Audio-Visual Multiple-Speaker Tracking for Robot Perception

Yutong Ban

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques et Informatique**

Arrêté ministériel : 25 Mai 2016

Présentée par

Yutong Ban

Thèse dirigée par **Radu Horaud**
et codirigée par **Xavier Alameda-Pineda, Laurent Girin**

préparée au sein **INRIA Grenoble Rhône-Alpes**
et de **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Audio-Visual Multiple-Speaker Tracking for Robot Perception

Suivi Multi-Locuteurs avec des Informations Audio-Visuelles pour la Perception des Robots

Thèse soutenue publiquement le **10 May 2019**,
devant le jury composé de :

Dr. Radu Horaud
INRIA Grenoble Rhône-Alpes, Directeur de thèse
Dr. Xavier Alameda-Pineda
INRIA Grenoble Rhône-Alpes, Co-Encadrant de thèse
Prof. Dr. Andrea Cavallaro
Queen Mary University of London, Rapporteur
Prof. Dr. Laura Leal-Taixé
Technical University of Munich, Rapporteur
Prof. Dr. Jean-Luc Schwartz
CNRS, Gipsa-lab, Université Grenoble Alpes, Président
Dr. Sileye Ba
Dailymotion Paris, Examinateur

Abstract

Robot perception plays a crucial role in human-robot interaction (HRI). The perception system provides the robot with information of the surroundings and enables it to interact with people. In a conversational scenario, a group of people may chat in front of the robot and move freely. In such situations, robots are expected to understand where the people are, who is speaking, or what they are talking about. This thesis concentrates on answering the first two questions, namely speaker tracking and diarization. To that end, we use different modalities of the robot's perception system. Similar to seeing and hearing for humans, audio and visual information are critical cues for robots in a conversational scenario. Advancements in computer vision and audio processing in the last decade revolutionized robot perception abilities and enabled us to build joint audio-visual applications. In this thesis, we present the following contributions: we first develop a variational Bayesian framework for tracking multiple objects. The variational Bayesian framework provides closed-form tractable problem solutions, enabling an efficient tracking process. The framework is first applied to visual multiple-person tracking. The birth and death processes are built jointly to deal with the varying number of people in the scene. We then augment the framework by exploiting the complementarity of vision and robot motor information. On the one hand, the robot's active motion can be integrated into the visual tracking system to stabilize the tracking. On the other hand, visual information can be used to perform motor servoing. As a next step we combine audio and visual information in the framework and exploit the association between the acoustic feature frequency bins with tracked people, to estimate the smooth trajectories of people, and to infer their acoustic status (i.e. speaking or silent). To adapt the framework to applications with no vision information, we employ it to acoustic-only speaker localization and tracking. Online dereverberation techniques are first applied then followed by the tracking system. Finally, we propose a variant of the acoustic-only tracking model based on the von-Mises distribution, which is specifically adapted to directional data. All proposed methods are validated on datasets both qualitatively and quantitatively.

Résumé

La perception des robots joue un rôle crucial dans l'interaction homme-robot (HRI). Le système de perception fournit les informations au robot sur l'environnement, ce qui permet au robot de réagir en consequence. Dans un scénario de conversation, un groupe de personnes peut discuter devant le robot et se déplacer librement. Dans de telles situations, les robots sont censés comprendre où sont les gens, ceux qui parlent et de quoi ils parlent. Cette thèse se concentre sur les deux premières questions, à savoir le suivi et la diarisation des locuteurs. Nous utilisons différentes modalités du système de perception du robot pour remplir cet objectif. Comme pour l'humain, l'ouie et la vue sont essentielles pour un robot dans un scénario de conversation. Les progrès de la vision par ordinateur et du traitement audio de la dernière décennie ont révolutionné les capacités de perception des robots. Dans cette thèse, nous développons les contributions suivantes : nous développons d'abord un cadre variationnel bayésien pour suivre plusieurs objets. Le cadre bayésien variationnel fournit des solutions explicites, rendant le processus de suivi très efficace. Cette approche est d'abord appliqué au suivi visuel de plusieurs personnes. Les processus de créations et de destructions sont en adéquation avec le modèle probabiliste proposé pour traiter un nombre variable de personnes. De plus, nous exploitons la complémentarité de la vision et des informations du moteur du robot : d'une part, le mouvement actif du robot peut être intégré au système de suivi visuel pour le stabiliser ; d'autre part, les informations visuelles peuvent être utilisées pour effectuer l'asservissement du moteur. Par la suite, les informations audio et visuelles sont combinées dans le modèle variationnel, pour lisser les trajectoires et déduire le statut acoustique d'une personne: parlant ou silencieux. Pour experimenter un scenario où l'information visuelle est absente, nous essayons le modèle pour la localisation et le suivi des locuteurs basé sur l'information acoustique uniquement. Les techniques de déréverbération sont d'abord appliquées, dont le résultat est fourni au système de suivi. Enfin, une variante du modèle de suivi des locuteurs basée sur la distribution de von-Mises est proposée, celle-ci étant plus adaptée aux données directionnelles. Toutes les méthodes proposées sont validées sur des bases de données specifiques à chaque application.

ACKNOWLEDGMENT

First and foremost I would like to express my sincere gratitude to my supervisor Dr. Radu Horaud, for all the support and the encouragement during the three years. He is always willing to share his knowledge and experience with us. More importantly, he taught me how to become a researcher, to formulate and solve a research problem, and to publish and present the research work.

My sincere gratitude also goes to my thesis co-advisers, Dr. Xavier Alameda-Pineda and Dr. Laurent Girin. I will miss the time we derived equations together on white paper, as some were very impressive. I would particularly like to thank Xavier, for his availability every time I knocked on his door with questions.

I am also grateful to the other members of my thesis committees, Dr. Andrea Cavallaro, Dr. Laura Leal-Taixé, Dr. Jean-Luc Schwartz and Dr. Sileye Ba. It is my honor to have your presence at my defense. I would like to especially thank my thesis examiners Dr. Andrea Cavallaro and Dr. Laura Leal-Taixé. Thanks for devoting your time and effort to review my documents.

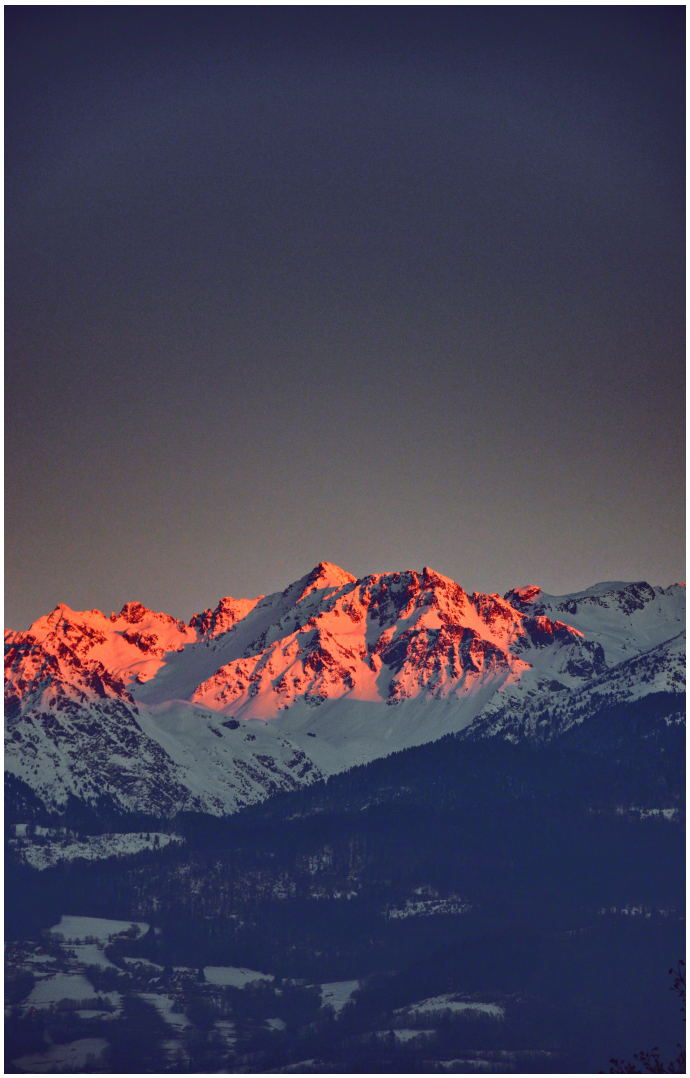
I would also like to thank other people that I collaborated with, Dr. Xiaofei Li, Dr. Sileye Ba, Dr. Georgios Evangelidis and Dr. Christine Evers. Collaborating with you expanded my horizon of the research world. I see many creative ideas come from your brilliant minds. I benefited a lot thanks to the inspirational discussions we had.

I would also like to thank the team engineers Guillaume S., Bastien, Fabien, and Quentin. Thanks to their work, we managed to run interesting demos on the robot. I would like to especially thank Soraya, for pushing me to write the thesis every day. Also, thanks to our secretary Nathalie Gillot for taking care of all the administrative work for us.

I truly enjoyed my time in the Perception team with all the lovely current and previous members: Israel, Vincent, Dionyssos, Stéphane, Benoit, Sylvain, Guillaume D., Yihong, Pablo, Simon, Mostafa, and others I forgot to mention.

I also appreciate the mountains around INRIA, they gave me a lot of good memories of Grenoble.

最后，我想感谢我的父母和家人，有你们提供的支持和环境，才能让我心无旁骛地投入到工作里。特别感谢我的未婚妻张雨昕，感谢你的陪伴和一次次的鼓励，我才得以顺利完成博士学业。



Shot from INRIA-06/02/2019 (f/5.6, 1/160, 105mm, ISO800)

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 GENERAL CONTEXT

Human-Robot Interaction (HRI) has been given growing attention in recent years. Owing to the rising advantage of machine intelligence technologies, the performance of the robot has been witnessing significant increase over the past decade. Robots, especially humanoid robots, have started to be employed in various public places, e.g. to welcome people at the reception, or to assist patients for medical care at hospitals. In such scenarios, people may talk directly to the robot or chat between themselves. Understanding the conversation and then taking part in it is therefore a crucial task. It not only demands the ability to solve single-modality-related tasks such as visual object detection or audio sound source localization, but also requires the knowledge to combine the multisensory data. Seeing and hearing are the two most essential abilities for humans to take part in conversations. Robots are also expected to use the two modalities for conversation. However, due to the different nature of the two modalities, combining audio and visual information is challenging. While visual information contains rich information about the environment, the visual field-of-view is limited. In contrast to vision, audio information is omnidirectional and it contains a speaker's voice information. However, audio is temporally sparse, because it is available only when the person speaks. In this thesis, we focus on two major problems: (i) where the person is in the scenario (ii) who speaks when. We first concentrate on tracking people using visual information. Secondly, we combine the robot motor information with the visual tracking system to extend the perception field-of-view and compensate the robot's ego-motion. Thirdly, we exploit the complementary nature of audio and visual information. We learn a mapping from image pixels to acoustic binaural features. This mapping allows us to track the speaking person using audiovisual cues, and jointly detect the speaking activeness of each person in a unified model. Moreover, for devices such as Amazon echo, no cameras are available in such cases. We thus concentrate on localizing and tracking the speaking person using only a microphone array in a reverberate environment, and further investigate online dereverberation and tracking techniques.

(a) (b) (c)

Figure 1.1: Robots in different scenarios

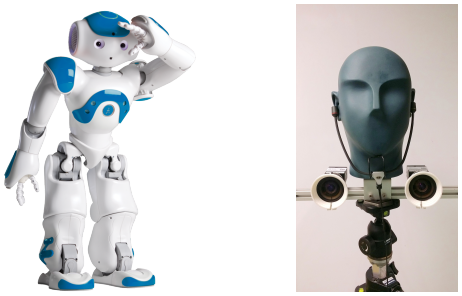
1.2 SCIENTIFIC CONTEXT AND MOTIVATION

This work was established at INRIA (French national research institute of computer science and automation), the Perception team, directed by Dr. Radu Horaud. The main research of the team focuses on joint audio visual applications and human activity analysis. One of the research scenarios is the *Cocktail party* scenario, where people chat and move freely in the crowded indoor environment. Such tasks are highly challenging for several reasons. At first, simultaneously talking speakers, reverberations and a noisy acoustic environment make it difficult to hear and understand a speaker. Moreover, solving problems in such a scenario requires the skills for extracting useful information from the crowded visual scene. Thirdly, once both audio and visual cues are extracted, combining multi-sensory data itself is a difficult task.

The work was established as a part of the Vision and Hearing in Action (VHIA) project, supported by ERC advanced Grant. The project targets audio visual perception and action applications on ego-centric robot platforms. Therefore, in this context, tracking a varying number of speaking people using both audio and visual cues is an essential part of the project.

The thesis was conducted under supervision by Dr. Radu Horaud. He was very patient to guide me during my thesis and gave me many kind and useful advices. Also I benefit great help from my co-supervisor Dr. Sileye Ba from 2015 to 2016, then followed by the co-supervision from Dr. Xavier Alameda-Pineda and Dr. Laurent Girin since 2016. It was so lucky for me to have a chance to collaborate with them.

This work also benefits a lot from two robot platforms in the team, namely NAO and Popeye. Both robots are equipped with audio visual sensors for perception. NAO is an autonomous, programmable humanoid robot platform developed by Aldebaran Robotics (former SoftBank Robotics). It is equipped with a pair of stereo cameras, four microphones and several joint motors for moving head, arms and legs. The popeye robot is an audio visual sensor platform which is designed to mimic the human audio-visual sensing system. The audio visual sensors are co-located on an acoustic dummy head. It consists of a pair of high resolution stereo cameras and six high quality microphones, which allows us to obtain high quality audio visual recordings.



(a) NAO robot (b) Popeye robot

Figure 1.2: Robot platforms used during this thesis

1.3 CONTRIBUTIONS

The contributions of this thesis are as follows:

- An on-line variational Bayesian model for multi-person tracking from visual observations provided by person detectors is proposed. This results in a variational expectation-maximization (VEM) algorithm with closed-form expressions both for the posterior distributions of the latent variables and for the estimation of the model parameters. A stochastic process for person birth is also developed to deal with the time-varying number of persons. The proposed method is benchmarked using the MOT16 challenge dataset.

- Multi-person tracking with a robotic platform is one of the cornerstones of human-robot interaction. Challenges arise from occlusions, appearance changes and a time-varying number of people. Furthermore, the final system is constrained by the hardware platform: low computational capacity and limited field-of-view. Therefore, we propose a method to simultaneously track a time-varying number of persons in three-dimensions and perform visual servoing. The complementary nature of the tracking and visual servoing enables the system to: (i) track several persons while compensating for large ego-movements and (ii) visually control the robot to keep a selected person of interest within the field of view. The variational Bayesian formulation allows us to effectively solve the inference problem through the use of closed-form solutions. More importantly, this leads to a computationally efficient procedure that runs at 10 FPS.

- In the context of conversational scenarios, we address the problem of tracking multiple speakers via the fusion of visual and auditory information. We propose to exploit the complementary nature of these two modalities in order to accurately estimate smooth trajectories of the tracked persons, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status – either speaking or silent – of each tracked person along with time. We propose to cast the problem at hand into a generative audio-visual fusion (or association) model

formulated as a latent-variable temporal graphical model. This may well be viewed as the problem of maximizing the posterior joint distribution of a set of continuous and discrete latent variables given the past and current observations, which is intractable. We propose a variational inference model which amounts to approximate the joint distribution with a factorized distribution. The solution takes the form of a closed-form expectation-maximization procedure. We describe in detail the inference algorithm, evaluate its performance and compare it with several baseline methods. These experiments show that the proposed audio-visual tracker performs well in informal meetings involving a time-varying number of people.

- We also address the problem of acoustic online multiple-speaker localization and tracking in a reverberant environment. We propose to use the direct-path relative transfer function (DP-RTF) – a feature that robustly encodes the inter-channel direct-path information against reverberation, hence well suited for reliable localization. A complex Gaussian mixture model (CGMM) is then used, such that each component weight represents the probability that an active speaker is present at a corresponding candidate source direction. Exponentiated gradient descent is used to update these weights online by minimizing a combination of negative log-likelihood and entropy. The latter imposes sparsity over the number of audio sources since in practice only a few speakers are simultaneously active. The outputs of this online localization process are then used as observations within a Bayesian filtering process whose computation is made tractable via an instance of variational expectation-maximization. Birth processes and speaker activity detection are used to account for the intermittent nature of speech. The method is thoroughly evaluated using several datasets.

- We propose to extend the variational Bayesian formulas to use the von Mises distribution. The von Mises distribution is a circular distribution which allows us to model audio-source directions of arrival (DOAs) with circular random variables. Since the DOAs are directional data, the circular random variables better fit the internal geometry of DOAs. This leads to a multi-target Kalman filter formulation which is intractable because of the combinatorial explosion of associating observations to state variables over time. We introduce a variational approximation of the filter's posterior distribution and we infer a variational expectation maximization (VEM) algorithm which is computationally efficient. We also propose an audio-source birth method that favors smooth source trajectories and which is used both to initialize the number of active sources and to detect new sources. We perform experiments with a recently released dataset comprising several moving sources as well as a moving microphone array.

1.4 DATASETS

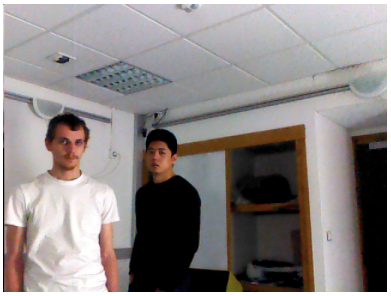
MOT Challenge Dataset Tracking the pedestrians has been studied for years in the tracking community. The MOT16 dataset thus provides a standard benchmark for tracking pedestrians, which is important for most applications in the computer vision domain. The dataset contains seven training sequences and seven evaluation sequences, which involves various settings such as the static camera or moving camera, indoor scenario or outdoor scenarios. The MOT16 dataset has been fully annotated. Moreover, the annotations not only comprise the labeled boxes but also provides different object classes and the visibility score of each object, allowing the community to evaluate their methods efficiently.

AVDIAR Dataset AVDIAR dataset [53] is created for audio-visual tracking and diarization. This dataset is challenging in terms of audio-visual analysis. There are usually several participants involved in informal conversations while wandering around. They are between two and four meters away from the audio-visual recording device. They take speech turns and often there are speech overlaps. Also, the speakers in the scenario are not necessarily facing the camera. The dataset is fully annotated. The visual annotations contains the centers, widths and heights of two bounding boxes for each person and in each video frame, a face bounding box. An identity (a number) is associated with each person through the entire dataset. The audio annotations comprise the speech status of each person over time (speaking or silent), with a minimum speech duration of 0.2 seconds. The audio source locations correspond to the centers of the face bounding boxes.

NAO-MPVS Dataset NAO-MPVS Dataset is a dataset for tracking multiple people while the robot motion is controlled. It is recorded using the NAO robot. Ten different sequences have been recorded in a regular living room scenario with its usual light source and background, where various people were moving around. The recorded sequences are thus challenging because of illumination variations, occlusions, appearance changes, and people leaving the robot's field-of-view. The dataset was recorded under two different high-level control rules: (i) the robot should servo the first tracked person and (ii) the robot should sequentially change the pursued person every three seconds. Aside from the visual recordings, the dataset also recorded the robot motor motion, which allows users to combine motor and visual information.

LOCATA Challenge The IEEE-AASP challenge on acoustic source LOCalization And TrAcking (LOCATA) provides a standard benchmark for acoustic source localization and tracking. The dataset contains real-life scenarios for an indoor environment. The LOCATA challenge includes six different tasks. The tasks involve localizing and tracking single static speaker, multiple static speakers, single moving speaker and multiple moving speakers, using static/moving microphone setups. In addition, the audio data in the challenge are recorded using different microphone setups, e.g. linear microphones, robot head and spherical microphone arrays.

(a) MOT Challenge dataset



(b) NAO-MPVS dataset

(c) AVDIAR dataset

(d) LOCATA Challenge

Figure 1.3: Datasets used in this thesis

The Kinovis-MST Dataset The Kinovis multiple speaker tracking (Kinovis-MST) datasets contain live acoustic recordings of various moving speakers in a reverberant environment. The data were recorded in the Kinovis multiple-camera laboratory at INRIA Grenoble Rhône-Alpes. The data were recorded with four microphones embedded into the head of a NAO robot. As there is a fan located inside the robot head nearby the microphones, there is a fair amount of stationary and spatially correlated microphone noise. The recordings contain between one and three moving participants that speak naturally, hence the number of active speech sources varies over time. Ground-truth trajectories and speech activity information were obtained with the Kinovis's motion capture system. Participants were wearing optical markers placed on their heads. Any time a participant is silent, he/she hides his/her infrared marker, thus allowing speaking/silent annotations of the recordings.

1.5 Manuscript Structure

This manuscript is organized as follows. Chapter 2 presents the fundamental probabilistic modelization and the detailed derivations of the proposed variational tracking model. In Chapter 3, we present the model which is applied on visual multiple objects tracking, as well as its combination with visual motor servoing. In Chapter 4, we focus on combining both audio and visual information to jointly track the speaker and detect the speaking activity. In Chapter 5, we present an online dereveberation localization and sound source tracking algorithm, as well as a variant of the model using von-Mises distribution for

bearing only speaker tracking. Finally, we conclude this thesis by discussing limitations and potential future work in Chapter 6.

CHAPTER 2

VARIATIONAL BAYESIAN MULTIPLE OBJECTS TRACKING

2.1 INTRODUCTION

Multiple objects tracking (MOT), or multiple targets tracking (MTT), has been studied for over 50 years. The problem arose and driven by applications in aerospace. Now MOT is applied within various disciplines, including aerospace traffic tracking, surveillance, remote sensing, computer vision, autonomous vehicles and also audio source tracking. In this section, we focus on presenting the existing methods for multiple objects tracking.

In the literature, the existing techniques are summarized into three main categories [150], namely multiple hypothesis tracking (MHT), random finite sets (RFS) and the joint probabilistic data association filter (JPDAF). In the tracking community, a hypothesis is a possible collection of compatible tracks representing a number of estimated trajectories [150]. At each time step, a single hypothesis tracking (SHT) algorithm summarizes all the past measurements in a single possible hypothesis. However, a multiple-hypothesis tracking (MHT) algorithm may consider plural potential possible tracking solutions for the measurements received in the past. Joint probabilistic data association filter (JPDAF) [16, 17] is one of the representative methods of single hypothesis tracking. JPDAF first evaluates all the possible combination between the current measurements and tracks, then combines all possible current hypotheses into a single one to form a single composite hypothesis. In practice, when a set of new measurements arrive, a gating process is first done to eliminate measurements too far from the target. Then an association matrix is carried out describing that the measurements belong to existed track, false measurements or not-yet tracked targets. The best solution for the association matrix is then found by an optimization algorithm like Hungarian method [115]. For each target, the measurements with the best association hypothesis are then combined with the target state dynamics, often Kalman filter [65]. Based on the distribution to model the number of the false measurements, JPDAF-based methods can be divided into two categories based on the distribution used for number of false measurement, where the Parametric JPDAF uses

Poisson distribution and Nonparametric JPDAF uses diffuse prior. In practice, JPDAF suffers from two pre-defined assumptions: (i) the number of the targets is known and set in advance; (ii) at most one of the validated measurements can be originated to the targets. However, such assumptions are not always true for tracking.

Different from an SHT, a MHT tracker [22, 23] keeps multiple hypotheses about the origin of the received data and has much higher computation and memory requirements. It is hoped that measurements in more than one scan provide more accurate assignments than those in a single scan. The idea was first introduced by [125]. There are two versions of MHT, hypothesis-oriented MHT (HOMHT)[125, 17] and track-oriented MHT (TOMHT)[22, 72]. At each time step, the ancestors of each hypothesis are stored. The algorithm generates all possible association hypothesis between the past hypothesis and current measurements. A common feature of the MHT methods is an exponentially growing number of hypothesis, which necessitates a pruning process. The pruning process scans all the hypothesis with a sliding window and discards the low probability hypothesis. However, generating all possible hypotheses only to discard most of them is inefficient. In addition, some hypotheses contain the same track, which makes efficiency even lower. An efficient HOMHT was implemented by [31], which aims to generate only the best hypotheses, and limit as few unnecessary hypotheses as possible. An efficient version of TOMHT was formulated by [16], in which a hypothesized target is represented by a target tree. Then the best global hypothesis is determined by solving a binary programming problem.

Another important approach for multiple objects tracking is the Random Finite Set (RFS) [149]. The RFS approach represents the multiple objects state as a finite set of single targets. Then the multiple objects tracking is formulated as a dynamic multiple objects state estimation problem. The first moment of RFS is known as the Probability Hypothesis Density (PHD). The PHD is a non-negative function whose integral over the selected region gives the expected number of the elements. The PHD filter [102] is the most popular variants among different RFS. It propagates the first moment of RFS instead of propagating the filtering density, which makes the computation less expensive. Under the linear Gaussian multi-object model, the PHD recursion yields to a closed form solution, which is called the Gaussian Mixture PHD filter [148]. In order to reduce the growing number of components, further post-processing to eliminate negligible components and merging similar components needs to be applied [148]. Particle filtering methods are further combined with PHD filter [149] to deal with no linear problems. Better performance is obtained by the Cardinalized PHD (CPHD) [100, 155], which jointly estimates the PHD and the cardinality distribution. CPHD also admits a closed-form solution. However, since it needs to estimate the cardinality, it involves a higher computational cost. Up to now, the PHD filters above are able to give smooth trajectories of the tracked objects and deal with the varying number of objects. However, PHD filters perform data association implicitly. This means the label/identity of the tracked objects can be obtained only by some additional post-processing steps. To overcome that, some efforts were made by the Generalized Labeled Multi-Bernoulli Tracker (GLMB) [154, 153]. In GLMB, tracked objects are attached with an additional identity label.

This Chapter presents the proposed multiple-object tracking problem formulation and the variational approximation used for solving the problem. The proposed tracker stays in the category of SHT. Unlike JPDAF, it does not suffer from the constraint that one track can only associate with one measurement. The variational approximation provides closed-form efficient solver which provides smooth trajectories based on assigning measurements and state dynamics. It automatically solves the problem of exponentially growing components number. Also, the proposed variational solution differs from the solution of JPDAF.

The rest of the Chapter is organized as follows, section 2.2 describes the notations and the probabilistic model. The section 2.3 details the intractability of the classic approach then followed by introducing the proposed variational solution. Finally the Chapter is concluded in 2.5.

2.2 PROBABILISTIC MODEL

2.2.1 MATHEMATICAL DEFINITIONS AND NOTATIONS

Unless otherwise specified, uppercase letters denote random variables while lowercase letters denote their realizations, *e.g.* $p(X = x)$, where $p(\cdot)$ denotes either a probability density function (pdf) or a probability mass function (pmf). For the sake of conciseness we generally write $p(x)$. Vectors are written in slanted bold, *e.g.* $\boldsymbol{X}, \boldsymbol{x}$, whereas matrices are written in bold, *e.g.* $\mathbf{Y}, \mathbf{y}$. Let t denote the common frame index. Let N be the upper bound of the number of persons that can simultaneously be tracked at any time t, and let $n \in \{1 \dots N\}$ be the person index. Let $n = 0$ denote *nobody*. A t subscript denotes variable concatenation at time t, e.g. $\mathbf{X}_t = (\boldsymbol{X}_{t1}, \dots, \boldsymbol{X}_{tn}, \dots, \boldsymbol{X}_{tN})$, and the subscript $1:t$ denotes concatenation from 1 to t, e.g. $\mathbf{X}_{1:t} = (\mathbf{X}_1, \dots, \mathbf{X}_t)$.

Assume that we track one dimensional position. Let $X_{tn} \in \mathcal{X} \subset \mathbb{R}$, $Y_{tn} \in \mathcal{Y} \subset \mathbb{R}$ be two latent variables that correspond to the 1D position and 1D velocity of person n at t. Let $\boldsymbol{S}_{tn} = (X_{tn}^\top, Y_{tn}^\top)^\top \subset \mathbb{R}^2$ be the complete set of continuous latent variables at t, where $^\top$ denotes the transpose operator. In this chapter a person is characterized with the position and velocity.

We now define the observations. Let $\{o_{tm}\}_{m=1}^{M_t}$ be realizations of the random observed variables $\{O_{tm}\}_{m=1}^{M_t}$. An observation $o_{tm} \in \mathcal{V} \subset \mathbb{R}$ corresponds to a detected person position. Note that the number of observations at t, M_t vary over time. Let $\boldsymbol{o}_{1:t} = (\boldsymbol{o}_1, \dots, \boldsymbol{o}_t)$ denote the set of observations from 1 to t.

We now define the assignment variables. The explicit data association in the proposed model is realized by the assignment variable, namely Z_t. It is a discrete variable which indicates the association between the observation to the person state. e.g. $p(Z_{tm} = n)$ denotes the probability of assigning the observation m at t to person n. Moreover, the proposed model contains a clutter class, where $p(Z_{tm} = 0)$ are the probabilities of assigning the observation m to none of the persons, or to nobody.

2.2.2 THE FILTERING DISTRIBUTION

We remind that the objective of tracking is to estimate the positions and velocities of participants (multiple person tracking). The audio-visual multiple-person tracking problem is cast into the problems of estimating the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t})$ and of inferring the state variable $\mathbf{S}_t$.

The problem at hand can now be cast into the estimation of the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t})$, and further inference of $\mathbf{S}_t$ and $\boldsymbol{Z}_t$. We make two hypotheses, namely (i) that the observations at frame t only depend on the assignment and state variables at t, and (ii) that the prior distribution of the assignment variables is independent of all the other variables. By applying the Bayes rule together with these hypotheses, and ignoring terms that do not depend on $\mathbf{S}_t$ and $\boldsymbol{Z}_t$, one can then write the filtering distribution of $(\mathbf{s}_t, \boldsymbol{z}_t)$ as:

$$p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t}) \propto p(\boldsymbol{o}_t|\mathbf{s}_t, \boldsymbol{z}_t)p(\boldsymbol{z}_t)p(\mathbf{s}_t|\boldsymbol{o}_{1:t-1}), \tag{2.1}$$

where the factorization consists of three components, $p(\boldsymbol{o}_t|\mathbf{s}_t, \boldsymbol{z}_t)$ is the observed-data likelihood., $p(\boldsymbol{z}_t)$ is the prior distribution of the assignment variable and $p(\mathbf{s}_t|\boldsymbol{o}_{1:t-1})$ is the predictive distribution. Each component is detailed in the following section.

2.2.3 THE OBSERVED-DATA LIKELIHOOD

As already mentioned above (Section 2.2.1), an observation o_{tm} is the position of the detected person position, namely $o_{tm} \in \mathcal{O} \subset \mathbb{R}$. Since the velocity is not observed, a 1×2 projection matrix $\mathbf{P} = [1\ 0]$ is used to project $\boldsymbol{s}_{tn}$ onto $\mathcal{O}$. Assuming that the M_t visual observations $\{o_{tm}\}_{m=1}^{M_t}$ available at t are independent.

$$p(\boldsymbol{o}_t|\mathbf{s}_t, \boldsymbol{z}_t) = \prod_{m=1}^{M_t} p(o_{tm}|\mathbf{s}_t, Z_{tm}), \tag{2.2}$$

where the observed positions are drawn from the following distributions:

$$p(o_{tm}|\mathbf{s}_t, Z_{tm} = n) = \begin{cases} \mathcal{N}(o_{tm}; \mathbf{P}\boldsymbol{s}_{tn}, \Phi_{tm}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(o_{tm}; \text{vol}(\mathcal{O})) & \text{if } n = 0, \end{cases} \tag{2.3}$$

where $\Phi_{tm} \in \mathbb{R}$ correspond to the covariance quantifying the measurement error of the $\mathcal{U}(\cdot; \text{vol}(\cdot))$ is the uniform distribution with $\text{vol}(\cdot)$ being the support volume of the variable space. If an observation belongs to a person, it is considered to follow a Gaussian distribution centered at the person position. Otherwise, if an observation is generated from the noise, or to say it belongs to nobody, it is viewed as uniformly distributed in the observation space.

2.2.4 THE PRIOR DISTRIBUTION

The observation-to-person assignments are assumed to be a priori independent so that the probabilities in (2.1) factorize as:

$$p(\boldsymbol{Z}_t) = \prod_{m=1}^{M_t} p(Z_{tm}), \tag{2.4}$$

It makes sense to assume that these distributions do not depend on t and that they are uniform. The following notations are introduced: $\eta_{mn} = p(Z_{tm} = n) = 1/(N+1)$.

2.2.5 THE PREDICTIVE DISTRIBUTION

$$p(\mathbf{s}_t|\boldsymbol{o}_{1:t-1}) = \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) p(\mathbf{s}_{t-1}|\boldsymbol{o}_{1:t-1}) d\mathbf{s}_{t-1}. \tag{2.5}$$

Eq. (2.5) is the predictive distribution of $\mathbf{s}_t$ given the past observations, i.e. from 1 to $t-1$. The state dynamics in (2.5) is modeled with a linear-Gaussian first-order Markov process. Moreover, it is assumed that the dynamics are independent over speakers:

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1\,n}, \boldsymbol{\Lambda}_{tn}), \tag{2.6}$$

where $\boldsymbol{\Lambda}_{tn}$ is the dynamics covariance matrix and $\mathbf{D}$ is the state transition matrix. Here we consider a constant velocity model, $\mathbf{D}$ is thus given by:

$$\mathbf{D} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

As described in Section 2.3 below, an important feature of the proposed model is that the predictive distribution (2.5) at frame t is computed from the state dynamics model (2.6) and an approximation of the filtering distribution $p(\mathbf{s}_{t-1}|\boldsymbol{o}_{1:t-1})$ at frame $t-1$, which also factorizes across person. As a result, the computation of (2.5) factorizes across speakers as well.

2.3 VARIATIONAL INFERENCE

2.3.1 SOLUTION WITH STANDARD EXPECTATION MAXIMIZATION ALGORITHM

In this section we try to solve the problem using the expectation maximization (EM) [36] algorithm. The aim of the EM algorithm is to find maximum likelihood solutions for models having latent variables. The two latent variables in the proposed model $\boldsymbol{z}_t$ and $\mathbf{s}_t$ are estimated in the Expectation steps (E-steps), namely E-Z step and E-S step, while the parameters are estimated in M-steps.

To solve the model by EM algorithm, we need to make two assumptions: (i) we assume at $t-1$, the state of the each person is independent. (ii) the posterior distribution of each person's state $p(\mathbf{s}_{t-1\,n}|\boldsymbol{o}_{1:t-1})$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{s}_{t-1\,n}; \boldsymbol{\mu}_{t-1\,n}, \Gamma_{t-1\,n})$. With the two assumptions, all the probabilities in (2.1) are defined.

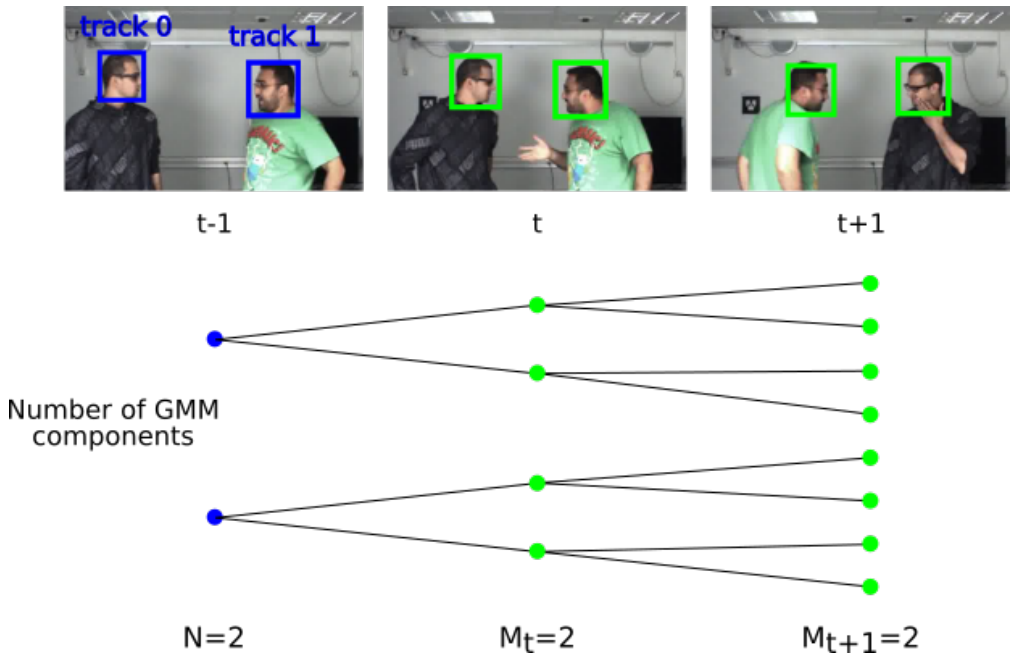


Figure 2.1: Illustration of growing number of GMM components in $p(\mathbf{s}_t|\boldsymbol{o}_{1:t})$ with EM solution. Green: detections; Blue: tracked person-state. Assume at $t-1$ there are two tracked person and each of them follows a Gaussian distribution. Then at $t+1$ the number of the GMM components in the filtering distribution $p(\mathbf{s}_t|\boldsymbol{o}_{1:t})$ will be $N \times M_t \times M_{t+1} = 8$.

§ E-Z step

The posterior of the assignment variable is obtained by marginalize the filtering distribution over the state variable $\mathbf{s}_t$.

$$\begin{aligned} p(\boldsymbol{z}_t|\boldsymbol{o}_{1:t}) &= \int p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t}) d\mathbf{s}_t \\ &= \int \prod_{m=1}^{M_t} \eta_{mn} \mathcal{N}(o_{tm}; \mathbf{P}\boldsymbol{s}_{tn}, \Phi_{tm})^{\delta(Z_{tm}=n)} \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1\,n}, \boldsymbol{\Lambda}_{tn}) d\mathbf{s}_t \end{aligned} \tag{2.7}$$

By driving the (2.7), we obtain

$$p(Z_{tm} = n|\boldsymbol{o}_{1:t}) = \alpha_{tmn} = \frac{\tau_{tmn}\eta_{mn}}{\sum_{i=0}^{N} \tau_{tmi}\eta_{mi}} \tag{2.8}$$

where

$$\tau_{tmn} = \begin{cases} \mathcal{N}(o_{tm}; \mathbf{PD}\boldsymbol{\mu}_{t-1\,n}, \mathbf{P}(\mathbf{D}\Gamma_{t-1\,n}\mathbf{D}^T + \boldsymbol{\Lambda}_{tn})\mathbf{P}^T + \Phi_{tm}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(o_{tm}; \text{vol}(\mathcal{O})) & \text{if } n = 0. \end{cases}$$

§ E-S step

The posterior distribution for the state variable is obtained by marginalizing the filtering distribution (2.1) over the assignment variable $\boldsymbol{z}_t$.

$$\begin{aligned} p(\mathbf{s}_t|\boldsymbol{o}_{1:t}) &= \int p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t}) d\boldsymbol{z}_t \\ &= \int \prod_{m=1}^{M_t} \eta_{mn} \mathcal{N}(o_{tm}; \mathbf{P}\boldsymbol{s}_{tn}, \Phi_{tm})^{\delta(Z_{tm}=n)} \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1\,n}, \boldsymbol{\Lambda}_{tn}) d\boldsymbol{z}_t \end{aligned} \quad (2.9)$$

By further deriving the equation (2.9), we obtain:

$$p(\boldsymbol{s}_{tn}|\boldsymbol{o}_{1:t}) = \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1\,n}, \mathbf{D}\Gamma_{t-1\,n}\mathbf{D}^T + \boldsymbol{\Lambda}_{tn}) \sum_{m=1}^{M_t} \alpha_{tmn} \mathcal{N}(o_{tm}; \mathbf{P}\boldsymbol{s}_{tn}, \Phi_{tm}) \quad (2.10)$$

As shown in (2.10), the posterior distribution of each person's position at time t is in the form of Gaussian mixture model (GMM), which contains M_t components. Assume the position of each person at initial time $t = 1$ follows the Gaussian distribution. Then as time step advances, the number of the GMM components in the filtering distribution grows exponentially. Such expression is difficult to implement in practice, which makes the problem intractable. A toy example is shown in Fig. 2.1. The M-steps in EM algorithm to estimate the parameters are similar with the M-steps in the proposed Variational EM algorithm, which will be shown in the next section.

2.3.2 VARIATIONAL INFERENCE

Since the direct estimation of the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t})$ is intractable. We thus overcome this problem via variational inference and associated EM closed-form solver [19, 133]. The variational inference has two main advantages: (i) it gives the intractable integral a closed-form solution; (ii) the variational approximation makes it possible to extend to more complex models and to use other probabilistic distributions other than Gaussian (*e.g.* von Mises distribution).

Let $\mathcal{F}$ be a set of pdfs over the latent variables $\mathbf{s}, \boldsymbol{z}$. For the parameter set $\boldsymbol{\theta}$, variational inference targets to seek the optimal member $q^\star$ in the variational family $\mathcal{F}$ that the Kullback-Leibler (KL) divergence from the true posterior $p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t})$ is minimized:

$$q^\star(\mathbf{s}_t, \boldsymbol{z}_t) = \underset{q(\boldsymbol{s}_t, \boldsymbol{z}_t) \in \mathcal{F}}{\text{argmin}} \; D_{KL}(q(\boldsymbol{s}_t, \boldsymbol{z}_t) \| p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t}, \boldsymbol{\theta})) \quad (2.11)$$

where the KL divergence is

$$D_{KL}(q(\boldsymbol{s}_t, \boldsymbol{z}_t) \| p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t}; \boldsymbol{\theta})) = -\int q(\mathbf{s}_t, \boldsymbol{z}_t) \ln \left\{ \frac{p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t})}{q(\mathbf{s}_t, \boldsymbol{z}_t)} \right\} d\mathbf{s}_t d\boldsymbol{z}_t \tag{2.12}$$

From the definition of the KL divergence we can obtain:

$$D_{KL}(q(\boldsymbol{s}_t, \boldsymbol{z}_t) \| p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t}; \boldsymbol{\theta})) = \ln p(\boldsymbol{o}_{1:t}; \boldsymbol{\theta}) - \mathcal{L}(q; \boldsymbol{\theta}) \tag{2.13}$$

where $\ln p(\boldsymbol{o}_{1:t}; \boldsymbol{\theta})$ is the marginal log-likelihood and $\mathcal{L}(q; \boldsymbol{\theta})$ is known as the lower bound, or variational free energy, defined as:

$$\mathcal{L}(q; \boldsymbol{\theta}) = \int q(\mathbf{s}_t, \boldsymbol{z}_t) \ln \left\{ \frac{p(\mathbf{s}_t, \boldsymbol{z}_t, \boldsymbol{o}_{1:t})}{q(\mathbf{s}_t, \boldsymbol{z}_t)} \right\} d\mathbf{s}_t d\boldsymbol{z}_t \tag{2.14}$$

where $\ln(p(\mathbf{s}_t, \boldsymbol{z}_t, \boldsymbol{o}_{1:t}))$ is the complete data log-likelihood. From (2.13) we can see that maximizing the lower bound $\mathcal{L}(q; \boldsymbol{\theta})$ by optimization with respect to the distribution $q(\mathbf{s}_t, \boldsymbol{z}_t)$ is equivalent to minimizing the KL divergence.

