Chapter 3

Audio-Visual Speaker Tracking

Volkan Kılıç and Wenwu Wang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.68146

Abstract

Target motion tracking found its application in interdisciplinary fields, including but not limited to surveillance and security, forensic science, intelligent transportation system, driving assistance, monitoring prohibited area, medical science, robotics, action and expression recognition, individual speaker discrimination in multi-speaker environments and video conferencing in the fields of computer vision and signal processing. Among these applications, speaker tracking in enclosed spaces has been gaining relevance due to the widespread advances of devices and technologies and the necessity for seamless solutions in real-time tracking and localization of speakers. However, speaker tracking is a challenging task in real-life scenarios as several distinctive issues influence the tracking process, such as occlusions and an unknown number of speakers. One approach to overcome these issues is to use multi-modal information, as it conveys complementary information about the state of the speakers compared to single-modal tracking. To use multi-modal information, several approaches have been proposed which can be classified into two categories, namely deterministic and stochastic. This chapter aims at providing multimedia researchers with a state-of-the-art overview of tracking methods, which are used for combining multiple modalities to accomplish various multimedia analysis tasks, classifying them into different categories and listing new and future trends in this field.

Keywords: audio-visual tracking, multi-speaker tracking, deterministic, stochastic approaches

1. Introduction

Speaker tracking aims at localizing the moving speakers in a scene by analysing the data sequences captured by sensors or arrays of sensors. It gained relevance in the past decades due to its widespread applications such as automatic camera steering in video conferencing

[1], individual speaker discriminating in multi-speaker environments [2], acoustic beamforming [3], audio-visual speech recognition [4], video indexing and retrieval [5], human-computer interaction [6], and surveillance and monitoring [7] in security applications. There are numerous challenges, which make speaker tracking a difficult task including, but not limited to, the estimation of the variable number of speakers and their states, and dealing with various conditions such as occlusions, limited view of cameras, illumination change and room reverberations [8–10].

Using multi-modal information is one way to address these challenges since more comprehensive observations for the state of the speakers can be collected in multi-modal tracking as compared to the single-modal case, and the collection of the multi-modal information can be achieved by sensors such as audio, video, thermal vision, laser-range finders and radio-frequency identification (RFID) [11–13]. Among these sensors, audio and video sensors are commonly used in speaker tracking compared to others, because of their easier installation, cheaper cost and more data-processing tools [9, 14, 15].

Earlier methods in speaker tracking employ either visual-only or audio-only data, and each modality offers some advantages but is also limited by some weaknesses [16, 17]. Tracking with only video [16–18] offers robust and accurate performance when the camera field of view covers the speakers. However, it degrades when the occlusion between speakers happens, when the speakers go out of the camera field of view, or any changes on illumination or target appearance have occurred. Although audio tracking [19–21] is not restricted by these limitations, it has a tendency to non-negligible-tracking errors because of intermittency of audio data. In addition, audio data may be corrupted by background noise and room reverberations. Nevertheless, the combination of audio and video data may improve the tracking performance when one of the modalities is missing or neither provides accurate measurements, as audio and visual modalities are often complementary to each other which can be exploited to further enhance their respective strengths and mitigate their weaknesses in tracking.

Previous techniques were focused on tracking a single person in a static and controlled environment. However, theoretical and algorithmic advances together with the increasing capability in computer processing have led to the emergence of more sophisticated techniques for tracking multiple speakers in dynamic and less controlled (or natural) environments [22–24]. In addition, the type of sensors used to collect the measurements is advanced from single- to multi-modal.

In the literature, there are many approaches for speaker tracking using multi-modal information, which can be categorized into two methods as one is deterministic and data-driven while the other is stochastic and model-driven [25, 26]. Deterministic approaches are considered as an optimization problem by minimizing a cost function, which needs to be defined appropriately. A representative method in this category is the mean-shift method [27, 28], which defines the cost function in terms of colour similarity measured by Bhattacharyya distance. The stochastic and model-driven approaches use a state-space approach based on the Bayesian framework as it is suitable for processing of multi-modal information [29]. Representative methods are the Kalman filter (KF) [30], extended KF (EKF) and particle filter (PF) [31]. The PF approach is more robust for non-linear and non-Gaussian models as compared with the KF

and EKF approaches since it easily approaches the Bayesian optimal estimate with a sufficiently large number of particles [11].

One challenge in the implementation of the PF to tracking problem is to choose an optimal number of particles [9, 32]. An insufficient number may introduce a particle impoverishment, while a larger number (than required) will lead to extra computational cost. Therefore, choosing the optimal number of particles is one of the issues that affect the performance of the tracker. To address this issue and to find the optimal number of particles for the PF to use, adaptive particle filtering (A-PF) approaches have been proposed in Refs. [9, 32–35]. Fox [34] proposed KLD sampling, which aims to bind the error introduced by the sample-based representations of the PF using the Kullback-Leibler divergence between maximum likelihood estimates (MLEs) of the states and the underlying distribution to optimize the number of particles. The KLD-sampling criterion is improved in Ref. [35] for the estimation of the number of particles, leading to an approach for adaptive propagation of the samples. Subsequent work [33] introduces the innovation error to estimate the number of particles by employing a twofold metric. The particles are removed by the first metric in case their distance to a neighbouring particle is smaller than a predefined threshold. The second metric is used to set the threshold on the innovation error in order to control the birth of the particles. These two thresholds need to be set before the algorithm is run. A new approach is proposed in Refs. [9, 32], which estimates noise variance besides the number of particles in an adaptive manner. Different from other existing adaptive approaches, adaptive noise variance is employed in this method for the estimation of the optimal number of particles based on tracking error and the area occupied by the particles in the image.

One assumption in the traditional PF used in multi-speaker tracking is that the number of speakers is known and invariant during the tracking. In practice, the presence of the speakers may change in a random manner, resulting in time-varying number of speakers. To deal with the unknown and variable number of speakers, the theory of random finite sets (RFSs) has been introduced, which allows multi-speaker filtering by propagation of the multi-speaker posterior [36–39]. However, the computational complexity of RFS grows exponentially as the number of speakers increases since the complexity order of the RFS is $O(M^{\Xi})$ where M is the number of measurements and Ξ is the number of speakers. The PHD filtering [40] approach is proposed to overcome this problem, as the first-order approximation of the RFS whose complexity scales linearly with the number of speakers since the complexity order of the PHD is $O(M\Xi)$. This framework has been found to be promising for multi-speaker tracking [36]. However, the PHD recursion involves multiple integrals that need to have closed-form solutions for implementation. So far, two analytic solutions have been proposed: Gaussian mixture PHD (GM-PHD) filter [41, 42] and sequential Monte Carlo PHD (SMC-PHD) filter [43, 44]. Applications of GM-PHD filter are limited by linear Gaussian systems, which lead us to consider SMC-PHD filter to handle non-linear/non-Gaussian problems in audio-visual tracking [15, 45].

Apart from the stochastic methodologies mentioned above, the mean-shift [28] is a deterministic and data-driven method, which focuses on target localization using representation of the target. The mean-shift easily convergences to peak of the function with a high speed and a small computational load. Moreover, as a non-parametric method, the solution of the mean

shift is independent from the features used to represent the targets. On the other hand, the performance of the mean-shift is degraded by occlusion or clutter as it searches the densest (most similar) region starting from the initial position in the region of interest. In this sense, the mean-shift trackers may fail easily in tracking small- and fast-moving targets as the region of interest may not cover the targets, which results in a track being lost after a complete occlusion. Also, it is formulated for single-target tracking, so it cannot handle a variable number of targets. Therefore, several methods [14, 15, 46–49] have been proposed by integrating both deterministic and stochastic approaches to benefit their respective strengths which will be discussed in Section 4.

2. Tracking modalities

2.1. Visual cues

Visual tracking is a challenging task in real-life scenarios, as the performance of a tracker is affected by the illumination conditions, occlusion by background objects and fast and complicated movements of the target [50, 51]. To address these problems, several visual features, that is, colour, texture, contour and motion [52], are employed in existing tracking systems.

Using colour feature is a very intuitive approach and commonly applied in target tracking as the information provided by colour helps to distinguish between targets and other objects. Several approaches can be found in the literature which employs colour information to track the target. In Ref. [53], a colour mixture model based on a Gaussian distribution is used for tracking and segmentation, while in Ref. [58], an adaptive mixture model is developed. Target detection and tracking can be easily maintained using colour information if the colour of the target is distinct from those of the background or other objects.

