RI, USA, 2000.

[34] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. *Robust Localization in Reverberant Rooms*. Springer, 2001.

[35] R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69. IEEE, 2005.

[36] B. S. Everitt and G. Dunn. *Applied Multivariate Data Analysis*, volume 2. Arnold London, 2001.

[37] M. F. Fallon and S. J. Godsill. Acoustic source localization and tracking of a time-varying number of speakers. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1409–1415, May 2012.

[38] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, 2003.

[39] C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. John Wiley & Sons, 2011.

[40] D. Fox. KLD-Sampling: Adaptive particle filters. In *Advances in Neural Information Processing Systems*, pages 713–720, 2001.

[41] D. Fox. KLD-Sampling: Adaptive particle filters and mobile robot localization. In *Advances in Neural Information Processing Systems*, 2001.

[42] D. Fox. Adapting the sample size in particle filters through KLD-Sampling. *International Journal of Robotics Research*, 22:985–1003, 2003.

[43] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer. Audiovisual person tracking with a mobile robot. In *Proceedings of International Conference on Intelligent Autonomous Systems*, pages 898–906, 2004.

[44] J.-J. Fuchs. On the application of the global matched filter to doa estimation with uniform circular arrays. *IEEE Transactions on Signal Processing*, 49(4):702–709, 2001.

[45] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore. Audio-visual speaker tracking with importance particle filters. In *Proceedings of IEEE International Conference on Image Processing*, volume 3, pages 25–28, Sep. 2003.

[46] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:601–616, Feb. 2007.

[47] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough. Kalman filters for audio-video source localization. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 118–121, Oct. 2005.

[48] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat. Vision and RFID data fusion for tracking people in crowds by a mobile robot. *Computer Vision and Image Understanding*, 114(6):641–651, 2010.

[49] S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration -a Statistical Model Based Approach*. Berlin, Germany: Springer-Verlag, 1998.

[50] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Proceedings IEE-F Radar and Signal Processing*, 140(2):107 –113, 1993.

[51] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti. Smart video surveillance: Exploring the concept of multiscale spatiotemporal tracking. *IEEE Signal Processing Magazine*, 22(2):38–51, Mar. 2005.

[52] M. Heuer, A. Al-Hamadi, B. Michaelis, and A. Wendemuth. Multi-modal fusion with particle filter for speaker localization and tracking. In *Proceedings of International Conference on Multimedia Technology*, pages 6450–6453, 2011.

[53] J. R. Hoffman and R. Mahler. Multitarget miss distance via optimal assignment. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans,*, 34(3):327–336, 2004.

[54] R. Hoseinnezhad, B.-N. Vo, B.-T. Vo, and D. Suter. Bayesian integration of audio and visual information for multi-target tracking using a CB-MeMBer filter. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 2300–2303. IEEE, 2011.

[55] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998.

[56] M. Isard and J. MacCormick. Bramble: A Bayesian multiple-blob tracker. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 2, pages 34–41, 2001.

[57] M. Jaward, L. Mihaylova, N. Canagarajah, and D. Bull. Multiple object tracking

using particle filters. In *Proceedings of IEEE Aerospace Conference*, pages 8–15, 2006.

[58] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, 2004.

[59] V. Kilic, M. Barnard, W. Wang, A. Hilton, and J. Kittler. Audio informed visual speaker tracking with SMC-PHD filter. In *Proceedings of IEEE International Conference on Multimedia and Expo*, July 2015.

[60] V. Kilic, M. Barnard, W. Wang, and J. Kittler. Adaptive particle filtering approach to audio-visual tracking. In *Proceedings of the 21st European Signal Processing Conference*, pages 1–5. IEEE, Sep. 2013.

[61] V. Kilic, M. Barnard, W. Wang, and J. Kittler. Audio constrained particle filter based visual tracking. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3627–3631, May 2013.

[62] V. Kilic, M. Barnard, W. Wang, and J. Kittler. Audio assisted robust visual tracking with adaptive particle filtering. *IEEE Transactions on Multimedia*, 17(2):1860–200, Feb. 2015.

[63] V. Kilic, X. Zhong, M. Barnard, W. Wang, and J. Kittler. Audio-visual tracking of a variable number of speakers with a random finite set approach. In *Proceedings of the 17th International Conference on Information Fusion*, pages 1–7. IEEE, July 2014.