Here we consider a way to restrict the variational family $\mathcal{F}$ of distribution $q(\mathbf{s}_t, \boldsymbol{z}_t)$, namely mean-field approximation, more precisely $p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t})$ is approximated with the following factorized form:

$$p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t}) \approx q(\mathbf{s}_t, \boldsymbol{z}_t) = q(\mathbf{s}_t) q(\boldsymbol{z}_t), \tag{2.15}$$

which implies

$$q(\mathbf{s}_t) = \prod_{n=1}^{N} q(\boldsymbol{s}_{tn}), \; q(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} q(z_{tm}), \tag{2.16}$$

where $q(Z_{tm} = n)$ are the variational posterior probabilities of assigning observation m to person n. The proposed variational approximation (2.15) amounts to break the conditional dependence of $\mathbf{S}$ and $\boldsymbol{Z}$ with respect to $\boldsymbol{o}_{1:t}$ which causes the computational intractability. This factorized approximation makes the calculation of $p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t})$ tractable. The optimal solution is given by an instance of the variational expectation maximization (VEM) algorithm [19, 133], which alternates between two steps until convergence:

- *Variational* E-step: find the optimal variational posterior distribution by $q^\star = \operatorname{argmax}_{q \in \mathcal{F}} \mathcal{L}(q; \boldsymbol{\theta}^\star)$, the approximate log-posterior distribution of each one of the latent variables is estimated by taking the expectation of the complete-data log-likelihood over the remaining latent variables, i.e. (2.17) and (2.18) below, and
- M-step: model parameters are estimated by maximizing the variational expected complete-data log-likelihood: $\boldsymbol{\theta}^\star = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q^\star; \boldsymbol{\theta})$.

In the case of the proposed model the latent variable log-posteriors write:

$$\log q^\star(\boldsymbol{s}_{tn}) = \mathbb{E}_{q(\boldsymbol{z}_t) \prod_{\ell \neq n} q(\boldsymbol{s}_{t\ell})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t})], \tag{2.17}$$

$$\log q^\star(z_{tm}) = \mathbb{E}_{q(\mathbf{s}_t) \prod_{\ell \neq m} q(z_{t\ell})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t})], \tag{2.18}$$

The expression of the optimal variational posterior distribution for each latent variable, $q^\star(\boldsymbol{s}_{tn})$ and $q^\star(z_{tm})$ can be obtained with (2.17) and (2.18). We now try to derive the two equations. But the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t})$ in (2.17) and (2.18) are not fully defined yet. The predictive distribution $p(\mathbf{s}_{t-1}|\boldsymbol{o}_{1:t-1})$ is incomplete. A remarkable consequence of the factorization (2.15) is that $p(\mathbf{s}_{t-1}|\boldsymbol{o}_{1:t-1})$ is replaced with $q^\star(\mathbf{s}_{t-1}) = \prod_{n=1}^{N} q(\boldsymbol{s}_{t-1\,n})$, consequently (2.5) becomes:

$$p(\mathbf{s}_t|\boldsymbol{o}_{1:t-1}) \approx \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_{n=1}^{N} q^\star(\boldsymbol{s}_{t-1\,n}) d\mathbf{s}_{t-1}. \tag{2.19}$$

It is now assumed that the variational posterior distribution $q^\star(\boldsymbol{s}_{t-1\,n})$ is Gaussian with mean $\boldsymbol{\mu}_{t-1\,n}$ and covariance $\boldsymbol{\Gamma}_{t-1\,n}$:

$$q^\star(\boldsymbol{s}_{t-1\,n}) = \mathcal{N}(\boldsymbol{s}_{t-1\,n}; \boldsymbol{\mu}_{t-1\,n}, \boldsymbol{\Gamma}_{t-1\,n}). \tag{2.20}$$

By substituting (2.20) into (2.19) and combining it with (2.6), the predictive distribution (2.5) becomes:

$$p(\mathbf{s}_t|\boldsymbol{o}_{1:t-1}) \approx \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1\,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}). \tag{2.21}$$

Note that the above distribution factorizes across persons. Now that all the factors in (2.1) have tractable expressions, A VEM algorithm can be applied. When we see the solution of E-S-step, we can found that based on the assumption of (2.20), the variational posterior distribution $q^\star(\boldsymbol{s}_{tn})$ also follows a Gaussian distribution (detailed in the next paragraph). Due to this recursiveness, we don't necessarily need to make the assumption of (2.20) at every time step. It is sufficient to make the assumption at initial time $t = 1$, i.e. $q^\star(\boldsymbol{s}_{1n}) = \mathcal{N}(\boldsymbol{s}_{1n}; \boldsymbol{\mu}_{1n}, \boldsymbol{\Gamma}_{1n})$, whose parameters may be easily initialized. In addition, for the most cases in this thesis, we use Gaussian distribution based models. Therefore we initialize $q^\star(\boldsymbol{s}_{1n})$ with a Gaussian distribution. But it is not always the case. In Section 5.6, we use a von-Mises distribution to initialize $q^\star(\boldsymbol{s}_{1n})$ since the filtering distribution in Section 5.6 is formulated with circular distribution.

2.4 VARIATIONAL EXPECTATION MAXIMIZATION

The proposed VEM algorithm iterates between an E-S-step, an E-Z-step, and an M-step on the following grounds.

§ E-S-step

The per-person variational posterior distribution of the state vector $q^\star(\boldsymbol{s}_{tn})$ is evaluated by developing (2.17). The complete-data likelihood $p(\mathbf{s}_t, \boldsymbol{z}_t|\boldsymbol{o}_{1:t})$ in (2.17) is the product of (2.2), (2.4) and (2.21). We thus first sum the logarithms of (2.2), of (2.4) and of (2.21).

Then we ignore the terms that do not involve $\boldsymbol{s}_{tn}$. Evaluation of the expectation over all the latent variables except $\boldsymbol{s}_{tn}$ yields the following Gaussian distribution:

$$q^{\star}(\boldsymbol{s}_{tn}) = \mathcal{N}(\boldsymbol{s}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}), \tag{2.22}$$

with:

$$\boldsymbol{\Gamma}_{tn} = \left(\sum_{m=1}^{M_t} \alpha_{tmn} \mathbf{P}^{\top} \Phi_{tm}^{-1} \mathbf{P} + \left(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n} \mathbf{D}^{\top} \right)^{-1} \right)^{-1}, \tag{2.23}$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn} \left(\sum_{m=1}^{M_t} \alpha_{tmn} \mathbf{P}^{\top} \Phi_{tm}^{-1} o_{tm} + \left(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n} \mathbf{D}^{\top} \right)^{-1} \mathbf{D}\boldsymbol{\mu}_{t-1\,n} \right), \tag{2.24}$$

where $\alpha_{tmn} = q^{\star}(Z_{tm} = n)$ is computed in the E-Z-step below. As mentioned before, once $q^{\star}(\boldsymbol{s}_{1n})$ is initialized with a Gaussian distribution, the variational posterior distribution $q^{\star}(\boldsymbol{s}_{tn})$ follows a Gaussian at each time frame t. The Gaussian distribution obtained mixes the previous prediction and the current observation automatically. This overcomes the problem of exponentially growing number of the components in the filtering distribution by the nature of the variational approximation.

§ E-Z-step

by developing (2.18), and following the same reasoning as above, we obtain the following closed-form expression for the variational posterior distribution of the visual assignment variable:

$$\alpha_{tmn} = q^{\star}(Z_{tm} = n) = \frac{\tau_{tmn}\eta_{mn}}{\sum_{i=0}^{N} \tau_{tmi}\eta_{mi}}, \tag{2.25}$$

where τ_{tmn} is given by:

$$\tau_{tmn} = \begin{cases} \mathcal{N}(o_{tm}; \mathbf{P}\boldsymbol{\mu}_{tn}, \Phi_{tm}) e^{-\frac{1}{2}\mathrm{tr}\left(\mathbf{P}^{\top} \Phi_{tm}^{-1} \mathbf{P} \boldsymbol{\Gamma}_{tn}\right)} & \text{if } 1 \leq n \leq N \\ \mathcal{U}(o_{tm}; \mathrm{vol}(\mathcal{O})) \qquad \text{if } n = 0. \end{cases}$$

§ M-steps

The entries of covariance matrix of the state dynamics, $\boldsymbol{\Lambda}_{tn}$, are the only parameters that need be estimated. To this aim, we develop $\mathbb{E}_{q^{\star}(\mathbf{s}_t) q^{\star}(\boldsymbol{z}_t)}[\log p(\mathbf{s}_t, \boldsymbol{z}_t | \boldsymbol{o}_{1:t})]$ and ignore the terms that do not depend on $\boldsymbol{\Lambda}_{tn}$. We obtain:

$$J(\boldsymbol{\Lambda}_{tn}) = \mathbb{E}_{q^{\star}(\boldsymbol{s}_{tn})} \big[\log \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1\,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^{\top} + \boldsymbol{\Lambda}_{tn}) \big],$$

which can be further developed as:

$$\begin{aligned} J(\boldsymbol{\Lambda}_{tn}) = \log |\mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^{\top} + \boldsymbol{\Lambda}_{tn}| + \mathrm{Tr}\big((\mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^{\top} + \boldsymbol{\Lambda}_{tn})^{-1} \\ \times \left((\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})(\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})^{\top} + \boldsymbol{\Gamma}_{tn} \right) \big). \end{aligned} \tag{2.26}$$

Hence, by differentiating (2.26) with respect to $\mathbf{\Lambda}_{tn}$ and equating to zero, we obtain:

$$\mathbf{\Lambda}_{tn} = \mathbf{\Gamma}_{tn} - \mathbf{D}\mathbf{\Gamma}_{t-1\,n}\mathbf{D}^\top + (\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})(\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1\,n})^\top. \tag{2.27}$$

Similarly, the M-Step for observation covariances corresponds to the estimation of Φ_{tm}. This is done by maximizing

$$J(\Phi_{tm}) = \mathbb{E}_{q^\star(\boldsymbol{s}_t)q^\star(Z_{tm})}\big[\log \mathcal{N}(o_{tm}, \mathbf{P}\boldsymbol{s}_{tn}, \Phi_{tm})^{\delta(Z_{tm}=n)}\big],$$

which is:

$$J(\Phi_{tm}) = \log|\Phi_{tm}| + \sum_{n=1}^{N} \alpha_{tmn}\mathrm{Tr}\big(\Phi_{tm}^{-1}\left(\mathbf{P}\Gamma_{tn}\mathbf{P}^\top + (o_{tm} - \mathbf{P}\boldsymbol{\mu}_{tn})(o_{tm} - \mathbf{P}\boldsymbol{\mu}_{tn})^\top\right)\big). \tag{2.28}$$

Differentiating $J(\Phi_{tm})$ with respect to Φ_{tm} and equating to zero gives:

$$\Phi_{tm} = \frac{1}{N}\sum_{n=1}^{N} \alpha_{tmn}\left(\mathbf{P}\Gamma_{tn}\mathbf{P}^\top + (o_{tm} - \mathbf{P}\boldsymbol{\mu}_{tn})(o_{tm} - \mathbf{P}\boldsymbol{\mu}_{tn})^\top\right) \tag{2.29}$$

2.5 CONCLUSION

In this Chapter, we first show the probabilistic formulation of the multiple-object tracking problem. Then we prove the solution using the standard EM algorithm is intractable because the number of components grows exponentially in the mixture model. Furthermore, a variational approximation is proposed to solve the problem. The variational approximation gives closed-form tractable solution which is simple to implement. Besides, the variational framework can be easily extended to more complex models. In the following chapters, the variants of the proposed model for different applications are presented. The applications include visual multiple-person tracking, audio-visual speaker tracking, and acoustic-only speaker tracking. One of the variants uses von Mises distribution for bearing-only acoustic speaker tracking, which shows the possibility of extending the proposed variational framework to use other distributions rather than the Gaussian distribution.

CHAPTER 3

VISUAL MULTIPLE OBJECT TRACKING

3.1 VISUAL MULTIPLE PEDESTRIAN TRACKING

3.1.1 INTRODUCTION

In this section we focus on tracking a varying number of objects using visual information. The problem of object tracking is ubiquitous in computer vision. While many object tracking methods are available, multiple-person tracking remains extremely challenging [91]. In addition to the difficulties related to single-object tracking (occlusions, self-occlusions, visual appearance variability, unpredictable temporal behavior, etc.), tracking a varying and unknown number of objects makes the problem more challenging, for the following reasons: (i) the observations associated with detectors need to be associated to objects being tracked, which includes the process of discarding detection errors, (ii) the number of objects is not known in advance and hence it must be estimated and updated over time, (iii) mutual occlusions (not present in single-tracking scenarios) must be robustly handled, and (iv) the number of objects varies over time and one has to deal with hidden states of varying dimensionality, from zero when there is no visible object, to a large number of detected objects. Note that in this case and if a Bayesian setting is being considered, as is often the case, an exact recursive filtering solution is intractable.

Several multiple-person tracking methods have been proposed within the trans-dimensional Markov chain model [58], where the dimensionality of the state-space is treated as a state variable. This allows to track a variable number of objects by jointly estimating the number of objects and their states. [66, 136, 160] exploited this framework for tracking a varying number of objects. The main drawback is that the states are inferred by means of a reversible jump Markov-chain Monte Carlo sampling, which is computationally expensive [57]. The random finite set framework proposed in [101, 98, 99] is also very popular, where the targets are modeled as realizations of a random finite set which is composed of an unknown number of elements. Because an exact solution to this model is computationally intensive, an approximation known as the probability hypothesis density (PHD) filter was proposed [96]. Further sampling-based approximations of random-set based filters

were subsequently proposed, e.g. [132, 30, 151]. These were exploited in [92] for tracking a time-varying number of active speakers using auditory cues and in [95] for multiple-target tracking using visual observations. Recently, conditional random fields have been introduced to address multiple-target tracking [159, 111, 60]. In this case, tracking is cast into an energy minimization problem. In radar tracking, popular multiple-target tracking methods are joint probabilistic data association (JPDA), and multiple hypothesis filters [15].

An interesting and less investigated framework for multiple-target tracking is the variational Bayesian class of models for tracking an unknown and varying number of persons. Although variational models are very popular in machine learning, their use for object tracking has been limited to tracking a fixed number of targets [145]. Variational Bayes methods approximate the joint a posteriori distribution of the complete set of latent variables by a separable distribution [134, 20]. In an online tracking scenario, where only past and current observations are available, this leads to approximating the filtering distribution. An interesting aspect of variational methods is that they yield closed-form expressions for the posterior distributions of the hidden variables and for the model parameters, thus enabling an intrinsically efficient filtering procedure implemented via a variational EM (VEM) algorithm. In this Chapter, we derive a variational Bayesian formulation for multiple-person tracking, and present results on the MOT 2016 challenge dataset [110]. The proposed method extends [6] in many apsects: (i) the assignment variables are included in the filtering equation and therefore the state variables and the assignment variables are jointly inferred, (ii) a temporal window is incorporated in the visibility process, leading to a tracker that is more robust to misdetections, (iii) death process allows to forget about *old* tracks and thus opens the door to large-scale processing, as needed in many realistic situations. Finally, full evaluation of the proposed tracker within the MOT 2016 challenge dataset assesses its performance against other state-of-the-art methods in a principled and systematic way. Examples of results obtained with our method and Matlab code are publicly available.[1]

The remainder of this section is organized as follows. Section 3.1.2 details the proposed Bayesian model and a variational solution is presented in Section 3.1.3. In Section 3.1.4, we depict the birth, visibility and death processes allowing to handle an unknown and varying number of persons. Section 3.1.5 presents benchmarking results.

3.1.2 PROBABILISTIC MODEL

We start by recalling the notations for visual object tracking. Let N be the maximum number of persons. The kinematic state of person $n \leq N$ is a random vector $\boldsymbol{S}_{tn} = (\boldsymbol{X}_{tn}^\top, \boldsymbol{Y}_{tn}^\top)^\top \in \mathbb{R}^6$, where $\boldsymbol{X}_{tn} \in \mathbb{R}^4$ is the person location and size, i.e., 2D image position, width and height, and $\boldsymbol{Y}_{tn} \in \mathbb{R}^2$ is the person velocity in the image plane. The multiple-person state random vector is denoted by $\mathbf{s}_t = (\mathbf{s}_{t1}^\top, \ldots, \mathbf{s}_{tN}^\top)^\top \in \mathbb{R}^{6N}$.

We assume the existence of a person detector, providing M_t localization observations at each time t. The m-th localization observation delivered by the detector at

[1] https://team.inria.fr/perception/research/ovbt/

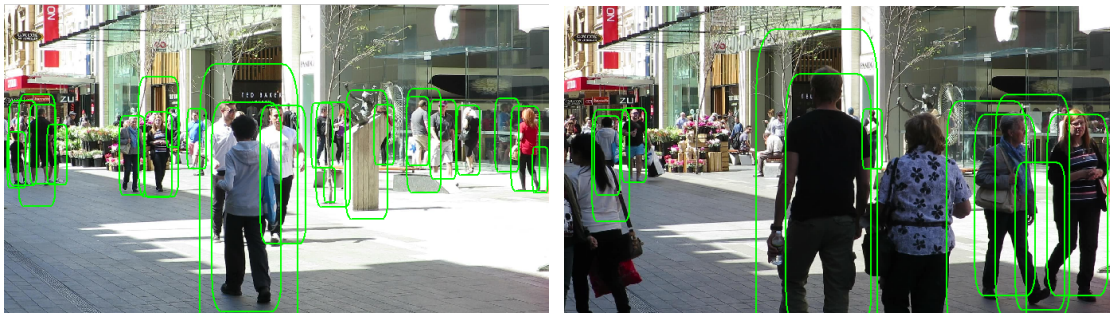


Figure 3.1: Examples of detected persons from the MOT 2016 dataset.

time t is denoted by $\boldsymbol{o}_{tm} \in \mathbb{R}^4$, and represents the location (2D position, width, height) of a person, e.g. Figure 3.1. The set of observations at time t is denoted by $\boldsymbol{o}_t = \{\boldsymbol{o}_{tm}\}_{m=1}^{M_t}$. Associated to $\boldsymbol{o}_{tm}$, there is a photometric description of the person appearance, denoted by $\mathbf{h}_{tm}$. This photometric observation is extracted from the bounding box of $\boldsymbol{o}_{tm}$. Altogether, the localization and photometric observations constitute the observations $\boldsymbol{o}_{tm} = (\boldsymbol{o}_{tm}, \mathbf{h}_{tm})$ used by our tracker. Definitions analogous to $\boldsymbol{o}_t$ hold for $\mathbf{h}_t = \{\mathbf{h}_{tm}\}_{m=1}^{M_t}$ and $\mathbf{o}_t = \{\boldsymbol{o}_{tm}\}_{m=1}^{M_t}$. The probability of a set of random variables is written as $p(\mathbf{o}_t) = p(\boldsymbol{o}_{t1}, \ldots, \boldsymbol{o}_{tM_t})$.

We also define an observation-to-person assignment (hidden) variable $\boldsymbol{z}_{tm}$, associated with each observation $\boldsymbol{o}_{tm}$. $Z_{tm} = n, n \in \{1 \ldots N\}$ means that $\boldsymbol{o}_{tm}$ is associated to person n. It is common that a detection corresponds to some clutter instead of a person. We cope with these false detections by defining a *clutter* target. In practice, the index $n = 0$ is assigned to this clutter target. Hence, the set of possible values for Z_{tm} is extended to $\{0\} \cup \{1 \ldots N\}$, and $Z_{tm} = 0$ means that observation $\boldsymbol{o}_{tm}$ has been generated by clutter and not by a person. The practical consequence of adding a clutter track is that the observations assigned to it play no role in the estimation of the parameters of the other tracks, thus leading to an estimation robust to outliers.

The online multiple-person tracking problem is cast into the estimation of the filtering distribution of the hidden variables given the causal observations $p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1} | \mathbf{o}_{1:t})$, where $\mathbf{o}_{1:t} = \{\mathbf{o}_1, \ldots, \mathbf{o}_t\}$. The filtering distribution used here is an extension of the basic model introduced in Chapter 2. Instead of estimating the $p(\boldsymbol{z}_t, \mathbf{s}_t | \mathbf{o}_{1:t})$ in the basic model, we jointly estimate $p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1} | \mathbf{o}_{1:t})$, which will further smooth the trajectories and assignments using the current observations $\mathbf{o}_t$. Importantly, we assume that the observations at time t only depend on the hidden and visibility variables at time t. The filtering distribution can be written as:

$$p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1} | \mathbf{o}_{1:t}) = \frac{p(\mathbf{o}_t | \boldsymbol{z}_t, \mathbf{s}_t) p(\boldsymbol{z}_t, \mathbf{s}_t | \boldsymbol{z}_{t-1}, \mathbf{s}_{t-1}) p(\mathbf{s}_{t-1}, \boldsymbol{z}_{t-1} | \mathbf{o}_{1:t-1})}{p(\mathbf{o}_t | \mathbf{o}_{1:t-1})} \quad (3.1)$$

The denominator of (3.1) only involves observed variables and therefore its evaluation is not necessary as long as one can normalize the expression arising from the numerator. Hence we focus on the two terms of the latter, namely the observation model $p(\mathbf{o}_t | \boldsymbol{z}_t, \mathbf{s}_t)$ and the dynamic distribution $p(\boldsymbol{z}_t, \mathbf{s}_t | \boldsymbol{z}_{t-1}, \mathbf{s}_{t-1})$.

The Observation Model. The joint observations are assumed to be independent and identically distributed:

$$p(\mathbf{o}_t|\boldsymbol{z}_t, \mathbf{s}_t) = \prod_{m=1}^{M_t} p(\boldsymbol{o}_{tm}|\boldsymbol{z}_{tm}, \mathbf{s}_t). \tag{3.2}$$

In addition, we make the reasonable assumption that, while localization observations depend both on the assignment variable and kinematic state, the appearance observations only depend on the assignment variable, that is the person identity, but not on his/her kinematic state. We also assume the localization and appearance observations to be independent given the hidden variables. Consequently, the observation likelihood of a single joint observation can be factorized as:

$$\begin{aligned} p(\boldsymbol{o}_{tm}|\boldsymbol{z}_{tm}, \mathbf{s}_t) &= p(\boldsymbol{o}_{tm}, \mathbf{h}_{tm}|\boldsymbol{z}_{tm}, \mathbf{s}_t) \\ &= p(\boldsymbol{o}_{tm}|\boldsymbol{z}_{tm}, \mathbf{s}_t) p(\mathbf{h}_{tm}|\boldsymbol{z}_{tm}). \end{aligned} \tag{3.3}$$

The localization observation model is defined depending on whether the observation is generated by clutter or by a person:

- If the observation is generated from clutter, namely $\boldsymbol{z}_{tm} = 0$, the variable $\boldsymbol{o}_{tm}$ follows an uniform distribution with probability density function $u(\boldsymbol{o}_{tm})$;
- If the observation is generated by person n, namely $\boldsymbol{z}_{tm} = n$, the variable $\boldsymbol{o}_{tm}$ follows a Gaussian distribution with mean $\mathbf{P}\mathbf{s}_{tn}$ and covariance $\mathbf{\Phi}$: $\boldsymbol{o}_{tm} \sim \mathcal{N}(\boldsymbol{o}_{tm}; \mathbf{P}\mathbf{s}_{tn}, \mathbf{\Phi})$

The linear operator $\mathbf{P}$ maps the kinematic state vectors onto the space of observations. For example, when $\mathbf{s}_{tn}$ represents the full-body kinematic state (full-body localization and velocity) and $\boldsymbol{o}_{tm}$ represents the full-body localization observation, $\mathbf{P}$ is a projection which, when applied to a state vector, only retains the localization components of the state vector. Finally, the full observation model is compactly defined by the following, where δ_{ij} stands for the Kronecker function:

$$p(\boldsymbol{o}_{tm}|\boldsymbol{z}_{tm} = n, \mathbf{s}_t) = u(\boldsymbol{o}_{tm})^{\delta_{0n}} \mathcal{N}(\boldsymbol{o}_{tm};\ \mathbf{P}\mathbf{s}_{tn}, \mathbf{\Phi})^{1-\delta_{0n}}. \tag{3.4}$$

The appearance observation model is also defined depending on whether the observations is clutter or not. When the observation is generated by clutter, it follows a uniform distribution with density function $u(\mathbf{h}_{tm})$. When the observation is generated by person n, it follows a Bhattacharya distribution with density defined by

$$b(\mathbf{h}_{tm}; \mathbf{h}_n) = \frac{1}{W_\lambda} \exp(-\lambda d_B(\mathbf{h}_{tm}, \mathbf{h}_n)) \tag{3.5}$$

where λ is a positive skewness parameter, $d_B(\cdot)$ is the Battacharya distance between histograms, $\mathbf{h}_n$ is the reference appearance model of person n. This gives the following compact appearance observation model:

$$p(\mathbf{h}_{tm}|\boldsymbol{z}_{tm} = n, \mathbf{s}_t) = u(\mathbf{h}_{tm})^{\delta_{0n}} b(\mathbf{h}_{tm}; \mathbf{h}_n)^{1-\delta_{0n}}. \tag{3.6}$$

The Predictive Distribution. Here we consider two hypotheses, firstly, we assume the at each time instance, assignment variable doesn't depends on the previous assignment. So we can factorize the the dynamic distribution into the observation-to-person prior distribution and the predictive distribution. Secondly, the kinematic state dynamics follow a first-order Markov chain, meaning that the state $\mathbf{s}_t$ only depends on state $\mathbf{s}_{t-1}$.

$$p(\boldsymbol{z}_t, \mathbf{s}_t | \boldsymbol{z}_{t-1}, \mathbf{s}_{t-1}) = p(\boldsymbol{z}_t) p(\mathbf{s}_t | \mathbf{s}_{t-1}). \tag{3.7}$$

The Observation-to-Person Prior Distribution. The joint distribution of the assignment variables can be factorized as:

$$p(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} p(\boldsymbol{z}_{tm}). \tag{3.8}$$

When observations are not yet available, the assignment variables $\boldsymbol{z}_{tm}$ are assumed to follow multinomial distributions defined as:

$$p(\boldsymbol{z}_{tm} = n) = a_{tn} \quad \text{with} \quad \sum_{n=0}^{N} a_{tn} = 1. \tag{3.9}$$

where a_{tn} represents the prior probability of observation $\boldsymbol{o}_{tm}$ to be generated from person n.

State Dynamics The kinematic state predictive distribution represents the probability distribution of the kinematic state at time t given the observations up to time $t-1$ $p(\mathbf{s}_t | \mathbf{s}_{t-1})$. The predictive distribution is mainly driven by the dynamics of person's kinematic states, which are modeled assuming that the person locations do not influence each other's dynamics, meaning that there is one first-order Markov chain for each person. Formally, this can be written as:

$$p(\mathbf{s}_t | \mathbf{s}_{t-1}) = \prod_{n=1}^{N} p(\mathbf{s}_{tn} | \mathbf{s}_{t-1n}). \tag{3.10}$$

For the model to be complete, $p(\mathbf{s}_{tn} | \mathbf{s}_{t-1,n})$ needs to be defined. The temporal evolution of the kinematic state $\mathbf{s}_{tn}$ is defined as:

$$p(\boldsymbol{s}_{tn} | \boldsymbol{s}_{t-1,n}) = \mathcal{N}(\boldsymbol{s}_{tn};\ \mathbf{D}\boldsymbol{s}_{t-1,n}, \boldsymbol{\Lambda}_n), \tag{3.11}$$

where $u(\boldsymbol{s}_{tn})$ is a uniform distribution over the motion state space, $\mathcal{N}$ is a Gaussian probability density function, $\mathbf{D}$ represents the dynamics transition operator, and $\boldsymbol{\Lambda}_n$ is a covariance matrix accounting for uncertainties on the state dynamics. The transition operator is defined as:

$$\mathbf{D} = \begin{pmatrix} & & \mathbb{I}_{2\times 2} \\ & \mathbb{I}_{4\times 4} & \\ & & \mathbf{0}_{2\times 2} \\ \mathbf{0}_{2\times 4} & & \mathbb{I}_{2\times 2} \end{pmatrix}$$

In other words, the dynamics of an existing person n, *either* follows a Gaussian with mean vector $\mathbf{D}\mathbf{s}_{t-1,n}$ and covariance matrix $\boldsymbol{\Lambda}_n$. The complete set of parameters of the proposed model is denoted with $\boldsymbol{\theta} = \big(\{\boldsymbol{\Phi}\},\ \{\boldsymbol{\Lambda}_n\}_{n=1}^{N}, \mathbf{A}_{1:t}\big)$, with $\mathbf{A}_t = \{a_{tn}\}_{n=0}^{N}$.

3.1.3 VARIATIONAL INFERENCE

Because of the combinatorial nature of the observation-to-person assignment problem, a direct optimization of the filtering distribution (3.1) with respect to the hidden variables is intractable. We propose to overcome this problem via a variational Bayesian inference method. The principle of this family of methods is to approximate the intractable filtering distribution $p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1}|\mathbf{o}_{1:t})$ by a separable distribution, e.g. $q(\boldsymbol{z}_t)q(\mathbf{s}_t)$. Since the filtering distribution we are targeting is different with the one in Chapter 2, we further make the variational factorization as:

$$p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1}|\mathbf{o}_{1:t}) \approx q(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1})q(\mathbf{s}_t)q(\mathbf{s}_{t-1}) \tag{3.12}$$

According to the variational Bayesian formulation [134, 20], given the observations and the parameters at the previous iteration $\boldsymbol{\theta}^\circ$, the optimal approximation has the following general expression:

$$\log q^\star(\boldsymbol{z}_t) = \mathbf{E}_{q(\mathbf{s}_t)q(\mathbf{s}_{t-1})q(\boldsymbol{z}_{t-1})} \left\{\log \widetilde{P}\right\}, \tag{3.13}$$

$$\log q^\star(\boldsymbol{z}_{t-1}) = \mathbf{E}_{q(\mathbf{s}_t)q(\mathbf{s}_{t-1})q(\boldsymbol{z}_t)} \left\{\log \widetilde{P}\right\}, \tag{3.14}$$

$$\log q^\star(\boldsymbol{s}_{tn}) = \mathbf{E}_{q(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1})q(\boldsymbol{s}_{t-1\,n})\prod_{m\neq n} q(\mathbf{s}_{tm})} \left\{\log \widetilde{P}\right\}, \tag{3.15}$$

$$\log q^\star(\boldsymbol{s}_{t-1\,n}) = \mathbf{E}_{q(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1})q(\mathbf{s}_{t,n})\prod_{m\neq n} q(\mathbf{s}_{t-1,m})} \left\{\log \widetilde{P}\right\}, \tag{3.16}$$

where, for simplicity, we used the notation $\widetilde{P} = p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1}|\mathbf{o}_{1:t}, \boldsymbol{\theta}^\circ)$. In our particular case, when these two equations are put together with the probabilistic model defined in (3.2), (3.7) and (3.10), the expression of $q(\boldsymbol{z}_t)$ is factorized further into:

$$\log q^\star(\boldsymbol{z}_{tm}) = \mathbf{E}_{q(\mathbf{s}_t)q(\mathbf{s}_{t-1})q(\boldsymbol{z}_{t-1})\prod_{j\neq k} q(\boldsymbol{z}_{tj})} \left\{\log \widetilde{P}\right\}, \tag{3.17}$$

Note that this equation leads to a finer factorization that the one we initially imposed. This behavior is typical of variational Bayes methods in which a very mild separability assumption can lead to a much finer factorization when combined with priors over hidden states and latent variables, i.e. (3.2), (3.7) and (3.10). The final factorization writes:

$$p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1}|\mathbf{o}_{1:t}) \approx \prod_{m=0}^{M_t} q(\boldsymbol{z}_{tm}) \prod_{m=0}^{M_{t-1}} q(\boldsymbol{z}_{t-1,m}) \prod_{n=0}^{N} q(\boldsymbol{s}_{tn})q(\boldsymbol{s}_{t-1\,n}). \tag{3.18}$$

Once the posterior distribution over the hidden variables is computed (see below), the optimal parameters are estimated using $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)$ with J defined as:

$$J(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = \mathbf{E}_{q(\boldsymbol{z},\mathbf{S})} \left\{\log p(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}, \mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{o}_{1:t}|\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)\right\}. \tag{3.19}$$

§ E-Z-Step

The estimation of $q(\boldsymbol{z}_{tm})$ is carried out by developing the expectation (3.17) which yields the following formula:

$$q^\star(\boldsymbol{z}_{tm} = n) = \alpha_{tmn} = \frac{\epsilon_{tmn} a_{tn}}{\sum_{m=0}^{N} \epsilon_{tmm} a_{tn}}, \tag{3.20}$$

and ϵ_{tmn} is defined as:

$$\epsilon_{tmn} = \begin{cases} u(\boldsymbol{o}_{tm})u(\mathbf{h}_{tm}) & n = 0, \\ \mathcal{N}(\boldsymbol{o}_{tm}, \mathbf{P}\boldsymbol{\mu}_{tn}, \boldsymbol{\Phi}) e^{-\frac{1}{2}\mathbf{Tr}\left(\mathbf{P}^\top (\boldsymbol{\Phi})^{-1} \mathbf{P} \boldsymbol{\Gamma}_{tn}\right)} b(\mathbf{h}_{tm}; \mathbf{h}_n) & n \neq 0, \end{cases} \tag{3.21}$$

where $\mathbf{Tr}(\cdot)$ is the trace operator and $\boldsymbol{\mu}_{tn}$ and $\boldsymbol{\Gamma}_{tn}$ are defined by (3.38) and (3.37) below. Intuitively, this approximation shows that the assignment of an observation to a person is based on spatial proximity between the observation localization and the person localization, and the similarity between the observation's appearance and the person's reference appearance.

§ E-S-Step

The estimation of $q^\star(\boldsymbol{s}_{tn})$ is derived from (3.15). Similarly to the previous posterior distribution, which boil down to the following formula:

$$q^\star(\boldsymbol{s}_{tn}) = \mathcal{N}(\boldsymbol{s}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}), , \tag{3.22}$$

where the mean vector $\boldsymbol{\mu}_{tn}$ and the covariance matrix $\boldsymbol{\Gamma}_{tn}$ are given by:

$$\boldsymbol{\Gamma}_{tn} = \Big(\sum_{m=0}^{M_t} \alpha_{tmn} \left(\mathbf{P}^\top (\boldsymbol{\Phi})^{-1} \mathbf{P} \right) + \boldsymbol{\Lambda}_n^{-1} \Big)^{-1}, \tag{3.23}$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn} \Big(\sum_{m=0}^{M_t} \alpha_{tmn} \mathbf{P}^\top (\boldsymbol{\Phi})^{-1} \boldsymbol{o}_{tm} + \boldsymbol{\Lambda}_n^{-1} \mathbf{D} \boldsymbol{\mu}_{t-1,n} \Big). \tag{3.24}$$

Similarly, for the estimation of the distribution

$$q^\star(\boldsymbol{s}_{t-1\,n}) = \mathcal{N}(\boldsymbol{s}_{t-1\,n}; \widehat{\boldsymbol{\mu}}_{t-1,n}, \widehat{\boldsymbol{\Gamma}}_{t-1,n}), \tag{3.25}$$

the mean and covariance are:

$$\widehat{\boldsymbol{\Gamma}}_{t-1,n} = \left(\mathbf{D}^\top \boldsymbol{\Lambda}_n^{-1} \mathbf{D} + \boldsymbol{\Gamma}_{t-1,n} \right)^{-1} \tag{3.26}$$

$$\widehat{\boldsymbol{\mu}}_{t-1,n} = \widehat{\boldsymbol{\Gamma}}_{t-1,n} \left(\mathbf{D}^\top \boldsymbol{\Lambda}_n^{-1} \boldsymbol{\mu}_{t,n} + \boldsymbol{\Gamma}_{t-1,n}^{-1} \boldsymbol{\mu}_{t-1,n} \right). \tag{3.27}$$

We note that the variational approximation of the kinematic-state distribution reminds the Kalman filter solution of a linear dynamical system with mainly one difference: in our formulation, (3.38) and (3.37), the means and covariances are computed by weighting the observations with α_{tmn}, i.e. (3.38) and (3.37).

§ M-step

Once the posterior distribution of the hidden variables is estimated, the optimal parameter values can be estimated via maximization of J defined in (3.19). Concerning the parameters of the a priori observation-to-object assignment $\mathbf{A}_t$ we compute:

$$J(a_{tn}) = \sum_{m=1}^{M_t} \alpha_{tmn} \log(a_{tn}) \quad \text{s.t.} \quad \sum_{n=0}^{N} a_{tn} = 1, \tag{3.28}$$

and we trivially obtain:

$$a_{tn} = \frac{\sum_{m=1}^{M_t} \alpha_{tmn}}{\sum_{m=0}^{N} \sum_{m=1}^{M_t} \alpha_{tmm}}. \tag{3.29}$$

The observation covariance $\mathbf{\Phi}$ and the state covariances $\mathbf{\Lambda}_n$ can be estimated during the M-step. However, in our current implementation estimates for $\mathbf{\Phi}$ and $\mathbf{\Lambda}_n$ are instantaneous, i.e., they are obtained only from the observations at time t (see the experimental section for details).

3.1.4 PERSON-BIRTH, -VISIBILITY AND -DEATH PROCESSES

Tracking a time-varying number of targets requires procedures to create tracks when new targets enter the scene and to delete tracks when corresponding targets leave the visual scene. In this Chapter, we propose a statistical-test based birth process that creates new tracks and a hidden Markov model (HMM) based visibility process that handles disappearing targets. In this section we present the inference model based on the stochastic birth-process.

§ Birth Process

The principle of the person birth process is to search for consistent trajectories in the history of observations associated to clutter. It is executed at the start of the tracking to initialize a latent variable for each detected person, as well as at any time t to detect new persons. The birth process considers B consecutive visual frames. At t, with $t > B$, we consider the set visual observations assigned to $n = 0$ from $t-B$ to t, namely observations whose posteriors (3.20) are maximized for $n = 0$ (at initialization all the observations are in this case). We then build observation sequences from this set, namely sequences of the form $(\tilde{\boldsymbol{o}}_{m_{t-B}}, \ldots, \tilde{\boldsymbol{o}}_{m_t})_{\tilde{n}} \in \mathcal{B}$, where m_t indexes the set of observations at t assigned to $n = 0$ and $\tilde{n}$ indexes the set $\mathcal{B}$ of all such sequences. Notice that the birth process only uses the bounding-box center, width and size, $\boldsymbol{o}$, and that the descriptor $\mathbf{h}$ is not used. Hence the birth process is only based on the smoothness of an observed sequence of bounding boxes. Let's consider the marginal likelihood of a sequence $\tilde{n}$, namely:

$$\begin{aligned}\mathcal{L}_{\tilde{n}} &= p((\tilde{\boldsymbol{o}}_{m_{t-B}}, \ldots, \tilde{\boldsymbol{o}}_{m_t})_{\tilde{n}}) \\ &= \int \ldots \int p(\tilde{\boldsymbol{o}}_{m_{t-B}} | \boldsymbol{s}_{t-B\,\tilde{n}}) \ldots p(\tilde{\boldsymbol{o}}_{m_t} | \boldsymbol{s}_{t\,\tilde{n}}) \\ &\times p(\boldsymbol{s}_{t\,\tilde{n}} | \boldsymbol{s}_{t-1\,\tilde{n}}) \ldots p(\boldsymbol{s}_{t-B+1\,\tilde{n}} | \boldsymbol{s}_{t-B\,\tilde{n}}) p(\boldsymbol{s}_{t-B\,\tilde{n}}) d\boldsymbol{s}_{t-B:t\,\tilde{n}},\end{aligned} \tag{3.30}$$

where $\boldsymbol{s}_{t,\tilde{n}}$ is the latent variable already defined and $\tilde{n}$ indexes the set $\mathcal{B}$. All the probability distributions in (3.30) were already defined, namely (3.11) and (3.4), with the exception of $p(\boldsymbol{s}_{t-B,\tilde{n}})$. Without loss of generality, we can assume that the latter is a normal distribution centered at $\tilde{\boldsymbol{o}}_{m_t}$ and with a large covariance. Therefore, the evaluation of (3.30) yields a closed-form expression for $\mathcal{L}_{\tilde{n}}$. A sequence $\tilde{n}$ generated by a person is likely to be smooth and hence $\mathcal{L}_{\tilde{n}}$ is high, while for a non-smooth sequence the marginal likelihood is low. A

newborn person is therefore created from a sequence of observations $\tilde{n}$ if $\mathcal{L}_{\tilde{n}} > \tau$, where τ is a user-defined parameter. As just mentioned, the birth process is executed to initialize persons as well as along time to add new persons. In practice, in (3.30) we set B=3 and hence, from t=1 to t=4 all the observations are initially assigned to $n = 0$. Finally, a new person is added by setting $q^{\star}(\boldsymbol{s}_{t,\tilde{n}}; \boldsymbol{\mu}_{t,\tilde{n}}, \boldsymbol{\Gamma}_{t,\tilde{n}})$ with $\boldsymbol{\mu}_{t,\tilde{n}} = [\tilde{\boldsymbol{o}}_{m_t}^{\top}, \mathbf{0}_2^{\top}]^{\top}$, and $\boldsymbol{\Gamma}_{tn}$ is set to the value of a birth covariance matrix (see (3.22)). Also, the reference appearance model for the new person is defined as $\mathbf{h}_{t,\tilde{n}} = \mathbf{h}_{t,m_t}$.

§ Visibility Process

A tracked person is said to be visible at time t whenever there are observations associated to that person, otherwise the person is considered not visible. Instead of deleting tracks, as classical for death processes, our model labels tracks without associated observations as *sleeping*. In this way, we keep the possibility to awake such sleeping tracks whenever their reference appearance highly matches an observed appearance.

We denote the n-th person visibility (binary) variable by V_{tn}, meaning that the person is visible at time t if $V_{tn} = 1$ and 0 otherwise. We assume the existence of a transition model for the hidden visibility variable V_{tn}. More precisely, the visibility state temporal evolution is governed by the transition matrix, $p(V_{tn} = j|V_{t-1,n} = i) = \pi_v^{\delta_{ij}}(1 - \pi_v)^{1-\delta_{ij}}$, where π_v is the probability to remain in the same state. To enforce temporal smoothness, the probability to remain in the same state is taken higher than the probability to switch to another state.

The goal now is to estimate the visibility of all the persons. For this purpose we define the visibility observations as $\nu_{tn} = a_{tn}$, being 0 when no observation is associated to person n. In practice, we need to filter the visibility state variables V_{tn} using the visibility observations ν_{tn}. In other words, we need to estimate the filtering distribution $p(V_{tn}|\nu_{1:tn})$ which can be written as:

$$p(V_{tn} = v_{tn}|\nu_{1:t}) = \frac{p(\nu_{tn}|v_{tn}) \sum_{v_{t-1,n}} p(v_{tn}|v_{t-1,n}) p(v_{t-1,n}|\nu_{1:t-1,n})}{p(\nu_{tn}|\nu_{1:t-1,n})}, \tag{3.31}$$

where the denominator corresponds to integrating the numerator over v_{tn}. In order to fully specify the model, we define the visibility observation likelihood as:

$$p(\nu_{tn}|v_{tn}) = (\exp(-\lambda\nu_{tn}))^{v_{tn}}(1 - \exp(-\lambda\nu_{tn}))^{1-v_{tn}} \tag{3.32}$$

Intuitively, when ν_{tn} is high, the likelihood is large if $v_{tn} = 1$ (person is visible). The opposite behavior is found when ν_{tn} is small. Importantly, at each frame, because the visibility state is a binary variable, its filtering distribution can be straightforwardly computed. We found this rather intuitive strategy to be somewhat "shaky" over time even taking the Markov dependency into account. This is why we enriched the visibility observations to span over multiple frames $\nu_{tn} = \sum_{l=0}^{L} a_{t+ln}$, so that if $v_{tn} = 1$, the likelihood is large when ν_{tn} is high and therefore the target is visible in one or more neighboring

frames. This is the equivalent of the hypothesis testing spanning over time associated to the birth process.

§ Death Process

The idea of the person-visibility process arises from encouraging track consistency when a target disappears and appears back in the field of view. However, a tracker that remembers *all* the tracks that have been previously seen is hardly scalable. Indeed, the memory resources required by a system that remembers all previous appearance templates grows indefinitely with new appearances. Therefore, one must discard *old* information to facilitate the scalability of the method to large datasets containing sequences with several dozens of different people involved. In addition to alleviating the memory requirements, this also reduces the computational complexity of the tracker. This is the motivation of including a death process into the proposed variational framework. Intuitively one would like to discard those tracks that have not been seen during several frames. In practice, we found that discarding those tracks that are not visible for ten consecutive frames yields a good trade-off between complexity, resource demand and performance. Setting this parameter for a different dataset should not be chimeric, since the precise interpretation of the meaning of it is straightforward.