Another approach for tracking is contour-based where shape matching or contour-evolution techniques [54] are used to track the target contour. Active models like snakes, geodesic-active contours, B-splines or meshes [55] can be employed to represent the contours. Occlusion of the target by other objects is the common problem in tracking. This problem can be addressed by detecting and tracking the contour of the upper body [56] rather than tracking the contour of the whole bodies, which leads to the detection of a new person as the upper bodies are often distinguishable from back and front view for different people.

Texture is another cue defined as a measure for surface intensity variation. Properties like smoothness and regularity can be quantified by the texture [57–59]. The texture feature is used with Gabor wavelet in Ref. [60]. The Gabor filters can be employed as orientation and scale-tunable edge and line detectors, and the statistics of these micro-features are mostly used to characterize the underlying texture information in a given region [61]. For improved detection and recognition, local patterns of image have gained attention recently. Local patterns are used in several application areas such as image classification and face detection since they offer promising results. In Ref. [62], the local binary patterns (LBPs) method is used to create a type of texture descriptor based on a grey-scale-invariant texture measure. Such a measure is tolerant to illumination changes.

Another cue used in tracking, particularly in indoor environments, is motion which is an explicit cue of human presence. One way to extract this cue is to apply foreground detection algorithms. A simple method for foreground detection is to compute the difference of two consecutive frames which gives the moving part of the image. Although it has been used in multi-modal-tracking systems [63], it fails when the person remains stationary since the person is considered part of background after some time.

The scale-invariant feature transform (SIFT) proposed in Ref. [64] has found wide use in tracking applications. SIFT uses local features to transform the image data into scale-invariant coordinates. Distinctive invariant features are extracted from images to provide matching between several views of an object. The SIFT feature is invariant to scaling, translation, clutter, rotation, occlusion and lighting which makes it robust to changes in three-dimensional (3D) viewpoint and illumination, and the presence of noise. Even a single feature has high matching rate in a large database because the SIFT features are generally distinctive. On the other hand, non-rigid targets [65] in noisy environments degrade the SIFT matching rate and recognition performance.

So far, several visual cues were introduced, and among them colour cues have been used more commonly in tracking applications due to their easy implementation and low complexity. Colour information can be used in the calculation of the histogram of possible targets at the initialization step as reference images which can be used in detection and tracking of the target. There are two common colour histogram models, RGB or HSV [66] in the literature and HSV is more preferable since it is observed to be more robust to illumination variation [9].

2.2. Audio cues

There are a variety of audio information that could be used in audio tracking such as sound source localization (SSL), time-delay estimation (TDE) and the direction of arrival (DOA) angle.

The audio source localization methods can be divided into three categories [67], namely steered beamforming, super-resolution spectral estimation and time-delay estimation. Beamformer-based source localization offers comparatively low resolution and needs a search over a highly non-linear surface [20]. Also, it is computationally expensive which may be limited in real-time applications. Super-resolution spectral estimation methods are not well suited for locating a moving speaker since it is under the assumption that the speaker location is fixed for a number of frames [68]. However, the location of a moving speaker may change considerably over time. In addition, these methods are not robust to modelling errors caused by room reverberation and mostly have high computational cost [20, 69]. The time-delay of arrival (TDOA)-based location estimators use the relative time delay between the wave-front arrivals at microphone positions in order to estimate the location of the speaker. As compared with the other two methods, the TDOA-based approach has advantages in the following two aspects. The first one is its computational efficiency and the second one its direct connection to the speaker location.

The problem of DOA estimation is similar to that of the TDOA estimation. To estimate the DOA, the TDOA needs to be determined between the sensor elements of the microphone

array. Estimation of source locations mainly depends on the quality of the DOA measurements. In the literature, several DOA estimation techniques such as the MUSIC algorithm [70] and the coherent signal subspace (CSS) [71] have been proposed. The main differences between them are the way of dealing with reverberation, background noise and movement of the sources [20]. The following three factors influence the quality of the DOA estimation. The spectral content of the speech segment is considered as the first one which is used for derivation of the DOAs. The reverberation level of the room is the second one which causes outlier in the measurements because of the reflections from the objects and walls. The positions of the microphone array to the speakers and the number of simultaneous sources in the field are considered the third factor.

3. Audio-visual speaker tracking

Speaker tracking is a fundamental part of multimedia applications which plays a critical role to determine the speaker trajectories and analyse the behaviour of speakers. Speaker tracking can be accomplished with the use of audio-only, visual-only or audio-visual information.

Audio-only information based approaches for speaker tracking have been presented in [19, 20, 37, 72–74]. An audio-based fusion scheme was proposed in Ref. [20] to detect multiple speakers where the locations from multiple microphone arrays are estimated and fused to determine the state of the same speaker. Separate KFs are employed for all the individual microphone arrays for the location estimation. To deal with motion of the speaker and measurement uncertainty, the probabilistic data association technique is used with an interacting model.

One issue in Ref. [20] is that it cannot deal with the tracking problem for a time-varying number of speakers. Ma et al. [37, 72] proposed an approach based on random finite set to track an unknown and time-varying number of speakers. The RFS theory and SMC implementation are used to develop the Bayesian RFS filter, which tracks the time-varying number of speakers and their states. The random finite set theory can deal with a time-varying number of speakers; however, the maximum number of speakers that can be handled is limited as its computational complexity increases exponentially with the number of speakers. In that sense, a cardinalized PHD (CPHD) filter is proposed in Ref. [74], which is the first-order approximation of the RFS, to reduce the computational cost caused by the number of speakers. The positions of the speakers are estimated using TDOA measurements from microphone pairs by asynchronous sensor fusion with the CPHD filter.

A time-frequency method and the PHD filter are used in Ref. [73] to localize and track simultaneous speakers. The location of multiple speakers is estimated based on the time-frequency method, which uses an array of three microphones, then the PHD filter is employed to the localization results as post-processing to handle miss-detection and clutters.

Speaker tracking with multi-modal information has also gained attention, and many approaches have been proposed in the past decade using audio-visual information [2, 6, 23, 29, 75–81],

providing the complementary characteristics of each modality. The differences among these existing works arise from the overall objective such as tracking either single or multiple speakers and the specific detection/tracking framework.

Audio-visual measurements are fused by graphical models in Ref. [23] to track a moving speaker in a cluttered and noisy environment. Audio and video observations are used jointly by computing their mutual dependencies. The model parameters are learnt using the expectation-maximization algorithm from a sequence of audio-visual data.

A hierarchical Kalman filter structure was proposed in Refs. [2, 80] to track people in a three-dimensional space using multiple microphones and cameras. Two independent local Kalman filters are employed for audio and video streams, and then the outputs of these two local filters are combined under one global Kalman filter.

Unlike [2, 80], particle filters are used in Ref. [81] to estimate the predictions from audio- and video-based measurements and audio-visual information fusion is performed at the feature level. In other words, the independent particle coordinates from the features of both modalities are fused for speaker tracking. These works [2, 23, 80, 81] have focused on the single-speaker case which cannot directly address the tracking problem for multiple speakers.

Two multi-modal systems are introduced in Ref. [75] for the tracking of multiple persons. A joint probabilistic data association filter is employed to detect speech and determine active speaker positions. Two systems are performed for visual features where a particle filter is applied first using foreground, colour, upper body detection and person region cues from multiple camera images and the latter is a blob tracker using only a wide-angle overhead view. Then, acoustic and visual tracks are integrated using a finite state machine. Unlike [75], a particle filtering framework is proposed in Ref. [29, 77] which incorporates the audio and visual detections into the particle filtering framework using an observation model. It has the capability to track multiple people jointly with their speaking activity based on a mixed-state dynamic graphical model defined on a multi-person state space. Another particle filter based multi-modal fusion approach is proposed in Ref. [78] where a single speaker can be identified in the presence of multiple visual observations. Gaussian mixtures model was adopted to fuse multiple observations and modalities. Compared to [29, 75, 77, 78], particle filtering framework is not used in Ref. [6]; instead, hidden Markov model based iterating decoding scheme is used to fuse audio and visual cues for localization and tracking of persons.

In Refs. [14, 76, 79], the Bayesian framework is used to handle the tracking problem for a varying number of speakers. The particle filter is used in Ref. [76], and observation likelihoods based on both audio and video measurements are formulated to use in the estimation of the weights of the particles, and then the number of people is calculated using the weights of these particles. The RFS theory based on multi-Bernoulli approximations is employed in Ref. [79] to integrate audio and visual cues with sequential Monte Carlo implementation. The nature of the random finite set formulation allows their framework to deal with the tracking problem for a varying number of targets. Sequential Monte Carlo implementation (or particle filter) of PHD filter is used in Ref. [14] where audio and visual modalities are fused in the steps of particle filter rather than using any data fusion algorithms. Their work substantially differs from existing works in AV

multi-speaker tracking with respect to the capabilities for dealing with multiple speakers, simultaneous speakers, and unknown and time-varying number of speakers.