[64] K. Kim and L. S. Davis. Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In *Proceedings of European Conference on Computer Vision*, pages 98–109. Springer, 2006.

[65] G. Kitagawa. Monte Carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[66] O. Lanz. Approximate bayesian multibody tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1436–1449, Sep. 2006.

[67] G. Lathoud. *Spatio-temporal Analysis of Spontaneous Speech with Microphone Arrays*. PhD thesis, EPFL University, Lausanne, Switzerland, 2006.

[68] G. Lathoud, J. Bourgeois, and J. Freudenberger. Sector-based detection for hands-free speech enhancement in cars. *EURASIP Journal on Applied Signal Processing*, 2006:169–184, 2006.

[69] G. Lathoud and M. Magimai-Doss. A sector-based, frequency-domain approach to detection and localization of multiple speakers. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 265–268, Mar. 2005.

[70] G. Lathoud, J. M. Odobez, and D. Gatica-Perez. AV16.3: An audio-visual corpus for speaker localization and tracking. In *Proceedings of the Machine Learning in Medical Imaging Workshop, S. Bengio and H. Bourlard Eds*, pages 182–195. Springer Verlag, 2005.

[71] M. Legg and S. Bradley. A combined microphone and camera calibration technique with application to acoustic imaging. *IEEE Transactions on Image Processing*, 22(10):4028–4039, Oct. 2013.

[72] I. Leichter, M. Lindenbaum, and E. Rivlin. Mean shift tracking with multiple reference color histograms. *Computer Vision and Image Understanding*, 114(3):400–408, 2010.

[73] P. Li. An adaptive binning color model for mean shift tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(9):1293–1299, 2008.

[74] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2953–2960, Jun. 2009.

[75] Z. Li, Q. L. Tang, and N. Sang. Improved mean shift algorithm for occlusion pedestrian tracking. *Electronics Letters*, 44(10):622–623, 2008.

[76] Y. Lim and J. Choi. Speaker selection and tracking in a cluttered environment with audio and visual information. *IEEE Transactions on Consumer Electronics*, 55(3):1581–1589, 2009.

[77] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

[78] Q. Liu, Y. Rui, A. Gupta, and J. J. Cadiz. Automating camera management for lecture room environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 442–449. ACM, 2001.

[79] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[80] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley. Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach. *IEEE Transactions on Signal Processing*, 54(9):3291–3304, Sep. 2006.

[81] E. Maggio and A. Cavallaro. Hybrid particle filter and mean shift tracker with adaptive transition model. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 221–224, Mar. 2005.

[82] E. Maggio and A. Cavallaro. *Video Tracking: Theory and Practice*. John Wiley & Sons, 2011.

[83] E. Maggio, M. Taj, and A. Cavallaro. Efficient multitarget visual tracking using random finite sets. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1016–1027, Aug. 2008.

[84] R. Mahler. PHD filters of higher order in target number. *IEEE Transactions on Aerospace and Electronic Systems*, 43(4):1523–1543, Oct. 2007.

[85] R. P. S. Mahler. Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1152–1178, 2003.

[86] R. P. S. Mahler. Statistics 101 for multisensor, multitarget data fusion. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):53–64, 2004.

[87] R. P. S. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Boston, MA, USA: Artech House, 2007.

[88] B. S. Manjunath and W.-Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[89] L. A. McGee and S. F. Schmidt. *Discovery of the Kalman Filter as a Practical Tool for Aerospace and Industry*. NASA, Technical Report NASA-TM-86847, 1985.

[90] S. J. McKenna, Y. Raja, and S. Gong. Tracking colour objects using adaptive mixture models. *Image and Vision Computing*, 17(3):225–231, 1999.

[91] A. S. Micilotta, E.-J. Ong, and R. Bowden. Real-time upper body detection and 3D pose estimation in monoscopic images. In *Proceedings of European Conference on Computer Vision*, pages 139–150. Springer, 2006.

[92] A. Milan, K. Schindler, and S. Roth. Challenges of ground truth evaluation of multi-target tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 735–742, 2013.

[93] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R Casas, J. Turmo, L. Cristoforetti, F. Tobia, et al. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3):389–407, 2007.

[94] J. Mullane, B.-N. Vo, M. D. Adams, and B.-T. Vo. A random-finite-set approach to Bayesian SLAM. *IEEE Transactions on Robotics*, 27(2):268–282, Jan. 2011.