3.1.5 EXPERIMENTS

We evaluated the performance of the proposed variational multiple-person tracker on the MOT 2016 dataset challenge [110]. This dataset is composed of seven training videos and seven test videos. Importantly, we use the detections that are provided with the dataset. Because multiple-person tracking intrinsically implies track creation (birth), deletion (death), target identity maintenance, and localization, evaluating multiple-person tracking techniques is a non-trivial task. Many metrics have been proposed, e.g. [126, 137, 138, 90].

We adopt the metrics used by the MOT 2016 benchmark, namely [138]. The main tracking measures are: the *multiple-object tracking accuracy* (MOTA), that combines false positives (FP), missed targets (FN), and identity switches (ID); the *multiple-object tracking precision* (MOTP), that measures the alignment of the tracker output bounding box with the ground truth; the false alarm per frame (FAF); the ratio of mostly tracked trajectories (MT); the ratio of mostly lost trajectories (ML) and the number of track fragmentations (Frag).

Fig 3.2 shows sample images of all test videos: They contain three sequences recorded with static cameras (MOT16-01, MOT16-03 and MOT16-08), which contain very crowded scenes and thus are very challenging, and five sequences with large camera motions, both translations and rotations, which make the data even more difficult to process.

As explained above, we use the public pedestrian detections provided within the MOT16 challenge. These static detections are complemented in two different ways. First, we extract velocity observations by means of a simple optical-flow based algorithm that looks

(a) MOT16-01 (b) MOT16-03 (c) MOT16-06

(d) MOT16-08 (e) MOT16-12 (f) MOT16-14

Figure 3.2: Samples images from the MOT 16 test sequences.

for the most similar region of the next temporal frame within the neighborhood of the original detection. Therefore, the observations operator P is the identity matrix, project the entire state variable into the observation space. Second, the appearance feature vector is the concatenation of joint color histograms of three regions of the torso in HSV space.

The proposed variational model is governed by several parameters. Aiming at providing an algorithm that is dataset-independent and that features a good trade-off between flexibility and performance, we set the observation covariance matrix $\mathbf{\Phi}$ and the state covariance matrix $\mathbf{\Lambda}_n$ automatically from the detections. More precisely, both matrices are imposed to be diagonal; for $\mathbf{\Phi}$, the variances of the horizontal position, of the width, and of the horizontal speed are $1/3$, $1/3$ and $1/6$ of the detected width. The variances of the vertical quantities are built analogously. The rationale behind this choice is that we consider that the true detection lies, more or less, within the width and height of the detected bounding box. Regarding $\mathbf{\Lambda}_n$, the diagonal entries are 1, 1 and $1/2$ of the detected width, and vertical quantities are defined analogously. Furthermore, in order to eliminate arbitrary false detections, we set $L = 5$ in the birth process. Finally, for sequences in which the size of the bounding boxes is roughly constant, we discarded those detections that were too large or too small.

Examples of the tracking results for all the test sequences except MOT16-07 are shown in Figure 3.3, while six frames from MOT16-07 are shown in Figure 3.4. In all figures, the red boxes represent our tracking result and the numbers within the boxes are the tracking indexes. Generally speaking, on one hand the variational model is crucial to properly associate detections with trajectories. On the other hand, the birth and visibility processes play a role when tracked objects appear and disappear. Regarding Figure 3.4, it contains 54 tracks recorded by a moving camera in a sequence of 500 frames. It is a very challenging tracking task, not only because the density of pedestrians is quite high, but also because significant camera motion makes the person trajectories to be both rough and discontinuous. One drawback of the proposed approach is that partially consistent false detections could lead to the creation of a false track, therefore tracking an inexis-

(a) MOT16-01 (b) MOT16-03 (c) MOT16-06

(d) MOT16-08 (e) MOT16-12 (f) MOT16-14

Figure 3.3: Sample results on several sequences of MOT16 datasets, red bounding boxes represents the tracking results, and the number inside each box is the track index.

Figure 3.4: Sample results on the sequence MOT16-07, encoded as in the previous figure.

tent pedestrian. On the positive side, the main advantage of the proposed model is that the probabilistic combination of the dynamic and appearance models can decrease the probability of switching the identities of two tracks.

Table 3.1 reports the performance of the proposed algorithm, which is referred to as OVBT (online variational Bayesian tracker), over the seven test sequences of the MOT 2016 challenge. The results obtained with OVBT are available on the MOT 2016 webpage.[2] One can notice that our method provides high precision (MOTP) but low accuracy (MOTA), meaning that some tracks were missed (mostly due to misdetections). This is consistent with a rather low MT measure. This behavior was more extreme when the visibility process did not include any observation aggregation over time. Indeed, we observed that considering multiple observations within the visibility process leads to better performance (for all sequences and almost all measures). In table 3.2 we report the comparison between the proposed method and the benchmark methods. The proposed OBVT

[2] `https://motchallenge.net/results/MOT16/`

Table 3.1: Evaluation of the proposed multiple-person tracking method with different features on the seven sequences of the MOT16 test dataset.

Sequence	**MOTA**	**MOTP**	**FAF**	**MT**	**ML**	**FP**	**FN**	**ID Sw**	**Frag**
MOT16-01	23.9	71.4	1.5	13.0%	39.1%	696	4,137	35	89
MOT16-03	46.9	75.7	4.1	17.6%	20.3%	6,173	48,631	689	1,184
MOT16-06	32.7	73.2	0.5	3.6%	58.4%	562	7,073	124	183
MOT16-07	33.6	73.3	2.2	9.3%	35.2%	1,077	9,605	158	272
MOT16-08	24.6	78.4	1.7	3.2%	41.3%	1,066	11,402	150	177
MOT16-12	32.8	76.7	0.9	10.5%	52.3%	766	4,749	63	80
MOT16-14	18.1	74.5	1.6	2.4%	61.6%	1,177	13,866	102	155
Over All	38.4± 8.8	75.4	1.9	7.5%	47.3%	11,517	99,463	1,321	2,140

Table 3.2: Benchmark of several methods on the MOT 16 test using the public detector.
↑: the higher the better; ↓: the lower the better

Tracker	**MOTA (↑)**	**MOTP (↑)**	**FAF (↓)**	**MT (↑)**	**ML (↓)**	**FP (↓)**	**FN (↓)**	**ID Sw (↓)**	**Frag(↓)**
JPDA_m [59]	26.2±6.1	76.3	0.6	4.1%	67.5%	3,689	130,549	**365 (12.9)**	**638 (22.5)**
SMOT [38]	29.7±7.3	75.2	2.9	5.3%	47.7%	17,426	107,552	3,108 (75.8)	4,483 (109.3)
DP_NMS [120]	32.2±9.8	76.4	**0.2**	5.4%	62.1%	**1,123**	121,579	972 (29.2)	944 (28.3)
CEM [112]	33.2±7.9	75.8	1.2	**7.8%**	54.4%	6,837	114,322	642 (17.2)	731 (19.6)
TBD [55]	33.7±9.2	**76.5**	1.0	7.2%	54.2%	5,804	112,587	2,418 (63.2)	2,252 (58.9)
OVBT	**38.4±8.6**	75.4	1.9	7.5%	**47.3%**	11,517	**99,463**	1,321 (29.1)	2,140 (47.1)

achieved the best performance by getting the highest MOTA score.

3.2 Tracking with Visually Controlled Head Movements

3.2.1 Introduction

In this section, the visual tracking algorithm is applied on a humanoid robot. The algorithm is further combined with robot motor movement to compensate the robot's ego-motion and perform visual servoing.

Robots are currently on the verge of sharing many common spaces with humans. Exemplar scenarios are the front desk of a hotel, museum guides, elder assistance or entertainment for children. In all these situations, and many others, the robotic platform is required to interact with people and, as part of its low-level behavioral skills, to perform person tracking and visual servoing. In practice this means that the robot is supposed to keep track of the locations of the people in the scene and, once a person of interest has been chosen, to keep that person within its visual field of view.

Visual servoing, i.e. robot control based on visual information, has been a well studied problem [43, 103]. Several methods were developed targeting different applications,

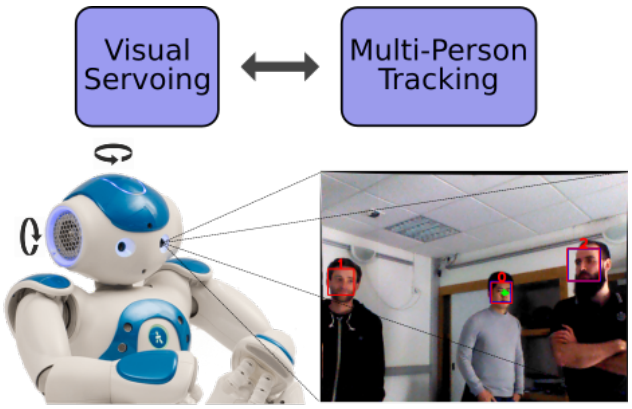


Figure 3.5: Schematic overview of the system. The visual servoing module estimates the optimal robot commands and the expected impact of the tracked positions. The multi-person tracking module refines the positions of the persons with the new observations and the information provided by the visual servoing.

such as grasping [114], mobile robot navigation [49] or autonomous aerial vehicle guidance [109]. In this section we are interested in visual servoing using a robot head, commonly referred to as head-eye coordination, which was studied in a wide range of applicative scenarios involving *a single* object/person of interest [33, 32, 140, 105, 118, 124, 1, 4, 144] and based on methodologies such as detect and pursuit, image feature tracking, or Kalmann filtering.

However, the vast majority of everyday situations consist of several persons. Clearly, not al these persons are of interest for the HRI task at hand. Nevertheless, the robot should be able to jointly perform multiple person tracking and visual servoing of one or a few persons. Compared to single-person methods, the presence of many people remains challenging. First, computationally cheap face/person detection algorithms often deliver noisy detections or even fail to provide a consistent sequence of bounding boxes. Second, even under the hypothesis of high-quality face/person detections, we are still left with the task of correctly associating these detections over time. For instance, [128] proposed to use an EKF for each tracked person, often leading to bad detection-to-person assignments and thus estimating roaming tracks. More computationally demanding algorithms exist, such as particle filtering, *e.g.* [130], but their use in real time applications is rather limited. Third, in real life scenarios, people will continuously appear and disappear from the field-of-view of the robot, and it is highly desirable to robustly track persons that disappear and reappear later on.

There is a plethora of methodologies that address the multiple-object (or person) tracking problem. For example, [47] tackled the problem by combining a sparse representation-based appearance model with a sliding window, and [28] proposed and aggregated local flow descriptor and a dynamic graphical model that is optimized off-line. To the best of our knowledge, most of the existing methods are not designed to deal with controlled camera motions, even if some of them are partially robust to ego-motion, since they need to extract feature points from the background in order to estimate camera motions [29]. It is not straightforward to take ego-motion information into account in case it is avail-

able. As we experimentally show, this can cause a huge drop on tracking performance, in particular when addressing the complex scenarios just mentioned.

In order to overcome this issue, we propose to embed visual servoing into the multi-person tracker, as schematically shown in Figure 3.5. Visual servoing requires the estimation of a Jacobian matrix that maps observed image features onto motor velocities, which in turn requires 3-D information. We do this by combining a person detector with a calibrated camera pair mounted onto the robot head. The estimated motor velocities are then explicitly taken into account by the person tracker itself. The latter is formulated as a Bayesian filtering method and we propose to use a variational approximation [7], [12]. Indeed, this solution is particularly efficient from a computational point of view and hence it is preferred over more standard sampling methods for the following advantages: (i) it is able to handle a number of persons that varies over time, and (ii) it is robust to disappearing/reappearing persons.

To summarize, we propose a joint multi-person tracking and visual servoing method that is able to simultaneously estimate the three-dimensional position of a time-varying number of people and to encompass the effect of the robot's motion on this estimation. This complements both visual servoing, by leveraging current methods from single-object tracking to multiple-object tracking, and by explicitly taking ego motion into account. We propose a Bayesian filter and its variational approximation allowing to effectively solve the inference problem of the filtering distribution while keeping a reasonably low computational load (the overall system works at 10 FPS). Third, we report a large experimental study on the NAO multi-person visual servoing (NAO-MPVS) dataset showing, not only that the addressed scenarios are challenging, but also that including the impact of the robot's motion into the probabilistic tracking framework is of utmost importance for the performance of the system. [3]

The remaining of the section is organized as follows. Section 3.2.2 presents the probabilistic tracker which is interleaved with the visual servoing module detailed in Section 3.2.4. The system architecture is described in Section 3.2.5. The experimental protocol and the results are reported in Section 3.2.6. Section 3.3 concludes the Chapter.

3.2.2 PROBABILISTIC MODEL

We adopt the probabilistic multiple person tracking formulation recently proposed in [7]. Let N_t denote the number of persons at time t. Let $\mathbf{X}_{tn} \in \mathbb{R}^3$ and $\mathbf{W}_{tn} \in \mathbb{R}^2$ denote the position of person n at t and its bounding box (width and height), respectively. Making use of the three-dimensional locations in the joint tracking-servoing method has two prominent advantages. First, it leads to a more stable tracker that is also more robust to object occlusions. Second, it allows us to compute the Jacobian associated to the visual servoing in closed form, and therefore the expected effect of the robot motion into the observed scene can be computed without any prior knowledge about the persons to be

[3] Supplemental material for this Chapter can be found at `https://team.inria.fr/perception/mot-servoing/`

tracked (see Section 3.2.4). This is a crucial advantage over existing methods, since it allows to encompass the effect of the robot control in the tracking framework and therefore *to infer the persons' locations by taking the robot motion into account*.

Aiming to privilege smooth trajectories, we track the velocity and the bounding box of each person in addition to his/her position. More formally, the tracking state variable is a concatenation of three variables: $\boldsymbol{s}_{tn} = [\mathbf{X}_{tn}^\top, \mathbf{W}_{tn}^\top, \dot{\mathbf{X}}_{tn}^\top]^\top$. These variables are expressed in the coordinate frame of the camera pair. Below we describe the probabilistic model (Section 3.2.2) from which we derive the filtering distribution (Section 3.2.3) based on a variational Bayes approximation [135]. The birth process allowing to take into account new disappearing/reappearing persons is detailed in Section 5.6.4.

In this section, we briefly summarize the the probabilistic model consists of two main components. On one hand, the tracking state dynamics delineates the probabilistic behavior of the state variable over time. On the other hand, the observation model associates the state variable at current t, $\boldsymbol{s}_{tn}$, to the observations. Such an association is modeled by assignment variables $\{Z_{tm}\}_{m=1}^{M_t}$, namely $Z_{tm} = n, n \in \{1, \ldots, N_t\}$, observation k at time t is assigned to person n. The observation model is the same as the one defined in 3.1.2, we will only detail the state dynamics which involves the robot's active motion.

§ The state dynamics

The state dynamics models the temporal evolution of the state variable. We make two hypotheses. Firstly, we assume that at each time instance, the assignment variable $\boldsymbol{Z}_t$ doesn't depend on $\boldsymbol{Z}_{t-1}$. Therefore, we can factorize the dynamic distribution into the observation-to-person prior distribution and the predictive distribution. Secondly, the state dynamics follow a first-order Markov chain, meaning that $\boldsymbol{s}_{tn}$ only depends on $\boldsymbol{s}_{t-1n}$:

$$p(\boldsymbol{z}_t, \boldsymbol{s}_t | \boldsymbol{z}_{t-1}, \boldsymbol{s}_{t-1}) = \prod_{n=1}^{N} p(\boldsymbol{s}_{tn} | \boldsymbol{s}_{t-1n}) \prod_{m=1}^{M_t} p(Z_{tm}). \tag{3.33}$$

The two modeling choices that need to be done are: the prior probability of the assignment variable Z_{tm} and the dynamic model. A priori, there is no reason to believe that one person is more prone to generate observations than another one, hence we set $p(Z_{tm}) = a_{tm} = \frac{1}{N_t+1}$, for all k. Regarding the dynamics of $\boldsymbol{s}_t$, we propose a transition model that takes the robot's motion explicitly into account. Indeed, let $\mathbf{C}_{tn}$ denote the expected $\mathbf{X}_{t-1n}$ due to the motion of the robot (see Section 3.2.4). Importantly, $\mathbf{C}_{tn}$ can be computed in closed-form thanks to the proposed formulation. We concatenate $\mathbf{C}_{tn}$ with a 5-dimensional vector of zeros (that would correspond to the expected shift of the bounding box and the velocity), and construct a 8-dimensional vector that for the sake of simplicity we will also denote with $\mathbf{C}_{tn}$. The explicit computation of $\mathbf{C}_{tn}$, described in detail in Section 3.2.4, allows us to better predict when a person appears, disappears, or reappears in the field of view. Notice that, due to the potentially large appearance variation, the use of the geometric proprioceptive information may become crucial for the

tracking performance. More formally, we model the transition probability with a Gaussian distribution defined as:

$$p(\boldsymbol{s}_{tn}|\boldsymbol{s}_{t-1n}) = \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1n} + \mathbf{C}_{tn}, \Lambda_n), \tag{3.34}$$

where $\mathbf{C}_{tn}$ is a translation associated to the effect of the controlled robot motion (see Section 3.2.4 for details), Λ_n models the uncertainty over the dynamics of the n-th source, and $\mathbf{D}$ is the following matrix:

$$\mathbf{D} = \begin{pmatrix} \mathbf{I}_3 & \mathbf{0} & \mathbf{I}_3 \\ \mathbf{0} & \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_3 \end{pmatrix}.$$

Due to the different state dimension, the matrix $\mathbf{D}$ in this section is a bit different than the previous section. But the matrix still represents a first order model on the dynamics of the person. In other words, the bounding box and the velocity do not change, while the position changes according to the previous velocity.

3.2.3 VARIATIONAL INFERENCE

In order to merge all the previous observations together with the current information gathered at time t, we write the *filtering distribution of the hidden random variables*:

$$p(\boldsymbol{z}_t, \boldsymbol{s}_t|\mathbf{o}_{1:t}) \propto p(\mathbf{o}_t|\boldsymbol{z}_t, \boldsymbol{s}_t)p(\boldsymbol{z}_t, \boldsymbol{s}_t|\mathbf{o}_{1:t-1}), \tag{3.35}$$

where the second term is the so-called predictive distribution, which is related to the filtering distribution at time $t-1$ by:

$$p(\boldsymbol{z}_t, \boldsymbol{s}_t|\mathbf{o}_{1:t-1}) = p(\boldsymbol{z}_t) \int p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1})p(\boldsymbol{s}_{t-1}|\mathbf{o}_{1:t-1})d\boldsymbol{s}_{t-1}$$

Since (3.35) does not accept a computationally tractable closed-form expression, we choose to use a variational approximation [135]. If properly designed, such approximations have the prominent advantage of deriving into closed-form updates for the a posterior (filtering) probabilities. Concisely, variational approximations consist on imposing a partition over the hidden variables, in our case:

$$p(\boldsymbol{z}_t, \boldsymbol{s}_t|\mathbf{o}_{1:t}) \approx \prod_{n=1}^{N_t} q(\boldsymbol{s}_{tn}) \prod_{m=1}^{M_t} q(Z_{tm}), \tag{3.36}$$

and then finding the optimal distributions $q(\boldsymbol{s}_{tn})$ and $q(Z_{tm})$ in the Kullback-Leibler distance sense.

Compared with the equations obtained in the previous section, the difference after combining the robot active motion appears on the derivation of E-S step. The expression of the optimal posterior distribution of the assignment variable $q^\star(\boldsymbol{z}_{tm})$ is same as (3.20).

The a posterior distribution for $\boldsymbol{s}_{tn}$ also turns out to be a Gaussian distribution with mean μ_{tn} and covariance Γ_{tn} given by:

$$\boldsymbol{\Gamma}_{tn} = \left(\left(\boldsymbol{\Lambda}_n + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top \right)^{-1} + \sum_{m=0}^{M_t} \alpha_{tmn}\mathbf{P}^\top \boldsymbol{\Phi}^{-1}\mathbf{P} \right)^{-1}, \tag{3.37}$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn} \left(\left(\boldsymbol{\Lambda}_n + \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top \right)^{-1} \left(\mathbf{D}\boldsymbol{\mu}_{t-1n} + \mathbf{C}_{tn} \right) + \sum_{m=0}^{M_t} \alpha_{tmn}\mathbf{P}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{o}_{tm} \right) \tag{3.38}$$

where $\boldsymbol{\mu}_{t-1n}$ is the expected position of person n in the previous time step. The velocity due to the robot motion $\mathbf{C}_{tn}$ is automatically included in the equation (3.38). These two steps are commonly iterated a few times at every time step. Remarkably, this strategy can also be used to learn the parameters of the model, for which we would then be required to derive the so-called M-step. The reader is referred to [7] for an exhaustive discussion.

3.2.4 VISUALLY-CONTROLLED HEAD MOVEMENTS

In this section we detail the visual servoing model allowing the robot to focus its attention on targets of interest. In order to simplify the discussion we remove the temporal and person indices t and n. The objective of the visual servoing module is to compute the required motor velocity to bring the person of interest to the center of the image. Therefore we need the Jacobian linking the image space to the motor space. Since such relationship is difficult to model, classically one models the motor-to-image Jacobian and then computes the inverse. In our case, we pass by the three-dimensional world and compute the Jacobian as the composite of a world-to-image Jacobian and a motor-to-world Jacobian.

§ World-to-image Jacobian

We consider the coordinate system associated to the left camera (at the initial head's position) to be the world's coordinate system. This is an arbitrary choice that can be replaced with any other static coordinate system with a simple rigid transformation. The non-linear mapping between world-coordinate and image-coordinate is:

$$\mathbf{V} = \mathbf{K}\frac{1}{X_3}\mathbf{X}, \tag{3.39}$$

where $\mathbf{X} = (X_1, X_2, X_3)^\top$, $\mathbf{V} = (V_1, V_2)^\top$ and $\mathbf{K} \in \mathbb{R}^{2\times 3}$ is the matrix of intrinsic parameters of the pinhole camera model. The Jacobian of this transformation writes:

$$\dot{\mathbf{V}} = \underbrace{\mathbf{K} \begin{pmatrix} 1/X_3 & 0 & -X_1/X_3^2 \\ 0 & 1/X_3 & -X_2/X_3^2 \end{pmatrix}}_{\mathbf{J}_{wi}(\mathbf{X})} \dot{\mathbf{X}}, \tag{3.40}$$

where $\dot{\mathbf{X}}$ is the velocity vector at $\mathbf{X}$, and $\mathbf{V}$ models the velocity as seen in the left camera image.

§ Motor-to-world Jacobian

In order to compute the Jacobian relating the velocity at $\mathbf{X}$ and the motor velocity, we first recall that in a general, the velocity of a three-dimensional point when the coordinate system is subject to a rigid motion, namely a rotation $\omega = [\omega_x, \omega_y, \omega_z]^\top$ and a translation $\mathbf{u} = [u_1, u_2, u_3]^\top$, can be expressed as:

$$\dot{\mathbf{X}} = \omega \times \mathbf{X} + \mathbf{u} = \begin{pmatrix} \mathbf{Sk}(\omega) & \mathbf{u} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix}, \tag{3.41}$$

where $\mathbf{Sk}(\omega)$ is the skew symmetric twist matrix representing the exterior product by the three-dimensional vector ω.

In our case, ω and $\mathbf{u}$ depend on the motor yaw and pitch rotation velocities $\dot{\alpha}$ and $\dot{\beta}$ respectively. As shown in the literature [63], for a rotation velocity $\dot{\alpha}$, the velocity at $\mathbf{X}$ can be expressed as:

$$\dot{\mathbf{X}} = \begin{pmatrix} -\mathbf{Sk}(\mathbf{X}) & \mathbf{I}_3 \end{pmatrix} \begin{pmatrix} \omega_1 \\ \mathbf{u}_1 \end{pmatrix} \dot{\alpha}, \tag{3.42}$$

where the values of ω_1 and $\mathbf{u}_1$ are acquired through a calibration phase (see Section 3.2.5).

The effect of the pitch is quite similar, with the only difference that, since we first apply the yaw rotation and then the pitch rotation, one has to take into account the effect of $\dot{\beta}$ after the rotation induced by α. Formally we write:

$$\dot{\mathbf{X}} = \begin{pmatrix} -\mathbf{Sk}(\mathbf{X}) & \mathbf{I}_3 \end{pmatrix} \begin{pmatrix} \mathbf{R}\omega_2 \\ -\mathbf{Sk}(\mathbf{R}\omega_2)\mathbf{t} + \mathbf{R}\mathbf{u}_2 \end{pmatrix} \dot{\beta}, \tag{3.43}$$

where ω_2 and $\mathbf{u}_2$ are obtained through the calibration and $\mathbf{R}$ and $\mathbf{t}$ are the rotation and translation vectors associated to the yaw state α. In all, the motor-to-world Jacobian writes:

$$\dot{\mathbf{X}} = \underbrace{\begin{pmatrix} -\mathbf{Sk}(\mathbf{X}) & \mathbf{I}_3 \end{pmatrix} \mathbf{L}(\alpha)}_{\mathbf{J}_{mw}(\mathbf{X})} \begin{pmatrix} \dot{\alpha} \\ \dot{\beta} \end{pmatrix}, \tag{3.44}$$

where $\mathbf{L}(\alpha) \in \mathbb{R}^{6\times 2}$ is a matrix that implicitly depends on the calibration parameters ω_1, ω_2, $\mathbf{u}_1$ and $\mathbf{u}_2$.

Importantly, since this equation is true for any point in the scene $\mathbf{X}$, it can be applied to estimate predicted people's current position from the previous time step, *i.e.* $\mathbf{D}\mu_{t-1n}$. By doing this, we compute the velocity of the person due to the robot's motion. In other words, at time t, the n-th person will not be around position $\mathbf{D}\mu_{t-1n}$, but close to $\mathbf{D}\mu_{t-1n} + \mathbf{J}_{mw}(\mathbf{D}\mu_{t-1n})\binom{\dot{\alpha}}{\dot{\beta}}$. This is the value given to the translation due to the robot control:

$$\mathbf{C}_{tn} = \mathbf{J}_{mw}(\mathbf{D}\mu_{t-1n}) \begin{pmatrix} \dot{\alpha} \\ \dot{\beta} \end{pmatrix}. \tag{3.45}$$

§ Joint Jacobian

The joint motor-to-image Jacobian is the product of the Jacobians above:

$$\dot{\mathbf{V}} = \mathbf{J}_{wi}(\mathbf{X})\mathbf{J}_{mw}(\mathbf{X}) \begin{pmatrix} \dot{\alpha} \\ \dot{\beta} \end{pmatrix} = \mathbf{J}(\mathbf{X}) \begin{pmatrix} \dot{\alpha} \\ \dot{\beta} \end{pmatrix}. \tag{3.46}$$

To summarize, we are interested in two Jacobian operators. First, the inverse of the motor-to-image Jacobian maps the desired image shift $\Delta\mathbf{V}$ into motor velocities:

$$\begin{pmatrix} \dot{\alpha}_c \\ \dot{\beta}_c \end{pmatrix} = \gamma \mathbf{J}^{-1}(\mathbf{X}_s)\Delta\mathbf{V}, \tag{3.47}$$

where $\mathbf{X}_s$ is the servo position in three-dimension, $0 < \gamma < 1$ is a scale factor and $\dot{\alpha}_c$ and $\dot{\beta}_c$ are the yaw and pitch velocities to control the robot. Second, we can estimate the impact of these motor velocities onto the people's position by computing $\mathbf{C}_{tn}$ using (3.45) with $\dot{\alpha}_c$ and $\dot{\beta}_c$.

3.2.5 SYSTEM AND ARCHITECTURE

The proposed joint multi-person tracking and visual servoing system is implemented on top of the *NAOLab* middleware, which utilises the synchronisation strategy to find temporal matches between proprioceptive and perceptive information. A motor-camera calibration procedure is used to estimate the spatial relationship between the motor and the camera coordinate systems. At the end of the section, implementation details of the whole system is introduced.

§ NAOLab

There are several reasons to use a middleware architecture. First, algorithm implementations can be platform-independent and thus easily portable. Second, the use of external computational resources is transparent. Third, prototyping is much faster. For all these reasons, we developed a remote and modular layer-based middleware architecture named NAOLab.

NAOLab consists of 4 layers: drivers, shared memory, synchronization engine and application programming interface (API). Each layer is divided into 3 modules devoted to vision, audio and proprioception respectively. The first layer is platform-dependent and interfaces the sensors and actuators through the network using serialized data structures. The second layer implements a common shared memory that provides a concurrent interface to deserialize data from the robot sensors and implements an event-based control for robot command. The third layer is dedicated to synchronize the audio, video and proprioception data, so that the joint tracking-servoing system handles temporally coherent information. The last layer of NAOLab provides a general programming interface in C++ or Matlab to handle the robot's sensor data and manage its actuators.

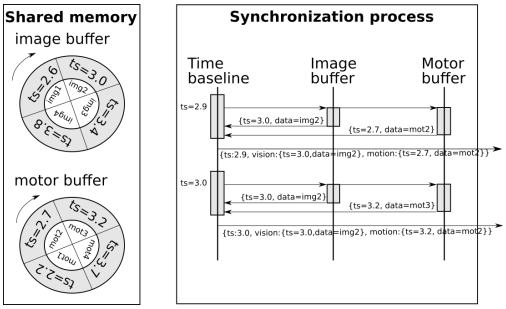


Figure 3.6: Robot data synchronization with NAOLab. The shared buffers contain time-stamped data. During the synchronization process, the nearest pairs of data are associated together regarding the time chosen time baseline.

§ Synchronization engine of NAOLab

The synchronization is implemented in the third *NAOLab* layer thanks to a circular data buffer (initialized to a fixed maximum size). The synchronization engine exploits these circular buffers together with the robot clock, and builds packages containing audio, visual and proprioception data whose corresponding time-stamps are close to each other. Figure 3.6 depicts the synchronization process for the multi-person tracking and visual servoing system (without audio involved), with a time baseline of 0.1 s and a buffer size of four packages.

As illustrated in Figure 3.7, the robot produces vision and proprioception data at different sampling rates. Each type of data is grabbed by a dedicated parallel process (drivers) who *publishes* the serialized data into the shared memory. After synchronization, the joint tracking and visual servoing module is able to request data from the shared memory or send motor-control commands to the motion drivers.

§ Motor-camera calibration

As previously discussed, the motor-to-world Jacobian required for the visual servoing depends on four parameters obtained through calibration: ω_1, ω_2, $\mathbf{u}_1$ and $\mathbf{u}_2$. In order to do that, we first notice that when the robot's head rotates from α_0 to α_i, there is an extrinsic rotation matrix $\mathbf{Q}_{0\to i}$ that can be expressed as a function of ω_1 and $\mathbf{u}_1$:

$$\begin{aligned}\mathbf{Q}_{0\to i} = \mathbf{I}_4 &+ \sin(\alpha_i - \alpha_0)\begin{pmatrix}\mathbf{S}(\omega_\mathbf{1}) & \mathbf{u}_1\\ \mathbf{0} & 0\end{pmatrix}\\ &+ (1-\cos(\alpha_i-\alpha_0))\begin{pmatrix}\mathbf{S}(\omega_\mathbf{1}) & \mathbf{u}_1\\ \mathbf{0} & 0\end{pmatrix}^2.\end{aligned} \tag{3.48}$$

At the same time, thanks to the cameras, the external matrix can be estimated with visual information. Indeed, the images of a static chessboard are recorded before and after the rotation, and by manually detecting the chessboard in the image, one can estimate the extrinsic matrix $\tilde{\mathbf{Q}}_{0\to i}$. Based on the previous equation and on the properties of the

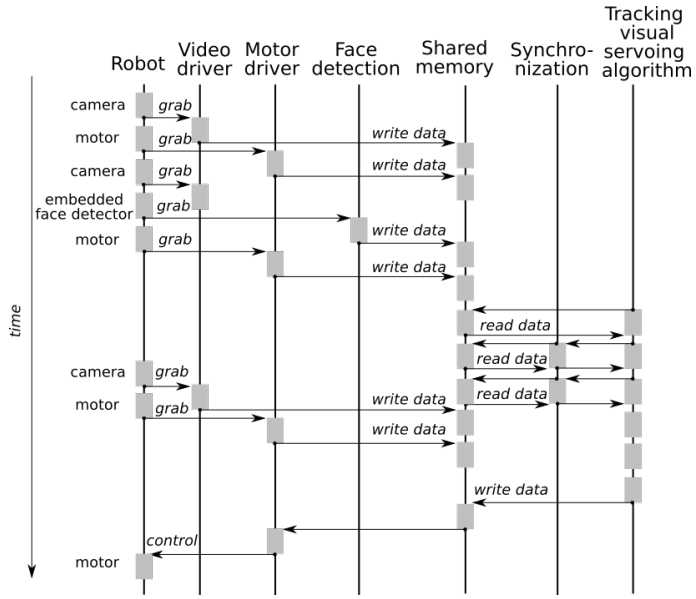


Figure 3.7: Data temporal flow chart: the drivers published serialized data into the shared memory. After synchronization, the joint tracking and servoing algorithm requests the data from which computes the appropriate motor control command, sent to the motor drivers through the shared memory.

trigonometric functions one can write:

$$J(\omega_{\mathbf{1}}, \mathbf{u}_1) = \sum_i \left\| 2\sin(\alpha_i - \alpha_0) \begin{pmatrix} \mathbf{Sk}(\omega_{\mathbf{1}}) & \mathbf{u}_1 \\ \mathbf{0} & 0 \end{pmatrix} - \tilde{\mathbf{Q}}_{0\to i} + \tilde{\mathbf{Q}}_{i\to 0} \right\|_F^2$$

where $\|.\|_F$ is the Frobenius norm. This cost function is then minimized to find the optimum values for the calibration parameters ω_1 and $\mathbf{u}_1$. The analogous procedure is repeated for the calibration parameters ω_2 and $\mathbf{u}_2$.

§ Implementation details

The overall system is implemented in C++, within the middleware framework described in Section 3.2.5. For the sake of reproducibility, we use the face detector and descriptor built-in on NAO, *i.e.* provided by NAO's API. The geometric observations, $\mathbf{g}_{tk}$ are face bounding boxes (image position, width and height). The position of the bouding box from the left and right camera images is combined by means of epipolar geometry, and triangulation to recover 3D face position. The face appearance descriptor is based on color histograms. Importantly, the detector and descriptor can be replaced or combined with other techniques thanks to the flexibility of the proposed probabilistic model for tracking. The detection and description of faces runs at 10 frames per second (FPS). Since the joint tracking-servoing computational load is less than 70 ms per time step, we are able to provide an on-line implementation of the joint multi-person tracking and visual servoing system.

The proposed variational model is governed by several parameters. Aiming at providing an algorithm that is dataset-independent and that features a good trade-off between

flexibility and performance, we set the observation covariance matrix Σ and the state covariance matrix Λ_n automatically from the detections. More precisely, both matrices are imposed to be diagonal; for Σ, the variances of the three-dimensional position, of the width, and of the horizontal (resp. vertical) speed are 1/2, 1/2 and 1/4 of the average detected width (resp. height). The rationale behind this choice is that we consider that the true detection lies approximately within the width and height of the detected bounding box. Regarding Λ_n , the diagonal entries are half of the tracked width and 5 times of motor speed. The window length chosen for the birth process is $T_{new} = 4$.

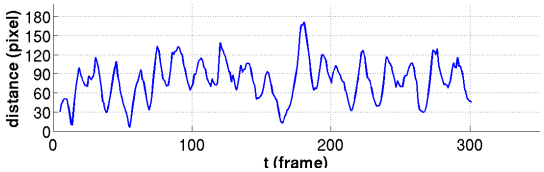
3.2.6 EXPERIMENTS

The proposed joint multi-person tracking and visual servoing system is evaluated on a series of scenarios using the NAO robot. Both left and right cameras provide VGA images, which are 640×480 pixels. Ten different sequences have been recorded in a regular living room scenario with its usual lighting source and background, where various people were moving around. The recorded sequences are thus challenging because of illumination variations, occlusions, appearance changes, and people leaving the robot field-of-view. We tried two different high-level control rules: (i) the robot should servo the first tracked person and (ii) the robot should sequentially change the pursued person every three seconds. The sequences with the servoing-tracking results are publicly available [4]. The sequences are named with the following scheme: NAO-MPVS-NS-P, which stands for NAO Multi-Person Visual Servoing, N is the number of people present in the sequence (although not constantly visible), S defines the strategy when $N > 1$ ("F" for following the first tracked person and "J" for jumping every three seconds) and P for the trial. For instance NAO-MPVS-1-1, is the first trial of a scenario involving one person, while NAO-MPVS-2J-3 is the third trial of a scenario involving two people and the control rule set to "jumping". In the following, we provide both quantitative and qualitative evaluation of both the visual servoing and of the multi-person tracking.

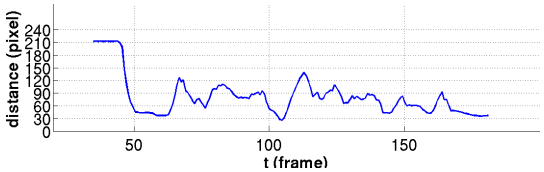
§ Visual servoing

Figure 3.8 shows the distance in pixels from the tracked person to the left camera image center over time, for three different sequences of the dataset, all under the servoing strategy of following the first tracked person. We can clearly see the oscillation due to the lag between the person's motion and the control response. Remarkably, shortly after each of the person's movements, the servoing mechanism position back the person in the image center. Indeed, after a few seconds the distance between the tracked person and the image center has decreased to below 30 pixels. Furthermore, if we compute the average distance for all frames of all sequences (*i.e.* almost $2,000$ frames), we obtain an average distance of 80.1 pixels, indicating that the proposed system is able to approximately maintain the person's face at the image center.

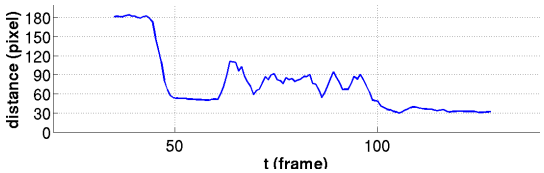
[4] `https://team.inria.fr/perception/mot-servoing/`



(a) NAO-MPVS-1-3



(b) NAO-MPVS-2F-2



(c) NAO-MPVS-2F-3

Figure 3.8: The distance between tracked person and left camera image center (in pixels) over time (in frames) for three different sequences.

Qualitatively speaking, Figure 3.9 shows four frames of the most challenging sequence in the dataset, NAO-MPVS-3F-1. This sequence involves three people, among which the tracked one passes behind the other two. Each of the frames shows the marked-up left camera image together with a bird-view representation of the tracked scene. While in the marked-up image we can see the face detection (blue), the tracked bounding box with the tracking ID (red), the target motion due to the robot control (magenta) and the target's self-motion (green), in the bird-view we can see the tracking ID and the trajectories. We can observe different prominent characteristics of the proposed system. Firstly, the ability to separate the image motion due to the robot control, from the image motion due to the natural movements of the target allows for the estimation of a smooth trajectory in the three-dimensional space. Secondly, the algorithm is able to keep a rough estimate of the positions of the targets that are out of the field of view, and even more important, to correctly re-assign the identity to a re-appearing person thanks to the cooperation of the state dynamics and appearance model. Thirdly, the capacity of the system to create a new track when a new person appears in the field-of-view thanks to the birth process. Finally, robustness to identify switches, even with illumination and appearance changes, occlusions and the robot's self-motion.

§ Multi-person tracking

We have also evaluated the impact of the visual servoing from the multi-person tracking perspective. Aiming to this, we compared the performance of the system when using/discarding the image-motion due to the robot control. In more detail, we manually annotated the position of the persons in three different sequences of increasing complexity (NAO-MPVS-1-1, -2J-1 and -3J-1) and we computed the following standard multi-person

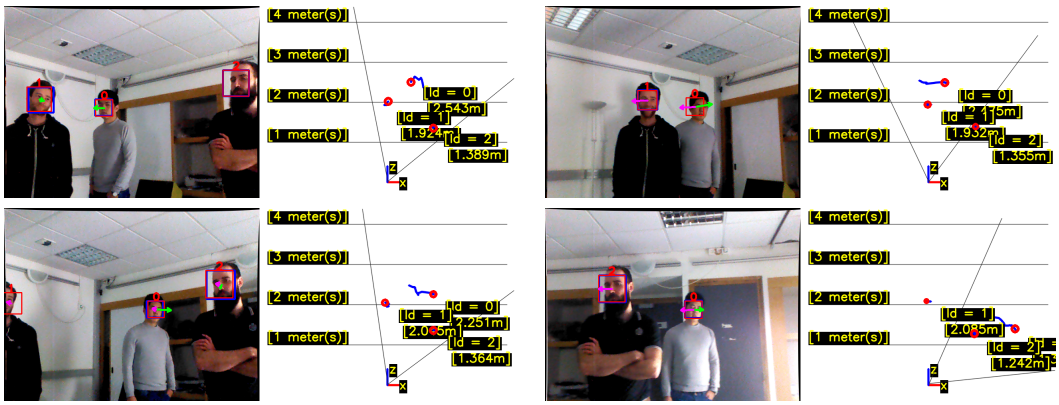


Figure 3.9: Left: robot left camera view, red bounding boxes represent the three-dimensional tracking results projected on the image, blue bounding boxes represents three-dimensional face-detection, green arrows represent people's self-velocity, magenta arrows represent the velocity due to the robot control. Right: scenario bird-view, red circles represent current tracking positions and the blue lines represent tracked people's trajectories. Example results are from NAO-MPVS dataset sequence NAO-MPVS-3F-1

Table 3.3: Comparison of tracking result w/o and w control by MOT metrics on three sequences with increasing complexity.
↑ : the higher the better ↓ : the lower the better

Sequence	Ctrl	MOTA(↑)	MOTP(↑)	FP(↓)	FN(↓)	IDs(↓)
NAO-MPVS-1-1	w/o	92.1	67.2	9	9	0
	w/	91.3	68.7	10	10	0
NAO-MPVS-2J-1	w/o	52.8	67.1	93	207	2
	w/	81.6	68.0	30	88	0
NAO-MPVS-3J-1	w/o	35.8	62.3	159	433	19
	w/	63.1	62.1	83	268	0
Overall	w/o	48.8	65.0	261	649	21
	w/	**73.1**	**65.3**	**123**	**366**	**0**

tracking evaluation metrics [139]: *multiple-object tracking accuracy* (MOTA), *multiple-object tracking precision* (MOTP), *false positives* (FP), *false negatives* (FN) and *identity switches* (ID). While for MOTA and MOTP are the higher the better, for the rest are the lower the better. Table 3.3 reports all these measures with (w/) and without (w/o) using the impact of the robot control (Ctrl) on the targets' position, *i.e.* $\mathbf{C}_{tn}$.

In light of the results, we can see that indeed NAO-MPVS-1-1 is an easy sequence. Indeed, the only person to be tracked does not perform large movements. It is therefore not surprising that (i) the performance measures are very high and (ii) there is no much performance difference when adding the impact of the control variable. When the complexity of the scenario increases (more people to track, larger movements) the proposed tracking framework including motor information leads to higher accuracy and less FP/FN/IDs results. This difference is specially remarkable in the case of the NAO-MPVS-3J-1 sequence, showing that tracking based only on the appearance and the position of people is not sufficient when multiple people need to be tracked, while at the same time the robot performs some movements. We also notice that the MOTP measure is not strongly affected by the information provided by the robot control, and this is expected. Indeed, MOTP measures the tracking precision in terms of how much do the bounding boxes of

the detected positives overlap with their assigned true positives. In other words, if a detected positive is too far from all true positives, it counts as a FP, but is not computed as a precision error. This confirms our hypothesis that the use of $\mathbf{C}_{tn}$ is crucial to correct large deviations of the tracking estimates due to the motor control; And at the same time result shows that it is not specially helpful to refine these tracking estimates. In other words, the use of $\mathbf{C}_{tn}$ is complementary to developing precise tracking methodologies which are able to provide very accurate bounding box localization once the large corrections due to the motor-control are applied.