4. Tracking algorithms

In this section, a brief review of tracking algorithms is presented which covers the following topics: Bayesian statistical methods, visual and audio-visual algorithms and non-linear filtering approaches.

Recall that in Section 1, tracking methods are either stochastic and model-driven or deterministic and data-driven [25].

The stochastic approaches are based on the Bayesian framework which uses a state-space approach [82]. Representative methods in this category are the Kalman filter (KF) [30], extended Kalman filter (EKF) [83, 84] and particle filter (PF) [11]. The PF approach is more robust as compared to the KF and EKF approaches as it can approach the Bayesian optimal estimate with a sufficiently large number of particles [11]. It has been widely applied to speaker tracking problems [29, 76, 81]. The PF is used to fuse object shapes and audio information in Refs. [29, 81]. In Ref. [76], independent audio and video observation models are fused for simultaneous tracking and detection of multiple speakers. However, one challenge in PF is to choose an appropriate number of particles. While an insufficient number may lead to particle impoverishment (i.e. loss of diversity among the particles), a larger number (than required) will induce additional computational cost. Therefore, the performance of the tracker depends on the number of particles that are estimated as an optimal value.

The PHD filter [85] is another stochastic method based on the finite-set statistics (FISST) theory, which propagates the first-order moment of a dynamic point process. The PHD filter is used in many application areas after its proposal and some applications with speaker tracking are reported in Refs. [37, 73]. It has an advantage over other Bayesian approaches such as Kalman and PF filters, in that the number of targets does not need to be known in advance since it is estimated in each iteration. The issue in the PHD filter is that it is prone to estimation error in the number of speakers in the case of low signal-to-noise ratio [36]. The reason is that the PHD filter restricts the propagation of multi-target posterior to the first-order distribution moment, resulting in loss of information for higher order cardinality. To address this issue, the cardinality distribution is also propagated with PHD distribution in the cardinalized PHD (CPHD) filter which improves the estimation of the target number [36, 86] and state of the speakers [74]. However, additional distribution for cardinality requires extra computational load, which makes the CPHD computationally more expensive than the PHD filter. Moreover, the spawning of new targets is not modelled explicitly in the CPHD filter.

As a deterministic and data-driven method, the mean-shift [28] uses representation of the target for localization, which is based on minimizing an appropriate cost function. In that sense, a similarity function is defined in Ref. [32] to reduce the state estimation problem to a search in the region of interest. To obtain fast localization, a gradient optimization method is

performed. The mean-shift works under the assumption that the representation of the target is sufficiently distinct from the background which may not be always true. Although the mean-shift is an efficient and robust approach, in occlusion and rapid motion scenarios [87, 88], it may fail when the target is out of the region of interest, in other words, the search area.

Many approaches have been proposed in the literature to address these problems in mean-shift tracking, which can be categorized into two groups. One group [87, 89–91] improves the mean-shift tracking by, for example, introducing adaptive estimation of the search area, iteration number and bin number. In the other group, the mean-shift algorithm is combined with other methods such as particle filter [46–49]. The stochastic and deterministic approaches are integrated under the same framework in many studies. Particle filtering (stochastic) is integrated with a variation approach (deterministic) in Ref. [25] where the 'switching search' algorithm is run for all the particles. In this algorithm, the momentum of the particles is compared with a pre-determined threshold value, and if it is smaller than the threshold, the deterministic search is run; otherwise, the particles are propagated in terms of a stochastic motion model.

The particle filtering and mean-shift are combined in Ref. [48] under the name of mean-shift embedded particle filter (MSEPF). It is inspired by Sullivan and Rittscher [25], but the mean shift is used as a variational method. It is aimed to integrate the advantages of the particle filtering and mean-shift method. The MSEPF has a capability to track the target with a small number of particles as the mean-shift search concentrates on the particles around local modes (maxima) of the observation. To deal with the possible changes in illuminations, a skin colour model is used and updated for every frame. As an observation model, colour and motion cues are employed. To use a multi-cue observation model, the mean-shift analysis is modified and applied to all the particles. Resampling (selective resampling) is, then, applied when the effective sample size is too small. The mean-shift and particle filtering methods are used independently in Ref. [46]. The estimated positions of the target obtained by these two methods are compared using the Bhattacharyya distance at every iteration and the best value is chosen as the estimated position of the target to avoid the algorithm from being trapped to a local maximum, and thus finding the true maximum beyond the local one.

A hybrid particle with a mean-shift tracker is proposed in Ref. [92] which works in a similar manner to that in Ref. [48]. Alternatively, [92] uses the original application of the mean-shift and performs the mean-shift process on all the particles to reach the local maxima. Moreover, an adaptive motion model is used to deal with manoeuvring targets, which have a high speed of movement. The kernel particle filter is proposed in Ref. [93] where small perturbations are added to the states of the particles after the mean-shift iteration to prevent the gradient ascent from being stopped too early in the density. Kernel radius is calculated adaptively every iteration and this method is applied to multiple target tracking using multiple hypotheses which are then evaluated and assigned to possible targets. An adaptive mean-shift tracking with auxiliary particles is proposed in Ref. [49]. As long as the conditions are met, such as the target remaining in the region of interest, and there are no serious distractions, the mean-shift is used as the main tracker. When sudden motions or distractions are detected by the motion estimator, auxiliary particles are introduced to support the mean-shift tracker. As the mean shift may diverge from the target and converge on the background, background/foreground

feature selection is applied to minimize the tracking error. Even though this study is inspired by Sullivan and Rittscher [25], where the main tracker is a particle filter, in Ref. [49], the main tracker is the mean-shift. In addition, the switched trackers are used to handle sudden movements, occlusion and distractions. Moreover, to maintain tracking even when the target appearance is affected by illumination or view point, the target model is updated online.

In the literature, several frameworks have been proposed to combine the mean-shift and particle filters. However, it is still required to have an explicitly designed framework for a variable number of targets. Both the mean-shift and particle filter were derived for tracking only a single target. To address this issue, the PHD filter is found as a promising solution as it is originally designed for multi-target tracking. However, the PHD filter does not have closed-form solutions as the recursion of the PHD filter includes multi-dimensional integrals. To derive analytical solution of the PHD filter, the particle filter or sequential Monte Carlo (SMC) implementation [44] is introduced which leads to SMC-PHD filtering. In Ref. [14], the mean-shift is integrated with standard SMC-PHD filtering, aiming at improving computational efficiency and estimation accuracy of the tracker for a variable number of targets.

Besides the tracking methods explained so far, speaker tracking with multi-modal usage introduces a problem which is known as data association. Each measurement coming from multi-modality needs to be associated with an appropriate target. Data association methods are divided into two classes [94]. Unique neighbour is the first data association, and a representative method in this class is multiple hypothesis tracking (MHT). Here, each existing track is associated with one of the measurements. All-neighbours data association belongs to the second class which uses all the measurements for updating the entire track estimate, for example, the joint probabilistic data association (JPDA). In MHT, the association between a target state and the measurements is maintained by multiple hypotheses. However, the required number of hypotheses increases exponentially over time [95]. In JPDA, separate Gaussian distributions for each target [96] are used to approximate the posterior target distribution which results in an extra computational cost. Data association algorithms in target-tracking applications with Bayesian methods and the PHD filter can be found in [20, 97–100]. However, it is found that classical data association algorithms are computationally expensive which lead to the fusion of multi-modal measurements inside the proposed framework [8, 9, 29, 73, 80, 81, 83]. As in Refs. [8, 9], audio and visual modalities are fused in the steps of the visual particle filter.

Among the methods explained above, the PF, RFS, PHD filter and mean-shift are the main methods discussed throughout this chapter and the main concepts of the methods are presented below.

4.1. Particle filtering

The PF became widely used tools in tracking after being proposed by Isard et al. [31] due to its ability to handle non-linear and non-Gaussian problems. The main idea of the PF is to represent a posterior density by a set of random particles with associated weights, and then compute estimates based on these samples and weights [101]. The principle of the particle filter is illustrated in **Figure 1**. Ten particles are initialized with equal weights in the first step.