[95] M. R. Naphade, T. Kristjansson, B. Frey, and T. S. Huang. Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In *Proceedings of International Conference on Image Processing*, pages 536–540. IEEE, 1998.

[96] Q. Nguyen and J. Choi. Localization and tracking for simultaneous speakers based on time-frequency method and probability hypothesis density filter. In *Pro-*

ceedings of IEEE International Conference on Robotics and Biomimetics, pages 2866–2871, Dec. 2011.

[97] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[98] K. Panta, B. Vo, S. Singh, and A. Doucet. Probability hypothesis density filter versus multiple hypothesis tracking. In *Signal Processing, Sensor Fusion, and Target Recognition XIII*, volume 5429, pages 284–295, 2004.

[99] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proceedings of the IEEE*, 92(3):495–513, 2004.

[100] N. T. Pham, W. Huang, and S. H. Ong. Tracking multiple objects using probability hypothesis density filter and color measurements. In *Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1511–1514, 2007.

[101] N. T. Pham, W. Huang, and S. H. Ong. Tracking multiple speakers using CPHD filter. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 529–532, 2007.

[102] M. K. Pitt, R. d. S. Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

[103] G. Potamianos, C. Neti, and S. Deligne. Joint audio-visual speech processing for recognition and enhancement. In *Proceedings of International Conference on Audio-Visual Speech Processing*, 2003.

[104] I. Potamitis, H. Chen, and G. Tremoulis. Tracking of multiple moving speakers with multiple microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 12(5):520–529, Sep. 2004.

[105] I. Potamitis and G. Kokkinakis. Speech separation of multiple moving speakers using multisensor multitarget techniques. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 37(1):72–81, 2007.

[106] Y. Raja, S. J. McKenna, and S. Gong. Segmentation and tracking using colour mixture models. In *Proceedings of Asian Conferenece on Computer Vision*, pages 607–614. Springer, 1998.

[107] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo. A metric for performance evaluation of multi-target tracking algorithms. *IEEE Transactions on Signal Processing*, 59(7):3452–3457, Jul. 2011.

[108] N. Roman and D. Wang. Binaural tracking of multiple moving sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):728–739, May 2008.

[109] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008.

[110] Y. Rui, D. Florencio, W. Lam, and J. Su. Sound source localization for circular arrays of directional microphones. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 93–96. IEEE, 2005.

[111] A. Sanin, C. Sanderson, and B. C. Lovell. Shadow detection: A survey and comparative evaluation of recent methods. *Pattern Recognition*, 45(4):1684–1695, 2012.

[112] S. Sarkka, A. Vehtari, and J. Lampinen. Rao-blackwellized Monte Carlo data association for multiple target tracking. In *Proceedings of the Seventh International Conference on Information Fusion*, pages 583–590, 2004.

[113] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, Mar. 1986.

[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*, 56(8):3447–3457, Aug. 2008.

[115] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In

Proceedings of International Conference on Robotics and Automation, volume 2, pages 1665–1670. IEEE, 2001.

[116] C. Shan, T. Tan, and Y. Wei. Real-time hand tracking using a mean shift embedded particle filter. *Pattern Recognition*, 40(7):1958–1970, 2007.

[117] C. Shan, Y. Wei, T. Tan, and F. Ojardias. Real time hand tracking by combining particle filtering and mean shift. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 669–674, 2004.

[118] S. T. Shivappa, B. D. Rao, and M. M. Trivedi. Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation. *IEEE Journal of Selected Topics in Signal Processing*, 4:882–894, Oct. 2010.

[119] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[120] L. Sigal, S. Sclaroff, and V. Athitsos. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 152–159, 2000.

[121] A. Smith, A. Doucet, N. Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer Science & Business Media, 2013.

[122] A. Soto. Self adaptive particle filter. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1398–1403, 2005.

[123] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 evaluation. In *Multimodal Technologies for Perception of Humans*, pages 3–34. Springer, 2008.

[124] N. Strobel, S. Spors, and R. Rabenstein. Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, 18(1):22–31, Jan. 2001.

[125] J. Sullivan and J. Rittscher. Guiding random particles by deterministic search. In *Proceedings of Eighth International Conference on Computer Vision*, volume 1, pages 323–330. IEEE, 2001.

[126] M. Taj. Surveillance performance evaluation initiative (spevi) audiovisual people dataset. Internet: http://www.eecs.qmul.ac.uk/∼andrea/spevi.html, 2007 [Aug. 24, 2014].