Overall the proposed joint multiple-person tracking and visual servoing framework leads to promising results even in sequences which contain large and frequent robot motions under challenging illumination conditions. Remarkably, the method is able to systematically keep the right person identity for all three sequences. This feature is highly desirable for numerous applications and critically depends upon the use of motor information during tracking.

3.3 CONCLUSION

This Chapter proposes a variational Bayesian solution to the visual multiple-target problem. The birth process and the visibility process are jointly proposed to deal with the varying number of people. The proposed method is benchmarked on MOT Challenge 2016. In addition, the variational tracker is implemented and applied on a real robot. The tracking system is further combined with visual servoing which is able to continuously estimate the three-dimensional position of a time-varying number of people and to encompass the effect of the robot's ego-motion. In addition, the proposed probabilistic formulation and a variational approximation allowing to effectively solve the inference problem while keeping a reasonably low computational cost (the overall systems works at 10 FPS). Furthermore, thanks to the motor information, the system can separate people's self-motion from the robot's ego-motion, leading to more robust tracking capabilities. The experimental study on the NAO Multi-Person Visual Servoing dataset confirms our hypothesis that including the robot's ego-motion into the tracking probabilistic framework is of utmost importance for the performance of the system. The possible future work of this part might be: (i) the calibration of other motors (*e.g.* robot's leg-joint), thus compensating for the full ego-motion and (ii) the combination of audio information to construct a tracking system based on audio-visual information, thus able to track outside the camera field-of-view.

CHAPTER 4

AUDIO-VISUAL TRACKING OF MULTIPLE SPEAKERS

4.1 INTRODUCTION

In this chapter we address the problem of tracking multiple speakers via the fusion of visual and auditory information [50, 61, 116, 68, 129, 67]. We propose to exploit the complementary nature of these two modalities in order to accurately estimate the position of each person at each time step, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status, either speaking or silent, of each tracked person. We propose to cast the problem at hand into a generative audio-visual fusion (or association) model formulated as a latent-variable temporal graphical model. We propose a tractable solver via a variational approximation.

We are particularly interested in tracking people involved in informal meetings and social gatherings. In this type of scenarios, participants wander around, cross each other, move in and out the camera field of view, take speech turns, etc. Acoustic room conditions, e.g. reverberation, and overlapping audio sources of various kinds drastically deteriorate or modify the microphone signals. Likewise, occluded persons, lighting conditions and middle-range camera-viewing complicate the task of visual processing. It is therefore impossible to gather reliable and continuous flows of visual **and** audio observations. Hence one must design a fusion and tracking method that is able to deal with intermittent visual and audio data.

We propose a multi-speaker tracking method based on a dynamic Bayesian model that fuses audio and visual information over time from their respective observations spaces. This may well be viewed as a generalization of single-observation and single-target Kalman filtering – which yields an exact recursive solution – to multiple-observations and -targets, which makes the recursive solution intractable. We propose a variational approximation of the posterior distribution over the continuous variables (positions and velocities of tracked persons) and discrete variables (observation-to-person associations) at each time

step, given all the past and present audio and visual observations. The approximation of this joint distribution with a factorized distribution makes the tracking problem tractable: the solution takes the form of a closed-form expectation maximization (EM) procedure.

In general, multiple object tracking consists of the temporal estimation of the kinematic state of each object, i.e. position and velocity. In computer vision, local descriptors are used to better discriminate between objects, e.g. person detectors/descriptors based on hand-crafted features [7] or on deep neural networks [8]. If the tracked objects emit sounds, their states can be inferred as well using sound-source localization techniques combined with tracking. These techniques are often based on the estimation of the sound's direction of arrival (DOA) using a microphone array, e.g. [89]. DOA estimation can be carried out either in the temporal domain [2], or in the spectral (Fourier) domain [40]. However, spectral-domain DOA estimation methods are more robust than temporal-domain methods, in particular in the presence of background noise and reverberation [77, 79].

Via proper camera-microphone calibration, audio and visual observations can be aligned such that a DOA corresponds to a 2D location in the image plane. In this Chapter we adopt the audio-visual alignment method of [35] which learns a mapping, from a vector space spanned by multichannel spectral features (or audio features, in short) to the image plane, as well as the inverse of this mapping. This allows us to exploit the richness of representing acoustic signals in the short-time Fourier domain [56] and to extract noise- and reverberation-free audio features [77].

We propose to represent the audio-visual fusion problem via two sets of independent variables, i.e. visual-feature-to-person and audio-feature-to-person sets of assignment variables. An interesting characteristic of this way of doing is that the proposed tracking algorithm can indifferently use visual features, audio features, or a combination of both, and choose independently for every target and at every time step. Indeed, audio and visual information are rarely available simultaneously and continuously. Visual information suffers from limited camera field-of-view, occlusions, false positives, missed detections, etc. Audio information is often corrupted by room acoustics, environmental noise and overlapping acoustic signals. In particular speech signals are sparse, non-stationary and are emitted intermittently, with silence intervals between speech utterances. Hence a robust audio-visual tracking must explicitly take into account the temporal sparsity of the two modalities and this is exactly what is proposed in this Chapter.

We use the AVDIAR dataset [53] to evaluate the performance of the proposed audio-visual tracker. We use the MOT (multiple object tracking) metrics to quantitatively assess method performance. In particular the tracking accuracy (MOTA), which combines false positives, false negatives, identity switches and compares them with the ground-truth trajectories, is a commonly used score to assess the quality of a multiple person tracker.[1] We use the MOT metrics to compare our method with two recently proposed audio-visual tracking methods [68, 67] and with a visual tracker [7]. An interesting outcome of the proposed method is that speaker diarization, i.e. who speaks when, can be coarsely inferred

[1] `https://motchallenge.net/`

from the tracking output, thanks to the audio-feature-to-person assignment variables. The speaker diarization results obtained with our method are compared with two other methods [147, 53] based on the diarization error rate (DER) score.

The remainder of the Chapter is organized as follows. Section 4.2 describes the related work. Section 4.3 describes in detail the proposed formulation. Section 4.4 describes the proposed variational approximation and Section 4.5 details the variational expectation-maximization procedure. The algorithm implementation is described in Section 4.6. Tracking results and comparisons with other methods are reported in Section 4.7. Finally, Section 4.8 draws a few conclusions. Supplemental materials are available on our website.[2]

4.2 Related Work

In computer vision, there is a long history of multiple object tracking methods. While these methods provide interesting insights concerning the problem at hand, a detailed account of existing visual trackers is beyond the scope of this Chapter. Several audio-visual tracking methods were proposed in the recent past, e.g. [26, 50, 61, 116]. These papers proposed to use approximate inference of the filtering distribution using Markov chain Monte Carlo particle filter sampling (MCMC-PF). These methods cannot provide estimates of the accuracy and merit of each modality with respect to each tracked person. Sampling and distribution estimation are performed in parameter space but no statistics are gathered in the observations spaces.

More recently, audio-visual trackers based on particle filtering (PF) and probability hypothesis density (PHD) filters were proposed, e.g. [68, 129, 67, 85, 83, 121]. In [67] DOAs of audio sources to guide the propagation of particles and combined the filter with a mean-shift algorithm to reduce the computational complexity. Some PHD filter variants were proposed to improve tracking performance [85, 87, 83]. The method of [68] also used DOAs of active audio sources to give more importance to particles located around DOAs. Along the same line of thought, [67] proposed a mean-shift sequential Monte Carlo PHD (SMC-PHD) algorithm that used audio information to improve the performance of a visual tracker. This implies that the persons being tracked must emit acoustic signals continuously and that multiple-source audio localization is reliable enough for proper audio-visual alignment.

PF- and PHD-based tracking methods are computationally efficient but their inherent limitation is that they are unable to associate observations to tracks. Hence they require an external post-processing mechanism that provides associations. Also, in the case of PF-based filtering, the number of tracked persons must be set in advance. Moreover, both PF- and PHD-based trackers provide non-smooth trajectories since the state dynamics are not explicitly enforced. In contrast, the proposed variational formulation embeds association variables within the model, uses a birth process to estimate the initial number of persons and to add new ones along time, and an explicit dynamic model yields smooth trajectories.

[2] https://team.inria.fr/perception/research/variational_av_tracking/

Another limitation of the methods proposed in [50, 116, 67, 85, 83, 121, 123] is that they need as input a continuous flow of audio and visual observations. To some extent, this is also the case with [68, 67, 84], where only the audio observations are supposed to be continuous. All these methods showed good performance in the case of the AV16.3 dataset [73] in which the participants spoke simultaneously and continuously – which is somehow artificial. The AV16.3 dataset was recorded in a specially equipped meeting room using a large number of cameras to guarantee that frontal views of the participants were always available. This contrasts with the AVDIAR dataset which was recorded with one sensor unit composed of two cameras and six microphones. The AVDIAR scenarios are composed of participants that take speech turns while they look at each other, hence they speak intermittently and they do not always face the cameras.

Recently, we proposed an audio-visual clustering method [51] and an audio-visual speaker diarization method [53]. The weighted-data clustering method of [51] analyzed a short time window composed of several audio and visual frames and hence it was assumed that the speakers were static within such temporal windows. Binaural audio features were mapped onto the image plane and were clustered with nearby visual features. There was no dynamic model that allowed to track speakers. The audio-visual diarization method [53] used an external multi-object visual tracker that provided trajectories for each tracked person. The audio-feature-space to image-plane mapping [35] was used to assign audio information to each tracked person at each time step. Diarization itself was modeled with a binary state variable (speaking or silent) associated with each person. The diarization transition probabilities (state dynamics) were hand crafted, with the assumption that the speaking status of a person was independent of all the other persons. Because of the small number of state configurations, i.e. $\{0,1\}^N$ (where N is the maximum number of tracked persons), the MAP solution could be found by exhaustively searching the state space. In Section 4.7.8 we use the AVDIAR recordings to compare our diarization results with the results obtained with [53].

The variational inference method proposed may well be viewed as a multimodal generalization of [7]. We show that the model of [7] can be extended to deal with observations living in completely different mathematical spaces. Indeed, we show that two (or several) different data-processing pipelines can be embedded and treated on an equal footing in the proposed formulation. Special attention is given to audio-visual alignment and to audio-to-person assignments: (i) we learn a mapping from the space of audio features to the image plane, as well as the inverse of this mapping, which are integrated in the proposed generative approach, and (ii) we show that the additional assignment variables due to the audio modality do not affect the complexity of the algorithm. Absence of observed data of any kind or erroneous data are carefully modeled: this enables the algorithm to deal with intermittent observations, whether audio, visual, or both. This is probably one of the most prominent features of the method, in contrast with most existing audio-visual tracking methods which require continuous and simultaneous flows of visual and audio data.

This Chapter is an extended version of [13] and of [14]. The probabilistic model and its variational approximation were briefly presented in [13] together with preliminary results

obtained with three AVDIAR sequences. Reverberation-free audio features were used in [14] where it was shown that good performance could be obtained with these features when the audio mapping was trained in one room and tested in another room. With respect to these two papers. we provide detailed descriptions of the proposed formulation, of the variational expectation maximization solver and of the implemented algorithm. We explain in detail the birth process, which is crucial for track initialization and for detecting potentially new tracks at each time step. We experiment with the entire AVDIAR dataset and we benchmark our method with the state-of-the-art multiple-speaker audio-visual tracking methods [68, 67] and with [7]. Moreover, we show that our tracker can be used for speaker diarization.

4.3 Proposed Model

4.3.1 Mathematical Definitions and Notations

In this Chapter, video and audio data are assumed to be synchronized, and let t denote the common frame index. Since the speakers are tracked on the image plane, we thus let $\boldsymbol{X}_{tn} \in \mathcal{X} \subset \mathbb{R}^2$, $\boldsymbol{Y}_{tn} \in \mathcal{Y} \subset \mathbb{R}^2$ and $\boldsymbol{W}_{tn} \in \mathcal{W} \subset \mathbb{R}^2$ be three latent variables that correspond to the 2D position, 2D velocity and 2D size (width and height) of person n at t. Typically, $\boldsymbol{X}_{tn}$ and $\boldsymbol{W}_{tn}$ correspond to the center and size of a bounding box of a person while $\boldsymbol{Y}_{tn}$ is the velocity of $\boldsymbol{X}_{tn}$. Let $\boldsymbol{S}_{tn} = (\boldsymbol{X}_{tn}^\top, \boldsymbol{W}_{tn}^\top, \boldsymbol{Y}_{tn}^\top)^\top \subset \mathbb{R}^6$ be the complete set of continuous latent variables at t, where $^\top$ denotes the transpose operator. Without loss of generality, in this section a person is characterized with the bounding box of her/his head and the center of this bounding box is assumed to be the location of the corresponding speech source.

We now define the observations. Let $\{\boldsymbol{f}_{tm}\}_{m=1}^{M_t}$ and $\{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$ be realizations of the visual and audio random observed variables $\{\boldsymbol{F}_{tm}\}_{m=1}^{M_t}$ and $\{\boldsymbol{G}_{tk}\}_{k=1}^{K_t}$, respectively. A visual observation, $\boldsymbol{f}_{tm} = (\boldsymbol{v}_{tm}^\top, \boldsymbol{u}_{tm}^\top)^\top$, corresponds to the bounding box of a detected face and it is the concatenation of the bounding-box center, width and height, $\boldsymbol{v}_{tm} \in \mathcal{V} \subset \mathbb{R}^4$, and of a feature vector $\boldsymbol{u}_{tm} \in \mathcal{H} \subset \mathbb{R}^d$ that describes the photometric content of that bounding box, i.e. a d-dimensional face descriptor (Section 4.7.3). An audio observation, $\boldsymbol{g}_{tk}$, corresponds to an inter-microphone spectral feature, where k is a frequency sub-band index. Let's assume that there are K sub-bands, that $K_t \leq K$ sub-bands are *active* at t, i.e. with sufficient energy, and that there are L frequencies per sub-band. Hence, $\boldsymbol{g}_{tk} \in \mathbb{R}^{2L}$ corresponds to L complex-valued Fourier coefficients which are represented by their real and imaginary parts. In practice, the inter-microphone features $\{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$ contain audio-source localization information and are obtained by applying the multi-channel audio processing method described in detail below (Section 4.7.2). Note that both the number of visual and of audio observations at t, M_t and K_t, vary over time. Let $\mathbf{o}_{1:t} = (\mathbf{o}_1, \ldots, \mathbf{o}_t)$ denote the set of observations from 1 to t, where $\mathbf{o}_t = (\mathbf{f}_t, \mathbf{g}_t)$.

We now define the assignment variables of the proposed latent variable model. There is an assignment variable (a discrete random variable) associated with each observed variable. Namely, let A_{tm} and B_{tk} be associated with $\boldsymbol{f}_{tm}$ and with $\boldsymbol{g}_{tk}$, respectively, e.g.

$p(A_{tm} = n)$ denotes the probability of assigning visual observation m at t to person n. Note that $p(A_{tm} = 0)$ and $p(B_{tk} = 0)$ are the probabilities of assigning visual observation m and audio observation k to none of the persons, or to nobody. In the visual domain, this may correspond to a false detection while in the audio domain this may correspond to an audio signal that is not uttered by a person. There is an additional assignment variable, C_{tk} that is associated with the audio generative model described in Section 4.3.4. The assignment variables are jointly denoted with $\boldsymbol{Z}_t = (\boldsymbol{A}_t, \boldsymbol{B}_t, \boldsymbol{C}_t)$.

4.3.2 THE FILTERING DISTRIBUTION

We remind that the objective is to estimate the positions and velocities of participants (multiple person tracking) and, possibly, to estimate their speaking status (speaker diarization). The audio-visual multiple-person tracking problem is cast into the problems of estimating the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ and of inferring the state variable $\mathbf{S}_t$. Subsequently, speaker diarization can be obtained from audio-feature-to-person information via the estimation of the assignment variables $\boldsymbol{B}_{tk}$ (Section 4.6.3).

We reasonably assume that the state variable $\mathbf{S}_t$ follows a first-order Markov model, and that the visual and audio observations only depend on $\mathbf{S}_t$ and $\boldsymbol{Z}_t$. By applying Bayes rule, one can then write the filtering distribution of $(\mathbf{s}_t, \boldsymbol{z}_t)$ as:

$$p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t}) \propto p(\mathbf{o}_t|\mathbf{s}_t, \boldsymbol{z}_t)p(\boldsymbol{z}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{o}_{1:t-1}), \tag{4.1}$$

with:

$$p(\mathbf{o}_t|\mathbf{s}_t, \boldsymbol{z}_t) = p(\mathbf{f}_t|\mathbf{s}_t, \boldsymbol{a}_t)p(\mathbf{g}_t|\mathbf{s}_t, \boldsymbol{b}_t, \boldsymbol{c}_t), \tag{4.2}$$

$$p(\boldsymbol{z}_t|\mathbf{s}_t) = p(\boldsymbol{a}_t)p(\boldsymbol{b}_t)p(\boldsymbol{c}_t|\mathbf{s}_t, \boldsymbol{b}_t), \tag{4.3}$$

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \int p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})d\mathbf{s}_{t-1}. \tag{4.4}$$

Eq. (4.2) is the joint (audio-visual) observed-data likelihood. Visual and audio observations are assumed independent conditionally to $\mathbf{S}_t$, and their distributions will be detailed in Sections 4.3.3 and 4.3.4, respectively.[3] Eq. (4.3) is the prior distribution of the assignment variable. The observation-to-person assignments are assumed to be a priori independent so that the probabilities in (4.3) factorize as:

$$p(\boldsymbol{a}_t) = \prod_{m=1}^{M_t} p(a_{tm}), \tag{4.5}$$

$$p(\boldsymbol{b}_t) = \prod_{k=1}^{K_t} p(b_{tk}), \tag{4.6}$$

$$p(\boldsymbol{c}_t|\mathbf{s}_t, \boldsymbol{b}_t) = \prod_{k=1}^{K_t} p(c_{tk}|\boldsymbol{s}_{tn}, B_{tk} = n). \tag{4.7}$$

[3] We will see that $\mathbf{G}_t$ depends on $\mathbf{X}_t$ but depends neither on $\mathbf{W}_t$ nor on $\mathbf{Y}_t$, and $\mathbf{F}_t$ depends on $\mathbf{X}_t$ and $\mathbf{W}_t$ but not on $\mathbf{Y}_t$.

It makes sense to assume that these distributions do not depend on t and that they are uniform. The following notations are introduced: $\eta_{mn} = p(A_{tm} = n) = 1/(N+1)$ and $\rho_{kn} = p(B_{tk} = n) = 1/(N+1)$. The probability $p(c_{tk}|\boldsymbol{s}_{tn}, B_{tk} = n)$ is discussed below (Section 4.3.4).

Eq. (4.4) is the predictive distribution of $\mathbf{s}_t$ given the past observations, i.e. from 1 to $t-1$. The state dynamics in (4.4) is modeled with a linear-Gaussian first-order Markov process. Moreover, it is assumed that the dynamics are independent over speakers:

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{s}_{t-1\,n}, \boldsymbol{\Lambda}_{tn}), \tag{4.8}$$

where $\boldsymbol{\Lambda}_{tn}$ is the dynamics' covariance matrix and $\mathbf{D}$ is the state transition matrix, given by:

$$\mathbf{D} = \begin{pmatrix} \mathbf{I}_{4\times4} & \mathbf{I}_{2\times2} \\ \mathbf{0}_{2\times4} & \mathbf{I}_{2\times2} \end{pmatrix}.$$

As described in Section 4.4 below, an important feature of the proposed model is that the predictive distribution (4.4) at frame t is computed from the state dynamics model (4.8) and an approximation of the filtering distribution $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$ at frame $t-1$, which also factorizes across speaker. As a result, the computation of (4.4) factorizes across speakers as well.

4.3.3 The Visual Observation Model

As already mentioned above (Section 4.3.1), a visual observation $\boldsymbol{f}_{tm}$ consists of the center, width and height of a bounding box, namely $\boldsymbol{v}_{tm} \in \mathcal{V} \subset \mathbb{R}^4$, as well as of a feature vector $\boldsymbol{u}_{tm} \in \mathcal{H} \subset \mathbb{R}^d$ describing the region inside the bounding box. Since the velocity is not observed, a 4×6 projection matrix $\mathbf{P}_f = (\mathbf{I}_{4\times4}\ \mathbf{0}_{4\times2})$ is used to project $\boldsymbol{s}_{tn}$ onto $\mathcal{V}$. Assuming that the M_t visual observations $\{\boldsymbol{f}_{tm}\}_{m=1}^{M_t}$ available at t are independent, and that the appearance of a person is independent of his/her position in the image, the visual likelihood in (4.2) is defined as:

$$p(\mathbf{f}_t|\mathbf{s}_t, \boldsymbol{a}_t) = \prod_{m=1}^{M_t} p(\boldsymbol{v}_{tm}|\mathbf{s}_t, a_{tm}) p(\boldsymbol{u}_{tm}|\mathbf{h}, a_{tm}), \tag{4.9}$$

where the observed bounding-box centers, widths, heights, and feature vectors are drawn from the following distributions:

$$p(\boldsymbol{v}_{tm}|\mathbf{s}_t, A_{tm} = n) = \begin{cases} \mathcal{N}(\boldsymbol{v}_{tm}; \mathbf{P}_f \boldsymbol{s}_{tn}, \boldsymbol{\Phi}_{tm}) & \text{if } 1 \le n \le N \\ \mathcal{U}(\boldsymbol{v}_{tm}; \text{vol}(\mathcal{V})) & \text{if } n = 0, \end{cases} \tag{4.10}$$

$$p(\boldsymbol{u}_{tm}|\mathbf{h}, A_{tm} = n) = \begin{cases} \mathcal{B}(\boldsymbol{u}_{tm}; \boldsymbol{h}_n) & \text{if } 1 \le n \le N \\ \mathcal{U}(\boldsymbol{u}_{tm}; \text{vol}(\mathcal{H})) & \text{if } n = 0, \end{cases} \tag{4.11}$$

where $\mathbf{\Phi}_{tm} \in \mathbb{R}^{4\times4}$ is a covariance matrix quantifying the measurement error in the bounding-box center and size, $\mathcal{U}(\cdot;\text{vol}(\cdot))$ is the uniform distribution with $\text{vol}(\cdot)$ being the support volume of the variable space, $\mathcal{B}(\cdot;\boldsymbol{h})$ is the Bhattacharya distribution with parameter $\boldsymbol{h}$, and $\mathbf{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N) \in \mathbb{R}^{d\times N}$ is a set of prototype feature vectors that model the appearances of the N persons.

4.3.4 The Audio Observation Model

It is well established in the multichannel audio signal processing literature that inter-microphone spectral features encode sound-source localization information [35, 40, 77]. Therefore, observed audio features, $\mathbf{g}_t = \{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$ are obtained by considering all pairs of a microphone array. Audio observations depend neither on $\boldsymbol{w}_t$ (size of the bounding box) nor on $\boldsymbol{y}_t$ (velocity). Hence one can replace $\boldsymbol{s}$ with $\boldsymbol{x} = \mathbf{P}_g\boldsymbol{s}$ in the equations below, with $\mathbf{P}_g = (\mathbf{I}_{2\times2}\ \mathbf{0}_{2\times4})$. By assuming independence across frequency sub-bands (indexed by k), the audio likelihood in (4.2) can be factorized as:

$$p(\mathbf{g}_t|\mathbf{s}_t, \boldsymbol{b}_t, \boldsymbol{c}_t) = \prod_{k=1}^{K_t} p(\boldsymbol{g}_{tk}|\boldsymbol{x}_{tb_{tk}}, b_{tk}, c_{tk}). \tag{4.12}$$

While the inter-microphone spectral coefficients $\boldsymbol{g}_{tk}$ contain localization information, in complex acoustic environments there is no explicit function that maps source locations onto inter-microphone spectral features. Moreover, this mapping is non-linear. We therefore make recourse to modeling this relationship via learning a regression function. We propose to use the piecewise-linear regression [34] which belongs to the mixture of experts (MOE) class of models. For that purpose we consider a training set of audio features and their associated source locations, $\mathcal{T} = \{(\boldsymbol{g}_i, \boldsymbol{x}_i)\}_{i=1}^{I}$ and let $(\boldsymbol{g}, \boldsymbol{x}) \in \mathcal{T}$. The joint probability of $(\boldsymbol{g}, \boldsymbol{x})$ writes:

$$p(\boldsymbol{g}, \boldsymbol{x}) = \sum_{r=1}^{R} p(\boldsymbol{g}|\boldsymbol{x}, C = r)p(\boldsymbol{x}|C = r)p(C = r). \tag{4.13}$$

Assuming Gaussian variables, we have $p(\boldsymbol{g}|\boldsymbol{x}, C = r) = \mathcal{N}(\boldsymbol{g}|\mathbf{L}_r\boldsymbol{x} + \boldsymbol{l}_r, \mathbf{\Sigma}_r)$, $p(\boldsymbol{x}|C = r) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\nu}_r, \mathbf{\Omega}_r)$, and $p(C = r) = \pi_r$, where matrix $\mathbf{L}_r \in \mathbb{R}^{2L\times2}$ and vector $\boldsymbol{l}_r \in \mathbb{R}^{2L}$ characterize the r-th affine transformation that maps the space of source locations onto the space spanned by inter-microphone sub-band spectral features, $\mathbf{\Sigma}_r \in \mathbb{R}^{2L\times2L}$ is the associated covariance matrix, and $\boldsymbol{x}$ is drawn from a Gaussian mixture model with R components, each component r being characterized by π_r, $\boldsymbol{\nu}_r \in \mathbb{R}^2$ and $\mathbf{\Omega}_r \in \mathbb{R}^{2\times2}$. The parameter set of this model is:

$$\{\mathbf{L}_r, \boldsymbol{l}_r, \mathbf{\Sigma}_r, \boldsymbol{\nu}_r, \mathbf{\Omega}_r, \pi_r\}_{r=1}^{r=R}. \tag{4.14}$$

These parameters can be estimated via a closed-form EM procedure from a training dataset, e.g. $\mathcal{T}$ (please consult [34, 35] and Section 4.7.2 below for more details). One should notice that there is a parameter set for each sub-band k, $1 \leq k \leq K$, hence there

are K models that need be trained in our case. It follows that (4.12) writes:

$$p(\boldsymbol{g}_{tk}|\boldsymbol{x}_{tn}, B_{tk}=n, C_{tk}=r) = \begin{cases}\mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr}\boldsymbol{x}_{tn}+\boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{g}_{tk}; \text{vol}(\mathcal{G})) & \text{if } n = 0.\end{cases} \tag{4.15}$$

The right-hand side of (4.7) can now be written as:

$$p(C_{tk}=r|\boldsymbol{x}_{tn}, B_{tk}=n) = \frac{\pi_r \mathcal{N}(\boldsymbol{x}_{tn}; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)}{\sum_{i=1}^{R} \pi_i \mathcal{N}(\boldsymbol{x}_{tn}; \boldsymbol{\nu}_i, \boldsymbol{\Omega}_i)}. \tag{4.16}$$

4.4 VARIATIONAL INFERENCE

Direct estimation of the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ is intractable. Consequently, evaluating expectations over this distribution is intractable as well. We overcome this problem via variational inference and associated EM closed-form solver [19, 133]. More precisely $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ is approximated with the following factorized form:

$$p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t}) \approx q(\mathbf{s}_t, \boldsymbol{z}_t) = q(\mathbf{s}_t)q(\boldsymbol{z}_t), \tag{4.17}$$

which implies

$$q(\mathbf{s}_t) = \prod_{n=1}^{N} q(\boldsymbol{s}_{tn}),\ q(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} q(a_{tm}) \prod_{k=1}^{K} q(b_{tk}, c_{tk}), \tag{4.18}$$

where $q(A_{tm}=n)$ and $q(B_{tk}=n, C_{tk}=r)$ are the variational posterior probabilities of assigning visual observation m to person n and audio observation k to person n, respectively. The proposed variational approximation (4.17) amounts to break the conditional dependence of $\mathbf{S}$ and $\boldsymbol{Z}$ with respect to $\mathbf{o}_{1:t}$ which causes the computational intractability. Note that the visual, $\boldsymbol{A}_t$, and audio, $\boldsymbol{B}_t$, $\boldsymbol{C}_t$, assignment variables are independent, that the assignment variables for each observation are also independent, and that B_{tk} and C_{tk} are conditionally dependent on the audio observation. This factorized approximation makes the calculation of $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ tractable. The optimal solution is given by an instance of the variational expectation maximization (VEM) algorithm [19, 133], which alternates between two steps:

- *Variational* E-step: the approximate log-posterior distribution of each one of the latent variables is estimated by taking the expectation of the complete-data log-likelihood over the remaining latent variables, i.e. (4.19), (4.20), and (4.21) below, and
- M-step: model parameters are estimated by maximizing the variational expected complete-data log-likelihood.

In the case of the proposed model the latent variable log-posteriors write:

$$\log q^\star(\boldsymbol{s}_{tn}) = \mathbb{E}_{q(\boldsymbol{z}_t)\prod_{\ell\neq n} q(\boldsymbol{s}_{t\ell})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})], \tag{4.19}$$

$$\log q^\star(a_{tm}) = \mathbb{E}_{q(\mathbf{s}_t)\prod_{\ell\neq m} q(a_{t\ell})\prod_k q(b_{tk},c_{tk})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})], \tag{4.20}$$

$$\log q^\star(b_{tk}, c_{tk}) = \mathbb{E}_{q(\mathbf{s}_t)\prod_m q(a_{tm})\prod_{\ell\neq k} q(b_{t\ell},c_{t\ell})}[\log p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})]. \tag{4.21}$$

A remarkable consequence of the factorization (4.17) is that $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$ is replaced with $q^\star(\mathbf{s}_{t-1}) = \prod_{n=1}^N q^\star(\boldsymbol{s}_{t-1\,n})$, consequently (4.4) becomes:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) \approx \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_{n=1}^N q(\boldsymbol{s}_{t-1\,n}) d\mathbf{s}_{t-1}. \tag{4.22}$$

It is now assumed that the variational posterior distribution $q^\star(\boldsymbol{s}_{t-1\,n})$ is Gaussian with mean $\boldsymbol{\mu}_{t-1\,n}$ and covariance $\boldsymbol{\Gamma}_{t-1\,n}$:

$$q^\star(\boldsymbol{s}_{t-1\,n}) = \mathcal{N}(\boldsymbol{s}_{t-1\,n}; \boldsymbol{\mu}_{t-1\,n}, \boldsymbol{\Gamma}_{t-1\,n}). \tag{4.23}$$

By substituting (4.23) into (4.22) and combining it with (4.8), the predictive distribution (4.22) becomes:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) \approx \prod_{n=1}^N \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1\,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1\,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}). \tag{4.24}$$

Note that the above distribution factorizes across persons. Now that all the factors in (4.1) have tractable expressions, A VEM algorithm can be applied.

4.5 VARIATIONAL EXPECTATION MAXIMIZATION

The proposed VEM algorithm iterates between an E-S-step, an E-Z-step, and an M-step on the following grounds.

§ E-S-step

the per-person variational posterior distribution of the state vector $q^\star(\boldsymbol{s}_{tn})$ is evaluated by developing (4.19). The complete-data likelihood $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$ in (4.19) is the product of (4.2), (4.3) and (4.24). We thus first sum the logarithms of (4.2), of (4.3) and of (4.24). Then we ignore the terms that do not involve $\boldsymbol{s}_{tn}$. Evaluation of the expectation over all the latent variables except $\boldsymbol{s}_{tn}$ yields the following Gaussian distribution:

$$q^\star(\boldsymbol{s}_{tn}) = \mathcal{N}(\boldsymbol{s}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}), \tag{4.25}$$

with:

$$\boldsymbol{\Gamma}_{tn} = \Bigg(\underbrace{\sum_{m=1}^{M_t} \alpha_{tmn} \mathbf{P}_f^\top \boldsymbol{\Phi}_{tm}^{-1} \mathbf{P}_f}_{\#1} + \underbrace{\sum_{k=1}^{K} \sum_{r=1}^{R} \beta_{tknr} \mathbf{P}_g^\top \mathbf{L}_{kr}^\top \boldsymbol{\Sigma}_{kr}^{-1} \mathbf{L}_{kr} \mathbf{P}_g}_{\#2} + \underbrace{\left(\boldsymbol{\Lambda}_{tn} + \mathbf{D} \boldsymbol{\Gamma}_{t-1\,n} \mathbf{D}^\top \right)^{-1}}_{\#3} \Bigg)^{-1}, \tag{4.26}$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn} \Bigg(\sum_{m=1}^{M_t} \alpha_{tmn} \mathbf{P}_f^\top \boldsymbol{\Phi}_{tm}^{-1} \boldsymbol{v}_{tm} + \sum_{k=1}^{K} \sum_{r=1}^{R} \beta_{tknr} \mathbf{P}_g^\top \mathbf{L}_{kr}^\top \boldsymbol{\Sigma}_{kr}^{-1} (\boldsymbol{g}_{kr} - \boldsymbol{l}_{kr}) + \left(\boldsymbol{\Lambda}_{tn} + \mathbf{D} \boldsymbol{\Gamma}_{t-1\,n} \mathbf{D}^\top \right)^{-1} \mathbf{D} \boldsymbol{\mu}_{t-1\,n} \Bigg), \tag{4.27}$$

where $\alpha_{tmn} = q^\star(A_{tm} = n)$ and $\beta_{tknr} = q^\star(B_{tk} = n, C_{tk} = r)$ are computed in the E-Z-step below. A key point is that, because of the recursive nature of the formulas above, it is sufficient to make the Gaussian assumption at $t = 1$, i.e. $q^\star(\boldsymbol{s}_{1n}) = \mathcal{N}(\boldsymbol{s}_{1n}; \boldsymbol{\mu}_{1n}, \boldsymbol{\Gamma}_{1n})$, whose parameters may be easily initialized. It follows that $q^\star(\boldsymbol{s}_{tn})$ is Gaussian at each frame.

We note that both (4.26) and (4.27) are composed of three terms: the first term (#1), second second term (#2) and third term (#3) of (4.26) correspond to the visual, audio, and model dynamics contributions to the precision, respectively. Remind that covariance $\boldsymbol{\Phi}_{tm}$ is associated with the visual observed variable in (4.10). Matrices $\mathbf{L}_{kr}$ and vectors $\boldsymbol{l}_{kr}$ characterize the piecewise affine mappings from the space of person locations to the space of audio features, and covariances $\boldsymbol{\Sigma}_{kr}$ capture the errors that are associated with both audio measurements and the piecewise affine approximation in (4.15). A similar interpretation holds for the three terms of (4.27).

§ E-Z-step

by developing (4.20), and following the same reasoning as above, we obtain the following closed-form expression for the variational posterior distribution of the visual assignment variable:

$$\alpha_{tmn} = q^\star(A_{tm} = n) = \frac{\tau_{tmn} \eta_{mn}}{\sum_{i=0}^{N} \tau_{tmi} \eta_{mi}}, \tag{4.28}$$

where τ_{tmn} is given by:

$$\tau_{tmn} = \begin{cases} \mathcal{N}(\boldsymbol{v}_{tm}; \mathbf{P}_f \boldsymbol{\mu}_{tn}, \boldsymbol{\Phi}_{tm}) e^{-\frac{1}{2}\text{tr}\left(\mathbf{P}_f^\top \boldsymbol{\Phi}_{tm}^{-1} \mathbf{P}_f \boldsymbol{\Gamma}_{tn}\right)} \\ \times \mathcal{B}(\boldsymbol{u}_{tm}; \boldsymbol{h}_n) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{v}_{tm}; \text{vol}(\mathcal{V})) \mathcal{U}(\boldsymbol{u}_{tm}; \text{vol}(\mathcal{H})) & \text{if } n = 0. \end{cases}$$

Similarly, for the variational posterior distribution of the audio assignment variables, developing (4.21) leads to:

$$\beta_{tknr} = q^\star(B_{tk} = n, C_{tk} = r) = \frac{\kappa_{tknr} \rho_{kn} \pi_r}{\sum_{i=0}^{N} \sum_{j=1}^{R} \kappa_{tkij} \rho_{ki} \pi_j}, \tag{4.29}$$

where κ_{tknr} is given by:

$$\kappa_{tknr} = \tag{4.30}$$
$$\begin{cases} \mathcal{N}(\boldsymbol{g}_{tk}; \mathbf{L}_{kr} \mathbf{P}_g \boldsymbol{\mu}_{tn} + \boldsymbol{l}_{kr}, \boldsymbol{\Sigma}_{kr}) e^{-\frac{1}{2}\text{tr}\left(\mathbf{P}_g^\top \mathbf{L}_{kr}^\top \boldsymbol{\Sigma}_{kr}^{-1} \mathbf{L}_{kr} \mathbf{P}_g \boldsymbol{\Gamma}_{tn}\right)} \\ \times \mathcal{N}(\tilde{\boldsymbol{x}}_{tn}; \boldsymbol{\nu}_r, \boldsymbol{\Omega}_r)) & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\boldsymbol{g}_{tk}; \text{vol}(\mathcal{G})) & \text{if } n = 0. \end{cases}$$

To obtain (4.30), an additional approximation is made. Indeed, the logarithm of (4.16) is part of the complete-data log-likelihood and the denominator of (4.16) contains a weighted sum of Gaussian distributions. Taking the expectation of this term is not tractable because of the denominator. Based on the dynamical model (4.8), we replace the state variable $\boldsymbol{x}_{tn}$ in (4.16) with a "naive" estimate $\tilde{\boldsymbol{x}}_{tn}$ predicted from the position and velocity inferred at $t-1$: $\tilde{\boldsymbol{x}}_{tn} = \boldsymbol{x}_{t-1\,n} + \boldsymbol{y}_{t-1\,n}$.

§ M-step

The entries of covariance matrix of the state dynamics, $\boldsymbol{\Lambda}_{tn}$ is estimated in M-step. The obtained expression is the same as (2.26).

4.6 ALGORITHM IMPLEMENTATION

The VEM procedure above will be referred to as VAVIT which stands for *variational audio-visual tracking*, and pseudo-code is shown in Algorithm 3. In theory, the order in which the two expectation steps are executed is not important. In practice, the issue of initialization is crucial. In our case, it is more convenient to start with the E-Z step rather than with the E-S step because the former is easier to initialize than the latter (see below). We start by explaining how the algorithm is initialized at $t = 1$ and then how the E-Z-step is initialized at each iteration. Next, we explain in detail the birth process. An interesting feature of the proposed method is that it allows to estimate who speaks when, or speaker diarization, which is then explained in detail.

Input: visual observations $\mathbf{f}_{1:t} = \{\mathbf{v}_{1:t}, \xi_{1:t}\}$;
audio observations $\mathbf{g}_{1:t}$;
Output: Parameters of $q(\mathbf{s}_{1:t})$: $\{\boldsymbol{\mu}_{1:t,n}, \boldsymbol{\Gamma}_{1:t,n}\}_{n=0}^{N}$ (the estimated position of each person n is given by the two first entries of $\boldsymbol{\mu}_{1:t,n}$);
Person speaking status for $1:t$
Initialization (see Section 4.6.1);
for $t = 1$ *to end* **do**
  Gather visual and audio observations at frame t;
  Perform voice activity detection;
  Initialization of E-Z step (see Section 4.6.1);
  for $iter = 1$ *to* N_{iter} **do**
    E-Z-step (vision):
    for $m \in \{1, ..., M_t\}$ **do**
      for $n \in \{0, ..., N_t\}$ **do**
        Evaluate $q(A_{tm} = n)$ with (4.28);
      end
    end
    E-Z-step (audio):
    for $k \in \{1, ..., K_t\}$ **do**
      for $n \in \{0, ..., N_t\}$ *and* $r \in \{1, ..., R\}$ **do**
        Evaluate $q(B_{tk} = n, C_{tk} = r)$ with (4.30);
      end
    end
    E-S-step:
    for $n \in \{1, ..., N_t\}$ **do**
      Evaluate $\boldsymbol{\Gamma}_{tn}$ and $\boldsymbol{\mu}_{tn}$ with (4.26) and (4.27);
    end
    M-step: Evaluate $\boldsymbol{\Lambda}_{tn}$ with (2.26);
  end
  Perform birth (see Section 4.6.2);
  Output the results;
end

Algorithm 1: Variational audio-visual tracking algorithm.

4.6.1 INITIALIZATION

At $t = 1$ one must provide initial values for the parameters of the distributions (4.25), namely $\boldsymbol{\mu}_{1n}$ and $\boldsymbol{\Gamma}_{1n}$ for all $n \in \{1 \dots N\}$. These parameters are initialized as follows. The means are initialized at the image center and the covariances are given very large values, such that the variational distributions $q(\boldsymbol{s}_{1n})$ are non-informative. Once these parameters are initialized, they remain constant for a few frames, i.e. until the birth process is activated (see Section 4.6.2 below).

As already mentioned, it is preferable to start with the E-Z-step than with the E-S-step

because the initialization of the former is straightforward. Indeed, the E-S-step (Section 4.5) requires current values for the posterior probabilities (4.28) and (4.30) which are estimated during the E-Z-step and which are both difficult to initialize. Conversely, the E-Z-step only requires current mean values, $\boldsymbol{\mu}_{tn}$, which can be easily initialized by using the model dynamics (4.8), namely $\boldsymbol{\mu}_{tn} = \mathbf{D}\boldsymbol{\mu}_{t-1n}$.

4.6.2 BIRTH PROCESS

The speaker birth is We now explain in detail the birth process, which is executed at the start of the tracking to initialize a latent variable for each detected person, as well as at any time t to detect new persons. The birth process considers B consecutive visual frames. At t, with $t > B$, we consider the set visual observations assigned to $n = 0$ from $t - B$ to t, namely observations whose posteriors (4.28) are maximized for $n = 0$ (at initialization all the observations are in this case). We then build observation sequences from this set, namely sequences of the form $(\tilde{\boldsymbol{v}}_{m_{t-B}}, \ldots, \tilde{\boldsymbol{v}}_{m_t})_{\tilde{n}} \in \mathcal{B}$, where m_t indexes the set of observations at t assigned to $n = 0$ and $\tilde{n}$ indexes the set $\mathcal{B}$ of all such sequences. Notice that the birth process only uses the bounding-box center, width and size, $\boldsymbol{v}$, and that the descriptor $\boldsymbol{u}$ is not used. Hence the birth process is only based on the smoothness of an observed sequence of bounding boxes. Let's consider the marginal likelihood of a sequence $\tilde{n}$, namely:

$$\begin{aligned}\mathcal{L}_{\tilde{n}} &= p((\tilde{\boldsymbol{v}}_{m_{t-B}}, \ldots, \tilde{\boldsymbol{v}}_{m_t})_{\tilde{n}}) \\ &= \int \ldots \int p(\tilde{\boldsymbol{v}}_{m_{t-B}} | \boldsymbol{s}_{t-B\,\tilde{n}}) \ldots p(\tilde{\boldsymbol{v}}_{m_t} | \boldsymbol{s}_{t\,\tilde{n}}) \\ &\times p(\boldsymbol{s}_{t\,\tilde{n}} | \boldsymbol{s}_{t-1\,\tilde{n}}) \ldots p(\boldsymbol{s}_{t-B+1\,\tilde{n}} | \boldsymbol{s}_{t-B\,\tilde{n}}) p(\boldsymbol{s}_{t-B\,\tilde{n}}) d\boldsymbol{s}_{t-B:t\,\tilde{n}},\end{aligned} \tag{4.31}$$

where $\boldsymbol{s}_{t,\tilde{n}}$ is the latent variable already defined and $\tilde{n}$ indexes the set $\mathcal{B}$. All the probability distributions in (4.31) were already defined, namely (4.8) and (4.10), with the exception of $p(\boldsymbol{s}_{t-B,\tilde{n}})$. Without loss of generality, we can assume that the latter is a normal distribution centered at $\tilde{\boldsymbol{v}}_{m_t}$ and with a large covariance. Therefore, the evaluation of (4.31) yields a closed-form expression for $\mathcal{L}_{\tilde{n}}$. A sequence $\tilde{n}$ generated by a person is likely to be smooth and hence $\mathcal{L}_{\tilde{n}}$ is high, while for a non-smooth sequence the marginal likelihood is low. A newborn person is therefore created from a sequence of observations $\tilde{n}$ if $\mathcal{L}_{\tilde{n}} > \tau$, where τ is a user-defined parameter. As just mentioned, the birth process is executed to initialize persons as well as along time to add new persons. In practice, in (4.31) we set B=3 and hence, from t=1 to t=4 all the observations are initially assigned to $n = 0$.