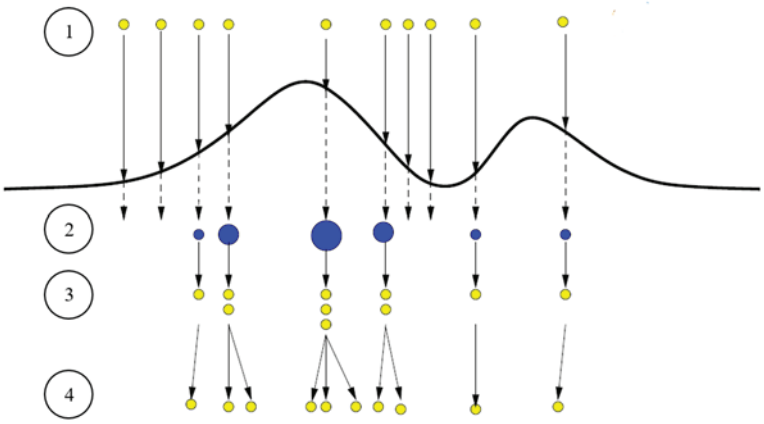

Figure 1. Steps of the particle filter. The first step is particle initialization with equal weights. The particles are weighted in the second step. After a resampling step is performed in the third step, the particles are distributed to predict the next state in the fourth step. This figure is adapted from Ref. [102].

In the second step, the particles are weighted based on given measurements, and as a result, some particles require small weights while others require larger weights represented by the size of the particles. The state distribution is represented by these weighted particles. Then, a resampling step is performed which selects the particles with large weights to generate a set of new particles with equal weights in the third step. In step four, these new particles are distributed again to predict the next state. This loop continues from steps two through four until all the observations are exhausted.

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

4.2. Random finite set and PHD filtering

The generic PF is designed for single-target tracking. Multi-target tracking is more complicated than single-target tracking as it is necessary to jointly estimate the number of targets and the state of the targets. One multi-target tracking scenario is illustrated in **Figure 2a**, where five targets exist in state space (bottom plane) given at the previous time with eight measurements in observation space (upper plane). In this scenario, the number of measurements is larger than the number of targets due to clutter or noise. When the targets are passed to the current time, the number of targets becomes three and two targets no longer exist.

In that sense, the variable number of targets and noisy measurements need to be handled for reliable tracking in multi-target case. The RFS approach [36] is an elegant solution to address this issue. The basic idea behind the RFS approach is to treat the collection of targets as a set-valued state called the multi-target state and the collection of measurements as a set-valued observation, called multi-observation. So, the problem of estimating multiple targets in the presence of clutter and uncertainty is handled by modelling these set-valued entities as

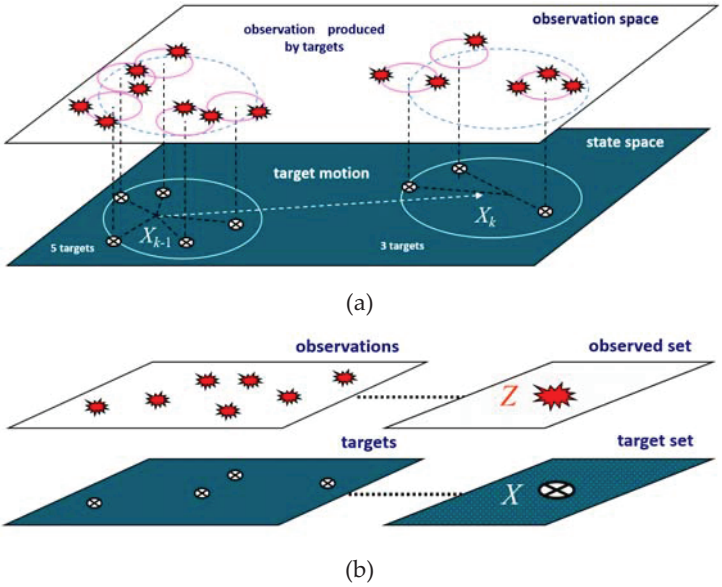

Figure 2. An illustration of the RFS theory in a multi-target tracking application. One possible multi-target tracking scenario is given in (a), and (b) represents the RFS approach to multi-target tracking. The figures are adapted from Ref. [103].

random finite sets [41]. The point here is to generalize the tracking problem from single target to multiple targets.

Figure 2b illustrates the RFS approach where all the targets are collected in one target set and all the measurements are considered as one measurement set. The RFS propagates the full multi-target posterior for multi-target filtering. The state model of the RFS incorporates individual target dynamics which are target birth, target spawn and target death. In addition, the observation model of the RFS incorporates the measurement likelihood as target detection uncertainty (miss-detection) and clutter (false alarm). These incorporations are implemented by assigning hypotheses, and all possible associations between hypotheses and measurement/targets need to be repeated at every time step, resulting in increased computational cost in the case of a high number of targets and measurements.

To alleviate the computational cost, the PHD filter is introduced which is a computationally cheaper alternative to the RFS. The PHD filter is the first-order approximation of the RFS and propagates only the first-order moments instead of the full multi-target posterior [44, 104]. The PHD filter function is denoted as the intensity $v(x)$ whose integral on any region of the state space gives the expected number of targets. The peaks of the PHD function point the highest local concentration of the expected number of targets, which can be used to provide estimates of individual targets [36]. The PHD filter is illustrated in **Figure 3** by a simple example [36] which corresponds to Eq. (1)

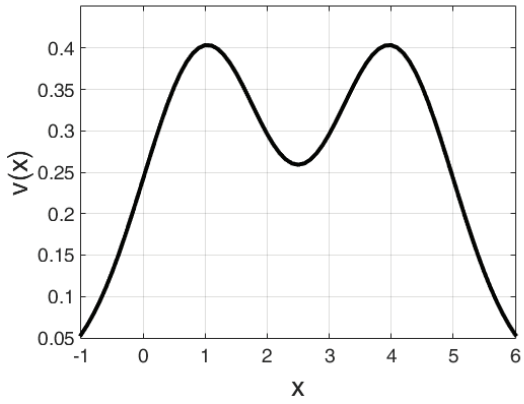

Figure 3. A simple example for the PHD filter. This figure is adapted from Ref. [36].

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

Figure 3 is plotted for Eq. (1) with $\sigma = 1$, $a = 1$ and $b = 4$. The peaks of $v(x)$ is near the target locations $x = 1$ and $x = 4$.

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

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

4.3. Mean-shift tracking

Different from stochastic approaches such as the PF, RFS and PHD filter, the mean-shift is a deterministic method [28]. The mean-shift can be defined as a simple iterative procedure that shifts each data point to the average of data points in its neighbourhood [105].

Common application areas are clustering [106], mode seeking [107], image segmentation [108] and tracking [109]. Simple implementation of the mean-shift method is illustrated in **Figure 4** where the purpose is to find the densest region of the distributed balls. The first step is to select an initial point with the region of interest as shown in **Figure 4a** where the circle indicates the region of interest centred on the initial point. In **Figure 4b**, the centre of the mass is calculated using the balls inside the region of interest. To get the distance and direction for shifting the initial point, the mean-shift vector is calculated in **Figure 4c**. The initial point is shifted to a new point together with the region of interest in **Figure 4d**. The centre of the mass is calculated again using the balls inside the region of interest which leads to new mass point in **Figure 4e**. The mean-shift vector is calculated to obtain the direction and distance for shifting and the region of interest is shifted to a new point as illustrated in **Figures 4f** and **g**, respectively. This iteration continues until the mean-shift method reaches the densest point in **Figure 4h**.