[127] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides. Audiovisual active speaker tracking in cluttered indoors environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38:799–807, Jun. 2008.

[128] F. Talantzis, A. Pnevmatikakis, and L. C. Polymenakos. Real time audio-visual person tracking. In *Proceedings of the Eighth IEEE Workshop on Multimedia Signal Processing*, pages 243–247, 2006.

[129] W. Tao, H. Jin, and Y. Zhang. Color image segmentation based on mean shift and normalized cuts. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(5):1382–1389, 2007.

[130] M. M. Trivedi, H. S. Kohsia, and I. Mikic. Dynamic context capture and distributed video arrays for intelligent spaces. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 35(1):145–163, Jan. 2005.

[131] Y. Ukrainitz and B. Sarel. Mean shift theory and applications. Internet: http://www.wisdom.weizmann.ac.il/∼vision/courses/2004-2/files/mean-shift/mean-shift.ppt [Oct. 24, 2014].

[132] V. Verma, S. Thrun, and R. Simmons. Variable resolution particle filter. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 976–984, 2003.

[133] J. Vermaak, M. Gangnet, A. Blake, and P. Perez. Sequential Monte Carlo fusion of sound and vision for speaker tracking. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 1, pages 741–746, 2001.

[134] B.-N. Vo and W.-K. Ma. A closed-form solution for the probability hypothesis density filter. In *Proceedings of the Eighth International Conference on Information Fusion*, volume 2, pages 856–863. IEEE, 2005.

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

[136] B.-N. Vo, S. Singh, and W. K. Ma. Tracking multiple speakers using random sets. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 357–360, 2004.

[137] H. Wang and M. Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(4):823–831, Aug. 1985.

[138] J. Wang and Y. Yagi. Discriminative mean shift tracking with auxiliary particles. In *Proceedings of Asian Conference on Computer Vision*, pages 576–585. Springer, 2007.

[139] J. Wang and Y. Yagi. Adaptive mean-shift tracking with auxiliary particles. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(6):1578–1589, Dec. 2009.

[140] Y. Wang, Z. Jing, and S. Hu. Data association for PHD filter based on MHT. In *Proceedings of the 11th International Conference on Information Fusion*, pages 1–8, 2008.

[141] Y. Wang and O. Lee. Active mesh-a feature seeking and tracking image sequence representation scheme. *IEEE Transactions on Image Processing*, 3(5):610–624, Sep. 1994.

[142] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of Tenth IEEE International Conference on Computer Vision*, volume 2, pages 1800–1807, 2005.

[143] M. Wölfel and J. W. McDonough. Combining multi-source far distance speech recognition strategies: Beamforming, blind channel and confusion network combination. In *Proceedings of interspeech*, pages 3149–3152, Sep. 2005.

[144] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 951–958, Jun. 2006.

[145] J. Wu, S. Hu, and Y. Wang. Adaptive multifeature visual tracking in a probability-hypothesis-density filtering framework. *Signal Processing*, 93(11):2915–2926, Nov. 2013.

[146] C. Yang, R. Duraiswami, and L. Davis. Efficient mean-shift tracking via a new similarity measure. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 176–183, 2005.

[147] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song. Recent advances and trends in visual tracking: A review. *Neurocomputing*, 74(18):3823–3831, 2011.

[148] A. Yilmaz, X. Li, and M. Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1531–1536, 2004.

[149] X. Yuan and S. Z. Li. Half quadratic analysis for mean shift: with extension to a sequential data mode-seeking method. In *Proceedings of International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[150] S. Zhong and F. Hao. Hand tracking by particle filtering with elite particles mean shift. In *Proceedings of Japan-China Joint Workshop on Frontier of Computer Science and Technology*, pages 163–167. IEEE, Dec. 2008.

[151] X. Zhong and A. B. Premkumar. Particle filtering approaches for multiple acoustic source detection and 2-D direction of arrival estimation using a single acoustic vector sensor. *IEEE Transactions on Signal Processing*, 60:4719–4733, Sept. 2012.

[152] S. K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13(11):1491–1506, 2004.

[153] X. Zhou, Y. F. Li, and B. He. Entropy distribution and coverage rate-based birth intensity estimation in GM-PHD filter for multi-target visual tracking. *Signal Processing*, 94:650–660, 2014.

[154] O. Zoidi, A. Tefas, and I. Pitas. Visual object tracking based on local steering kernels and color histograms. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(5):870–882, May 2013.