4.6.3 SPEAKER DIARIZATION

Speaker diarization consists of assigning temporal segment of speech to persons [3]. We introduce a binary variable χ_{tn} such that $\chi_{tn} = 1$ if person n speaks at time t and $\chi_{tn} = 0$ otherwise. Traditionally, speaker diarization is based on the following assumptions. First, it is assumed that speech signals are sparse in the time-frequency domain. Second, it is assumed that each time-frequency point in such a spectrogram corresponds to a single

speech source. Therefore, the proposed speaker diarization method is based on assigning time-frequency points to persons.

In the case of the proposed model, speaker diarization can be coarsely inferred from frequency sub-bands in the following way. The posterior probability that the speech signal available in the frequency sub-band k at frame t was uttered by person n, given the audio observation $\boldsymbol{g}_{tk}$, is:

$$p(B_{tk} = n|\boldsymbol{g}_{tk}) = \sum_{r=1}^{R} p(B_{tk} = n, C_{tk} = r|\boldsymbol{g}_{tk}), \tag{4.32}$$

where B_{tk} is the audio assignment variable and C_{tk} is the affine-mapping assignment variable defined in Section 4.3.4. Using the variational approximation (4.29), this probability becomes:

$$p(B_{tk} = n|\boldsymbol{g}_{tk}) \approx \sum_{r=1}^{R} q^{\star}(B_{tk} = n, C_{tk} = r) = \sum_{r=1}^{R} \beta_{tknr}, \tag{4.33}$$

and by accumulating probabilities over all the frequency sub-bands, we obtain the following:

$$\chi_{tn} = \begin{cases} 1 & \text{if } \frac{1}{K_t} \sum_{k=1}^{K_t} \sum_{r=1}^{R} \beta_{tknr} \geq \gamma \\ 0 & \text{otherwise,} \end{cases} \tag{4.34}$$

where γ is a user-defined threshold. Note that there is no dynamic model associated with diarization: χ_{tn} is estimated independently at each frame and for each person. More sophisticated diarization models can be found in [117, 53].

4.7 EXPERIMENTS

4.7.1 DATASET

We use the AVDIAR dataset [53] to evaluate the performance of the proposed audio-visual tracking method. This dataset is challenging in terms of audio-visual analysis. There are several participants involved in informal conversations while wandering around. They are in between two and four meters away from the audio-visual recording device. They take speech turns and often there are speech overlaps. They turn their faces away from the camera. The dataset is annotated as follows:[4] The visual annotations comprise the centers, widths and heights of two bounding boxes for each person and in each video frame, a face bounding box and an upper-body bounding box. An identity (a number) is associated with each person through the entire dataset. The audio annotations comprise the speech status of each person over time (speaking or silent), with a minimum speech

[4]Please consult `https://team.inria.fr/perception/avdiar/` for a detailed description of the dataset.

duration of 0.2 seconds. The audio source locations correspond to the centers of the face bounding boxes.

The dataset was recorded with a sensor composed of two cameras and six microphones, but only one camera is used in the experiments described below. The videos were recorded at 25 FPS. The frame resolution is of 1920×1200 pixels corresponding to a field of view of $97^\circ \times 80^\circ$. The microphone signals are sampled at 16000 Hz. The dataset was recorded into two different rooms, *living-room* and *meeting-room*, e.g. Fig. 4.1 and Fig. 4.2. These two rooms have quite different lighting conditions and acoustic properties (size, presence of furniture, background noise, etc.). Altogether there are 18 sequences associated with living-room (26928 video frames) and 6 sequences with meeting-room (6031 video frames). Additionally, there are two training datasets, $\mathcal{T}_1$ and $\mathcal{T}_2$ (one for each room) that contain input-output pairs of multichannel audio features and audio-source locations that allow to estimate the parameters (4.14) using the method of [35]. This yields a mapping between source locations in the image plane, $\boldsymbol{x}$, and audio features, $\boldsymbol{g}$. Audio feature extraction is described in detail below.

4.7.2 AUDIO FEATURES

The STFT (short-time Fourier transform) [56] is applied to each microphone signal using a 16 ms Hann window (256 audio samples per window) and with an 8 ms shift (50% overlap), leading to 128 frequency bins and to 125 audio FPS. Inter-channel features are then computed using [79]. These features – referred to as *direct-path relative transfer function (DP-RTF) features* – are robust both against background noise and reverberation, hence they do not depend on the room acoustic properties as they encode the direct path from the audio source to the microphones. The audio features are averaged over five audio frames in order to properly align them with the video frames. The feature vector is then split into $K = 16$ sub-bands, each sub-band being composed of $L = 8$ frequencies; sub-bands with low energy are disregarded. This yields the set of audio observations at t, $\{\boldsymbol{g}_{tk}\}_{k=1}^{K_t}$, $K_t \leq K$ (see Section 4.3.4).

4.7.3 VISUAL PROCESSING

Because in AVDIAR people do not necessarily face the camera, face detection is not very robust. Instead we use a body-pose detector [24] from which we infer a full-body bounding-box and a head bounding-box. We use the person re-identification CNN-based method [163] to extract en embedding from the full-body bounding-box. This yields the features vectors $\{\boldsymbol{u}_{tm}\}_{m=1}^{M_t} \subset \mathbb{R}^{2048}$ (Section 4.3.3). Similarly, the center, width and height of the head bounding-box yield the observations $\{\boldsymbol{v}_{tm}\}_{m=1}^{M_t} \subset \mathbb{R}^4$ at each frame t.

4.7.4 EXPERIMENTAL SETTINGS

One interesting feature of the proposed tracking is its flexibility in dealing with visual data, audio data or visual and audio data. Moreover, the algorithm is able to automatically

switch from unimodal to multimodal. In order to quantitatively assess the performance and merits of each one of these variants we used two configurations:

- *Full camera field of view (FFOV)*: The entire horizontal field of view of the camera, i.e. 1920 pixels, or 97°, is being used, such that visual and audio observations, **if any**, are simultaneously available, and
- *Partial camera field of view (PFOV)*: The horizontal field of view is restricted to 768 pixels (or 49°) and there are two *blind* strips (576 pixels each) on its left- and right-hand sides; the *audio field of view* remains unchanged, 1920 pixels, or 97°.

The PFOV configuration allows us to test scenarios in which a participant may leave the camera field of view and still be heard. Notice that since ground-truth annotations are available for the full field of view, it is possible to assess the performance of the tracker using audio observations only, as well as to analyse the behavior of the tracker when it switches from audio-only tracking to audio-visual tracking.

4.7.5 Evaluation Metrics

We used standard multi-object tracking (MOT) metrics to quantitatively evaluate the performance of the proposed tracking algorithm. The multi-object tracking accuracy (MOTA) is the most commonly used metrics for MOT. It is a combination of false positives (FP), false negatives (FN; aka missed track), and identity switches (IDs), and is defined as:

$$\text{MOTA} = 100\left(1 - \frac{\sum_t(\text{FP}_t + \text{FN}_t + \text{IDs}_t)}{\sum_t \text{GT}_t}\right), \tag{4.35}$$

where GT stands for the ground-truth person trajectories, as annotated in the AVDIAR dataset. After comparison with GT trajectories, each estimated trajectory can be classified as mostly tracked (MT), partially tracked (PT) and mostly lost (ML). If a trajectory is covered by a correct estimation at least 80% of the time, it is considered as MT. Similarly, it is considered as ML if it is covered less than 20%. In our experiments, MT and ML scores represent the percentage of trajectories which are considered as mostly tracked and mostly lost respectively. In addition, the number of track fragmentations (FM) counts how many times the estimated trajectories are discontinuous (whereas the corresponding GT trajectories are continuous).

In our experiments, the threshold of overlap to consider that a ground truth is covered by an estimation is set to 0.1. In the PFOV configuration, we need to evaluate the audio-only tracking, i.e. the speakers are in the blind areas. As mentioned before, audio localization is less accurate than visual localization. Therefore, for evaluating the audio-only tracker we relax by a factor of two the expected localization accuracy with respect to the audio-visual localization accuracy.

Table 4.1: MOT scores for the living-room sequences (full camera field of view)

Method	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	FM(↓)	MT(↑)	ML(↓)
AS-VA-PF [68]	10.37	44.64 %	43.95%	732	918	20%	7.5 %
AV-MSSMC-PHD [67]	18.96	8.13 %	72.09%	581	486	17.5%	52.5%
OBVT [7]	96.32	1.77%	1.79%	80	131	92.5%	0%
VAVIT (proposed)	96.03	1.85%	2.0%	86	152	92.5%	0%

Table 4.2: MOT scores for the meeting-room sequences (full camera field of view)

Method	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	FM(↓)	MT(↑)	ML(↓)
AS-VA-PF [68]	62.43	18.63%	17.19%	297	212	70.59 %	0%
AV-MSSMC-PHD [67]	28.48	0.93%	69.68%	155	60	0 %	52.94%
OBVT [7]	98.50	0.25%	1.11%	25	10	100.00%	0%
VAVIT (proposed)	98.16	0.38%	1.27%	32	15	100.00%	0%

Table 4.3: MOT scores for the living-room sequences (partial camera field of view)

Method	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	FM(↓)	MT(↑)	ML(↓)
AS-VA-PF [68]	17.82	36.86%	42.88%	1722	547	32.50%	7.5%
AV-MSSMC-PHD [67]	20.61	5.54%	72.45%	989	471	12.5%	40%
OBVT [7]	66.39	0.48%	32.95%	129	203	45%	7.5%
VAVIT (proposed)	69.62	8.97%	21.18%	152	195	70%	5%

Table 4.4: MOT scores for the meeting-room sequences (partial camera field of view)

Method	MOTA(↑)	FP(↓)	FN(↓)	IDs(↓)	FM(↓)	MT(↑)	ML(↓)
AS-VA-PF [68]	29.04	23.05%	45.19 %	461	246	29.41%	17.65%
AV-MSSMC-PHD [67]	26.95	1.05%	70.62%	234	64	5.88%	52.94%
OBVT [7]	64.24	0.43%	35.18%	24	25	36.84%	15.79%
VAVIT (proposed)	65.27	5.07%	29.5%	26	26	47.37%	10.53%

4.7.6 BENCHMARKING WITH BASELINE METHODS

To quantitatively evaluate its performance, we benchmarked the proposed method with two state-of-the-art audio-visual tracking methods. The first one is the audio-assisted video adaptive particle filtering (AS-VA-PF) method of [68], and the second one is the sparse audio-visual mean-shift sequential Monte-Carlo probability hypothesis density (AV-MSSMC-PHD) method of [67]. [68] takes as input a video and a sequence of sound locations. Sound locations are used to reshape the typical Gaussian noise distribution of particles in a propagation step, then uses the particles to weight the observation model. [67] uses audio information to improve the performance and robustness of a visual SMC-PHD filter. Both methods show good performance in meeting configurations, e.g. the AV16.3 dataset [73]: the recordings used a circular microphone array placed on a table and located at the center of the room, as well as several cameras fixed on the ceiling. The scenarios associated with AV16.3 are somehow artificial in the sense that *the participants speak simultaneously and continuously*. This stays in contrast with the AVDIAR recordings where people take speech turns in informal conversations.

Since both [68] and [67] require input from a multiple sound-source localization (SSL) algorithm, the multi-speaker localization method proposed in [79] is used to provide input

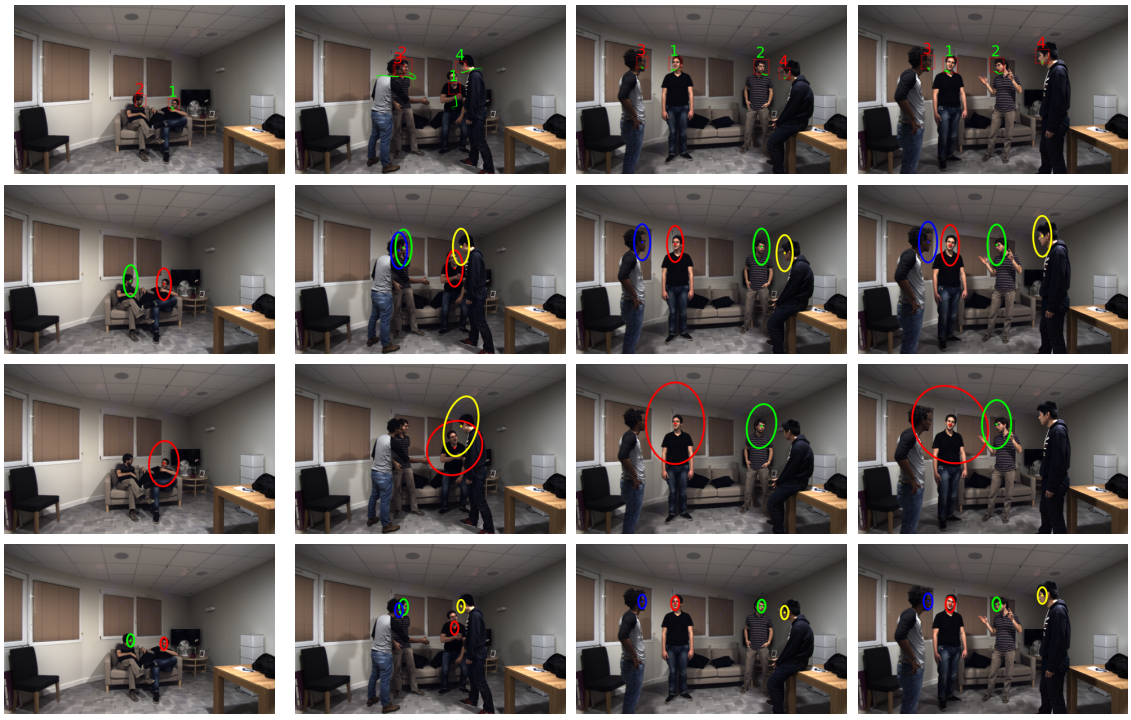


Figure 4.1: Four frames sampled from Seq13-4P-S2M1. First row: green digits denote speakers while red digits denote silent participants. Second, third and fourth rows: visual, audio, and dynamic contours of constant densities (covariances), respectively, of each tracked person. The tracked persons are color-coded: green, yellow, blue, and red.

to [68] and [67].[5] We also compare the proposed method with a visual multiple-person tracker, more specifically the *online Bayesian variational tracker* (OBVT) of [7], which is based on a similar variational inference as the one presented in this chapter. In [7] visual observations were provided by color histograms. In our benchmark, for the sake of fairness, the proposed tracker and [7] share the same visual observations (Section 4.7.3).

The MOT scores obtained with these methods as well as the proposed method are reported in Table 4.1, Table 4.2, Table 4.3 and Table 4.4. The symbols $\uparrow$ and $\downarrow$ indicate higher the better and lower the better, respectively. The tables report results obtained with the meeting-room and living-room sequences and for the two configurations mentioned above: full and partial camera fields of view, respectively. The most informative metric is MOTA (MOT accuracy) and one can easily see that both [7] and the proposed method outperform the other two methods. The poorer performance of both [68] and [67] for all the configurations is generally explained by the fact that these two methods assume that audio and visual observations are simultaneously available. In particular, [68] is not robust against visual occlusions, which leads to poor IDs (identity switches) scores.

The AV-MSSMC-PHD method [67] uses audio information in order to count the num-

[5]The authors of [68] and [67] kindly provided their software packages.

Figure 4.2: Four frames sampled from Seq19-2P-S1M1. The camera field of view is limited to the central strip. Whenever the participants are outside the central strip, the tracker entirely relies on audio observations and on the model's dynamics.

ber of speakers. The algorithm detects multiple speakers whenever multiple audio sources are detected. In practice, the algorithm rarely finds multiple speakers and in most of the cases it only tracks one speaker. This explains why both FN (false negatives) and IDs (identity switches) scores are high, i.e. Tables 4.1, 4.2, and 4.3.

One can notice that in the case of FFOV, [7] and the proposed method yield similar results in terms of MOT scores: they both exhibit low FP, FN and IDs scores and, consequently, high MOTA scores. Moreover, they have very good MT, PT and ML scores (out of 40 sequences 37 are mostly tracked, 3 are partially tracked, and none is mostly lost). As expected, the inferred trajectories are more accurate for visual tracking (whenever visual observations are available) than for audio-visual tracking: indeed, the latter fuses visual and audio observations which slightly degrades the accuracy because audio localization is less accurate than visual localization.

As for the PFOV configuration (Table 4.3 and Table 4.4), the proposed algorithm yields the best MOTA scores both for the meeting and for the living rooms. Both [68] and [67] have difficulties when visual information is not available, e.g. the left- and right-hand blind strips on both sides of the restricted field of view: both these algorithms fail to track speakers when they walk outside the visual field of view. While [67] is able to detect a speaker when it re-enters the visual field of view, [68] is not. Obviously, the tracking algorithm of [7] fails in the absence of visual observations.

4.7.7 AUDIO-VISUAL TRACKING EXAMPLES

We now provide and discuss results obtained with three recordings, one FFOV sequence, Seq13-4P-S2-M1 (Fig. 4.1) and two PFOV sequences, Seq19-2P-S1M1 (Fig. 4.2) and Seq22-1P-S0M1 (Fig. 4.3).[6] These sequences are challenging in terms of audio-visual tracking: participants are seated, then they stand up or they wander around. Some participants take speech turns and interrupt each other, while other participants remain silent.

The first row of Fig. 4.1 shows four frames sampled from a video recording with two then four participants, labeled 1, 2, 3, and 4. Green digits designate participants detected as speakers and red digits correspond to participants detected as listeners. The second row shows ellipses of constant density (visual covariances), i.e. the inverse of the precision #1 in (5.36). Notice that in the second frame the detection of person 3, who turns his back to the camera, was missed. The third row shows the audio covariances, i.e. the inverse of the precision #2 in (5.36). The audio covariances are much larger than the visual ones since audio localization is less accurate than visual localization. There are two distinct audio sources close to each other that are correctly detected, localized and assigned to persons 1 and 4 and therefore it is still possible to assign audio activities to both 1 and 4. The fourth row shows the contribution of the dynamic model to the covariance, i.e. the inverse of the precision #3 in (5.36). Notice that these "dynamic" covariances are small, in comparison with the "observation" covariances, which reflects a smooth trajectory and ensures tracking continuity when audio or visual observations are either weak or totally absent. Fig. 4.2 shows a tracking example with a PFOV (partial camera field of view) configuration. In this case, audio and visual observations are barely available simultaneously. The independence of the visual and audio observation models and their fusion within the same dynamic model guarantees robust tracking results.

Fig. 4.3 shows the ground-truth trajectory of a person and the trajectories estimated with the audio-visual tracker [68], with the visual tracker [7], and with the proposed method. The ground-truth trajectory corresponds to a sequence of bounding-box centers. Both [68] and [7] failed to estimate a correct trajectory. Indeed, [68] requires simultaneous availability of audio-visual data while [7] cannot track outside the visual field of view. Notice the dangled trajectory obtained with [68] in comparison with the smooth trajectories obtained with variational inference, i.e. [7] and proposed.

4.7.8 SPEAKER DIARIZATION RESULTS

As already mentioned in Section 4.6.3, speaker diarization information can be extracted from the output of the proposed VAVIT algorithm. Notice that, while audio diarization is an extremely well investigated topic, audio-visual diarization has received much less attention. In [117] it is proposed an audio-visual diarization method based on a dynamic Bayesian network that is applied to video conferencing. The method assumes that participants take speech turns, which is an unrealistic hypothesis in the general case. The diarization method of [113] requires audio, depth and RGB data. More recently, [53]

[6] https://team.inria.fr/perception/research/variational_av_tracking/

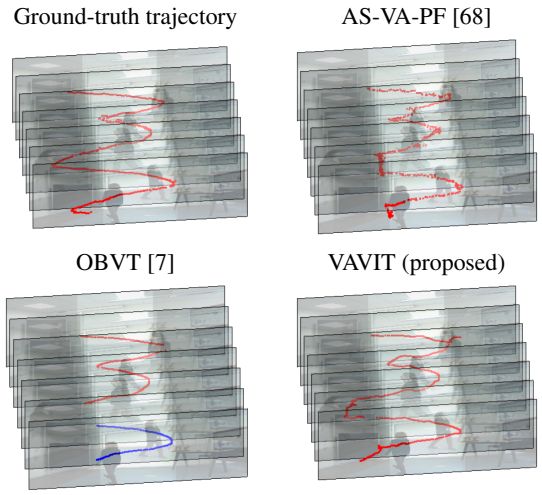


Figure 4.3: Trajectories associated with a tracked person under the PFOV configuration. The ground-truth trajectory corresponds to the center of the bounding-box of the head. The trajectory of [68] dangles. Both [68] and [7] fail to track outside the camera field of view. In the case of OBVT, there is an identity switch, from "red" (before the person leaves the visual field of view) to "blue" (after the person re-enters in the visual field of view).

proposed a Bayesian dynamic model for audio-visual diarization that takes as input fused audio-visual information. Since diarization is not the main objective of this work, we only compared our diarization results with [53], which achieves state of the art results, and with the diarization toolkit of [147] which only considers audio information.

The diarization error rate (DER) is generally used as a quantitative measure. As for MOT, DER combines FP, FN and IDs scores. The NIST-RT evaluation toolbox[7] is used. The results obtained with these two methods and with ours are reported in Table 4.5, with both the full field-of-view and partial field-of-view configurations (FFOV and PFOV). The proposed method performs better than the audio-only baseline method [147]. In comparison with [53], the proposed method performs slightly less well despite the lack of a diarization dynamic model. Indeed, [53] estimates diarization within a temporal model that takes into account both diarization dynamics and audio activity at each time step, whereas our method is only based on audio activity at each time step.

The ability of the proposed audio-visual tracker to perform diarization is illustrated with the FFOV sequence Seq13-4P-S2-M1 (Fig. 4.1) and with the PFOV sequence Seq19-2P-S1M1 (Fig. 4.2), e.g. Fig. 4.4 and Fig. 4.5, respectively.

[7] `https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation`

Table 4.5: DER (diarization error rate) scores obtained with the AVDIAR dataset.

Sequence	DiarTK [147]	[53]	Proposed (FFOV)	Proposed (PFOV)
Seq01-1P-S0M1	43.19	3.32	1.64	1.86
Seq02-1P-S0M1	49.9	-	2.38	2.09
Seq03-1P-S0M1	47.25	-	6.59	14.65
Seq04-1P-S0M1	32.62	9.44	4.96	10.45
Seq05-2P-S1M0	37.76	-	29.76	30.78
Seq06-2P-S1M0	56.12	-	14.72	15.83
Seq07-2P-S1M0	41.43	-	42.36	37.56
Seq08-3P-S1M1	31.5	-	38.4	48.86
Seq09-3P-S1M1	52.74	-	38.26	68.81
Seq10-3P-S1M1	56.95	-	54.26	54.04
Seq12-3P-S1M1	63.67	17.32	44.67	47.25
Seq13-4P-S2M1	47.56	29.62	43.45	43.17
Seq15-4P-S2M1	62.53	-	41.49	64.38
Seq17-2P-S1M1	17.24	-	16.53	15.63
Seq18-2P-S1M1	35.05	-	19.55	20.58
Seq19-2P-S1M1	38.96	-	26.47	27.84
Seq20-2P-S1M1	43.58	35.46	38.24	44.3
Seq21-2P-S1M1	32.22	20.93	25.87	25.9
Seq22-1P-S0M1	23.53	4.93	2.79	3.32
Seq27-3P-S1M1	46.05	18.72	47.07	54.75
Seq28-3P-S1M1	30.68	-	23.54	31.77
Seq29-3P-S1M0	38.68	-	30.74	35.92
Seq30-3P-S1M1	51.15	-	49.71	57.94
Seq32-4P-S1M1	41.51	30.20	46.25	43.03
Overall	42.58	18.88	28.73	33.36

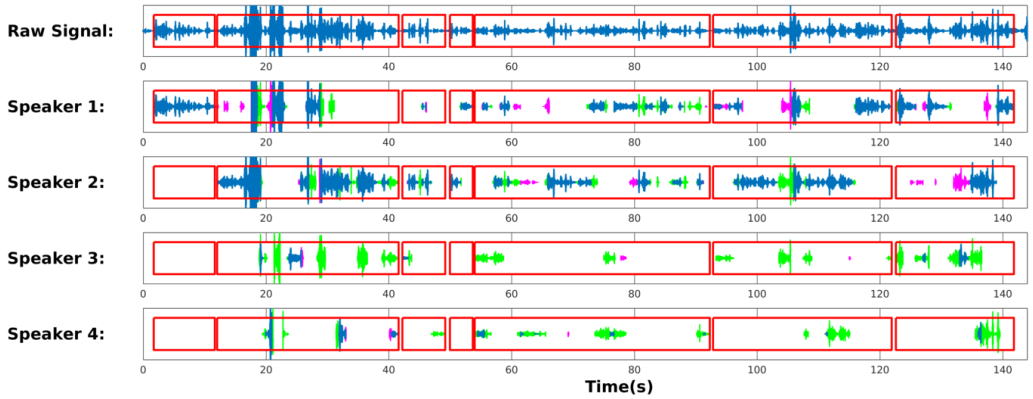


Figure 4.4: Diarization results obtained with Seq13-4P-S2M1 (FFOV). The first row shows the audio signal recorded with one of the microphones. The red boxes show the result of the voice activity detector which is applied to all the microphone signals prior to tracking. For each speaker, correct detections are shown in blue, missed detections are shown in green, and false positives are shown in magenta

4.8 Conclusions

We addressed the problem of tracking multiple speakers using audio and visual data. It is well known that the generalization of single-person tracking to multiple-person tracking is computationally intractable and a number of methods were proposed in the past. Among

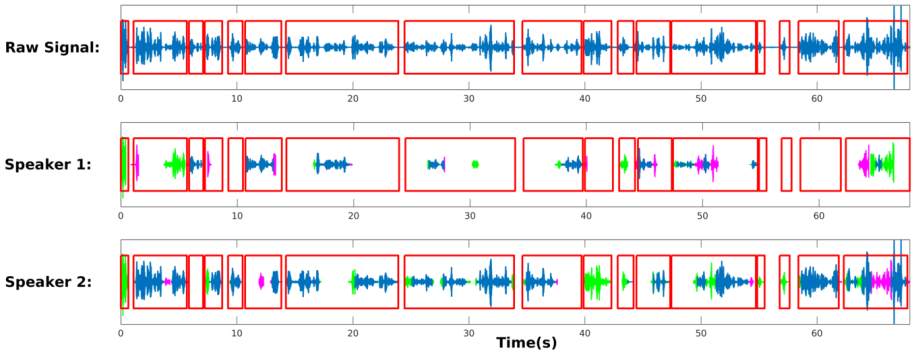


Figure 4.5: Diarization results obtained with Seq19-2P-S1M1 (PFOV).

these methods, sampling methods based on particle filtering (PF) or on PHD filters have recently achieved the best tracking results. However, these methods have several drawbacks: (i) the quality of the approximation of the filtering distribution increases with the number of particles, which also increases the computational burden, (ii) the observation-to-person association problem is not explicitly modeled and a post-processing association mechanism must be invoked, and (iii) audio and visual observations must be available simultaneously and continuously. Some of these limitations were recently addressed both in [68] and in [67], where audio observations were used to compensate the temporal absence of visual observations. Nevertheless, people speak with pauses and hence audio observations are rarely continuously available.

In contrast, we proposed a variational approximation of the filtering distribution and we derived a closed-form variational expectation-maximization algorithm. The observation-to-person association problem is fully integrated in our model, rather than as a post-processing stage. The proposed VAVIT algorithm is able to deal with intermittent audio or visual observations, such that one modality can compensate the other modality when one of them is missing, is noisy or is too weak. Using the MOT scores we show that the proposed method performs better than the baseline PF-based method [68].

CHAPTER 5

ACOUSTIC LOCALIZATION AND TRACKING OF MULTIPLE SPEAKERS

5.1 INTRODUCTION

The localization and tracking of multiple speakers in real world environments are very challenging tasks, in particular in the presence of reverberation and ambient noise and of natural conversations, e.g. short sentences, speech pauses and frequent speech turns among speakers. Methods based on time differences of arrival (TDOAs) between microphones, such as generalized cross-correlation [70], are typically used for single-speaker localization, e.g.[27]. In the case of multiple speakers, beamforming-based methods, e.g. steered-response power (SRP) [37], and subspace methods, e.g. multiple signal classification (MUSIC) [64], are widely used. The W-disjoint orthogonality (WDO) principle [161] assumes that the audio signal is dominated by a single audio source in small regions of the time-frequency (TF) domain. This assumption is particularly valid in the case of speech signals. Applying the short-time Fourier transform (STFT), or any other TF representation, inter-channel localization features, such as the interaural phase differences (IPDs) [161], can be extracted. In [161], multiple-speaker localization is based on the histogram of inter-channel features, which is suitable only in the case where there is no wrapping of phase measures. In [104], a Gaussian mixture model (GMM) is used as a generative model of the inter-channel features of multiple speakers, with each GMM representing one speaker, and each GMM component representing one candidate inter-channel time delay. An expectation maximization (EM) algorithm iteratively estimates the component weights and assigns the features to their corresponding candidate time delays. This method overcomes the phase ambiguity problem by jointly considering all frequencies in the likelihood maximization procedure. After maximizing the likelihood, the azimuth of each speaker is given by the component that has the highest weight in the corresponding GMM. The complex-valued version of IPD, i.e. the pair-wise relative phase ratio (PRP), is used in [41]. Instead of setting one GMM for each speaker, a single complex Gaussian mixture model (CGMM) is used for all speakers with each component

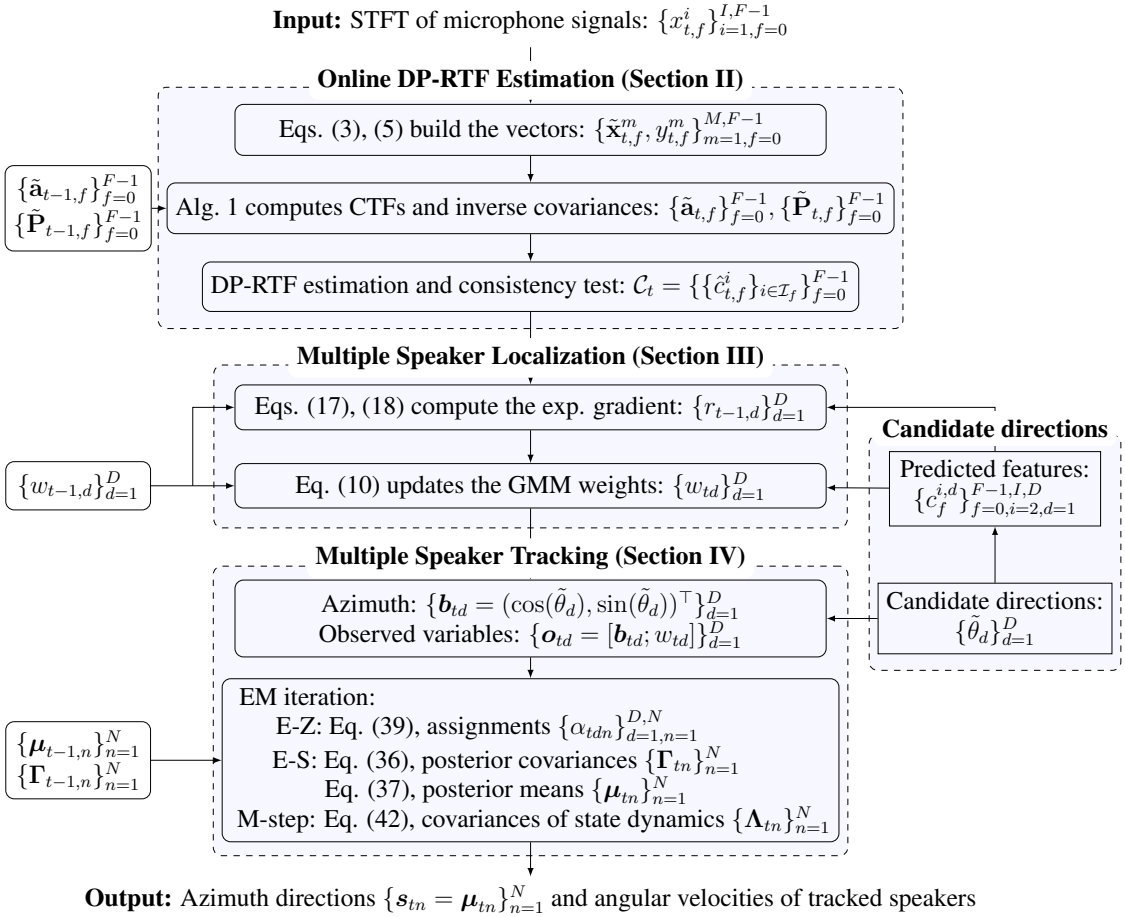


Figure 5.1: Flowchart of the proposed multiple-speaker localization and tracking methodology.

representing one candidate speaker location. After maximizing the likelihood of the PRP features, with an EM algorithm, the weight of each component represents the probability that there is an active speaker at the corresponding candidate location. Therefore, for an unknown number of speakers, counting and localization of active speakers can be jointly carried out by selecting components with large weights.

The inter-channel features and associated localization methods mentioned above assume a direct-path propagation model: hence, they perform poorly in reverberant environments. To overcome this limitation, several TDOA estimators based on system identification were proposed in [62, 39, 42, 71]. In [78] it is proposed to use the DP-RTF as a TF-domain inter-channel localization feature robust against reverberation. The estimation of the DP-RTF is based on the identification of the room impulse response (RIR) in the STFT-domain, i.e. the convolutive transfer function (CTF) [5, 141]. Overall, the method of [78] combines the merits of robust TDOA estimators [62, 39, 42, 71] and of the WDO assumption mentioned above.

To localize moving speakers, one-stage methods such as SRP and MUSIC can be directly used using frame-wise spatial spectrum estimators. In contrast, methods based on inter-channel features require to assign frame-wise features to speakers in an adaptive/recursive way, e.g. the smoothed histogram method of [119]. Similar to [41], [131] uses one CGMM for each predefined speaker; the model is plugged into a recursive EM (REM) algorithm in order to update the mixture's weights.

Speaker tracking methods are generally based on Bayesian inference which combines localization with dynamic models in order to estimate the posterior probability distribution of audio-source directions, e.g. [127, 45, 13]. Kalman filtering and particle filtering were used in [82] and in [146], respectively, for tracking a single audio source. In order to address the problem of multiple speakers, possibly with unknown and time-varying number of speakers, additional discrete latent variables are needed, i.e. observation-to-speaker assignments, as well as speaker-birth and -death processes, e.g. [7], [52]. Sampling-based methods were widely used, e.g. extended particle filtering [48, 143, 25], or sequential Monte Carlo implementation of the probability hypothesis density (PHD) filter [152, 93]. However, the computational burden of sampling-based methods can be prohibitive in practice. Under some assumptions, the multiple-target tracking GMM-PHD filter of [148] has an analytical solution and is computationally efficient: it was adopted for multiple-speaker tracking in [45].

In this chapter we propose a method for the simultaneous localization and tracking of multiple moving speakers (please refer to Figure 5.1 for a method overview). We the following original contributions:

- Since we deal with moving speakers or, more generally, with moving audio sources, DP-RTF features are computed using the *online* CTF estimation framework presented in [81], based on recursive least squares (RLS), rather than using the *batch* CTF estimation of [78] which assumes static audio sources. The online RLS algorithm has a faster convergence rate than the least mean squares (LMS) algorithms described in [62, 39], which is important when dealing with moving sources.

- A crucial ingredient of multiple speaker localization is to properly assign acoustic features, i.e. DP-RTFs, to audio-source directions. We adopt the maximum-likelihood formulation of [41]. We propose to use exponentiated gradient (EG) [69] to update the source directions from their current estimated values. The EG-based recursive estimator proposed below is better suited for moving sources/speakers than the batch estimator proposed in [78].

- The problem of multiple speaker tracking is computationally intractable because the number of possible associations between acoustic features and sources/speakers grows exponentially with time. In this chapter we adopt a Bayesian variational approximation of the posterior filtering distribution which leads to an efficient VEM algorithm. In order to deal with a varying number of speakers, we propose a birth process which allows to initialize new speakers at any time.

In this chapter, the proposed online localization method is an extended version of [81] which has proposed an online DP-RTF method that has been combined with REM to estimate the source directions. In this chapter, while we keep the DP-RTF method of [81] we propose to use EG. The advantages of using EG instead of REM are described in detail in Section 5.3.

The chapter is organized as follows (please refer to Figure 5.1). Section 5.2 presents the online DP-RTF estimation method. Section 5.3 describes the EG-based speaker localization method and Section 5.4 describes the variational approximation of the tracker and the associated VEM algorithm. Section 5.6.5 presents an empirical evaluation of the method based on experiments performed with real audio recordings. Section 5.7 concludes the paper. Supplemental materials are available on our website.[1]

5.2 RECURSIVE MULTICHANNEL DP-RTF ESTIMATION

5.2.1 RECURSIVE LEAST SQUARES

To accord with the acoustic signal processing literature, the notations in this Chapter are slightly different from the ones in the previous Chapters. For the sake of clarity, we first consider the noise-free single-speaker case. In the time domain $x^i(\tau) = a^i(\tau) \star s(\tau)$ is the i-th microphone signal, $i = 1, \ldots, I$, where τ is the time index, $s(\tau)$ is the source signal, $a^i(\tau)$ is the RIR from the source to the i-th microphone, and $\star$ denotes the convolution. Applying the STFT and using the CTF approximation, for each frequency index $f = 0, \ldots, F-1$ we have:

$$x^i_{t,f} = a^i_{t,f} \star s_{t,f} = \sum_{q=0}^{Q-1} a^i_{q,f} s_{t-q,f}, \tag{5.1}$$

where $x^i_{t,f}$ and $s_{t,f}$ are the STFT coefficients of the corresponding signals, and the CTF $a^i_{t,f}$ is a sub-band representation of $a^i(\tau)$. Here, the convolution is executed with respect to the frame index t. The number of CTF coefficients Q is related to the reverberation time of the RIR. The first CTF coefficient $a^i_{0,f}$ mainly consists of the direct-path information, thence the DP-RTF is defined as the ratio between the first CTF coefficients of two channels: $a^i_{0,f}/a^r_{0,f}$, where channel r is the reference channel.

Based on the cross-relation method [157], using the CTF model of one microphone pair (i, j), we have: $x^i_{t,f} \star a^j_{t,f} = x^j_{t,f} \star a^i_{t,f}$. This can be written in vector form as:

$$\mathbf{x}^{i\top}_{t,f} \mathbf{a}^j_f = \mathbf{x}^{j\top}_{t,f} \mathbf{a}^i_f, \tag{5.2}$$

with $\mathbf{a}^i_f = (a^i_{0,f}, \ldots, a^i_{Q-1,f})^\top$, where $^\top$ denotes matrix/vector transpose, and $\mathbf{x}^i_{t,f} = (x^i_{t,f}, \ldots, x^i_{t-Q+1,f})^\top$. The CTF vector involving all channels is defined as $\mathbf{a}_f = (\mathbf{a}^{1\top}_f, \ldots, \mathbf{a}^{I\top}_f)^\top$. There is a total of $I(I-1)/2$ distinct microphone pairs, indexed by (i, j) with $i =$

[1] https://team.inria.fr/perception/research/multi-speaker-tracking/

$1, \ldots, I-1$ and $j = i+1, \ldots, I$. For each pair, we construct a cross-relation equation in terms of $\mathbf{a}_f$. For this aim, we define:

$$\mathbf{x}_{t,f}^{ij} = [\underbrace{0, \ldots, 0}_{(i-1)Q}, \mathbf{x}_{t,f}^{j\top}, \underbrace{0, \ldots, 0}_{(j-i-1)Q}, -\mathbf{x}_{t,f}^{i\top}, \underbrace{0, \ldots, 0}_{(I-j)Q}]^\top. \tag{5.3}$$

Then, for each pair (i, j), we have:

$$\mathbf{x}_{t,f}^{ij\top} \mathbf{a}_f = 0. \tag{5.4}$$

Let's assume, for simplicity, that the reference channel is $r = 1$. To avoid the trivial solution $\mathbf{a}_f = \mathbf{0}$ of (5.4), we constrain the first CTF coefficient of the reference channel to be equal to 1. This is done by dividing both sides of (5.4) by $a_{0,f}^1$ and by moving the first entry of $\mathbf{x}_{t,f}^{ij}$, denoted by $-y_{t,f}^{ij}$, to the right side of (5.4), which rewrites as:

$$\tilde{\mathbf{x}}_{t,f}^{ij\top} \tilde{\mathbf{a}}_f = y_{t,f}^{ij}, \tag{5.5}$$

where $\tilde{\mathbf{x}}_{t,f}^{ij}$ is $\mathbf{x}_{t,f}^{ij}$ with the first entry removed, and $\tilde{\mathbf{a}}_f$ is the relative CTF vector:

$$\tilde{\mathbf{a}}_f = \left(\frac{\tilde{\mathbf{a}}_f^{1\top}}{a_{0,f}^1}, \frac{\mathbf{a}_f^{2\top}}{a_{0,f}^1}, \ldots, \frac{\mathbf{a}_f^{I\top}}{a_{0,f}^1} \right)^\top, \tag{5.6}$$

where $\tilde{\mathbf{a}}_f^1 = (a_{1,f}^1, \ldots, a_{Q-1,f}^1)^\top$ denotes $\mathbf{a}_f^1$ with the first entry removed. For $i = 2, \ldots, I$, the DP-RTFs appear in (5.6) as the first entries of $\frac{\mathbf{a}_f^{i\top}}{a_{0,f}^1}$. Therefore, the DP-RTF estimation amounts to solving (5.5).