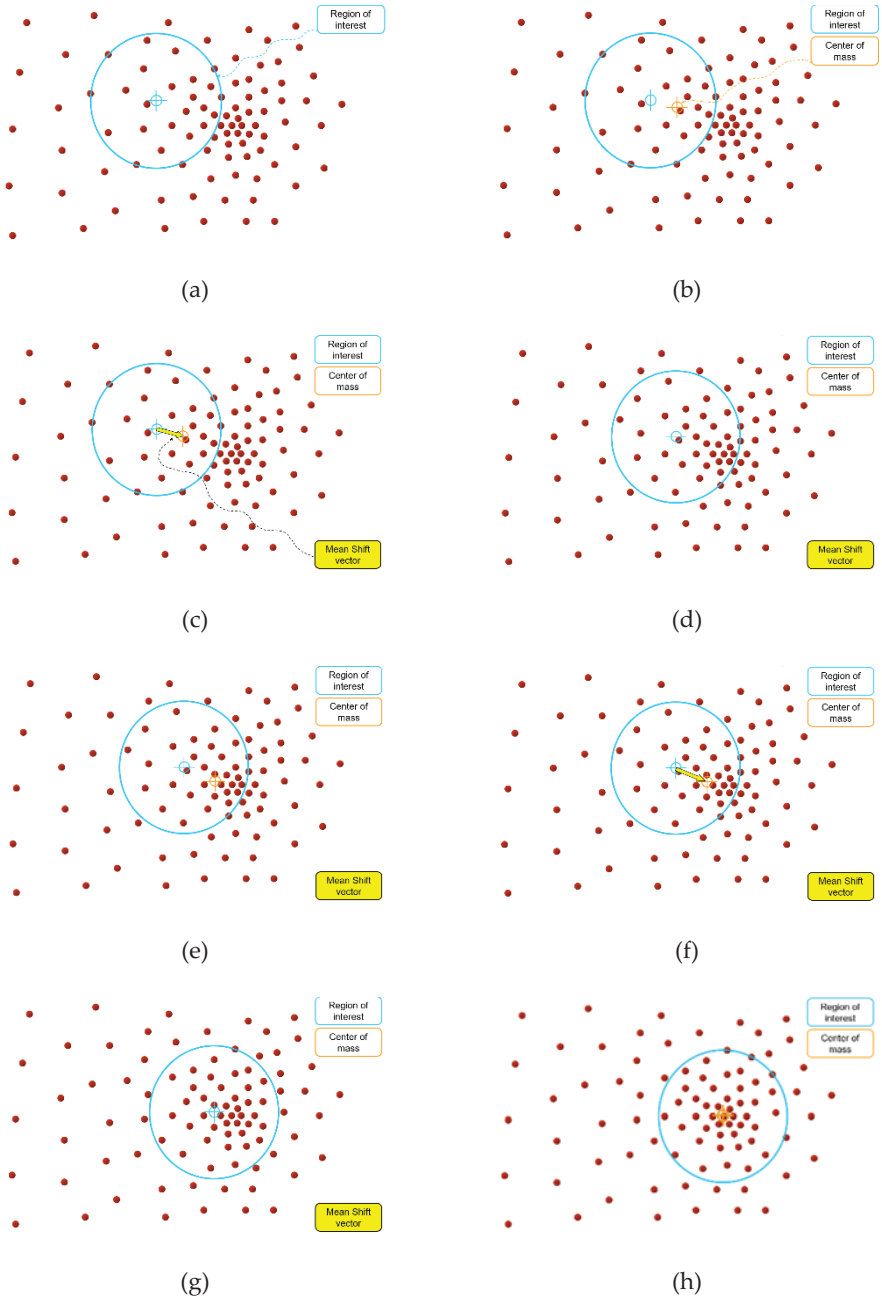

Figure 4. Simple descriptions of the mean-shift process. These figures are adapted from Ref. [110].

5. Relevant datasets

In order to perform a quantitative evaluation of the audio-visual tracker, both audio and video sequences are required. In that sense, several datasets are presented in the literature that combine multiple audio and video sources for tracking.

The augmented multi-party interaction (AMI) [111] corpus includes 100 h of meetings, which were recorded in English using three different rooms. Natural conversations are included in some of the meetings, and many others, in particular those using a scenario in which the participants play different roles in a design team, are also reasonably natural. The number of speakers in the natural conversations varies from three to five. In one artificial meeting, four speakers are involved, taking four pre-arranged roles (as industrial designer, interface designer, marketing and project manager). Other artificial meetings also appear in the AMI corpus, such as a film club scenario. Generally, the speakers are mostly static or with small movements. In addition, calibration information is not available which is required for 3D tracking as it is needed to project the coordinates from the two-dimensional (2D) image into 3D space.

CLEAR (CLassification of Events, Activities and Relationships) is the next dataset created for people identification, activities, human-human interaction and relevant scenarios [112]. Recordings are captured with multiple users in realistic meeting rooms equipped with a multitude of audio-visual sensors. The rooms have five calibrated cameras, and four of them are mounted to the corners of the room while the last panoramic camera is mounted to the ceiling of the room. All cameras are synchronized with the audio streams collected by the linear microphone array placed on the walls. In most scenarios, the speakers are generally still and seated around the table. They speak one by one.

Another dataset is SPEVI (Surveillance performance evaluation initiative) [113] created for single- and multi-modal people detection and tracking. Sequences are captured by a video camera and two linear microphone arrays. The SPEVI dataset has three sequences. The sequences motinas_Room160 and motinas_Room105 are captured in rooms with reverberation. The sequence motinas_Chamber is captured in a reduced reverberation room. In this dataset, audio signals were recorded with linear microphone arrays and the calibration information is not available.

One of the most challenging datasets that can be used for the evaluation of audio-visual tracking algorithm is AV16.3 corpus which is developed by the IDIAP research institute [114]. The corpus AV16.3 involves various scenarios where subjects are moving and speaking at the same time whilst being recorded by three calibrated video cameras and two circular eight-element microphone arrays.

Recordings in the AV16.3 involve challenging scenarios such as object initialization, partial and total occlusion, overlapped speech, illumination change, close and far locations, variable number of objects, and small and large angular separations. Circular microphone arrays were used to record the audio signals at 16 kHz and video sequences were captured at 25 Hz. The recordings of audio and video were performed independently from each other. Each video

frame is a colour image of 288 × 360 pixels and some sequences are annotated to get the ground truth (GT) speaker position which allows one to measure the accuracy of each tracker and to compare the performance of the algorithms. In addition, it provides calibration information of the cameras and challenging scenarios like occlusions and moving speakers.

The most recently released dataset is 'S3A speaker tracking with Kinect2' [115, 116] which uses a Kinect for Windows V2.0 for recording the visual data and dummy head for recording the audio data. It contains four sequences in a studio where people are talking and walking slowly around a dummy head which is located at the centre of the room. Different from other cameras, Kinect sensor provides in-depth information besides the colour which helps to extract the 3D position of the speaker without using additional view of the scene. In addition, annotated data are provided which can be used as ground truth data to estimate the performance of the tracker.

6. Performance metrics

Several metrics have been proposed to evaluate the performance of tracking methods in the literature. In this section, four metrics are introduced.

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

The multiple object tracking (MOT) metric is the next metric which was proposed in Ref. [117]. It is defined with MOT precision (MOTP) and MOT accuracy (MOTA) quantities. The precision is measured with the MOTP using a pre-defined threshold value

$$MOTP = \frac{\sum_{i,k} d_k^i}{\sum_k c_k} \tag{3}$$

where d_k^i is the distance between the ith object and its corresponding hypothesis and c_k is the number of matches between the objects and hypotheses for time frame k.

Tracking errors are measured with the MOTA which covers the false positives, false negatives and mismatches. If the error is greater than the threshold value, it is assumed that the false positive and false negative count if the speaker is not tracked with the accuracy measured by the threshold. Mismatches are the case where the speaker identity is switched [117]

$$MOTA = 1 - \frac{\sum_k (m_k + fp_k + mm_k)}{\sum_k g_k} \tag{4}$$

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

The next metric is the trajectory-based measures (TBMs) proposed in Refs. [118, 119], where the performance is measured based on trajectory. It categorizes the trajectories as mostly

tracked (MT), mostly lost (ML) and partially tracked (PT). MT is defined as if the tracker follows at least 80% of its ground truth (GT) trajectory. If the tracker follows less than 20% of its GT, it is called ML. If the followed trajectory is between 20 and 80% of the GT trajectory, it is called PT. Also, track fragmentation (Frag) is defined as the total number of times that GT is interrupted. Identity switches (IDs) are computed by calculating change in GT identity.

OSPA-T (Optimal Subpattern Assignment for Tracks) [120] is the last performance metric designed for the evaluation of multi-speaker tracking systems. It is an improved version of the OSPA metric [121] by extending it for tracking management evaluation. To transfer the cardinality error into the state error, a penalty value is used in the OSPA. So its performance evaluation includes both source number estimation and speaker position estimation:

$$e_{\mathrm{OSPA}}(\hat{\mathcal{X}}_k, \mathcal{X}_k) = \min_{\pi \in \Pi_{\hat{\Xi}_k, \Xi_k}} \sqrt[a]{\frac{1}{\Xi_k}\left(\sum_{i=1}^{\hat{\Xi}_k} \overline{d}^{(c)}(\hat{\mathbf{x}}_{i,k}, \mathbf{x}_{\pi_i,k})^a + c^a(\Xi_k - \hat{\Xi}_k)\right)} \tag{5}$$

where $\hat{\mathcal{X}}_k = \{\hat{\mathbf{x}}_{1,k}, \ldots, \hat{\mathbf{x}}_{\hat{\Xi}_k,k}\}$ is an estimation of the ground-truth state set $\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{\Xi_k,k}\}$ and $\Pi_{\hat{\Xi}_k,\Xi_k}$ is the set of maps $\pi : 1, \ldots, \hat{\Xi}_k \to 1, \ldots, \Xi_k$. The state cardinality estimation $\hat{\Xi}_k$ may not be the same as the ground truth Ξ_k. The OSPA error defined in Eq. (5) is for $\hat{\Xi}_k \leq \Xi_k$. If $\Xi_k < \hat{\Xi}_k$, then $e_{\mathrm{OSPA}}(\hat{\mathcal{X}}_k, \mathcal{X}_k) = e_{\mathrm{OSPA}}(\mathcal{X}_k, \hat{\mathcal{X}}_k)$. The function $\overline{d}^{(c)}$ is denoted as $\min(c, \overline{d}(\cdot))$. Here, c is defined as the cut-off value in order to weight the penalties for cardinality and localization errors. Additionally, the metric order is defined by a which determines the sensitivity to outliers. The OSPA-T metric differs from other metrics since it considers not only the position estimation of the speaker but also the estimation of the number of speakers in the evaluation of the tracking results. As OSPA-T measures the error based on these two terms, state (position estimation) and cardinality (number of speaker estimation), it causes ambiguities about how much error is contributed from each term to the final error. In addition to the x_1 and x_2 variables of the state vector, the scale variable, s, may be considered in the evaluation. However, this will cause more ambiguities in the contributions of the terms to the final error and deteriorate the reliability of the metric.

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

7. Experimental results and analysis

In this chapter, six trackers are included to cover the recent paradigms. The trackers are restricted to the ones either for which access to the source code has been permitted or tracker performance has been reported on commonly used datasets.

To deal with the tracking problem for unknown and time-varying number of speakers, Kılıç et al. [14] propose to use particle PHD (SMC-PHD) filter. DOA information is employed as an audio cue and it is integrated with video data under SMC-PHD filter framework. Audio data are used to determine when to propagate and re-allocate surviving, spawned and born particles based on their types. The particles are concentrated around the DOA line, which is drawn from the centre of the microphone array to the estimated speaker position by audio information.

As a baseline algorithm, the visual SMC-PHD (V-SMC-PHD) filter, which uses colour information as a visual cue, is compared with the audio-visual SMC-PHD (AV-SMC-PHD) to see the advantage of using multi-modal information in challenging tracking scenarios like occlusion. Sequence 24 from AV16.3 dataset is run for V-SMC-PHD and AV-SMC-PHD, and tracking results are given in **Figure 5**.

The first row shows the results of V-SMC-PHD filter which fails to track after occlusion. Also, it shows poor performance before the occlusion in terms of the detection of the speakers. It is reported in Ref. [14] that the AV-SMC-PHD filter tracks the speakers more accurately and shows better performance than the V-SMC-PHD filter in terms of accuracy and ability for re-detection of the speakers after lost.

The same experiments are repeated for three-speaker case using Sequence 45 camera #3 from the AV16.3 dataset and the results are given in **Figure 6**. It is reported in Ref. [14] that AV-SMC-

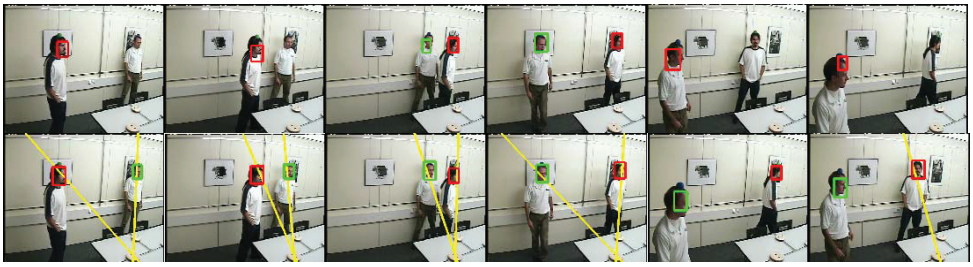

Figure 5. AV16.3, sequence 24 camera #1: occlusions with two speakers [14]. Performance of the V-SMC-PHD filter is shown in the first row. The second row is given for the AV-SMC-PHD filter.

Figure 6. AV16.3, sequence 45 camera #3: occlusions with three speakers [14]. The tracking results of the V-SMC-PHD and the AV-SMC-PHD filters are shown in the first and second rows, respectively.

PHD filter has better capability in detecting and following all the speakers even after the occlusions.

To improve the estimation accuracy of the AV-SMC-PHD filter, [12] integrates the mean-shift method in order to shift the particles to a local maximum of the distribution function which drives particles closer the speaker position. The generic mean-shift algorithm is modified for multiple-speaker case and applied after the audio contribution to the particles, and this algorithm is named as AVMS-SMC-PHD filter.

Even though the integration of the mean-shift improves the estimation accuracy, applying the mean-shift process to all the particles introduces extra computational cost [12]. To address this problem, [12] proposes a sparse sampling scheme which chooses sparse particles and runs the mean-shift method only on those particles rather than all the particles which results in a significant reduction in computational cost. This method is named as sparse-AVMS-SMC-PHD filter. Another tracking algorithm is given in Ref. [122], which uses the merits of dictionary learning for multi-speaker tracking. It is tested using some sequences (seq24, seq25 and seq30) of the AV16.3 dataset.

The results of these five trackers on sequences of AV16.3 are given in **Table 1** and the OSPA-T metric is used for comparison. The tracker in Ref. [122] outperforms the V-SMC-PHD; however, the AVMS-SMC-PHD shows better performance than the others.

These tracking results are compared with those of [123] which uses the PHD filter for tracking and reports the results only for seq24 cam1 and cam2 in terms of Wasserstein distance. **Table 2** shows the results of six trackers.

		Tracking algorithm #1 [122]	**V SMC-PHD [14]**	**AV SMC-PHD [14]**	**AVMS SMC-PHD [14]**	**Sparse AVMS SMC-PHD [14]**
seq24	cam1	22.28	27.12	17.71	13.93	14.50
	cam2	17.60	25.91	19.83	14.97	15.35
	cam3	28.18	24.32	18.94	14.12	15.72
seq25	cam1	21.49	25.84	19.13	15.72	17.17
	cam2	19.17	25.66	18.47	13.93	15.39
	cam3	29.35	29.99	21.61	17.07	17.62
seq30	cam1	35.98	35.60	25.22	16.65	19.27
	cam2	28.40	24.97	19.37	14.86	16.16
	cam3	34.60	37.64	25.31	19.29	19.67
seq45	cam1	NA	48.68	29.46	22.95	23.40
	cam2	NA	39.24	29.47	21.47	23.16
	cam3	NA	39.09	28.43	22.43	23.80
Average		**26.34**	**32.01**	**22.75**	**17.28**	**18.43**

Table 1. Comparison results of the tracking algorithms for the AV16.3 dataset using the OSPA-T metric [14].

seq24	Tracking algorithm #1 [122]	Tracking algorithm #2 [123]	V SMC-PHD [14]	AV SMC-PHD [14]	AVMS SMC-PHD [14]	Sparse AVMS SMC-PHD [14]
cam1	9.02	7.20	16.96	7.94	6.67	7.45
cam2	6.40	4.80	19.17	7.59	5.24	5.73
Average	**7.71**	**6.00**	**18.06**	**7.76**	**5.96**	**6.59**

Table 2. Tracking algorithms are compared in terms of mean Wasserstein distance (in pixel) [14].

Among six trackers, the AVMS-SMC-PHD outperforms the other trackers in terms of the average accuracy.

The trackers of [14] are also tested in different datasets. One sequence from each AMI and CLEAR dataset is used to test the trackers. **Figure 7** shows the results of V-SMC-PHD and AV-SMC-PHD for a sequence of the AMI dataset. In this dataset, the speakers talk one by one. Hence, one DOA line is drawn per time instance. Since the speakers remain still, the visual trackers do not fail to track the speakers.

Other sequence is UKA_20060726 from the CLEAR dataset where the speakers talk one by one and mostly sit around the table. The performance of visual and audio-visual trackers is given in **Figure 8**.