Equation (5.5) is defined for one microphone pair and for one frame. In batch mode, the terms $\tilde{\mathbf{x}}_{t,f}^{ij\top}$ and $y_{t,f}^{ij}$ of this equation can be concatenated accross microphone pairs and frames to construct a least square formulation. For online estimation, we would like to update the $\tilde{\mathbf{a}}_f$ using the current frame t. For notational convenience, let $m = 1, \ldots, M$ denote the index of a microphone pair, where $M = I(I-1)/2$. Then let the superscript ^{ij} be replaced with ^{m}. The fitting error of (5.5) is

$$e_{t,f}^m = y_{t,f}^m - \tilde{\mathbf{x}}_{t,f}^{m\top} \tilde{\mathbf{a}}_f. \tag{5.7}$$

At the current frame t, for the microphone pair m, RLS aims to minimize the error

$$J_{t,f}^m = \sum_{t'=1}^{t-1} \sum_{m'=1}^{M} \lambda^{t-t'} |e_{t',f}^{m'}|^2 + \sum_{m'=1}^{m} |e_{t,f}^{m'}|^2, \tag{5.8}$$

which sums up the fitting error of all the microphone pairs for the past frames and the microphone pairs up to m for the current frame. The forgetting factor $\lambda \in (0, 1]$ gives a lower weight to older frames, whereas all microphone pairs have the same weight at each frame. To minimize $J_{t,f}^m$, we set its complex derivative with respect to $\tilde{\mathbf{a}}_f^*$ to zero, where

Input: $\tilde{\mathbf{x}}_{t,f}^m$, $y_{t,f}^m$, $m = 1, \ldots, M$
Initialization: $\tilde{\mathbf{a}}_{t,f}^0 \leftarrow \tilde{\mathbf{a}}_{t-1,f}^M$, $\mathbf{P}_{t,f}^0 \leftarrow \lambda^{-1}\mathbf{P}_{t-1,f}^M$
for $m = 1$ to M **do**
  $e_{t,f}^m = y_{t,f}^m - \tilde{\mathbf{x}}_{t,f}^{m\top}\tilde{\mathbf{a}}_{t,f}^{m-1}$
  $\mathbf{g} = \mathbf{P}_{t,f}^{m-1}\tilde{\mathbf{x}}_{t,f}^{m*}/(1 + \tilde{\mathbf{x}}_{t,f}^{m\top}\mathbf{P}_{t,f}^{m-1}\mathbf{x}_{t,f}^{m*})$
  $\mathbf{P}_{t,f}^m = \mathbf{P}_{t,f}^{m-1} - \mathbf{g}\tilde{\mathbf{x}}_{t,f}^{m\top}\mathbf{P}_{t,f}^{m-1}$
  $\tilde{\mathbf{a}}_{t,f}^m = \tilde{\mathbf{a}}_{t,f}^{m-1} + e_{t,f}^m\mathbf{g}$
end for
Output: $\tilde{\mathbf{a}}_{t,f}^M$, $\mathbf{P}_{t,f}^M$

Algorithm 2: RLS at frame t

* denotes complex conjugate. We obtain an estimate of $\tilde{\mathbf{a}}_f$ at frame t for microphone pair m as:

$$\tilde{\mathbf{a}}_{t,f}^m = \mathbf{R}_{t,f}^{m\,-1} r_{t,f}^m, \tag{5.9}$$

with

$$\mathbf{R}_{t,f}^m = \sum_{t'=1}^{t-1}\sum_{m'=1}^{M} \lambda^{t-t'}\tilde{\mathbf{x}}_{t',f}^{m'*}\tilde{\mathbf{x}}_{t',f}^{m'\top} + \sum_{m'=1}^{m} \tilde{\mathbf{x}}_{t,f}^{m'*}\tilde{\mathbf{x}}_{t,f}^{m'\top},$$

$$r_{t,f}^m = \sum_{t'=1}^{t-1}\sum_{m'=1}^{M} \lambda^{t-t'}\tilde{\mathbf{x}}_{t',f}^{m'*}y_{t',f}^{m'} + \sum_{m'=1}^{m} \tilde{\mathbf{x}}_{t,f}^{m'*}y_{t,f}^{m'}.$$

It can be seen that the covariance matrix $\mathbf{R}_{t,f}^m$ is computed based on the rank-one modification, thence its inverse, denoted by $\mathbf{P}_{t,f}^m$, can be computed using the Sherman-Morrison formula, without the need of matrix inverse. The recursion procedure is summarized in Algorithm 2, where $\mathbf{g}$ is the *gain vector*. The current frame t is initialized with the previous frame $t-1$. At the first frame, we initialize $\tilde{\mathbf{a}}_{1,f}^0$ as zero, and $\mathbf{P}_{1,f}^0$ as the identity. At each frame, all microphone pairs are related to the same CTF vector that corresponds to the current speaker direction, hence all microphone pairs should be simultaneously used to estimate the CTF vector of the current frame. In batch mode, this can be easily implemented by concatenating the microphone pairs. However, in RLS, to satisfy the rank-one modification of the covariance matrix, we need to process the microphone pairs one by one as shown in (5.8) and Algorithm 2. At the end of the iterations over all microphone pairs, $\tilde{\mathbf{a}}_{t,f}^M$ is the "final" CTF estimation for the current frame, and is used for speaker localization. The DP-RTF estimates, denoted as $\tilde{c}_{t,f}^i,\ i = 2, \ldots, I$, are obtained from $\tilde{\mathbf{a}}_{t,f}^M$. Note that implicitly we have $\tilde{c}_{t,f}^1 = 1$.

5.2.2 MULTIPLE MOVING SPEAKERS

So far, the proposed online DP-RTF estimation method has been presented in the noise-free single-speaker case. The noisy multiple-speaker case was considered in [78], but

only for static speakers, i.e. batch mode, and in the two-channel case. We summarize the principles of this method and then explain in details the present online/multi-channel extension.

§ Estimation of the CTF vector

It is reasonable to assume that the CTF vector doesn't vary over a few consecutive frames and that only one speaker is active within a small region in the TF domain, due to the sparse representation of speech in this domain. Consequently, the CTF vector can be estimated over the current frame and a few past frames. An estimated CTF value, at each TF bin, is then assumed to be associated with only one speaker. The CTF vector computation in the case of multiple speakers can be carried out using the RLS algorithm, presented in Section 5.2.1, by adjusting the forgetting factor λ to yield a short memory.

We set the forgetting factor $\lambda = \frac{P-1}{P+1}$, where P is the number of frames being used. To efficiently estimate the CTF vector $\tilde{\mathbf{a}}_{t,f}^{M}$ of length $IQ-1$, we need $\rho \times (IQ-1)$ equations, where the factor ρ should be chosen in order to achieve a good tradeoff between the validity of the above assumptions and a robust estimate of $\tilde{\mathbf{a}}_{t,f}^{M}$. To guarantee that $\rho \times (IQ-1)$ equations are available, we need $P = \frac{\rho(IQ-1)}{I(I-1)/2} \approx \rho\frac{2Q}{I-1}$ frames. One may observe that the number of frames needed to estimate $\tilde{\mathbf{a}}_{t,f}^{M}$ decreases as the number of microphones increases.

§ Noise reduction

When noise is present, especially if the noise sources are temporally/spatially correlated, the CTF estimate can be contaminated. In addition, even in a low-noise case, many TF bins are dominated by noise due to the sparsity of speech spectra. To classify the speech frames and noise frames, and to remove the noise, we use the inter-frame spectral subtraction algorithm proposed in [76, 81].

The cross- and auto-power spectral density (PSD) between the convolution vector of the microphone signals, i.e. $\mathbf{x}_{t,f}^{i}$, and the current frame of the reference channel, i.e. $x_{t,f}^{1}$, is computed by averaging the cross- and auto-periodograms over frames. In the present work, we use recursive averaging:

$$\boldsymbol{\phi}_{t,f}^{i} = \beta\boldsymbol{\phi}_{t-1,f}^{i} + (1-\beta)\mathbf{x}_{t,f}^{i}x_{t,f}^{1\,*}, \quad i = 1,\ldots,I, \tag{5.10}$$

where β is a smoothing factor. The noise frames and speech frames are classified based on the minimum statistics [76] of the PSD of $x_{t,f}^{1}$, i.e. the first entry of $\boldsymbol{\phi}_{t,f}^{1}$. If the frames are well classified then the noise frames only include negligible speech power, due to the sparsity and non-stationarity of speech; the speech frames include noise power similar to the noise frames, due to the stationarity of noise. Therefore, inter-frame spectral subtraction can be performed as follows: for each speech frame, the cross- and auto-PSD of its nearest noise frame is subtracted from its cross- and auto-PSD, then its noise-free cross- and auto-PSD is obtained and denoted as $\hat{\boldsymbol{\phi}}_{t,f}^{i}$.

Instead of using $\mathbf{x}_{t,f}^i$, we use $\hat{\boldsymbol{\phi}}_{t,f}^i$ to construct (5.3). Correspondingly, we have a new formula (5.4), which is still valid, since it is equivalent to, with noise removed, taking the cross- and auto-PSD between both sides of the initial formula (5.4) and $x_{t,f}^1$. In the RLS process, only the speech frames (after spectral subtraction) are used, and the noise frames are skipped. A speech frame with a preceding noise frame is initialized with the latest speech frame.

§ Consistency test

In practice, a DP-RTF estimate can sometimes be unreliable. Possible reasons are that in a small frame region, (i) the CTF is time-varying due to a fast movement of the speakers, (ii) multiple speakers are present, (iii) only noise is present due to a wrong noise-speech classification, or (iv) only reverberation is present at the end of a speech occurrence. In [78], a consistency test was proposed to tackle this problem: If a small frame region indeed corresponds to one active speaker, the DP-RTFs estimated using different reference channels are consistent, otherwise the DP-RTFs are biased, with inconsistent bias values. In the present work, we use the first and second channels as references, we obtain the DP-RTF estimates $\tilde{c}_{t,f}^i$ (with $\tilde{c}_{t,f}^1 = 1$) and $\bar{c}_{t,f}^i$ (with $\bar{c}_{t,f}^2 = 1$), respectively. Then $\tilde{c}_{t,f}^i$ and $\bar{c}_{t,f}^i / \bar{c}_{t,f}^1$ are two estimates of the same DP-RTF $a_{0,f}^i / a_{0,f}^1$. To measure the similarity between these two estimates, we define the vectors $\mathbf{c}_{1,t,f}^i = (1, \tilde{c}_{t,f}^i)^\top$ and $\mathbf{c}_{2,t,f}^i = (1, \bar{c}_{t,f}^i / \bar{c}_{t,f}^1)^\top$, where the first entries are the DP-RTFs corresponding to $a_{0,f}^1 / a_{0,f}^1 = 1$. The similarity is the cosine of the angle between the two unit vectors:

$$d_{t,f}^i = \frac{|\mathbf{c}_{1,t,f}^{i\,H} \mathbf{c}_{2,t,f}^i|}{\sqrt{\mathbf{c}_{1,t,f}^{i\,H} \mathbf{c}_{1,t,f}^i \mathbf{c}_{2,t,f}^{i\,H} \mathbf{c}_{2,t,f}^i}}, \tag{5.11}$$

where ^{H} denotes conjugate transpose. If $d_{t,f}^i \in [0, 1]$ is larger than a threshold (which is fixed to 0.75 in this work) then the two estimates are consistent, otherwise they are simply ignored. Then, the two estimates are averaged and normalized as done in [78], resulting in a final complex-valued feature $\hat{c}_{t,f}^i$ whose module lies in the interval $[0, 1]$.

Finally, at frame t, we obtain a set of features $\mathcal{C}_t = \{\{\hat{c}_{t,f}^i\}_{i \in \mathcal{I}_f}\}_{f=0}^{F-1}$, where $\mathcal{I}_f \subseteq \{2, \ldots, I\}$ denotes the set of microphone indices that pass the consistency test. Note that $\mathcal{I}_f$ is empty if frame t is a noise frame at frequency f, or if no channel passes the consistency test. Each one of these features is assumed to be associated with only one speaker.

5.3 LOCALIZATION OF MULTIPLE MOVING SPEAKERS

In this section we describe the proposed frame-wise online multiple-speaker localizer. We start by briefly presenting the underlying complex Gaussian mixture model, followed by the recursive estimation of its parameters.

5.3.1 GENERATIVE MODEL FOR MULTIPLE-SPEAKER LOCALIZATION

In order to associate DP-RTF features from $\mathcal{C}_t$ with speakers and to localize each active speaker, we adopt the generative model proposed in [41]. Let $\mathcal{D} = \{\tilde{\theta}_1, \ldots, \tilde{\theta}_d, \ldots, \tilde{\theta}_D\}$ be a set of D candidate source *directions*, e.g. azimuth angles. An observed feature $\hat{c}^i_{t,f}$ (cf. Section 5.2), when emitted by a sound source located along the direction $\tilde{\theta}_d$, is assumed to be drawn from a complex-Gaussian distribution with mean $c^{i,d}_f$ and variance σ^2, i.e. $\hat{c}^i_{t,f}|d \sim \mathcal{N}_c(c^{i,d}_f, \sigma^2)$. The mean $c^{i,d}_f$ is the predicted feature at frequency f for channel i, and is precomputed based on direct-path propagation along azimuth $\tilde{\theta}_d$ to the microphones. The variance σ^2 is empirically set as a constant value. The marginal density of an observed feature $\hat{c}^i_{t,f}$ (taking into account all candidate directions) is a CGMM with each component corresponding to a candidate direction:

$$p(\hat{c}^i_{t,f}|\mathcal{D}) = \sum_{d=1}^{D} w_d \mathcal{N}_c(\hat{c}^i_{t,f}; c^{i,d}_f, \sigma^2), \tag{5.12}$$

where $w_d \geq 0$ is the prior probability (component weight) of the d-th component, with $\sum_{d=1}^{D} w_d = 1$. Let us denote the vector of weights with $\boldsymbol{w} = (w_1, \ldots, w_D)^\top$. Note that this vector is the only free parameter of the model.

Assuming that the observations in $\mathcal{C}_t$ are independent, the corresponding (normalized) negative log-likelihood function, as a function of w_d, is given by:

$$\mathcal{L}_t = -\frac{1}{|\mathcal{C}_t|} \sum_{\hat{c}^i_{t,f} \in \mathcal{C}_t} \log\Big(\sum_{d=1}^{D} w_d \mathcal{N}_c(\hat{c}^i_{t,f}; c^{i,d}_f, \sigma^2) \Big), \tag{5.13}$$

where $|\mathcal{C}_t|$ denotes the cardinality of $\mathcal{C}_t$. Once $\mathcal{L}_t$ is minimized, each weight w_d represents the probability that a speaker is active in the direction $\tilde{\theta}_d$. Therefore, sound source localization amounts to the minimization of $\mathcal{L}_t$. In addition, taking into account the fact that the number of actual active speakers is much lower than the number of candidate directions, an entropy term was proposed in [78] as a regularizer to impose a sparse solution for w_d. The entropy is defined as

$$H = -\sum_{d=1}^{D} w_d \log(w_d). \tag{5.14}$$

The concave-convex procedure [162] was adopted in [78], to minimize the objective function $\mathcal{L} + \gamma H$ w.r.t. $\boldsymbol{w}$, where $\mathcal{L}$ is the normalized negative log-likelihood of the DP-RTF features of all frames, i.e. batch mode optimization, and the positive scalar γ was used to control the tradeoff between likelihood minimization and imposing sparsity over the weights. In the batch mode, the weight vector $\boldsymbol{w}$ is shared across all frames. Hence this method is not suitable for moving speakers.

5.3.2 RECURSIVE PARAMETER ESTIMATION

We now describe a recursive method for updating the weight vector from $\boldsymbol{w}_{t-1}$ to $\boldsymbol{w}_t$, i.e. from frame $t-1$ to frame t, using the DP-RTF features at t. This can be formulated as the following online optimization problem [69]:

$$\boldsymbol{w}_t = \underset{\boldsymbol{w}}{\operatorname{argmin}} \ \chi(\boldsymbol{w}, \boldsymbol{w}_{t-1}) + \eta(\mathcal{L}_t + \gamma H), \tag{5.15}$$

$$\text{s.t.} \quad w_d > 0, \ \forall d \in \{1 \dots D\} \quad \text{and} \quad \sum_{d=1}^{D} w_d = 1, \tag{5.16}$$

where $\chi(\boldsymbol{a}, \boldsymbol{b})$ is a distance between $\boldsymbol{a}$ and $\boldsymbol{b}$. The positive scalar factor η controls the parameter update rate. To minimize (5.15), the derivative of the objective function w.r.t $\boldsymbol{w}$ is set to zero, yielding a set of equations with no closed-form solution. To speed up the computation, it is assumed that $\boldsymbol{w}_t$ is close to $\boldsymbol{w}_{t-1}$, thence the derivative of $\mathcal{L}_t + \gamma H$ at $\boldsymbol{w}$ can be approximated with the derivative of $\mathcal{L}_t + \gamma H$ at $\boldsymbol{w}_{t-1}$. This assumption is reasonable when parameter evolution is not too fast. As a result, when the distance $\chi(\boldsymbol{w}, \boldsymbol{w}_{t-1})$ is Euclidean, the objective function leads to gradient descent with a step length equal to η. Nevertheless, the constraints (5.16) lead to an inefficient gradient descent procedure. To obtain an efficient solver, we exploit the fact that the weights w_d are probability masses, hence we replace the Euclidean distance with the more suitable Kullback-Leibler divergence, *i.e.* $\chi(\boldsymbol{w}, \boldsymbol{w}_{t-1}) = \sum_{d=1}^{D} w_d \log \frac{w_d}{w_{t-1,d}}$, which results in the exponentiated gradient algorithm [69].

The partial derivatives of $\mathcal{L}_t$ and H w.r.t w_d at the point $w_{t-1,d}$ are computed with, respectively:

$$\begin{aligned} \frac{\partial(\mathcal{L}_t)}{\partial w_d}\Big|_{w_{t-1,d}} &= -\frac{1}{|\mathcal{C}_t|} \sum_{\hat{c}^i_{t,f} \in \mathcal{C}_t} \frac{\mathcal{N}_c(\hat{c}^i_{t,f}; c^{i,d}_f, \sigma^2)}{\sum_{d'=1}^{D} w_{t-1,d'} \mathcal{N}_c(\hat{c}^i_{t,f}; c^{i,d'}_f, \sigma^2)}, \\ \frac{\partial H}{\partial w_d}\Big|_{w_{t-1,d}} &= -(1 + \log(w_{t-1,d})), \quad \forall d \in \{1 \dots D\}. \end{aligned} \tag{5.17}$$

Then, the exponentiated gradient,

$$r_{t-1,d} = e^{-\eta\left(\frac{\partial(-\mathcal{L}_t)}{\partial w_d}\Big|_{w_{t-1,d}} + \gamma \frac{\partial H}{\partial w_d}\Big|_{w_{t-1,d}}\right)}, \quad \forall d \in \{1 \dots D\}, \tag{5.18}$$

is used to update the weights with:

$$w_{t,d} = \frac{r_{t-1,d} w_{t-1,d}}{\sum_{d'=1}^{D} r_{t-1,d'} w_{t-1,d'}}, \quad \forall d \in \{1 \dots D\}. \tag{5.19}$$

It is clear from (5.19) that the parameter constraints (5.16) are necessarily satisfied. The exponentiated gradient algorithm sequentially evaluates (5.17), (5.18) and (5.19) at each frame. At the first frame, the weights are initialized with the uniform distribution, namely $w_{1,d} = \frac{1}{D}$. When $\mathcal{C}_t$ is empty, such as during a silent period, the parameters are recursively updated with $w_{t,d} = (1 - \eta')w_{t-1,d} + \eta' \frac{1}{D}$.

The weight $\boldsymbol{w}_t$ as a function of $\tilde{\theta}_d$, i.e. $w_{t,d}$, exhibits a handful of peaks that could correspond to active speakers. The use of an entropy regularization term was shown to both suppress small spurious peaks, present without using the regularization term, and to sharpen the peaks corresponding to actual active speakers, thus allowing to better localize true speakers and to eliminate erroneous ones. In the case of moving speakers, a peak should shift along time from a direction $\tilde{\theta}_d$ to a nearby direction. Spatial smoothing of the weight function raises the weight values around a peak, which results in smoother peak jumps. In our experiments, spatial smoothing is carried out with $w_{t,d} = (w_{t,d} + 0.02w_{t,d-1} + 0.02w_{t,d+1})/1.04$. One may think that spatial smoothing and entropy regularization neutralize each other, but in practice it was found that their combination is beneficial.

5.3.3 Peak Selection and Frame-wise Speaker Localization

Frame-wise localization and counting of active speakers could be carried out by selecting the peaks of $\boldsymbol{w}_t(\tilde{\theta}_d)$ larger than a predefined threshold [78, 81]. However, peak selection does not exploit the temporal dependencies of moving speakers. Moreover, peak selection can be a risky process since a too high or too low threshold value may lead to undesirable missed detection or false alarm rates. In order to avoid these problems, we adopt a weighted-data Bayesian framework: all the candidate directions and the associated weights are used as observations by the multiple speaker tracking method described in Section 5.4 below. The localization results obtained with peak selection are compared with the localization results obtained with the proposed tracker in Section 5.6.5.

5.4 Multiple Speaker Tracking

Let N be the maximum number of speakers that can be simultaneously active at any time t, and let n be the speaker index. Moreover, let $n = 0$ denote *no speaker*. We now introduce the main variables and their notations. Upper case letters denote random variables while lower case letters denote their realizations.

Let $\boldsymbol{S}_{tn}$ be a latent (or state) variable associated with speaker n at frame t, and let $\mathbf{S}_t = (\boldsymbol{S}_{t1}, \ldots, \boldsymbol{S}_{tn}, \ldots, \boldsymbol{S}_{tN})$. $\boldsymbol{S}_{tn}$ is composed of two parts: the speaker direction and the speaker velocity. In this work, speaker direction is defined by an azimuth θ_{tn}. To avoid phase (circular) ambiguity we describe the direction with the unit vector $\boldsymbol{U}_{tn} = (\cos(\theta_{tn}), \sin(\theta_{tn}))^\top$. Moreover, let $V_{tn} \in \mathbb{R}$ be the angular velocity. Altogether we define a realization of the state variable as $\boldsymbol{s}_{tn} = [\boldsymbol{u}_{tn}; v_{tn}]$ where the notation $[\cdot;\cdot]$ stands for vertical vector concatenation.

Let $\mathbf{O}_t = (\boldsymbol{O}_{t1}, \ldots, \boldsymbol{O}_{td}, \ldots, \boldsymbol{O}_{tD})$ be the observed variables at frame t. Each realization $\boldsymbol{o}_{td}$ of $\boldsymbol{O}_{td}$ is composed of a candidate location, or azimuth $\tilde{\theta}_{td} \in \mathcal{D}$, and a weight w_{td}. The weight w_{td} is the probability that there is an active speaker in the direction $\tilde{\theta}_{td}$, namely (5.15). As above, let the azimuth be described by a unit vector $\boldsymbol{b}_{td} = (\cos(\tilde{\theta}_{td}), \sin(\tilde{\theta}_{td}))^\top$. In summary we have $\boldsymbol{o}_{td} = [\boldsymbol{b}_{td}; w_{td}]$. Moreover, let Z_{td} be a (latent) assignment variable

associated with each observed variable $\boldsymbol{O}_{td}$, such that $Z_{td} = n$ means that the observation indexed by d at frame t is assigned to active speaker $n \in \{0, \ldots, N\}$. Note that $Z_{td} = 0$ is a "fake" assignment – the corresponding observation is assigned to an audio source that is either background noise or any other source that has not yet been identified as an active speaker.

The problem at hand can now be cast into the estimation of the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t})$, and further inference of $\mathbf{s}_t$ and $\boldsymbol{z}_t$. In this work we make two hypotheses, namely (i) that the observations at frame t only depend on the assignment and state variables at t, and (ii) that the prior distribution of the assignment variables is independent of all the other variables. By applying the Bayes rule together with these hypotheses, and ignoring terms that do not depend on $\mathbf{s}_t$ and $\boldsymbol{z}_t$, the filtering distribution is proportional to:

$$p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t}) \propto p(\mathbf{o}_t|\mathbf{s}_t, \boldsymbol{z}_t)p(\boldsymbol{z}_t)p(\mathbf{s}_t|\mathbf{o}_{1:t-1}), \tag{5.20}$$

which contains three terms: the observation model, the prior distribution of the assignment variable and the predictive distribution over the sources state. We now characterize each one of these three terms.

§ Audio observation model

The audio observation model describes the distribution of the observations given speakers state and assignment. We assume the different observations are independent, conditioned on speakers state and assignment, which can be written as:

$$p(\mathbf{o}_t|\boldsymbol{z}_t, \mathbf{s}_t) = \prod_{d=1}^{D} p(\mathbf{o}_{td}|\boldsymbol{z}_t, \mathbf{s}_t). \tag{5.21}$$

Since the weights describe the confidence associated with each observed azimuth, we adopt the weighted-data GMM model of [54]:

$$p(\boldsymbol{b}_{td}|Z_{td} = n, \mathbf{s}_{tn}; w_{td}) = \begin{cases} \mathcal{N}(\boldsymbol{b}_{td}; \mathbf{M}\boldsymbol{s}_{tn}, \frac{1}{w_{td}}\boldsymbol{\Sigma}) & \text{if } n \in \{1, \ldots, N\} \\ \mathcal{U}(\text{vol}(\mathcal{B})) & \text{if } n = 0 \end{cases}, \tag{5.22}$$

where the matrix $\mathbf{M} = [\mathbf{I}_{2\times 2}, \mathbf{0}_{2\times 1}]$ projects the state variable onto the space of source directions and $\boldsymbol{\Sigma}$ is a covariance matrix (set empirically to a fixed value in the present study). Note that the weight plays the role of a precision: The higher the weight w_{td}, the more reliable the source direction $\boldsymbol{b}_{td}$. The case $Z_{td} = 0$ follows a uniform distribution over the volume of the observation space.

§ Prior distribution

The prior distribution of the assignment variable is independent over observations and is assumed to be uniformly distributed over all the speakers (including the case $n = 0$),

hence:

$$p(\boldsymbol{z}_t) = \prod_{d=1}^{D} p(Z_{td} = n) \quad \text{with} \quad \pi_{dn} = p(Z_{td} = n) = \frac{1}{N+1}. \tag{5.23}$$

§ Predictive distribution

The predictive distribution describes the relationship between the state $\mathbf{s}_t$ and the past observations up to frame t, $\mathbf{o}_{1:t-1}$. To calculate this distribution, we first marginalize $p(\mathbf{s}_t|\mathbf{o}_{1:t-1})$ over $\mathbf{s}_{t-1}$, writing:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1}) d\mathbf{s}_{t-1}, \tag{5.24}$$

where the two terms under the integral are the state dynamics and the marginal filtering distribution of the state variable at frame $t-1$, respectively. We model the state dynamics as a linear-Gaussian first-order Markov process, independent over the speakers, *i.e.* :

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}_{t-1,n}\boldsymbol{s}_{t-1,n}, \boldsymbol{\Lambda}_{tn}), \tag{5.25}$$

where Λ_{tn} is the dynamics' covariance matrix and $\mathbf{D}_{t-1,n}$ is the state transition matrix. Given the estimated azimuth $\theta_{t-1,n}$ and angular velocity $v_{t-1,n}$ at frame $t-1$, we have the following relation:

$$\begin{pmatrix} \cos(\theta_{tn}) \\ \sin(\theta_{tn}) \end{pmatrix} = \begin{pmatrix} \cos(\theta_{t-1,n} + v_{t-1,n}\Delta t) \\ \sin(\theta_{t-1,n} + v_{t-1,n}\Delta t) \end{pmatrix}, \tag{5.26}$$

where Δt is the time increment between two consecutive frames. Expanding (5.26) and assuming that the angular displacement $v_{t-1,n}\Delta t$ is small, the state transition matrix can be written as:

$$\mathbf{D}_{t-1,n} = \begin{pmatrix} 1 & 0 & -\sin(\theta_{t-1,n})\Delta t \\ 0 & 1 & \cos(\theta_{t-1,n})\Delta t \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.27}$$

In the following $\mathbf{D}_{t-1,n}$ is written as $\mathbf{D}$, only to lighten the equations.

5.4.1 Variational Expectation Maximization Algorithm

At this point, the standard solution to the calculation of the filtering distribution consists of using EM methodology. EM alternates between evaluating the expected complete-data log-likelihood and maximizing this expectation with respect to the model parameters. More precisely, the expectation writes:

$$J(\boldsymbol{\Theta}, \boldsymbol{\Theta}^o) = \mathbf{E}_{p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t}, \boldsymbol{\Theta}^o)} \left[\log p(\boldsymbol{z}_t, \mathbf{s}_t, \mathbf{o}_{1:t}|\boldsymbol{\Theta})\right], \tag{5.28}$$

where $\boldsymbol{\Theta}^o$ denotes the current parameter estimates and $\boldsymbol{\Theta}$ denotes the new estimates, obtained via maximization of (5.28). However, given the hybrid combinatorial-and-continuous nature of the latent space, such solution is intractable in practice, due to combinatorial explosion. We thus propose to use of a variational approximation to solve the problem efficiently. We inspire from [7] and propose the following factorization:

$$p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t}) \approx q(\boldsymbol{z}_t, \mathbf{s}_t) = q(\boldsymbol{z}_t) \prod_{n=0}^{N} q(\boldsymbol{s}_{tn}). \tag{5.29}$$

The optimal solution is then given by two E-steps, an E-S step for each individual state variable $\boldsymbol{S}_{tn}$ and an E-Z step for the assignment variable $\boldsymbol{Z}_t$:

$$\log q^\star(\boldsymbol{s}_{tn}) = \mathbf{E}_{q(\boldsymbol{z}_t)\prod_{m\neq n} q(\boldsymbol{s}_{tm})}[\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})], \tag{5.30}$$

$$\log q^\star(\boldsymbol{z}_t) = \mathbf{E}_{q(\mathbf{s}_t)}[\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})]. \tag{5.31}$$

It is easy to see that in order to compute (5.30) and (5.31), two elements are needed: the predictive distribution (5.24) and the marginal filtering distribution at $t-1$, $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$. Remarkably, as a consequence of the factorization (5.29), we can replace $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$ with $q^\star(\mathbf{s}_{t-1}) = \prod_{n=1}^{N} q^\star(\boldsymbol{s}_{t-1,n})$ in (5.24) and compute the predictive distribution as follows:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) \approx \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_{n=1}^{N} q^\star(\boldsymbol{s}_{t-1,n}) d\mathbf{s}_{t-1}. \tag{5.32}$$

This predictive distribution factorizes across speakers. Moreover, one prominent feature of the proposed variational approximation is that, if the posterior distribution at time $t-1$ $q^\star(\boldsymbol{s}_{t-1,n})$ is assumed to be a Gaussian, say

$$q^\star(\boldsymbol{s}_{t-1,n}) = \mathcal{N}(\boldsymbol{s}_{t-1,n}; \boldsymbol{\mu}_{t-1,n}, \boldsymbol{\Gamma}_{t-1,n}), \tag{5.33}$$

then (the approximation of) the predictive distribution (5.32) is a Gaussian. More specifically, the derivation of (5.32) leads to:

$$p(\mathbf{s}_{tn}|\mathbf{o}_{1:t-1}) = \mathcal{N}(\mathbf{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}). \tag{5.34}$$

Moreover, as we will see in the E-S-step below, the posterior distribution at time t, $q^\star(\boldsymbol{s}_{tn})$, is also a Gaussian.

§ E-S step

The computation of the variational posterior distribution $q^\star(\boldsymbol{s}_{tn})$, for all currently tracked speakers, is carried out by developing (5.30) as follows. We first exploit (5.20), (5.21), (5.23) and (5.34) to rewrite $\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})$ in (5.30) as a sum of individual log-probabilities. Then we eliminate all terms not depending on $\boldsymbol{s}_{tn}$ and we evaluate the expectation of the remaining terms. Because the terms not depending on $\boldsymbol{s}_{tn}$ were disregarded, the expectation is computed only with respect to $q^\star(\boldsymbol{z}_t)$. This nicely makes the computation of

$q^\star(\boldsymbol{s}_{tn})$ independent of the structure of $q^\star(\boldsymbol{s}_{tm})$ for $m \neq n$. Eventually, this yields a Gaussian distribution:

$$q^\star(\boldsymbol{s}_{tn}) = \mathcal{N}(\boldsymbol{s}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}), \tag{5.35}$$

with the following parameters:

$$\begin{aligned}\boldsymbol{\Gamma}_{tn} = &\Big(\Big(\sum_{d=1}^{D} \alpha_{tdn} w_{td}\Big)\mathbf{M}^\top \boldsymbol{\Sigma}^{-1}\mathbf{M} \\ &+ \Big(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top\Big)^{-1}\Big)^{-1}, \end{aligned}\tag{5.36}$$

$$\begin{aligned}\boldsymbol{\mu}_{tn} = &\boldsymbol{\Gamma}_{tn}\Big(\mathbf{M}^\top \boldsymbol{\Sigma}^{-1}\Big(\sum_{d=1}^{D} \alpha_{tdn} w_{td} \boldsymbol{b}_{td}\Big) \\ &+ \Big(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top\Big)^{-1}\mathbf{D}\boldsymbol{\mu}_{t-1,n}\Big), \end{aligned}\tag{5.37}$$

where $\alpha_{tdn} = q^\star(Z_{td} = n)$ is the variational posterior distribution of the assignment variable, which will be detailed in Section 5.6.3. Importantly, the first two entries of $\boldsymbol{\mu}_{tn}$ in (5.37) represent the estimated azimuth of speaker n. The "final" azimuth estimate at frame t is thus given by this subvector at the end of the VEM iterations. Since we use a unit-vector representation, we normalize this vector at each iteration of the algorithm. Finally, note that since we have shown that $q^\star(\boldsymbol{s}_{t-1,n})$ being Gaussian leads to $q^\star(\boldsymbol{s}_{tn})$ being Gaussian as well, it is sufficient to assume that $q^\star(\boldsymbol{s}_{1n})$ is Gaussian, namely at $t = 1$: $q^\star(\boldsymbol{s}_{1n}) = \mathcal{N}(\boldsymbol{s}_{1n}; \boldsymbol{\mu}_{1n}, \boldsymbol{\Gamma}_{1n})$.

§ E-Z step

Developing (5.31) with the same principles as above, one can easily find that the variational posterior distribution of the assignment variable factorizes as:

$$q(\boldsymbol{z}_t) = \prod_{d=1}^{D} q(z_{td}). \tag{5.38}$$

In addition, we obtain a closed-form expression for $q^\star(z_{td})$:

$$\alpha_{tdn} = q^\star(Z_{td} = n) = \frac{\rho_{tdn}\pi_{dn}}{\sum_{i=0}^{N} \rho_{tdi}\pi_{di}}, \tag{5.39}$$

where π_{dn} was defined in (5.23), and ρ_{tdn} is given by:

$$\rho_{tdn} = \begin{cases} \mathcal{N}(\boldsymbol{b}_{td}; \mathbf{M}\boldsymbol{\mu}_{tn}, \frac{1}{w_{td}}\boldsymbol{\Sigma}) & \\ \times e^{-\frac{1}{2}\mathrm{tr}\left(w_{td}\mathbf{M}^\top \boldsymbol{\Sigma}^{-1}\mathbf{M}\boldsymbol{\Gamma}_{tn}\right)} & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\mathrm{vol}(\mathcal{G})) & \text{if } n = 0. \end{cases} \tag{5.40}$$

§ M-step

The M-step estimates model parameters by maximize J in (5.28) with respect to the model parameters. Here we only estimate the covariance matrix of the state dynamics $\mathbf{\Lambda}_{tn}$. The detailed calculation please refer to (2.27).

5.4.2 SPEAKER-BIRTH PROCESS

A birth process is used to initialize new tracks, *i.e.* speakers that become active. We take inspiration from the birth process for visual tracking proposed in [7] and adapt it to audio tracking. The general principle is the following. In a short period of time, say from $t-L$ to t, with L being small (typically 3), we assume that at most one new (yet untracked) speaker becomes active. For each frame from $t-L$ to t, among the observations assigned to $n=0$ we select the one with the highest weight, and thus obtain an observation sequence $\tilde{\boldsymbol{o}}_{t-L:t}$. We then compute the marginal likelihood of this sequence according to our model, $\epsilon_0 = p(\tilde{\boldsymbol{o}}_{t-L:t})$. If these observations have been generated by a speaker that has not been detected yet (hypothese H_1), then they are assumed to be consistent with the model, *i.e.* exhibit smooth trajectories, and ϵ_0 will be high; otherwise, *i.e.* if they have been generated by background noise (hypothese H_0), they will be more randomly spread over the range of possible observations, and ϵ_0 will be low. Giving birth to a new speaker track amounts to setting a threshold ϵ_1 and deciding between the two hypotheses:

$$\epsilon_0 \underset{H_0}{\overset{H_1}{\gtrless}} \epsilon_1. \tag{5.41}$$

This process is applied continuously over time to detect new speakers. This includes speaker track initialization at $t=1$. Note that initially all the assignment variables are set to $n=0$ (background noise), namely $Z_{1d}=0, \forall d$.

As for the computaiotn of $p(\tilde{\boldsymbol{o}}_{t-L:t})$, we first rewrite it as the marginalization of the joint probability of the selected observations and the state trajectory $\hat{\boldsymbol{s}}_{t-L:t}$ of a potential speaker:

$$\epsilon_0 = \int p(\tilde{\boldsymbol{o}}_{t-L:t}, \hat{\boldsymbol{s}}_{t-L:t}) d\hat{\boldsymbol{s}}_{t-L:t}, \tag{5.42}$$

which, under the proposed model, is given by:

$$\epsilon_0 = \int \Big(\prod_{i=t-L+1}^{t} p(\tilde{\boldsymbol{o}}_i|\hat{\boldsymbol{s}}_i) p(\hat{\boldsymbol{s}}_i|\hat{\boldsymbol{s}}_{i-1}) \Big) p(\tilde{\boldsymbol{o}}_{t-L}|\hat{\boldsymbol{s}}_{t-L}) p(\hat{\boldsymbol{s}}_{t-L}) d\hat{\boldsymbol{s}}_{t-L:t}.$$

All the terms in the above equation have been defined during the derivation of our model except the marginal prior distribution of the state $p(\hat{\boldsymbol{s}}_{t-L})$, and all these terms are Gaussian. For the track-birth process, we just want to test if the trajectory of observations from $t-L$ to t is coherent, and we can define here $p(\hat{\boldsymbol{s}}_{t-L})$ as a non-informative distribution, such as a uniform distribution. In practice we choose a Gaussian distribution with a very

Input: audio observations $\boldsymbol{b}_{1:t}$;
for $t = 1$ *to end* **do**
  Gather observations at frame t;
  for $iter = 1$ *to* N_{iter} **do**
    E-Z-step:
    for $d \in \{1, ..., D\}$ **do**
      for $n \in \{0, ..., N\}$ **do**
        Evaluate $q(Z_{td} = n)$ with (5.39);
        ;
      end
    end
    E-S-step:
    for $n \in \{1, ..., N\}$ **do**
      Evaluate $\boldsymbol{\Gamma}_{tn}$ and $\boldsymbol{\mu}_{tn}$ with (5.36) and (5.37);
    end
    M-step: Evaluate $\boldsymbol{\Lambda}_{tn}$ with (2.27);;
  end
  Speaker-Birth Process (see Section 5.4.2);
  Detect speaker activity (see Section 5.4.3);
  for $n \in \{1, ..., N\}$ **do**
    if *the speaker* n *is detected as active* **then**
      Output the results $\boldsymbol{\mu}_{tn}$;
    end
  end
end

Algorithm 3: Variational EM acoustic tracking.

large covariance, to ensure a closed-form solution to (5.43). Due to room limitation, we do not present more details. Let us just mention that in practice we set $L = 3$, which enables efficient speaker birth detection.

5.4.3 Speaker Activity Detection

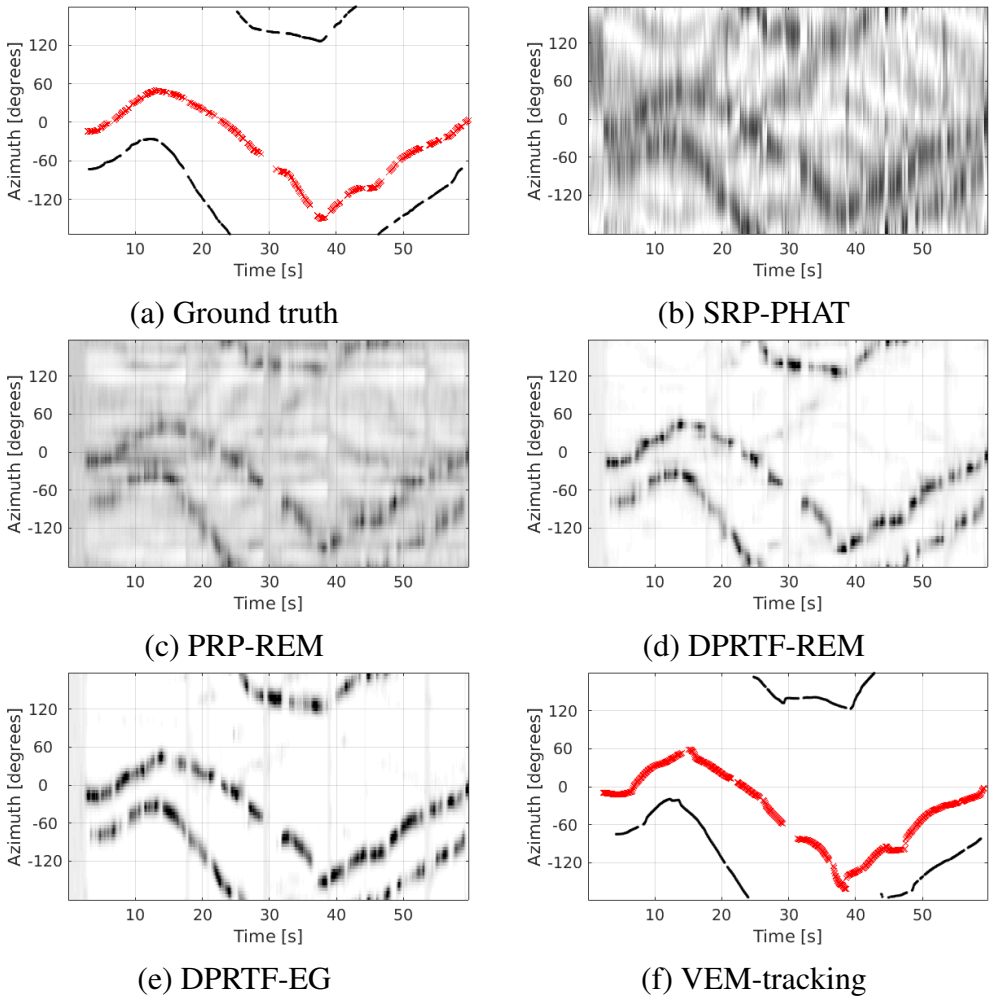
A very interesting feature of the proposed model is that, once speaker tracks have been estimated, the posterior distribution of the assignment variables $\boldsymbol{Z}_t$ can be used for speech activity detection, *i.e.* who are the active speakers at each frame, a task also referred to as *speaker diarization* in the multi-speaker context. This can be formalized as testing for each frame t and each speaker n between the two following hypotheses: H_1: Speaker n is active at frame t, and H_0: Speaker n is silent at frame t. In the present work, this is done by computing the following *weighted sum of weights*, averaged over a small number of frames L' to take into account speaker activity inertia, and comparing with a threshold δ,

a test formally written as:

$$\sum_{i=t-L'+1}^{t} \sum_{d=1}^{D} \alpha_{idn} w_i^d \underset{H_0}{\overset{H_1}{\gtrless}} \delta. \tag{5.43}$$

Overall, the variational EM tracking algorithm is described in Algorithm 3.

5.5 EXPERIMENTS



(a) Ground truth (b) SRP-PHAT

(c) PRP-REM (d) DPRTF-REM

(e) DPRTF-EG (f) VEM-tracking

Figure 5.2: Results of speaker localization and tracking for Recording 1 / Task 6 of LOCATA data. (a) Ground truth trajectory and voice activity (red for speaker 1, black for speaker 2). Intervals in the trajectories are speaking pauses. (b)-(e) One-dimensional heat maps as a function of time for the four tested localization methods. (f) Results for the proposed VEM-based tracker. Black and red colors demonstrate a successful tracking, *i.e.* continuity of the tracks despite of speech pauses.

5.5.1 EXPERIMENTAL SETUPS

§ Datasets

We tested and empirically validated our method with the LOCATA and the Kinovis-MST datasets. The LOCATA (a IEEE-AASP challenge for sound source localization and tracking) [88] data were recorded in the Computing Laboratory of the Department of Computer Science of Humboldt University Berlin. The room size is 7.1 m $\times$ 9.8 m $\times$ 3 m, with a reverberation time $T_{60} \approx 0.55$ s. We report the results of the development corpus for tasks #3 and #5 with a single moving speaker, and for tasks #4 and #6 with two moving speakers, each task comprising three recorded sequences.[2] There are twelve microphones arranged such as to form a spherical array and placed on the head of a NAO robot. We used two microphone configurations: four quasi-planar microphones, located on the top of the head, numbered 5, 8, 11, 12, and eight microphones numbered 1, 3, 4, 5, 8, 10, 11, 12. An optical motion capture system was used to provide ground-truth positions of the robot and of the speakers. The participants speak continuously during the entire recordings. However, speech pauses are inevitable and these pauses may last several seconds. Each participant has a head-mounted microphone. We applied the voice activity detector [80] to these microphone signals to obtain ground-truth voice activity information of each participant. The signal-to-noise ratio (SNR) is approximatively 23.4 dB.