The average error of the trackers for sequences IS1001a and UKA_20060726 is given in **Table 3** in terms of the OSPA-T metric. It is reported in Ref. [14] that there is no significant difference on the performance of the visual and audio-visual trackers since the speakers talk one by one. The audio-visual tracker runs as a visual tracker for the silent speakers, while it is more

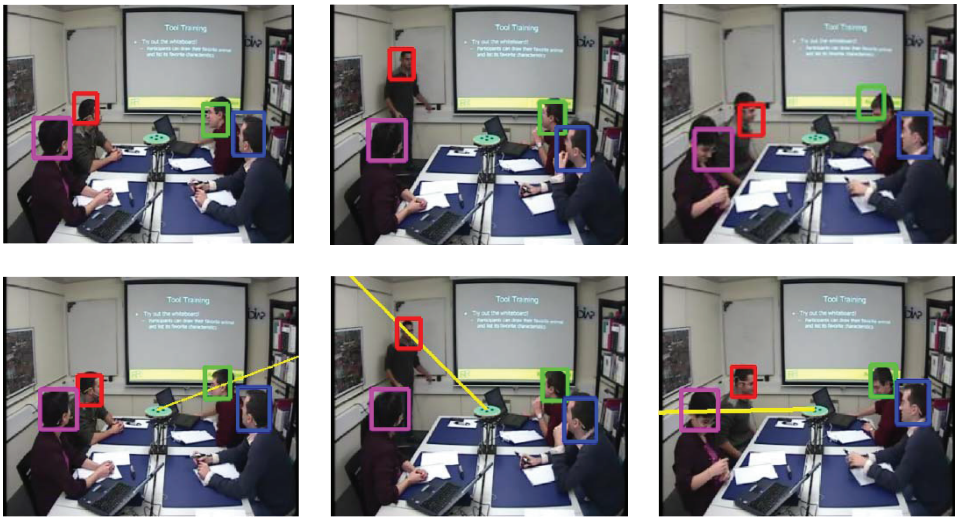

Figure 7. AMI dataset, sequence IS1001a. The first and second rows show the results of the V-SMC-PHD and the AV-SMC-PHD filter, respectively [14].

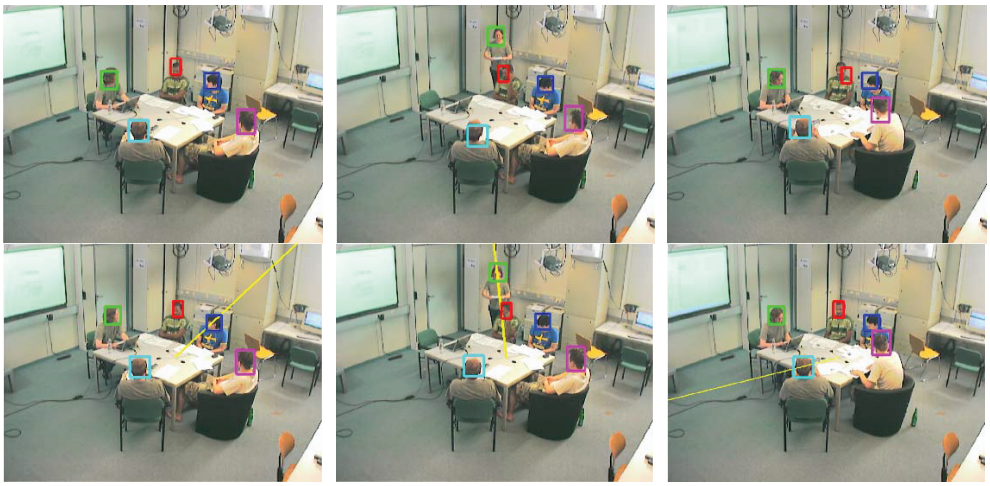

Figure 8. CLEAR dataset, sequence UKA_20060726. The first and second rows show the results of the V-SMC-PHD and the AV-SMC-PHD filters, respectively [14].

Sequences	V SMC-PHD [14]	AV SMC-PHD [14]	AVMS SMC-PHD [14]	Sparse AVMS SMC-PHD [14]
IS1001a	25.32	21.51	18.91	20.37
UKA_20060726	28.33	25.94	23.14	24.82

Table 3. Comparison results of the tracking algorithms for the AMI and CLEAR dataset.

effective for the talking speakers because of the additional information coming from audio modality.

8. Chapter summary

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

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

In order to perform a quantitative evaluation of the proposed algorithms, both audio and video sequences are required. Publicly available datasets such as AV16.3, CLEAR, AMI, SPEVI and S3A were introduced with the fundamental differences including physical setup, scenarios and challenges.

Moreover, performance metrics were analysed in order to see which aspects are considered more in the evaluation and impacts of these perspectives on the evaluation results.

Acknowledgements

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014307/2 and the MOD University Defence Research Collaboration in Signal Processing.

Author details

Volkan Kılıç[1]* and Wenwu Wang[2]

*Address all correspondence to: volkan.kilic@ikc.edu.tr

1 Izmir Katip Celebi University, Izmir, Turkey

2 University of Surrey, Guildford, UK

References

[1] Liu Q, et al. Automating camera management for lecture room environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2001

[2] Talantzis F, Pnevmatikakis A, Constantinides AG. Audio–visual active speaker tracking in cluttered indoors environments. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008;**38**(3):799–807

[3] Wölfel M, McDonough JW. Combining multi-source far distance speech recognition strategies: beamforming, blind channel and confusion network combination. In: INTERSPEECH; 2005

[4] Potamianos G, Neti C, Deligne S. Joint audio-visual speech processing for recognition and enhancement. In: AVSP 2003-International Conference on Audio-Visual Speech Processing; 2003

[5] Naphade MR, et al. Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In: ICIP 98. Proceedings of the 1998 International Conference on Image Processing. IEEE; 1998

[6] Shivappa ST, Rao BD, Trivedi MM. Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation. IEEE Journal of Selected Topics in Signal Processing. 2010;**4(5)**:882–894

[7] Hampapur A, et al. Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. IEEE Signal Processing Magazine. 2005;**22(2)**:38–51

[8] Kılıç V, et al. Audio constrained particle filter based visual tracking. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2013

[9] Kılıç V, et al. Audio assisted robust visual tracking with adaptive particle filtering. IEEE Transactions on Multimedia. 2015;**17**(2):186–200

[10] Katsaggelos AK, Bahaadini S, Molina R. Audiovisual fusion: Challenges and new approaches. Proceedings of the IEEE. 2015;**103(9)**:1635–1653

[11] Atrey PK, et al. Multimodal fusion for multimedia analysis: A survey. Multimedia Systems. 2010;**16(6)**:345–379

[12] Germa T, et al. Vision and RFID data fusion for tracking people in crowds by a mobile robot. Computer Vision and Image Understanding. 2010;**114(6)**:641–651

[13] Liu Q, et al. Identity association using PHD filters in multiple head tracking with depth sensors. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2016

[14] Kılıç V, et al. Mean-shift and sparse sampling-based SMC-PHD filtering for audio informed visual speaker tracking. IEEE Transactions on Multimedia. 2016;**18**(12):2417–2431

[15] Kilic V. Audio-visual tracking of multiple moving speakers [PhD thesis]. University of Surrey; 2016

[16] Smeulders AW, et al. Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;**36(7)**:1442–1468

[17] Liu Q, Zhao X, Hou Z. Survey of single-target visual tracking methods based on online learning. IET Computer Vision. 2014;**8(5)**:419–428

[18] Walia GS, Kapoor R. Human detection in video and images—a state-of-the-art survey. International Journal of Pattern Recognition and Artificial Intelligence. 2014;**28 (03)**:1455004

[19] Fallon MF, Godsill SJ. Acoustic source localization and tracking of a time-varying number of speakers. IEEE Transactions on Audio, Speech, and Language Processing. 2012;**20 (4)**:1409–1415

[20] Potamitis I, Chen H, Tremoulis G. Tracking of multiple moving speakers with multiple microphone arrays. IEEE Transactions on Speech and Audio Processing. 2004;**12(5)**:520–529

[21] Barnard M, Wang W. Audio head pose estimation using the direct to reverberant speech ratio. Speech Communication. 2016;**85**:98–108

[22] Lanz O. Approximate Bayesian multibody tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006;**28(9)**:1436–1449

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

[24] Walia GS, Kapoor R. Recent advances on multicue object tracking: A survey. Artificial Intelligence Review. 2016;**46(1)**:1–39

[25] Sullivan J, Rittscher J. Guiding random particles by deterministic search. In: ICCV 2001. Proceedings of the Eighth IEEE International Conference on Computer Vision. IEEE; 2001

[26] Zhou SK, Chellappa R, Moghaddam B. Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Transactions on Image Processing. 2004;**13 (11)**:1491–1506