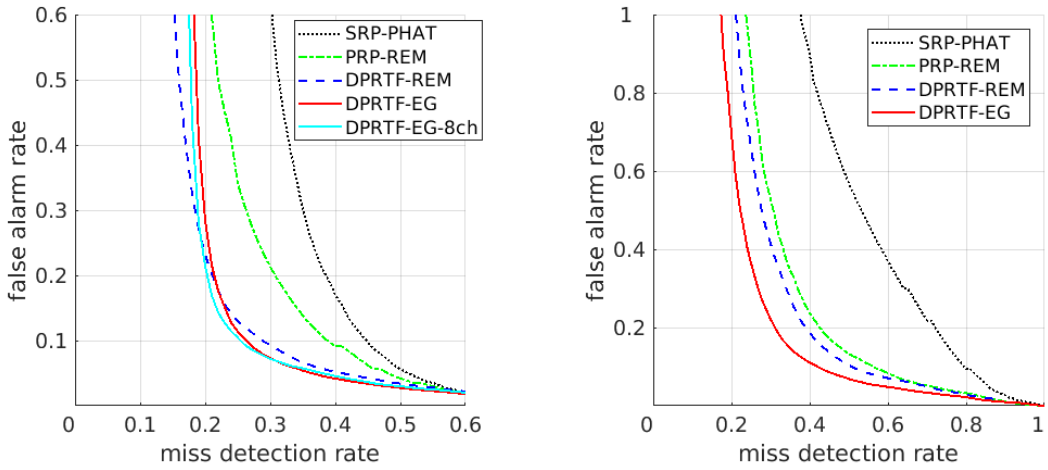
The Kinovis-MST dataset was recorded in the Kinovis multiple-camera laboratory at INRIA Grenoble.[3] The room size is 10.19 m $\times$ 9.87 m $\times$ 5.6 m, with $T_{60} \approx 0.53$ s. A v5 NAO robot with four microphones [75] was used. The geometric layout of the microphones is similar to the one of the robot used in LOCATA. The speakers were moving around the robot with a speaker-to-robot distance ranging between 1.5 m and 3.5 m. As with LOCATA, a motion capture system was used to obtain ground-truth trajectories of the moving participants and the location of the robot. Ten sequences were recorded with up to three participants, for a total length of about 357 s. The robot's head has built-in fans located nearby the microphones, hence the recordings contain a notable amount of stationary and spatially correlated noise with an SNR of approximatively 2.7 dB[75]. The participants behave more naturally than in the LOCATA scenarios, i.e. they take speech turns in a natural multi-party dialog. When one participant is silent, he/she manually hides the infrared marker located on his head to make it invisible to the motion capture system. This provides ground-truth speech activity information for each participant. This dataset and the associated annotations allow us to test the proposed tracking algorithm when the number of active speakers varies over time.

§ Parameter setting

For both datasets, we perform $360°$-wide azimuth estimation and tracking: $D = 72$ azimuth directions at every $5°$ in $[-175°, 180°]$ are used as candidate directions. The

[2]The results obtained with the proposed method were officially submitted to the LOCATA challenge and they will be available soon at `https://locata.lms.tf.fau.de/`.

[3]`https://kinovis.inria.fr/inria-platform/`



(a) LOCATA dataset (b) Kinovis-MST dataset

Figure 5.3: receiver operating characteristic (ROC) curve.

CGMM mean $c_f^{i,d}$ is the head-related transfer function (HRTF) ratio between two microphones, which are precomputed based on the direct-path propagation model for each candidate direction. In the Kinovis-MST dataset, the HRTFs have been measured to compute the CGMM means. For LOCATA, the TDOAs are computed based on the coordinate of microphones, which are then used to compute the phase of the CGMM means, while the magnitude of the CGMM means are set to a constant, e.g. 0.5, for all the frequencies. All the recorded signals are resampled to 16 kHz. The STFT uses the Hamming window with length of 16 ms and shift of 8 ms. The CTF length is $Q = 8$ frames. The RLS forgetting factor λ is computed using $\rho = 1$. The exponentiated gradient update factor is $\eta = 0.07$. The smoothing factor η' is set to 0.065. The entropy regularization factor is $\gamma = 0.1$. For the tracker, the covariance matrix is set to be isotropic $\mathbf{\Sigma} = 0.03\mathbf{I}_2$. The threshold giving birth to a new identity is $\tau_1 = 0.75$ and $L = 3$. To decide whether a person is speaking or is silent, $L' = 3$ frames are used, with a threshold $\delta = 0.15$. At each time instance, the VEM algorithm has 5 iterations. Corresponding to the STFT frame shift, i.e. 8 ms, the frame rate of the proposed system is 125 frames per second.

§ Comparison with Baseline Methods

The proposed method is evaluated both in "frame-wise localization" mode and in "tracker" mode. In the first mode, the frame-wise online localization module of Section 5.3 is applied without the tracker of Section 5.4. Instead, it is followed by the peak selection process described in [78]. This method is referred to as DP-RTF using EG (DPRTF-EG). In tracker mode, DPRTF-EG is directly followed by the proposed VEM tracker, without peak selection. It is then simply referred to as VEM-tracker. In that case, the directions of active speakers are given by the state variable, and the continuity of the speaker tracks is given by the assignment variable. We compare DPRTF-EG with several baseline meth-

ods:

- The standard beamforming-based localization method called SRP using phase transform (PHAT) (SRP-PHAT) [37]. The same STFT configuration and candidate directions are used for SRP-PHAT and for the proposed method. The steering vector for each candidate direction is derived from the HRTFs and TDOAs for the Kinovis-MST and LOCATA datasets, respectively. The frame-wise SRP is recursively smoothed with a smoothing factor set to 0.065.
- A method combining PRP features, CGMM model and parameter update using REM [131], referred to as PRP-REM. We also combine the DPRTF features and CGMM with REM (referred to as DPRTF-REM). This is to evaluate the proposed DP-RTF feature w.r.t. PRP, and the EG-based online parameters update method w.r.t. REM. For both baselines, the STFT and CGMM settings are the same as for the proposed method. The updating factor of REM is set to 0.065.

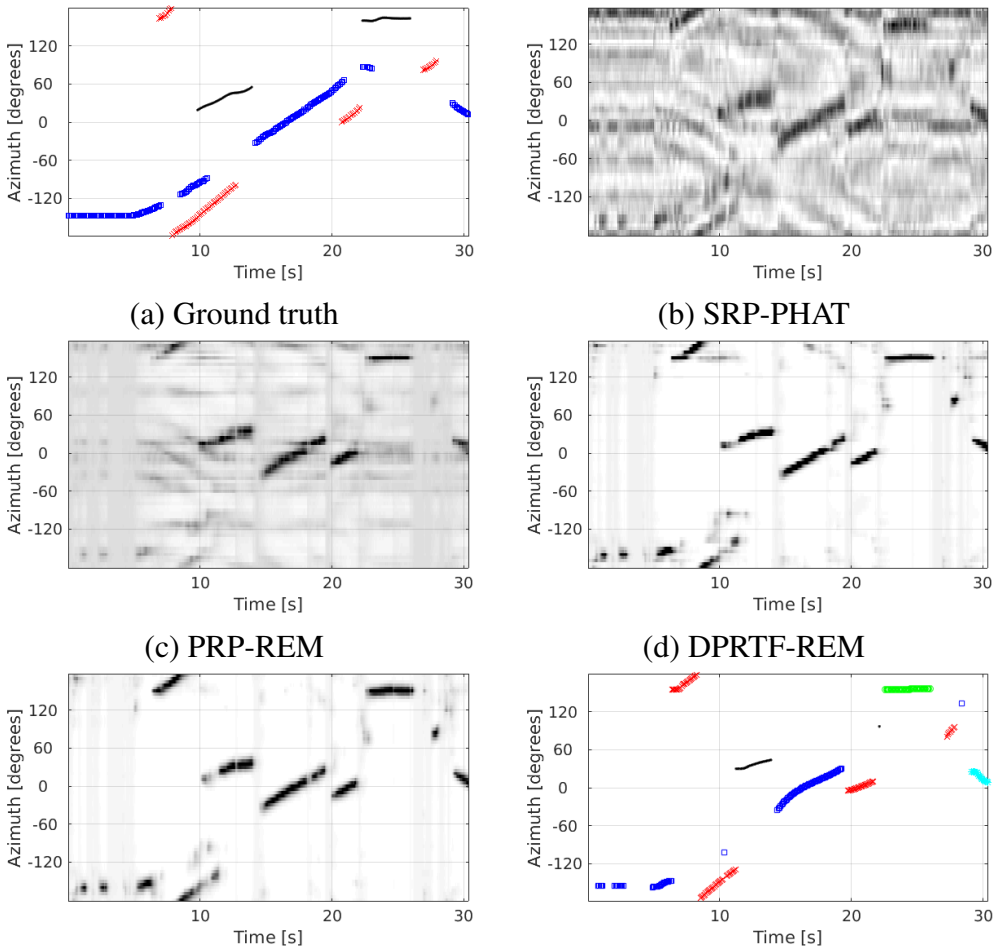
§ Evaluation Metrics

The detected speakers should be assigned to the actual speakers for performance evaluation. This is done using a greedy matching algorithm. First the azimuth difference for all possible detected-actual speaker pairs are computed, then the detected-actual speaker pair with the smallest difference is picked out as a matched pair. This procedure is iterated until the detected or actual speakers are all picked out. For each matched pair, the detected speaker is then considered to be successfully localized if the azimuth difference is not larger than $15°$. The absolute error is calculated for the successfully localized sources. The mean absolute error (MAE) is computed by averaging the absolute error of all speakers and frames. For the unsuccessful localizations, we count the miss detection (MD) (speaker active but not detected) and false alarms (FAs) (speaker detected but not active). Then the MD and FA rates are computed, using all the frames, as the percentage of the total MDs and FAs out of the total number of actual speakers, respectively. In addition to these localization metrics, we also count the identity switches (IDs) to evaluate the tracking continuity. ID is an absolute number. It represents the number of the identity changes in the tracks for a whole test sequence.

The computation time is measured with the real-time factor (RF), which is the processing time of a method divided by the length of the processed signal. Note that all the methods are implemented in MATLAB.

5.5.2 Results for LOCATA Dataset

For convenience, both the spatial spectrum of SRP-PHAT and the CGMM component weights profile will be referred to as heatmaps. Fig. 5.2 shows an example of a result obtained with a LOCATA sequence. Two speakers are moving and continuously speaking with short pauses. The SRP-PHAT heatmap (Fig. 5.2 (b)) is cluttered due to the non ideal beampattern of the microphone array and to the influence of reverberation and noise. For



(a) Ground truth (b) SRP-PHAT

(c) PRP-REM (d) DPRTF-REM

(e) DPRTF-EG (f) VEM-tracking

Figure 5.4: Results of speaker localization and tracking for one sequence of the Kinovis-MST dataset. (a) Ground truth trajectory and voice activity (red for speaker 1, black for speaker 2, blue for speaker 3). (b)-(e) One-dimensional heat maps as a function of time for the four tested localization methods. (f) Results for the proposed VEM-based tracker.

most of the time, SRP-PHAT has prominent response power for the true speaker directions. Localization of the most dominant speaker can be made by selecting the direction with the largest response power. However, it is difficult to correctly count the number of active speakers and localize less dominant speakers, since there exist a number of spurious peaks. PRP-REM (Fig. 5.2 (c)) exhibits a clearer heatmap compared to SRP-PHAT, but there exist some spurious trajectories as well, since the PRP features are contaminated by reverberation. DPRTF-REM (Fig. 5.2 (d)) removes most of the spurious trajectories, which illustrates the robustness of the proposed DP-RTF feature against reverberation. From Fig. 5.2 (e), it can be seen that the proposed EG algorithm further removes the interferences by applying the entropy regularization. In addition, the peak evolution is smoother compared with Fig. 5.2 (d), which is mainly due to the use of the spatial

smoothing. Fig. 5.2 (f) illustrates the result obtained with the proposed VEM tracker, with DPRTF-EG providing the observations. The proposed tracker gives smoother and cleaner results compared with the other methods. Even when the observations have a low weight, the tracker is still able to give the correct speaker trajectories. This is ensured by the second term in (5.37) which exploits the source dynamics model and continues to provide localization information even when $w_{t,d}$ (and/or α_{tdn}) becomes small. As a result, the tracker is able to preserve the identity of speakers in spite of the (short) speech pauses. In the presented sequence example, the estimated speaker identities are quite consistent with the ground truth.

To empirically evaluate the quality of the heatmaps provided by the localization methods, we computed the ROC curve (MD rate versus FA rate) for the LOCATA dataset by varying the peak selection threshold, for each tested method, Fig. 5.3. As already mentioned, in addition to using four microphones, we also tested an eight-microphone configuration, which is referred to as DPRTF-EG-8ch.

By analyzing the ROC curves, one notices that the methods based on DP-RTF perform better than SRP-PHAT and than PRP-REM, which is consistent with the heatmaps of Fig. 5.2: SRP-PHAT and PRP-REM are more sensitive to the presence of reverberations than the proposed methods. The performance of both DPRTF-REM and DPRTF-EG cannot be easily discriminated using the ROC curves. DPRTF-EG-8ch performs slightly better than DPRTF-EG, which means that the performance of the proposed method can be slightly improved by increasing the number of microphones. One may conclude that the proposed method is well suited when only a small number of microphones are available. With all methods, the FA rate can be trivially decreased to be close to 0 by increasing the peak selection threshold. However, the MD rate cannot be decreased to 0 even with a very small peak-selection threshold, since some speech frames that are actually present cannot be detected as the heatmap peaks due to the influence of noise and reverberation, and to a possible latency in the detection.

Table 5.1: Localization and tracking results for the LOCATA data.

	MD rate (%)	FA rate (%)	MAE (°)	IDs	RF
SRP-PHAT	39.2	18.6	5.2	-	0.06
PRP-REM	30.9	19.6	5.0	-	0.30
DPRTF-REM	23.3	15.2	4.6	-	0.97
DPRTF-EG	23.9	13.0	4.0	-	0.97
DPRTF-EG-8ch	22.7	13.2	4.1	-	3.03
VEM + EG	22.7	12.4	4.1	10	2.05
VEM + EG-8ch	22.9	11.0	3.2	6	4.05

For each curve, a good balance between FA rate and MD rate is achieved at the left-bottom corner, which can be detected as the point with the minimum distance to the origin. The average localization results corresponding to this optimal left-bottom point are summarized in Table 5.1 for each tested method. It can be seen that, besides MD and FA, the DPRTF-based methods achieve smaller MAE than SRP-PHAT and PRP-REM,

since the proposed DP-RTF features are robust against reverberation and thus leads to smaller biases for the heatmap peaks. DPRTF-EG has a higher MD rate than DPRTF-REM, while it also has lower FA rate, and a lower MAE, due to the effect of entropy regularization. With eight microphones, i.e. DPRTF-EG-8ch, MD is 1% smaller than the MD of DPRTF-EG, since the use of a non coplanar microphone setup provides more accurate localization than a coplanar setup. The proposed tracker performs the best in terms of MD and of FA. The tracker slightly reduces FA compared to DPRTF-EG. It also reduces the MD score since some correct speaker trajectories can be recovered even when the observations have (very) low weights, as explained above. In addition, the proposed tracker achieves quite consistent speaker ID estimation. For the whole LOCATA dataset, only ten identity switches were observed when using DPRTF-EG, and this number is reduced to six when using DPRTF-EG-8ch. The remaining identity switches are mainly due to speakers with crossing trajectories, a hard case for multiple audio-source tracking.

As for the computation time, SRP-PHAT has the smallest RF. Based on the fact that the RFs of DPRTF-REM and DPRTF-EG are identical, we can conclude that the REM algorithm and the proposed EG algorithm have comparable computational complexities. The RFs of PRP-REM, DPRTF-REM (or DPRTF-EG) and DPRTF-EG-8ch are different due to different computational complexities for feature estimation, more precisely due to the different dimensions of the vector to be estimated. The CTF identification used for DP-RTF estimation solves an RLS problem with the unknown CTF vector $\tilde{\mathbf{a}}_f \in \mathbb{C}^{(IQ-1)\times 1}$. Remind that I and Q denote the number of microphones and the CTF length, respectively. In the present work, we have set $I = 4/Q = 8$ for DPRTF-REM (or DPRTF-EG), $I = 8/Q = 8$ for DPRTF-EG-8ch. PRP is defined based on the narrow-band assumption, or equivalently based on the CTF with $Q = 1$, thence we have $I = 4/Q = 1$ for PRP-REM. The proposed localization method, i.e. DPRTF-EG with four microphones, has an RF smaller than one, which means it can be run in real time. The RF for the proposed tracker (VEM) is computed by the sum of the localization time and of the tracking time. For acoustic tracking, the tracker observes an DOA estimate every 8 ms. However, an 8 ms speaker motion is small. Thus in practice, the tracker uses one DOA estimate per 32 ms intervals, which leads to an RF of 2.05 for the four-channel (4ch) case and 4.05 for the eight-channel (8ch) case. The RF can be further improved by using less DOA estimates.

5.5.3 RESULTS FOR KINOVIS-MST DATASET

Fig. 5.4 shows an example of result for a Kinovis-MST sequence. Three participants are moving and intermittently speaking. It can be seen that, for many frames, the response power of SRP-PHAT and the CGMM component weights of PRP-REM corresponding to the true active speakers are not prominent, compared to the spurious trajectories. Again, DPRTF-REM and DPRTF-EG provide much better heatmaps, though they also miss some speaking frames, e.g. at the beginning of Speaker 3's trajectory (in blue). The possible reasons are i) the NAO robot (v5) has a relative strong ego-noise [75], and thus the signal-to-noise ratio of the recorded signals is relative low, and ii) the speakers are moving with a varying source-to-robot distance and the direct-path speech is contaminated by more

Table 5.2: Localization and tracking results for the Kinovis-MST dataset.

	MD rate (%)	FA rate (%)	MAE (°)	IDs	RF
SRP-PHAT	60.0	37.1	5.5	-	0.07
PRP-REM	40.3	23.1	5.1	-	0.32
DPRTF-REM	37.6	22.0	5.5	-	0.73
DPRTF-EG	31.4	19.5	5.3	-	0.73
VEM + EG	31.1	11.7	4.9	11	2.12

reverberations when the speakers are distant. Overall, DPRTF-REM and DPRTF-EG are able to monitor the moving, appearance, and disappearance of active speakers for most of the time, with a small time lag due to the temporal smoothing.

This kind of recording/scenario is very challenging for the tracking method, especially for speaker identity preservation, since the participants are intermittently speaking and moving. In a general manner, the proposed tracker achieves relatively good results, as illustrated in Fig. 5.4 (f). The tracked trajectories are smooth and clean. If the true trajectory of one speaker has an approximately constant direction, the tracker is able to re-identify the speaker even after a long silence thanks to the above-mentioned combination of observations and dynamics in (5.37), e.g. Speaker 1's trajectory in red. In the case that the speaker changes his/her movement when he/she is silent, the track can be lost. When the person speaks again, it is indeed difficult to re-idendify him/her based on the dynamics estimated before the silence period. The tracker may then prefer to give birth to a new speaker. This is illustrated by the black trajectory turning into green, and the blue trajectory turning into cyan in Fig. 5.4. Note that the silence periods are here much longer than in the LOCATA example of Fig. 5.2.

Fig 5.3 show the ROC curves for the Kinovis-MST dataset. Compared to the ROC curves for the LOCATA dataset, all the four localization methods have a worse ROC curve, especially along the MD rate axis, for the reasons mentioned above. Table 5.2 summarizes the localization and tracking results for the optimal bottom-left point of the ROC curves. It can be seen that, for the four localization methods, MAEs are quite close, namely the heatmap peaks have similar biases. Compared with the results for the LOCATA dataset, the advantage of the proposed tracker is more significant for the Kinovis-MST dataset, especially the FA rate is reduced by 7.8% relative to DPRTF-EG. The identity switches are mainly caused by speakers changing their movement while being silent, as discussed above. Compared to the LOCATA dataset, DPRTF-EG has smaller RF, since the Kinovis-MST dataset is more noisy and more noise frames are skipped in the RLS algorithm.

5.6 Acoustic Speaker Tracking Extension with Von-mises Distribution

5.6.1 Introduction

In this section we address the problem of simultaneously tracking several audio sources, namely the problem of estimating source trajectories from a sequence of observed features. Audio-source tracking is useful for a number of tasks such as audio-source separation, spatial filtering, speech enhancement and speech recognition, which in turn are essential for robust voice-based home assistants. Audio source tracking is difficult because audio signals are adversely affected by noise, reverberation and interferences between acoustic signals.

Single-source tracking methods are often based on observing the TDOAs between two microphones. Since the mapping between TDOAs and the space of source locations is non-linear, sequential Monte Carlo particle filters are used, e.g. [146, 156, 164, 122]., Alternatively, microphone arrays can be used to estimate DOAs of audio sources. The problem can then be cast into a linear dynamic model, e.g. the adaptive Kalman filter [18]. In this case source directions should however be modeled as circular random variables, e.g. the wrapped Gaussian distribution [142], or the von Mises distribution [44, 108].

Multiple-source tracking is more challenging since it raises additional difficulties: (i) the number of active audio sources is unknown and it varies over time, (ii) multiple DOAs must be simultaneously estimated, and (iii) DOA-to-source assignments must be computed over time. The problem of tracking an unknown number of sources is specifically addressed in the framework of random finite sets [97]. Since the probability density function (pdf) is computationally intractable, its first order approximation can be propagated in time using the PHD filter [97, 148, 86]. In [94] the PHD filter was applied to audio recordings to track multiple sources from TDOA estimates. In [46] the wrapped Gaussian distribution is incorporated within a PHD filter. A mixture of von Mises distributions is combined with a PHD filter in [107]. The drawback of PHD-based filters is twofold: they don't yield smooth trajectories since state dynamics are not explicitly enforced, and they cannot find observation-to-source associations without recourse to ad-hoc post-processing.

Multiple target tracking is also formulated as a variational approximation of Bayesian filtering [7]. Observation-to-target associations are modeled as discrete latent variables and their realizations are estimated within a compact and efficient VEM solver. Moreover, the problem of tracking a varying number of targets is addressed via track-birth and track-death processes. The variational approximation of [7] was recently extended to track multiple audio sources using a mixture of wrapped Gaussian distributions [74].

This section builds on [44], [7] and [74] and proposes to use the von Mises distribution to model DOAs with circular random variables. This leads to a multi-target Kalman filter which is intractable because of the combinatorial explosion of associating observations to state variables over time. We propose a variational approximation of the filter's posterior

distribution and we infer a VEM algorithm which is computationally efficient. We also propose an audio-source birth method that favors smooth source trajectories and which is used both to initialize the number of active sources and to detect new sources. We perform experiments with a recently released dataset comprising several moving sources as well as a moving microphone array.

The section is organized as follows. Section 5.6.2 describes the probabilistic model and Section 5.6.3 describes a variational approximation of the filtering distribution and the VEM algorithm. Section 5.6.4 briefly describes the source birth method. Experiments and comparisons with other methods are described in Section 5.6.5. Supplemental materials (mathematical derivations, software and videos) can be found at
`https://team.inria.fr/perception/research/audiotrack-vonm/`.

5.6.2 Probabilistic Model

Let N denote the unknown number of audio sources and the state $s_{tn} \in (-\pi, \pi]$ be the DOA of source $n \in \{1, \dots, N\}$ at t, and $\boldsymbol{s}_t = (s_{t1}, \dots, s_{tN})$. Furthermore, let $y_{tm} \in (-\pi, \pi]$ denote the m-th DOA observation at time t, where $m \in \{1, \dots, M_t\}$ and M_t is the number of observations at t. Each observation is accompanied by its corresponding confidence $\omega_{tm} \in [0, 1]$. Let z_{tm} denote the observation-to-source assignment variable, where $z_{tm} = n$ means that y_{tm} is an observation of source n, and $z_{tm} = 0$ means that y_{tm} is an observation of a dummy source.

Within a Bayesian model, multiple target tracking can be formulated as the estimation of the filtering distribution $p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t})$. We assume that the state variables s_{tn} follow a first-order Markov model, and that the observations depend only on the current state and on the assignment variables. Under these two hypotheses the posterior pdf is given by:

$$p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t}) \propto p(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{s}_t) p(\boldsymbol{z}_t) p(\boldsymbol{s}_t | \boldsymbol{y}_{1:t-1}), \tag{5.44}$$

where $p(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{s}_t)$ is the observation likelihood, $p(\boldsymbol{z}_t)$ is the prior pdf of the assignment variables and $p(\boldsymbol{s}_t | \boldsymbol{y}_{1:t-1})$ is the predictive pdf of the state variables.

§ Observation likelihood

Assuming that DOA estimates are independently and identically distributed (i.i.d.), the observation likelihood can be written as:

$$p(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{s}_t) = \prod_{m=1}^{M_t} p(y_{tm} | \boldsymbol{z}_t, \boldsymbol{s}_t). \tag{5.45}$$

The likelihood that a DOA corresponds to a source is modelled by a von Mises distribution [44], whereas the likelihood that a DOA corresponds to a dummy source is modelled by a uniform distribution:

$$p(y_{tm} | z_{tm} = n, s_{tn}) = \begin{cases} \mathcal{M}(y_{tm}; s_{tn}, \kappa_y \omega_{tm}) & n \neq 0 \\ \mathcal{U}(y_{tm}) & n = 0 \end{cases}, \tag{5.46}$$

where $\mathcal{M}(y\,;s,\kappa) = (2\pi I_0(\kappa))^{-1}\exp\{\kappa\cos(y-s)\}$ denotes the von Mises distribution with mean s and concentration κ, $I_p(\cdot)$ denotes the modified Bessel function of the first kind of order p, κ_y denotes the concentration of audio observations, and $\mathcal{U}(y_{tm}) = (2\pi)^{-1}$ denotes the uniform distribution along the support of the unit circle.

§ Prior pdf of the assignment variables

Assuming that the assignment variables are i.i.d., the prior pdf is given by:

$$p(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} p(Z_{tm} = n), \tag{5.47}$$

and we denote with $\pi_n = p(Z_{tm} = n)$, $\sum_{n=0}^{N}\pi_n = 1$, the prior probability that source n is associated with y_{tm}.

§ Predictive pdf of the state variables

The predictive pdf extrapolates information inferred in the past to the current time step using a dynamical model of the source motion, i.e.,

$$p(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}) = \int p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1})p(\boldsymbol{s}_{t-1}|\boldsymbol{y}_{1:t-1})d\boldsymbol{s}_{t-1}. \tag{5.48}$$

where $p(s_t|s_{t-1})$ denotes the prior pdf modelling the source motion and $p(s_{t-1}|y_{1:t-1})$ corresponds to the filtering distribution at time $t-1$. The source motion model is assumed source-independent and follows a von Mises distribution:

$$p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) = \prod_{n=1}^{N}\mathcal{M}(s_{tn}; s_{t-1,n}, \kappa_d), \tag{5.49}$$

where κ_d is the concentration of the state dynamics. $\Theta = \{\kappa_y, \kappa_d, \pi_0, \ldots, \pi_N\}$ denotes the set of model parameters.

As already mentioned in Section 5.6.1, the filtering distribution corresponds to a mixture model whose number of components grows exponentially along time, therefore solving (5.44) directly is computationally intractable. Below we infer a variational approximation of (5.44) which drastically reduces the explosion of the number of mixture components; consequently, it leads to a computationally tractable algorithm.

5.6.3 Variational Approximation and Algorithm

Since solving (5.44) is computationally intractable, we propose to approximate the conditional independence between the states and the assignment variables, more precisely

$$p(\boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t}) \approx q(\boldsymbol{s}_t)q(\boldsymbol{z}_t). \tag{5.50}$$

The proposed factorization leads to a VEM algorithm [21], where the posterior distribution of the two variables are found by two variational E-steps:

$$q^{\star}(\boldsymbol{z}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t})\Big), \tag{5.51}$$

$$q^{\star}(\boldsymbol{s}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t})\Big). \tag{5.52}$$

The model parameters Θ are estimated by maximizing the expected complete-data log-likelihood:

$$Q(\Theta, \tilde{\Theta}) = \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{ \log p(\boldsymbol{y}_t, \boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t-1}, \Theta)\Big\}. \tag{5.53}$$

where $\tilde{\Theta}$ are the old parameters. To illustrate the impact of the proposed approximation on the filtering distribution, we observe that (i) the posterior pdf of the assignment is observation-independent, and that (ii) the posterior pdf of the state variables is source-independent, i.e.,

$$q(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} q(z_{tm}), \quad q(\boldsymbol{s}_t) = \prod_{n=1}^{N} q(s_{tn}). \tag{5.54}$$

Therefore, the predictive pdf is also separable:

$$p(s_{tn}|\boldsymbol{y}_{1:t-1}) = \int p(s_{tn}|s_{t-1,n}) p(s_{t-1,n}|\boldsymbol{y}_{1:t-1}) \mathrm{d}s_{t-1,n}.$$

Moreover, assuming that the filtering pdf at time $t-1$ follows a von Mises distribution, i.e. $q(s_{t-1,n}) = \mathcal{M}(s_{t-1,n}; \mu_{t-1,n}, \kappa_{t-1,n})$, then the predictive pdf is approximately a von Mises distribution (see [44], [106, (3.5.43)]):

$$p(s_{tn}|\boldsymbol{y}_{1:t-1}) \approx \mathcal{M}(s_{tn}; \mu_{t-1,n}, \tilde{\kappa}_{t-1,n}), \tag{5.55}$$

where the predicted concentration parameter, $\tilde{\kappa}_{t-1,n}$, is:

$$\tilde{\kappa}_{t-1,n} = A^{-1}(A(\kappa_{t-1,n})A(\kappa_d)), \tag{5.56}$$

and where $A(a) = I_1(a)/I_0(a)$, and $A^{-1}(a) \approx (2a - a^3)/(1 - a^2)$. Using (5.51), (5.52) and (5.53), the filtering distribution is therefore obtained by iterating through three steps, i.e. the E-S, E-Z and M steps, provided below (detailed mathematical derivations can be found in the appendices).

§ E-S step

Inserting (5.44) and (5.48) in (5.52), $q^{\star}(s_{tn})$ reduces to the von Mises distribution, $\mathcal{M}(s_{tn}; \mu_{tn}, \kappa_{tn})$. The mean μ_{tn} and concentration κ_{tn} are given by:

$$\mu_{tn} = \tan^{-1} \left(\frac{\kappa_y \sum_{m=1}^{M_t} \alpha_{tmn}\omega_{tm} \sin(y_{tm}) + \tilde{\kappa}_{t-1,n} \sin(\mu_{t-1,n})}{\kappa_y \sum_{m=1}^{M_t} \alpha_{tmn}\omega_{tm} \cos(y_{tm}) + \tilde{\kappa}_{t-1,n} \cos(\mu_{t-1,n})} \right), \tag{5.57}$$

$$\kappa_{tn} = \left((\kappa_y)^2 \sum_{m=1}^{M_t} (\alpha_{tmn}\omega_{tm})^2 + \tilde{\kappa}_{t-1,n}^2 + 2(\kappa_y)^2 \sum_{m=1}^{M_t} \sum_{l=m+1}^{M_t} \alpha_{tmn}\omega_{tm}\alpha_{tln} w_{tl} \cos(y_{tm} - y_{tl}) + 2\kappa_y \tilde{\kappa}_{t-1,n} \sum_{m=1}^{M_t} (\alpha_{tmn}\omega_{tm} \cos(y_{tm} - \mu_{t-1,n})) \right)^{1/2}, \tag{5.58}$$

where $\alpha_{tmn} = q^{\star}(Z_{tm} = n)$ denotes the variational posterior probability of the assignment variable. Therefore, the expressibility of the posterior distribution as a product of von Misses propagates over time, and only needs to be assumed at $t = 1$. Please consult Appendix A.1 for more details.

§ E-Z step

By computing the expectation over $\boldsymbol{s}_t$ in (5.51), the following expression is obtained:

$$\alpha_{tmn} = q^{\star}(z_{tm} = n) = \frac{\pi_n \beta_{tmn}}{\sum_{l=0}^{N} \pi_l \beta_{tml}} \tag{5.59}$$

where β_{tmn} is given by (please consult Appendix A.2 for a detailed derivation):

$$\beta_{tmn} = \begin{cases} \omega_{tm}\kappa_y A(\omega_{tm}\kappa_y) \cos(y_{tm} - \mu_{tn}) & n \neq 0 \\ 1/2\pi & n = 0, \end{cases}$$

with $\kappa_{tmn} = \kappa_y \omega_{tm} A(\kappa_{tn})$ for $n > 0$.

§ M step

The parameter set Θ are evaluated by maximizing (5.53). The priors (5.47) are obtained using the conventional update rule [21]: $\pi_n \propto \sum_{m=1}^{M_t} \alpha_{tnm}$. The concentration

parameters, κ_y and κ_d, are evaluated using gradient descent, namely (please consult Appendix A.3):

$$\frac{\partial Q}{\partial \kappa_y} = \sum_{m,n=1}^{M_t,N} \alpha_{tnm}\omega_{tm}\Big(A(\kappa_{tn})\cos(\mu_{tn} - y_{tm}) - A(\kappa_y\omega_{tm})\Big)$$
$$\frac{\partial Q}{\partial \kappa_d} = \sum_{n=1}^{N}\Big(A(\kappa_{tn})\cos(\mu_{tn} - \mu_{t-1,n}) - A(\tilde{\kappa}_{t-1,n})\Big)B(\kappa_d).$$

5.6.4 Audio-Source Birth Process

We now describe in detail the proposed birth process which is essential to initialize the number of audio sources as well as to detect new sources at any time. The birth process gathers all the DOAs that were not assigned to a source, i.e. assigned to $n = 0$, at current frame t as well over the L previous frames ($L = 2$ in all our experiments). From this set of DOAs we build DOA/observation sequences (one observation per frame) and let $\hat{y}^j_{t-L:t}$ be such a sequence of DOAs, where j is the sequence index. We consider the marginal likelihood:

$$\tau_j = p(\hat{y}^j_{t-L:t}) = \int p(\hat{y}^j_{t-L:t}, s_{t-L:t})\mathrm{d}s_{t-L:t}. \tag{5.60}$$

Using (5.48) and the harmonic sum theorem, the integral (5.60) becomes (please consult Appendix A.4):

$$\tau_j = \prod_{l=0}^{L} \frac{I_0(\overline{\kappa}^j_{t-l})}{2\pi I_0(\kappa_y\hat{\omega}^j_{t-l})I_0(\hat{\kappa}^j_{t-l})}, \tag{5.61}$$

where $\hat{\omega}_t$ is the confidence associated with $\hat{y}_t$. The concentration parameters, $\overline{\kappa}^j_{t-l}$ and $\hat{\kappa}^j_{t-l+1}$, depend on the observations and are recursively computed for each sequence j:

$$\overline{\kappa}^j_{t-l} = \sqrt{(\hat{\kappa}^j_{t-l})^2 + (\kappa_y\hat{\omega}^j_{t-l})^2 + \hat{\kappa}^j_{t-l}\kappa_y\hat{\omega}^j_{t-l}\cos(\hat{y}^j_{t-l} - \hat{\mu}^j_{t-l})},$$
$$\hat{\mu}^j_{t-l+1} = \tan^{-1}\left(\frac{\hat{\kappa}^j_{t-l}\sin(\hat{\mu}^j_{t-l}) + \kappa_y\hat{\omega}^j_{t-l}\sin(\hat{y}^j_{t-l})}{\hat{\kappa}^j_{t-l}\cos(\hat{\mu}^j_{t-l}) + \kappa_y\hat{\omega}^j_{t-l}\cos(\hat{y}^j_{t-l})}\right),$$
$$\hat{\kappa}^j_{t-l+1} = A^{-1}(A(\tilde{\kappa}^j_{t-l})A(\kappa_d)).$$

The sequence j^* with the maximal marginal likelihood (5.61), namely $j^* = \mathrm{argmax}_j(\tau_j)$, is supposed to be generated from a not yet known audio source only if τ_{j^*} is larger than a threshold: a new source $\tilde{n}$ is created in this case and $q(s_{t\tilde{n}}) = \mathcal{M}(s_{t\tilde{n}}; \hat{\mu}_{tj^*}, \hat{\kappa}_{tj^*})$.

5.6.5 Experimental Evaluation

The proposed method was evaluated using the audio recordings from Task 6 of the IEEE-AASP LOCATA challenge development dataset [88], which involve multiple moving sound sources and moving microphone arrays. The online sound-source localization

Method	MD (%)	FA (%)	MAE (°)
vM-PHD [107]	33.4	9.5	4.5
GM-ZO [74]	27.0	10.8	4.7
GM-FO [74]	**22.3**	6.3	3.2
vM-VEM (proposed)	23.9	**5.9**	**2.6**

Table 5.3: Method evaluation with the LOCATA dataset.

method with peak detection proposed in the previous section was used to provide DOA estimates at each STFT frame. The evaluation protocol follows 5.5.1.

The observation-to-source assignment posteriors and the DOAs confidence weights are used to estimate voice-active frames:

$$\sum_{t'=t-D}^{t}\sum_{m=1}^{M_t}\alpha_{t'mn}\omega_{t'm} \underset{silent}{\overset{active}{\gtrless}} \delta \tag{5.62}$$

where $D = 2$ and $\delta = 0.025$ is a voice activity detection (VAD) threshold. Once an active source is detected, we output its trajectory.

The MAEs, MDs and FAs values, averaged over all recordings, are summarized in Table 5.3. We compared the proposed von Mises VEM algorithm (vM-VEM) with three multi-speaker trackers: the von Mises PHD filter (vM-PHD) [107] and two versions the multiple speaker tracker of [74] based on Gaussians models (GM). [74] uses a first-order dynamic model whose effect is to smooth the estimated trajectories. We compared with both first-order (GM-FO) and zero-order (GM-ZO) dynamics. The proposed vM-VEM tracker yields the lowest false alarm (FA) rate of 5.9% and MAE of 3.2, and the second lowest MD rate of 23.9%. The GM-FO variant of [74] yields an MD rate of 22.3% since it uses velocity information to smooth the trajectories. This illustrates the advantage of the von-Mises distribution when it is applied on the directional data(DOA). The proposed von-Mises model uses only a zero-order dynamics, however, it has already achieved the competitive performance with the Gaussian model using the first-order dynamics.

The results for recordings #1 and #2 in Task 6 are shown in Fig. 5.5. Note that the PHD-based filter method [107] has two caveats. First, observation-to-source assignments cannot be estimated (unless a post-processing step is performed), and second, we also observe the source trajectories are not smooth. This stays in contrast with the proposed method which explicitly represents assignments with discrete latent variables and estimates them iteratively with VEM. Moreover, the proposed method yields smooth trajectories similar with those estimated by [74].

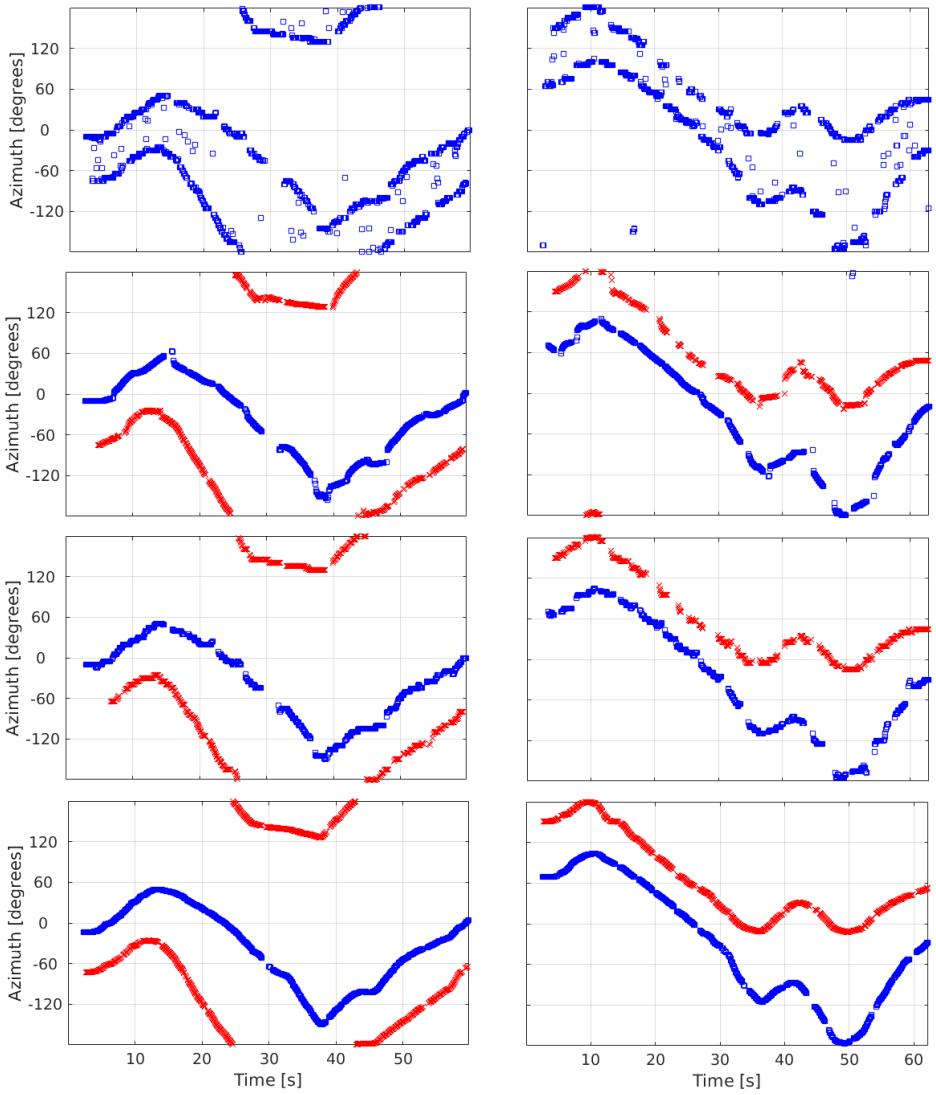


Figure 5.5: Results obtained with recordings #1 (left) and #2 (right) from Task 6 of the LOCATA dataset. Top-to-down: vM-PHD [107], GM-FO [74], vM-VEM (proposed) and ground-truth trajectories. Different colors represent different audio sources. Note that vM-PHD is unable to associate sources with trajectories.

5.7 Conclusion

In this Chapter, we proposed and combined i) a recursive DP-RTF feature estimation method, ii) an online multiple-speaker localization method, and iii) a multiple-speaker tracking method. The resulting framework provides online speaker counting, localization and consistent tracking (*i.e.* preserving speaker identity over a track despite intermittent speech production). The three algorithms are computationally efficient. In particular, the tracking algorithm implemented in the variational Bayesian framework yields a tractable solver under the form of VEM. Experiments with two datasets, recorded in

a realistic environment, verify that the proposed method is robust against reverberation and noise. Moreover, the tracker can efficiently track multiple moving speakers, detect whether there is speech or people are silent, as long as the motion associated with silent people is smooth. However, the tracking of the person from silent to active remains a difficult task. The combination of the proposed method with speaker identification will be addressed in the future.

The proposed VEM tracker based on Gaussain models can be easily adapted to work in tandem with any frame-wise localizer providing source location estimates and/or corresponding weights (and if no weights are provided by the localizer, the tracker can be applied with all weights set to one). Such characteristics makes the proposed tracker very flexible, and easily reusable by the audio processing research community.

We also proposed a multiple audio-source tracking method using the von Mises distribution and we further inferred a tractable solver based on a variational approximation of the posterior filtering distribution. Unlike the Gaussian distribution, the von Mises distribution explicitly models the circular variables associated with audio-source localization and tracking based on source DOAs. Using the recently released LOCATA dataset, we empirically showed that the proposed method compares favorably with two recent methods.

CHAPTER 6

CONCLUSION

6.1 SUMMARY

This thesis focuses on developing methodologies for robots perception. It includes several contributions. Firstly, we developed a variational Bayesian framework for multiple-object tracking. Object-birth and death process are jointly proposed with the model to deal with time-varying objects. The framework is very flexible and is simple to adapt to different tasks. The method was first applied on visual multiple-person tracking and was benchmarked on MOT16 dataset. We then employ the method on a robot platform. We further calibrated the robot head-joint motor with robot camera, which allows us to (i) compensate the robot's large ego-motion while tracking (ii) control the motor based on the visual information. Furthermore, using the same framework, we exploit the complementarity of audio and visual information, to accurately estimate smooth trajectories of the tracked persons, to deal with the partial or total absence of one of the modalities over short periods of time, and to estimate the acoustic status – either speaking or silent – of each tracked person along with time. In addition, we focus on online acoustic localization and tracking multiple speakers. We propose to online estimate the direct-path relative transfer function via exponentiated gradient descent (EG-DPRTF). The localization results are then fed into a Bayesian filtering framework. Birth process and the speaker activity detection are employed to account for the intermittent nature of speech. Moreover, we propose to extend the variational Bayesian formulas to use the von Mises distribution. Because the circular distribution better fits the directional data (DoAs) compared with Gaussian models. The formulas showed the extensibility of the proposed variational framework.

6.2 FUTURE RESEARCH DIRECTIONS

From this thesis, there are several possible future research directions:

- Chapter 3: A probabilistic model is presented for visual pedestrian tracking. However, the geometry information and appearance information are combined in a relatively simple way. As the deep learning became the most popular methodology, neural networks can be used for data association. It is possible to fuse the appearance and the geometry information into the same neural network and build an end-to-end framework for tracking. In the second section, we calibrated the robot head-joint motor with the robot camera to compensate active motion while tracking. Instead of calibrating with robot head-joint motor, we can put a gyroscope on the robot and calibrate the robot camera with the gyroscope. A gyroscope is a device which can detect the orientation and the angular velocity. Therefore most of the robot motion can be detected, which makes the vision application robust to different kinds of robot motion.
- Chapter 4: We focus on combining different modalities, especially . The data associations between acoustic feature frequency bins and images pixels are exploited. In the proposed work, such an association is used to track the active speaker, to detect the speaking activity of a person. However, assigning an audio frequency bin to a speaking person can do more tasks. One possible direction is to extend the proposed model to separate simultaneously emitting sound sources. Another possibility is to select a speaking person and beamform his/her voice.
- Chapter 5: We present an acoustic multiple-speaker tracking model. The information that data association uses is the speaker location information. However, with the development of deep learning techniques, features to identify different human voices became available. Incorporating such information into the model would be interesting. Moreover, in Chapter 5, we also presented a tracking model with von-Mises distribution. But due to some mathematical constraints, the proposed model uses only a zero-order dynamic model. However, using the variational framework, it is possible to extend the model to the first-order dynamic model with the same mathematical constraint, which will largely improve the tracking performance during speaker occlusions.

CHAPTER A

APPENDIX: VARIATIONAL VON-MISES TRACKING MODEL

A.1 DERIVATION OF THE E-S STEP

In order to obtain the formulae for the E-S step, we start from its definition in (5.52):

$$q^{\star}(\boldsymbol{s}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t})\Big). \tag{A.1}$$

We now use the decomposition in (5.44), to write:

$$q^{\star}(\boldsymbol{s}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{y}_t|\boldsymbol{s}_t, \boldsymbol{z}_t)\Big) p(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}). \tag{A.2}$$

Let us know develop the expectation:

$$
\begin{aligned}
&\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{y}_t|\boldsymbol{s}_t, \boldsymbol{z}_t) \\
&= \mathrm{E}_{q(\boldsymbol{z}_t)} \sum_{m=1}^{M_t} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}) \\
&= \sum_{m=1}^{M_t} \mathrm{E}_{q(z_{tm})} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}) \\
&= \sum_{m=1}^{M_t} \mathrm{E}_{q(z_{tm})} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}) \\
&= \sum_{m=1}^{M_t} \sum_{n=0}^{N} q(z_{tm} = n) \log p(y_{tm}|\boldsymbol{s}_t, z_{tm} = n) \\
&= \sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm} \log p(y_{tm}|s_{tn}, z_{tm} = n) \\
&= \sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm} \log \mathcal{M}(y_{tm}; s_{tn}, \omega_{tm}\kappa_y) \\
&\overset{\boldsymbol{s}_t}{=} \sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm}\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn}),
\end{aligned}
$$

where $\overset{\boldsymbol{s}_t}{=}$ denotes the equality up to an additive constant that does *not* depend on $\boldsymbol{s}_t$. Such a constant would become a multiplicative constant after the exponentiation in (A.2), and therefore can be ignored.

By replacing the developed expectation together with (5.55) we obtain:

$$
\begin{aligned}
q^\star(\boldsymbol{s}_t) \propto \exp\Big(\sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm}\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn}) \Big) \\
\prod_{n=0}^{N} \mathcal{M}(\boldsymbol{s}_{tn}; \mu_{t-1,n}, \tilde{\kappa}_{t-1,n}),
\end{aligned}
$$

which can be easily rewritten as:

$$
\begin{aligned}
q^\star(\boldsymbol{s}_t) \propto \prod_{n=0}^{N} \exp\Big(\sum_{m=1}^{M_t} \alpha_{tnm}\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn}) \\
+ \tilde{\kappa}_{t-1,n} \cos(s_{tn} - \mu_{t-1,n}) \Big).
\end{aligned}
$$

This is important since it demonstrates that the a posteriori distribution on $\boldsymbol{s}_t$ is separable on n and therefore independent for each speaker. In addition, it allows us to rewrite

the a posteriori distribution for each speaker, i.e. on s_{tn} as a von Mises distribution by using the harmonic addition theorem, thus obtaining

$$q^\star(\boldsymbol{s}_t) = \prod_{n=0}^{N} q^\star(s_{tn}) = \prod_{n=0}^{N} \mathcal{M}(s_{tn}; \mu_{tn}, \kappa_{tn}), \tag{A.3}$$

with μ_{tn} and κ_{tn} defined as in (14) and (15).

A.2 Derivation of the E-Z step

Similarly to the previous section, and in order to obtain the closed-form solution of the E-Z step, we start from its definition in (5.51):

$$q^\star(\boldsymbol{z}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t})\Big), \tag{A.4}$$

and we use the decomposition in (5.44) to write:

$$q^\star(\boldsymbol{z}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(\boldsymbol{y}_t | \boldsymbol{s}_t, \boldsymbol{z}_t)\Big) p(\boldsymbol{z}_t). \tag{A.5}$$

Since both the observation likelihood and the prior distribution are separable on z_{tm}, we can write:

$$q^\star(\boldsymbol{z}_t) \propto \prod_{m=1}^{M_t} \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(y_{tm} | \boldsymbol{s}_t, z_{tm})\Big) p(z_{tm}), \tag{A.6}$$

proving that the a posteriori distribution is also separable on m.

We can thus analyze the posterior of each z_{tm} separately, by computing $q^\star(z_{tm} = n)$:

$$q^\star(z_{tm} = n) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(y_{tm} | \boldsymbol{s}_t, z_{tm} = n)\Big) p(z_{tm} = n)$$

Let us first compute the expectation for $n \neq 0$:

$$\begin{aligned}
&\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(y_{tm} | \boldsymbol{s}_t, z_{tm} = n) \\
&= \mathrm{E}_{q(s_{tn})} \log p(y_{tm} | s_{tn}, z_{tm} = n) \\
&= \mathrm{E}_{q(s_{tn})} \log \mathcal{M}(y_{tm}; s_{tn}, \omega_{tm}\kappa_y) \\
&\overset{z_{tm}}{=} \int_0^{2\pi} q(s_{tn}) \omega_{tm} \kappa_y \cos(y_{tm} - s_{tn}) \mathrm{d}s_{tn} \\
&= \frac{\omega_{tm}\kappa_y}{2\pi I_0(\omega_{tm}\kappa_y)} \int_0^{2\pi} \exp\Big(\cos(s_{tn} - \mu_{tn})\Big) \cos(s_{tn} - y_{tm}) \mathrm{d}s_{tn} \\
&= \omega_{tm}\kappa_y A(\omega_{tm}\kappa_y) \cos(y_{tm} - \mu_{tn}),
\end{aligned}$$

where for the last line we used the following variable change $\bar{s} = s_{tn} - \mu_{tn}$ and the definition of I_1 and A.

The case $n = 0$ is even easier since the observation distribution is a uniform: $\mathrm{E}_{q(s_{tn})} \log p(y_{tm}|s_{tn}, z_{tm} = n) = \mathrm{E}_{q(s_{tn})} - \log 2\pi = -\log(2\pi)$.

By using the fact that the prior distribution on z_{tm} is denoted by $p(z_{tm} = n) = \pi_n$, we can now write the a posteriori distribution as $q^\star(z_{tm} = n) \propto \pi_n \beta_{tmn}$ with:

$$\beta_{tmn} = \begin{cases} \omega_{tm}\kappa_y A(\omega_{tm}\kappa_y)\cos(y_{tm} - \mu_{tn}) & n \neq 0 \\ 1/2\pi & n = 0 \end{cases},$$

and finally obtaining the results in (5.59) and (5.46).

A.3 DERIVATION OF THE M STEP

In order to derive the M step, we need first to compute the Q function in (5.53),

$$\begin{aligned} Q(\Theta, \tilde{\Theta}) &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{ \log p(\boldsymbol{y}_t, \boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t-1}, \Theta)\Big\} \\ &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{ \underbrace{\log p(\boldsymbol{y}_t|\boldsymbol{s}_t, \boldsymbol{z}_t, \Theta)}_{\kappa_y} + \\ &= + \underbrace{\log p(\boldsymbol{z}_t|\Theta)}_{\pi_n' s} + \underbrace{\log p(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}, \Theta)}_{\kappa_d}\Big\}, \end{aligned}$$

where each parameter is show below the corresponding term of the Q function. Let us develop each term separately.

A.3.1 OPTIMIZING κ_y

$$\begin{aligned} Q_{\kappa_y} &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{ \log \prod_{m=1}^{M_t} p(y_{tm}|\boldsymbol{s}_t, z_{tm})\Big\} \\ &= \sum_{m=1}^{M_t} \mathrm{E}_{q(\boldsymbol{s}_t)q(z_{tm})}\Big\{ \log p(y_{tm}|\boldsymbol{s}_t, z_{tm})\Big\} \\ &= \sum_{m=1}^{M_t} \mathrm{E}_{q(\boldsymbol{s}_t)} \sum_{n=0}^{N} \alpha_{tmn}\Big\{ \log p(y_{tm}|\boldsymbol{s}_t, z_{tm} = n)\Big\} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N} \alpha_{tmn}\mathrm{E}_{q(s_{tn})}\Big\{ \log \mathcal{M}(y_{tm}; s_{tn}, \omega_{tm}\kappa_y)\Big\} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N} \alpha_{tmn} \int_0^{2\pi} q(s_{tn})(\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn}) - \log(I_0(\omega_{tm}\kappa_y)))\mathrm{d}s_{tn} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N} \alpha_{tmn}\Big(\omega_{tm}\kappa_y \cos(y_{tm} - \mu_{tn})A(\kappa_{tn}) - \log(I_0(\omega_{tm}\kappa_y))\Big), \end{aligned}$$

and by taking the derivative with respect to κ_y we obtain:

$$\frac{\partial Q}{\partial \kappa_y} = \sum_{m=1}^{M_t}\sum_{n=0}^{N} \alpha_{tmn}\omega_{tm}\Big(\cos(y_{tm}-\mu_{tn})A(\kappa_{tn}) - A(\omega_{tm}\kappa_y)\Big),$$

which corresponds to what was announced in the manuscript.

A.3.2 Optimizing π_n's

$$\begin{aligned}
Q_{\pi_n} &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{\log\prod_{m=1}^{M_t} p(z_{tm})\Big\} \\
&= \sum_{m=1}^{M_t}\mathrm{E}_{q(z_{tm})}\Big\{\log p(z_{tm})\Big\} \\
&= \sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\Big\{\log p(z_{tm}=n)\Big\} \\
&= \sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\Big\{\log \pi_n\Big\}
\end{aligned}$$

This is the very same formulae that for any mixture model, and therefore the solution is standard and corresponds to the one reported in the manuscript.

A.3.3 Optimizing κ_d

$$\begin{aligned}
Q_{\kappa_d} &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{\log\prod_{n=1}^{N} p(s_{tn}|\boldsymbol{y}_{1:t-1})\Big\} \\
&= \sum_{n=1}^{N}\mathrm{E}_{q(s_{tn})}\Big\{\log\mathcal{M}(s_{tn};\mu_{t-1,n},\tilde{\kappa}_{t-1,n})\Big\} \\
&= \sum_{n=1}^{N}\mathrm{E}_{q(s_{tn})}\Big\{-\log I_0(\tilde{\kappa}_{t-1,n}) + \tilde{\kappa}_{t-1,n}\cos(s_{tn}-\mu_{t-1,n})\Big\} \\
&= \sum_{n=1}^{N} -\log I_0(\tilde{\kappa}_{t-1,n}) + \tilde{\kappa}_{t-1,n}\cos(\mu_{tn}-\mu_{t-1,n})A(\kappa_{tn}),
\end{aligned}$$

where the dependency on κ_d is implicit in $\tilde{\kappa}_{t-1,n} = A^{-1}(A(\kappa_{t-1,n})A(\kappa_d))$.

By taking the derivative with respect to κ_d we obtain:

$$\frac{\partial Q}{\partial \kappa_d} = \sum_{n=1}^{N}\Big(A(\kappa_{tn})\cos(\mu_{tn}-\mu_{t-1,n}) - A(\tilde{\kappa}_{t-1,n})\Big)\frac{\partial\tilde{\kappa}_{t-1,n}}{\partial\kappa_d}$$

with

$$\frac{\partial \tilde{\kappa}_{t-1,n}}{\partial \kappa_d} = \tilde{A}(A(\kappa_{t-1,n})A(\kappa_d))A(\kappa_{t-1,n})\frac{I_2(\kappa_d)I_0(\kappa_d) - I_1^2(\kappa_d)}{I_0^2(\kappa_d)},$$

where $\tilde{A}(a) = \mathrm{d}A^{-1}(a)/\mathrm{d}a = (2 - a^2 + a^4)/(1 - a^2)^2$.

By denoting the previous derivative as $B(\kappa_d) = \frac{\partial \tilde{\kappa}_{t-1,n}}{\partial \kappa_d}$, we obtain the expression in the manuscript.

A.4 DERIVATION OF THE BIRTH PROBABILITY

In this section we derive the expression for τ_j by computing the integral (5.60). Using the probabilistic model defined, we can write (the index j is omitted):

$$\begin{aligned}
&\int p(\hat{y}_{t-L:t}, s_{t-L:t})\mathrm{d}s_{t-L:t} \\
&\quad = \int \prod_{\tau=-L}^{0} p(\hat{y}_{t+\tau}|s_{t+\tau}) \prod_{\tau=-L+1}^{0} p(s_{t+\tau}|s_{t+\tau-1})p(s_{t-L})\mathrm{d}s_{t-L:t}
\end{aligned}$$

We will first marginalize s_{t-L}. To do that, we notice that if $p(s_{t-L})$ follows a von Mises with mean $\hat{\mu}_{t-L}$ and concentration $\hat{\kappa}_{t-L}$, then we can write:

$$\begin{aligned}
&p(\hat{y}_{t-L}|s_{t-L})p(s_{t-L}) \\
&\quad = \mathcal{M}(\hat{y}_{t-L}; s_{t-L}, \hat{\omega}_{t-L}\kappa_y)\mathcal{M}(s_{t-L}; \hat{\mu}_{t-L}, \hat{\kappa}_{t-L}) \\
&\quad = \mathcal{M}(s_{t-L}; \bar{\mu}_{t-L}, \bar{\kappa}_{t-L})\frac{I_0(\bar{\kappa}_{t-L})}{2\pi I_0(\hat{\omega}_{t-L}\kappa_y)I_0(\hat{\kappa}_{t-L})}
\end{aligned}$$

with

$$\begin{aligned}
\bar{\mu}_{t-L} &= \tan^{-1}\left(\frac{\hat{\omega}_{t-L}\kappa_y \sin \hat{y}_{t-L} + \hat{\kappa}_{t-L}\sin\hat{\mu}_{t-L}}{\hat{\omega}_{t-L}\kappa_y \cos \hat{y}_{t-L} + \hat{\kappa}_{t-L}\cos\hat{\mu}_{t-L}}\right), \\
\bar{\kappa}_{t-L}^2 &= (\hat{\omega}_{t-L}\kappa_y)^2 + \hat{\kappa}_{t-L}^2 + 2\hat{\omega}_{t-L}\kappa_y\hat{\kappa}_{t-L}\cos(\hat{y}_{t-L} - \hat{\mu}_{t-L}),
\end{aligned}$$

where we used the harmonic addition theorem.

Now we can effectively compute the marginalization. The two terms involving s_{t-L} are:

$$\begin{aligned}
&\int \mathcal{M}(s_{t-L+1}; s_{t-L}, \kappa_d)\mathcal{M}(s_{t-L}; \bar{\mu}_{t-L}, \bar{\kappa}_{t-L})ds_{t-L} \\
&\approx \mathcal{M}(s_{t-L+1}; \hat{\mu}_{t-L+1}, \hat{\kappa}_{t-L+1})
\end{aligned}$$

with

$$\begin{aligned}
\hat{\mu}_{t-L+1} &= \bar{\mu}_{t-L}, \\
\hat{\kappa}_{t-L+1} &= A^{-1}(A(\bar{\kappa}_{t-L})A(\kappa_d)).
\end{aligned}$$

Therefore, the marginalization with respect to s_{t-L} yields the following result:

$$\begin{aligned}
&\int p(\hat{y}_{t-L:t}, s_{t-L:t})\mathrm{d}s_{t-L:t} \\
&\quad = \int \prod_{\tau=-L}^{0} p(\hat{y}_{t+\tau}|s_{t+\tau}) \prod_{\tau=-L+1}^{0} p(s_{t+\tau}|s_{t+\tau-1})p(s_{t-L})\mathrm{d}s_{t-L:t} \\
&\quad = \frac{I_0(\bar{\kappa}_{t-L})}{2\pi I_0(\hat{\omega}_{t-L}\kappa_y)I_0(\hat{\kappa}_{t-L})} \int \prod_{\tau=-L+1}^{0} p(\hat{y}_{t+\tau}|s_{t+\tau})\times \\
&\qquad \prod_{\tau=-L+2}^{0} p(s_{t+\tau}|s_{t+\tau-1})p(s_{t-L+1})\mathrm{d}s_{t-L+1:t}.
\end{aligned}$$

Since we have already seen that $p(s_{t-L+1})$ is also a von Mises distribution, we can use the same reasoning to marginalize with respecto to s_{t-L+1}. This strategy yields to the recursion presented in the main text.

A.5 PUBLICATIONS AND SUBMISSIONS

Here are is the list of papers that have been published or submitted during my PhD.

JOURNAL SUBMISSIONS:

- [10] **Yutong Ban**, Xavier Alameda-Pineda, Christine Evers, Radu Horaud, Tracking Multiple Audio Sources with the von Mises Distribution and Variational EM, *IEEE Signal Processing Letters (March 2019)*.
- [11] **Yutong Ban**, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud, Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers, *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (September 2018)*.
- [74] Xiaofei Li*****, **Yutong Ban***, Laurent Girin, Xavier Alameda-Pineda, Radu Horaud, Online Localization and Tracking of Multiple Speakers in Reverberant Environments, *IEEE Journal on Selected Topics in Signal Processing (Accepted on February 2019) (* indicates the equally contributed authors)*.

CONFERENCE PAPERS:

- [14] **Yutong Ban**, Xiaofei Li, Xavier Alameda-Pineda, Laurent Girin, Radu Horaud, Accounting for Room Acoustics in Audio-Visual Multi-Speaker Tracking, In *IEEE International Conference on Acoustics, Speech and Signal Processing (***ICASSP***), Apr 2018, Calgary, Alberta, Canada*.
- [13] **Yutong Ban**, Laurent Girin, Xavier Alameda-Pineda, Radu Horaud, Exploiting the Complementarity of Audio and Visual Data in Multi-Speaker Tracking, *ICCV Workshop on Computer Vision for Audio-Visual Media, Oct 2017, Venezia, Italy*.
- [9] **Yutong Ban**, Xavier Alameda-Pineda, Fabien Badeig, Sileye Ba, Radu Horaud, Tracking a Varying Number of People with a Visually-Controlled Robotic Head, *IEEE/RSJ International Conference on Intelligent Robots and Systems (***IROS***), Sep 2017, Vancouver, Canada*.
- [12] **Yutong Ban**, Sileye Ba, Xavier Alameda-Pineda, Radu Horaud, Tracking Multiple Persons Based on a Variational Bayesian Model, *IEEE ECCV Workshop on Benchmarking Mutliple Object Tracking, Oct 2016, Amsterdam, Netherlands*.

PREPRINTS:

- [158] Yihong Xu, **Yutong Ban**, Xavier Alameda-Pineda, Radu Horaud, DeepMOT: A Differentiable Framework for Training Multiple Object Trackers, *arXiv preprint, 2019*.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

BIBLIOGRAPHY

[1] Don Joven Agravante, Jordi Pages, and François Chaumette. Visual servoing for the reem humanoid robot's upper body. In *IEEE International Conference on Robotics and Automation*, 2013.

[2] Xavier Alameda-Pineda and Radu Horaud. A geometric approach to sound source localization from time-delay estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(6):1082–1095, 2014.

[3] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[4] Marco Antonelli, Angel P Del Pobil, and Michele Rucci. Bayesian multimodal integration in a robot replicating human head and eye movements. In *IEEE International Conference on Robotics and Automation*, 2014.

[5] Yekutiel Avargel and Israel Cohen. System identification in the short-time Fourier transform domain with crossband filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1305–1319, 2007.

[6] Sileye Ba, Xavier Alameda-Pineda, Alessio Xompero, and Radu Horaud. An online variational Bayesian model for multi-person tracking from cluttered scenes. *Computer Vision and Image Understanding*, 2016.

[7] Sileye Ba, Xavier Alameda-Pineda, Alessio Xompero, and Radu Horaud. An online variational Bayesian model for multi-person tracking from cluttered scenes. *Computer Vision and Image Understanding*, 153:64–76, 2016.

[8] S. Bae and K. Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):595–610, March 2018.

[9] Yutong Ban, Xavier Alameda-Pineda, Fabien Badeig, Sileye Ba, and Radu Horaud. Tracking a varying number of people with a visually-controlled robotic head. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4144–4151. IEEE, 2017.

[10] Yutong Ban, Xavier Alameda-PIneda, Christine Evers, and Radu Horaud. Tracking multiple audio sources with the von mises distribution and variational em. *arXiv preprint arXiv:1812.08246*, 2018.

[11] Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Variational bayesian inference for audio-visual tracking of multiple speakers. *arXiv preprint arXiv:1809.10961*, 2018.

[12] Yutong Ban, Sileye Ba, Xavier Alameda-Pineda, and Radu Horaud. Tracking multiple persons based on a variational Bayesian model. In *European Conference on Computer Vision Workshops*, pages 52–67, Amsterdam, Netherlands, 2016.

[13] Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud. Exploiting the complementarity of audio and visual data in multi-speaker tracking. In *IEEE ICCV Workshop on Computer Vision for Audio-Visual Media*, pages 446–454, Venezia, Italy, October 2017.

[14] Yutong Ban, Xiaofei Li, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Accounting for room acoustics in audio-visual multi-speaker tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Alberta, Canada, April 2018.

[15] Y. Bar-Shalom, F. Daum, and J. Huang. The probabilistic data association filter: estimation in the presence of measurement origin and uncertainty. *IEEE Control System Magazine*, 29(6):82–100, 2009.

[16] Yaakov Bar-Shalom. Multitarget-multisensor tracking: advanced applications. *Norwood, MA, Artech House, 1990, 391 p.*, 1990.

[17] Yaakov Bar-Shalom, Peter K Willett, and Xin Tian. *Tracking and data fusion*. YBS publishing Storrs, CT, USA:, 2011.

[18] D Bechler, M Grimm, and K Kroschel. Speaker tracking with a microphone array using Kalman filtering. *Advances in Radio Science*, 1(B. 3):113–117, 2003.

[19] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[20] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[21] Christopher M Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[22] Samuel Blackman and Robert Popoli. Design and analysis of modern tracking systems (artech house radar library). *Artech house*, 1999.

[23] Samuel S Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):5–18, 2004.

[24] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, Hawaii, USA, 2017.

[25] Volkan Cevher, Rajbabu Velmurugan, and James H McClellan. Acoustic multi-target tracking using direction-of-arrival batches. *IEEE Transactions on Signal Processing*, 55(6):2810–2825, 2007.

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

[27] Jingdong Chen, Jacob Benesty, and Yiteng Huang. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on applied signal processing*, 2006:170–170, 2006.

[28] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *IEEE International Conference on Computer Vision*, 2015.

[29] Wongun Choi, Caroline Pantofaru, and Silvio Savarese. A general framework for tracking multiple people from a moving camera. *IEEE transactions on pattern analysis and machine intelligence*, 35(7), 2013.

[30] D.E. Clark and J. Bell. Convergence results for the particle PHD filter. *IEEE Transactions on Signal Processing*, 54(7):2652–2661, 2006.

[31] Ingemar J. Cox and Sunita L. Hingorani. An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 18(2):138–150, 1996.

[32] Armel Cretual and François Chaumette. Application of motion-based visual servoing to target tracking. *The International Journal of Robotics Research*, 20(11), 2001.

[33] Armel Crétual, François Chaumette, and Patrick Bouthemy. Complex object tracking by visual servoing based on 2d image motion. In *International Conference on Pattern Recognition*, volume 2, 1998.

[34] A. Deleforge, F. Forbes, and R. Horaud. High-dimensional regression with Gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911, 2015.

[35] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin. Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(4):718–731, 2015.

[36] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[37] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein. Robust localization in reverberant rooms. In Michael S Brandstein and Darren Ward, editors, *Microphone Arrays*, pages 157–180. Springer, 2001.

[38] Caglayan Dicle, Octavia I Camps, and Mario Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE international conference on computer vision*, pages 2304–2311, 2013.

[39] Simon Doclo and Marc Moonen. Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP Journal on Applied Signal Processing*, 2003:1110–1124, 2003.

[40] Yuval Dorfan and Sharon Gannot. Tree-based recursive expectation-maximization algorithm for localization of acoustic sources. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(10):1692–1703, 2015.

[41] Yuval Dorfan and Sharon Gannot. Tree-based recursive expectation-maximization algorithm for localization of acoustic sources. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(10):1692–1703, 2015.

[42] Tsvi G Dvorkind and Sharon Gannot. Time difference of arrival estimation of speech source in a noisy and reverberant environment. *Signal Processing*, 85(1):177–204, 2005.

[43] Bernard Espiau, François Chaumette, and Patrick Rives. A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation*, 8(3), 1993.

[44] Christine Evers, Emanuel AP Habets, Sharon Gannot, and Patrick A Naylor. DoA reliability for distributed acoustic tracking. *IEEE Signal Processing Letters*, 2018.

[45] Christine Evers, Alastair H Moore, Patrick A Naylor, Jonathan Sheaffer, and Boaz Rafaely. Bearing-only acoustic tracking of moving speakers for robot audition. In *IEEE International Conference on Digital Signal Processing (DSP)*, pages 1206–1210, 2015.

[46] Christine Evers and Patrick A Naylor. Acoustic SLAM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1484–1498, 2018.

[47] Loïc Fagot-Bouquet, Romaric Audigier, Yoann Dhome, and Frédéric Lerasle. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. In *European Conference on Computer Vision*, 2016.

[48] Maurice F Fallon and Simon 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, 2012.

[49] Chris Gaskett, Luke Fletcher, Alexander Zelinsky, et al. Reinforcement learning for visual servoing of a mobile robot. In *Australian Conference on Robotics and Automation*, 2000.

[50] 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(2):601–616, 2007.

[51] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud. EM algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2402–2415, 2016.

[52] Israel Gebru, Sileye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal Bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[53] Israel D. Gebru, Silèye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal Bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1086–1099, 2018.

[54] Israel Dejene Gebru, Xavier Alameda-Pineda, Florence Forbes, and Radu Horaud. Em algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2402–2415, 2016.

[55] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):1012–1025, 2014.

[56] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[57] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[58] P. J. Green. Trans-dimensional Markov chain Monte Carlo. In *Oxford Statistical Science Series*, pages 179–198, 2003.

[59] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 3047–3055, 2015.

[60] A. Heili, A. Lopez-Mendez, and J.-M Odobez. Exploiting long-term connectivity and visual motion in CRF-based multi-person tracking. *IEEE Transactions on Image Processing*, 23(7):3040–3056, 2014.

[61] T.M. Hospedales and S. Vijayakumar. Structure inference for Bayesian multisensory scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2140–2157, 2008.

[62] Yiteng Huang and Jacob Benesty. Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization. In *Adaptive Signal Processing*, pages 227–247. Springer, 2003.

[63] Seth Hutchinson, Gregory D Hager, and Peter I Corke. A tutorial on visual servo control. *IEEE Transactions on Robotics and Automation*, 12(5), 1996.

[64] Carlos T Ishi, Olivier Chatot, Hiroshi Ishiguro, and Norihiro Hagita. Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2027–2032, 2009.

[65] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[66] Z. Khan, T. Balch, and F. Dellaert. An MCMC-based particle filter for tracking multiple interacting targets. In *European Conference on Computer Vision*, pages 279–290, Prague, Czech Republic, 2004.

[67] Volkan Kılıç, Mark Barnard, Wenwu Wang, Adrian Hilton, and Josef Kittler. Mean-shift and sparse sampling-based SMC-PHD filtering for audio informed visual speaker tracking. *IEEE Transactions on Multimedia*, 18(12):2417, 2016.

[68] Volkan Kılıç, Mark Barnard, Wenwu Wang, and Josef Kittler. Audio assisted robust visual tracking with adaptive particle filtering. *IEEE Transactions on Multimedia*, 17(2):186–200, 2015.

[69] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[70] Charles Knapp and G Clifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.

[71] Konrad Kowalczyk, Emanuël AP Habets, Walter Kellermann, and Patrick A Naylor. Blind system identification using sparse learning for TDOA estimation of room reflections. *IEEE Signal Processing Letters*, 20(7):653–656, 2013.

[72] T Kuiren. Issues in the design of practical multitarget tracking algorithms. *Multitarget-multisensor tracking: advanced applications*, pages 43–87, 1990.

[73] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez. AV16.3: an audio-visual corpus for speaker localization and tracking. In *Machine Learning for Multimodal Interaction*, pages 182–195. Springer, 2004.

[74] Xiaofei Li, Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud. Online localization and tracking of multiple moving speakers in reverberant environment. *ArXiV preprint*, 2018.

[75] Xiaofei Li, Laurent Girin, Fabien Badeig, and Radu Horaud. Reverberant sound localization with a robot head based on direct-path relative transfer function. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2819–2826. IEEE, 2016.

[76] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot. Estimation of relative transfer function in the presence of stationary noise based on segmental power

spectral density matrix subtraction. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 320–324, 2015.

[77] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot. Estimation of the direct-path relative transfer function for supervised sound-source localization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2171–2186, 2016.

[78] Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot. Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1997–2012, 2017.

[79] Xiaofei Li, Laurent Girin, Radu Horaud, Sharon Gannot, Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot. Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(10):1997–2012, 2017.

[80] Xiaofei Li, Radu Horaud, Laurent Girin, and Sharon Gannot. Voice activity detection based on statistical likelihood ratio with adaptive thresholding. In *IEEE International Workshop on Acoustic Signal Enhancement*, pages 1–5, 2016.

[81] Xiaofei Li, Bastien Mourgue, Laurent Girin, Sharon Gannot, and Radu Horaud. Online localization of multiple moving speakers in reverberant environments. In *The Tenth IEEE Workshop on Sensor Array and Multichannel Signal Processing*, 2018.

[82] Zhiwei Liang, Xudong Ma, and Xianzhong Dai. Robust tracking of moving sound source using multiple model Kalman filter. *Applied acoustics*, 69(12):1350–1355, 2008.

[83] Yang Liu, Adrian Hilton, Jonathon Chambers, Yuxin Zhao, and Wenwu Wang. Non-zero diffusion particle flow smc-phd filter for audio-visual multi-speaker tracking. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[84] Yang Liu, Qinghua Hu, Zou Yuexian, and Wenwu Wang. Labelled non-zero particle flow for smc-phd filtering. *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[85] Yang Liu, Wenwu Wang, Jonathon Chambers, Volkan Kilic, and Adrian Hilton. Particle flow SMC-PHD filter for audio-visual multi-speaker tracking. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 344–353, 2017.

[86] Yang Liu, Wenwu Wang, and Volkan Kilic. Intensity particle flow smc-phd filter for audio speaker tracking. *LOCATA Workshop*, 2018.

[87] Yang Liu, Wenwu Wang, and Yuxin Zhao. Particle flow for sequential monte carlo implementation of probability hypothesis density. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4371–4375. IEEE, 2017.

[88] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann. The LOCATA challenge data corpus for acoustic source localization and tracking. In *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, July 2018.

[89] Anthony Lombard, Yuanhang Zheng, Herbert Buchner, and Walter Kellermann. TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1490–1503, 2011.

[90] W. Longyin, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, Ming-H. Yang, and S. Lyu. DETRAC filter multiple target tracker: A new benchmark and protocol for multi-object tracking. *arXiv:1511.04136*, 2015.

[91] W. Luo, J. Xing, X. Zhang, W. Zhao, and T.-K. Kim. Multiple object tracking: a review, 2015. arXiv:1409.761.

[92] W. K. Ma, B. N. Vo, and S. S. Singh. 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, 2006.

[93] Wing-Kin Ma, Ba-Ngu Vo, Sumeetpal S Singh, and Adrian 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, 2006.

[94] Yanna Ma and Akinori Nishihara. Efficient voice activity detection algorithm using long-term spectral flatness measure. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–18, 2013.

[95] 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, 2008.

[96] R. P. S. Mahler. A theoretical foundation for the Stein-Winter" probability hypothesis density (PHD)" multitarget tracking approach. Technical report, 2000.

[97] R. P. S. Mahler. Multitarget Bayes filtering via first-order multitarget moments. *IEEE Trans. Aerosp. Electron. Syst.*, 39(4):1152–1178, October 2003.

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

[99] R. P. S. Mahler. Statistics 102 for multisensor multitarget data fusion. *IEEE Selected Topics on Signal Processing*, 19(1):53–64, 2013.

[100] Ronald Mahler. Phd filters of higher order in target number. *IEEE Transactions on Aerospace and Electronic systems*, 43(4), 2007.

[101] Ronald P.S. Mahler. Multisource multitarget filtering: a unified approach. In *Aerospace/Defense Sensing and Controls*, pages 296–307. International Society for Optics and Photonics, 1998.

[102] Ronald PS Mahler. Multitarget bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic systems*, 39(4):1152–1178, 2003.

[103] Ezio Malis, Francois Chaumette, and Sylvie Boudet. 2 1/2 d visual servoing. *IEEE Transactions on Robotics and Automation*, 15(2), 1999.

[104] Michael I Mandel, Ron J Weiss, and Daniel PW Ellis. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394, 2010.

[105] Éric Marchand and François Chaumette. Feature tracking for visual servoing purposes. *Robotics and Autonomous Systems*, 52(1), 2005.

[106] Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

[107] Ivan Marković, Josip Ćesić, and Ivan Petrović. Von Mises mixture PHD filter. *IEEE Signal Processing Letters*, 22(12):2229–2233, 2015.

[108] Ivan Marković and Ivan Petrović. Bearing-only tracking with a mixture of von Mises distributions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 707–712. IEEE, 2012.

[109] Luis Mejias, Srikanth Saripalli, Pascual Campoy, and Gaurav S Sukhatme. Visual servoing of an autonomous helicopter in urban areas using feature tracking. *Journal of Field Robotics*, 23(3-4), 2006.

[110] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. In *arXiv:1603.00831 [cs]*, 2016.

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

[112] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):58–72, 2014.

[113] Vicente Peruffo Minotto, Claudio Rosito Jung, and Bowon Lee. Multimodal multi-channel on-line speaker diarization using sensor fusion through SVM. *IEEE Transactions on Multimedia*, 17(10):1694–1705, 2015.

[114] Amine Abou Moughlbay, Enric Cervera, and Philippe Martinet. Error regulation strategies for model based visual servoing tasks: Application to autonomous object grasping with nao robot. In *International Conference on Control Automation Robotics & Vision*, 2012.

[115] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

[116] S.M. Naqvi, Miao Yu, and J.A. Chambers. A multimodal approach to blind source separation of moving sources. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):895 –910, 2010.

[117] A. Noulas, G. Englebienne, and B. J. A. Krose. Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):79–93, 2012.

[118] Damir Omrčen and Aleš Ude. Redundant control of a humanoid robot head with foveated vision for object tracking. In *IEEE International Conference on Robotics and Automation*, 2010.

[119] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris. Real-time multiple sound source localization and counting using a circular microphone array. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2193–2206, 2013.

[120] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1201–1208. IEEE, 2011.

[121] Xinyuan Qian, Alessio Brutti, Maurizio Omologo, and Andrea Cavallaro. 3D audio-visual speaker tracking with an adaptive particle filter. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2896–2900, New-Orleans, Louisiana, 2017.

[122] Xinyuan Qian, Andrea Cavallaro, Alessio Brutti, and Maurizio Omologo. Locata challenge: speaker localization with a planar array. *LOCATA Workshop*, 2018.

[123] Xinyuan Qian, Alessio Xompero, Alessio Brutti, Oswald Lanz, Maurizio Omologo, and Andrea Cavallaro. 3d mouth tracking from a compact microphone array co-located with a camera. In *International Conference on Acoustics, Speech and Signal Processing*, 2018.

[124] Jartuwat Rajruangrabin and Dan O Popa. Robot head motion control with an emphasis on realism of neck–eye coordination during object tracking. *Journal of Intelligent & Robotic Systems*, 63(2), 2011.

[125] Donald Reid et al. An algorithm for tracking multiple targets. *IEEE transactions on Automatic Control*, 24(6):843–854, 1979.

[126] B. Ristic, B.-N. Vo, and D. Clark. Performance evaluation of multi-target tracking using the OSPA metric. In *IEEE International Conference on Information Fusion*, pages 1–7, Edinburgh, UK, 2010.

[127] Nicoleta Roman and DeLiang Wang. Binaural tracking of multiple moving sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):728–739, 2008.

[128] Junji Satake and Jun Miura. Robust stereo-based person detection and tracking for a person following robot. In *IEEE ICRA Workshop on People Detection and Tracking*, 2009.

[129] Niclas Schult, Thomas Reineking, Thorsten Kluss, and Christoph Zetzsche. Information-driven active audio-visual source localization. *PloS one*, 10(9), 2015.

[130] Dirk Schulz, Wolfram Burgard, Dieter Fox, and Armin B Cremers. Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In *IEEE International Conference on Robotics and Automation*, 2001.

[131] Ofer Schwartz and Sharon Gannot. Speaker tracking using recursive EM algorithms. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):392–402, 2014.

[132] H. Sidenbladh. Multi-target particle filtering for the probability hypothesis density. In *IEEE International Conference on Information Fusion*, pages 800–806, Tokyo, Japan, 2003.

[133] V. Smidl and A. Quinn. *The Variational Bayes Method in Signal Processing*. Springer, 2006.

[134] V. Smidl and A. Quinn. *The Variational Bayes Method in Signal Processing*. Springer, 2006.

[135] Václav Šmídl and Anthony Quinn. *The variational Bayes method in signal processing*. Signals and communication technology. Berlin: Springer, 2006.

[136] K. Smith, D. Gatica-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. In *IEEE Computer Vision and Pattern Recognition*, pages 962–969, San Diego, USA, 2005.

[137] K. Smith, D. Gatica-Perez, J.-M. Odobez, and S. Ba. Evaluating multi-object tracking. In *IEEE CVPR Workshop on Empirical Evaluation Methods in Computer Vision*, pages 36–36, San Diego, USA, 2005.

[138] R. Stiefelhagen, K. Bernardin, R. Bowers, J. S. Garofolo, D. Mostefa, and P. Soundararajan. CLEAR 2006 evaluation. In *First International Workshop on Classification of Events and Relationship, CLEAR 2006*. Springer, 2005.

[139] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The clear 2006 evaluation. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, 2006.

[140] Rustan Stolkin, Ionut Florescu, Morgan Baron, Colin Harrier, and Boris Kocherov. Efficient visual servoing with the abcshift tracking algorithm. In *IEEE International Conference on Robotics and Automation*, 2008.

[141] Ronen Talmon, Israel Cohen, and Sharon Gannot. Relative transfer function identification using convolutive transfer function approximation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):546–555, 2009.

[142] Johannes Traa and Paris Smaragdis. A wrapped Kalman filter for azimuthal speaker tracking. *IEEE Signal Processing Letters*, 20(12):1257–1260, 2013.

[143] Jean-Marc Valin, François Michaud, and Jean Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228, 2007.

[144] Lorenzo Vannucci, Nino Cauli, Egidio Falotico, Alexandre Bernardino, and Cecilia Laschi. Adaptive visual pursuit involving eye-head coordination and prediction of the target motion. In *IEEE-RAS International Conference on Humanoid Robots*, 2014.

[145] J. Vermaak, N.D. Lawrence, and P. Perez. Variational inference for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 773–780, Madison, USA, 2003.

[146] Jaco Vermaak and Andrew Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3021–3024, 2001.

[147] Deepu Vijayasenan and Fabio Valente. DiarTk: an open source toolkit for research in multistream speaker diarization and its application to meeting recordings. In *INTERSPEECH*, pages 2170–2173, Portland, OR, USA, 2012.

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

[149] B-N Vo, Sumeetpal Singh, and Arnaud Doucet. Sequential monte carlo methods for multitarget filtering with random finite sets. *IEEE Transactions on Aerospace and electronic systems*, 41(4):1224–1245, 2005.

[150] Ba-ngu Vo, Mahendra Mallick, Yaakov Bar-shalom, Stefano Coraluppi, Richard Osborne III, Ronald Mahler, and Ba-tuong Vo. Multitarget tracking. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–15, 2015.

[151] Ba-Ngu Vo, Sumeetpal Singh, and Arnaud Doucet. Random finite sets and sequential monte carlo methods in multi-target tracking. In *IEEE International Radar Conference*, pages 486–491, Huntsville, USA, 2003.

[152] Ba-Ngu Vo, Sumeetpal Singh, and Wing Kin Ma. Tracking multiple speakers using random sets. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 2, pages ii–357. IEEE, 2004.

[153] Ba-Ngu Vo, Ba-Tuong Vo, and Dinh Phung. Labeled random finite sets and the bayes multi-target tracking filter. *IEEE Transactions on Signal Processing*, 62(24):6554–6567, 2014.

[154] Ba-Tuong Vo and Ba-Ngu Vo. Labeled random finite sets and multi-object conjugate priors. *IEEE Transactions on Signal Processing*, 61(13):3460–3475, 2013.

[155] Ba-Tuong Vo, Ba-Ngu Vo, and Antonio Cantoni. Analytic implementations of the cardinalized probability hypothesis density filter. *IEEE Transactions on Signal Processing*, 55(7):3553–3567, 2007.

[156] Darren B Ward, Eric A Lehmann, and Robert C Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on speech and audio processing*, 11(6):826–836, 2003.

[157] Guanghan Xu, Hui Liu, Lang Tong, and Thomas Kailath. A least-squares approach to blind channel identification. *IEEE Transactions on signal processing*, 43(12):2982–2993, 1995.

[158] Yihong Xu, Yutong Ban, Xavier Alameda-Pineda, and Radu Horaud. Deepmot: A differentiable framework for training multiple object trackers. *arXiv preprint arXiv:1906.06618*, 2019.

[159] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2034–2041, Providence, USA, 2012.

[160] M. Yang, Y. Liu, L. Wen, and Z. You. A probabilistic framework for multitarget tracking with mutual occlusions. *IEEE Confenrence on Computer Vision and Pattern Recognition*, pages 1298 – 1305, 2014.

[161] Ozgur Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing,*, 52(7):1830–1847, 2004.

[162] Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

[163] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, Hawaii, USA, 2017.

[164] Xionghu Zhong and James R Hopgood. Particle filtering for TDOA based acoustic source tracking: Nonconcurrent multiple talkers. *Signal Processing*, 96:382–394, 2014.