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

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

[29] Gatica-Perez D, et al. Audio-visual speaker tracking with importance particle filters. In: ICIP 2003. Proceedings of the 2003 International Conference on Image Processing. IEEE; 2003

[30] Bar-Shalom Y. Tracking and Data Association. Academic Press Professional, Inc.; 1987

[31] Isard M, Blake A. Condensation—Conditional density propagation for visual tracking. International Journal of Computer Vision. 1998;**29(1)**:5–28

[32] Kılıç V, et al. Adaptive particle filtering approach to audio-visual tracking. In: Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European. IEEE; 2013

[33] Closas P, Fernández-Prades C. Particle filtering with adaptive number of particles. In: Aerospace Conference. IEEE; 2011

[34] Fox D. Adapting the sample size in particle filters through KLD-sampling. The International Journal of Robotics Research. 2003;**22(12)**:985–1003

[35] Soto A. Self Adaptive Particle Filter. In: IJCAI; 2005

[36] Mahler RP. Statistical Multisource-Multitarget Information Fusion. Boston, MA, USA: Artech House, Inc.; 2007

[37] Vo B.-N, Singh SS, Ma W-K. Tracking multiple speakers using random sets. In: ICASSP (2); 2004

[38] Kılıç V, et al. Audio-visual tracking of a variable number of speakers with a random finite set approach. In: 2014 17th International Conference on Information Fusion (FUSION). IEEE; 2014

[39] Tang X, et al. A multiple-detection probability hypothesis density filter. IEEE Transactions on Signal Processing. 2015;**63(8)**:2007–2019

[40] Mahler RP. Multitarget Bayes filtering via first-order multitarget moments. IEEE Transactions on Aerospace and Electronic Systems. 2003;**39(4)**:1152–1178

[41] Vo B-N, Ma W-K. A closed-form solution for the probability hypothesis density filter. In: 2005 7th International Conference on Information Fusion. IEEE; 2005

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

[43] Vo B-N, Singh S, Doucet A. Sequential Monte Carlo implementation of the PHD filter for multi-target tracking. In: Proceedings of the International Conference on Information Fusion; 2003

[44] Vo B-N, Singh S, Doucet A. Sequential Monte Carlo methods for multitarget filtering with random finite sets. IEEE Transactions on Aerospace and Electronic Systems. 2005;**41(4)**:1224–1245

[45] Kılıç V, et al. Audio informed visual speaker tracking with SMC-PHD filter. In: 2015 IEEE International Conference on Multimedia and Expo (ICME). IEEE; 2015

[46] Deguchi K, Kawanaka O, Okatani T. Object tracking by the mean-shift of regional color distribution combined with the particle-filter algorithms. In: ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition. IEEE; 2004

[47] Shan C, Tan T, Wei Y. Real-time hand tracking using a mean shift embedded particle filter. Pattern Recognition. 2007;**40(7)**:1958–1970

[48] Shan C, et al. Real time hand tracking by combining particle filtering and mean shift. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition. IEEE; 2004

[49] Wang J, Yagi Y. Adaptive mean-shift tracking with auxiliary particles. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009;**39(6)**:1578–1589

[50] Zoidi O, Tefas A, Pitas I. Visual object tracking based on local steering kernels and color histograms. IEEE Transactions on Circuits and Systems for Video Technology. 2013;**23 (5)**:870–882

[51] Dardari D, Closas P, Djurić PM. Indoor tracking: Theory, methods, and technologies. IEEE Transactions on Vehicular Technology. 2015;**64(4)**:1263–1278

[52] Yang C, Duraiswami R, Davis L. Efficient mean-shift tracking via a new similarity measure. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE; 2005

[53] Raja Y, McKenna SJ, Gong S. Segmentation and tracking using colour mixture models. In: Asian Conference on Computer Vision. Springer; 1998

[54] Yilmaz A, Li X, Shah M. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004;**26(11)**:1531–1536

[55] Wang Y, Lee O. Active mesh-a feature seeking and tracking image sequence representation scheme. IEEE Transactions on Image Processing. 1994;**3**(5):610–624

[56] Micilotta AS, Ong E-J, Bowden R. Real-time upper body detection and 3D pose estimation in monoscopic images. In: European Conference on Computer Vision. Springer; 2006

[57] Fergus R, Perona P, Zisserman A. Object class recognition by unsupervised scale-invariant learning. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE; 2003

[58] Shotton J, et al. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision. 2009;**81(1)**:2–23

[59] Winn J, Criminisi A, Minka T. Object categorization by learned universal visual dictionary. In: Tenth IEEE International Conference on Computer Vision (ICCV'05). Vol. 1. IEEE; 2005

[60] Manjunath BS, Ma W-Y. Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1996;**18(8)**:837–842

[61] Yang H, et al. Recent advances and trends in visual tracking: A review. Neurocomputing. 2011;**74(18)**:3823–3831

[62] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002;**24(7)**:971–987

[63] Perez P, Vermaak J, Blake A. Data fusion for visual tracking with particles. Proceedings of the IEEE. 2004;**92(3)**:495–513

[64] Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 2004;**60(2)**:91–110

[65] Ke Y, Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors. In: CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE; 2004

[66] Sigal L, Sclaroff S, Athitsos V. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2000

[67] Brandstein MS, Silverman HF. A practical methodology for speech source localization with microphone arrays. Computer Speech & Language. 1997;**11(2)**:91–126

[68] DiBiase JH, Silverman HF, Brandstein MS. Robust localization in reverberant rooms. In: Microphone Arrays. Springer; 2001. pp. 157–180

[69] Brandstein MS. A framework for speech source localization using sensor arrays. 1995

[70] Schmidt R. Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation. 1986;**34(3)**:276–280

[71] Wang H, Kaveh M. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1985;**33(4)**:823–831

[72] Ma W-K, et al. Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach. IEEE Transactions on Signal Processing. 2006;**54(9)**:3291–3304

[73] Nguyen Q, Choi J. Localization and tracking for simultaneous speakers based on time-frequency method and probability hypothesis density filter. In: 2011 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE; 2011

[74] Pham NT, Huang W, Ong S. Tracking multiple speakers using CPHD filter. In: Proceedings of the 15th ACM international conference on Multimedia. ACM; 2007

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

[76] Checka N, et al. Multiple person and speaker activity tracking with a particle filter. In: Proceedings (ICASSP'04) of the IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE; 2004

[77] Gatica-Perez D, et al. Audiovisual probabilistic tracking of multiple speakers in meetings. IEEE Transactions on Audio, Speech, and Language Processing. 2007;**15(2)**:601–616

[78] Heuer M, et al. Multi-modal fusion with particle filter for speaker localization and tracking. In: 2011 International Conference on Multimedia Technology (ICMT). IEEE; 2011

[79] Hoseinnezhad R, et al. Bayesian integration of audio and visual information for multi-target tracking using a CB-MeMBer filter. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2011

[80] Talantzis F, Pnevmatikakis A, Polymenakos LC. Real time audio-visual person tracking. In: 2006 IEEE Workshop on Multimedia Signal Processing. IEEE; 2006

[81] Vermaak J, et al. Sequential Monte Carlo fusion of sound and vision for speaker tracking. In: ICCV 2001. Proceedings of the Eighth IEEE International Conference on Computer Vision; 2001; IEEE

[82] Adams M, Vo B-N, Mahler R. Advances in probabilistic modeling: Applications of stochastic geometry [From the Guest Editors]. IEEE Robotics & Automation Magazine. 2014;**21(2)**:21–24

[83] Gehrig T, et al. Kalman filters for audio-video source localization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE; 2005

[84] McGee LA, Schmidt SF. Discovery of the Kalman filter as a practical tool for aerospace and industry. NASA, Moffett Field, CA, USA, NASATM-86847, 1985

[85] Mahler RP. "Statistics 101" for multisensor, multitarget data fusion. IEEE Aerospace and Electronic Systems Magazine. 2004;**19(1)**:53–64

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

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

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

[89] Collins RT. Mean-shift blob tracking through scale space. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE; 2003

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

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

[92] Maggio E, Cavallaro A. Hybrid particle filter and mean shift tracker with adaptive transition model. In: ICASSP (2). Citeseer; 2005

[93] Chang C, Ansari R, Khokhar A. Multiple object tracking with kernel particle filter. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE; 2005

[94] Blackman S, Popoli R. Design and Analysis of Modern Tracking Systems (Book). Norwood, MA: Artech House, 1999

[95] Panta K, et al. Probability hypothesis density filter versus multiple hypothesis tracking. In: Defense and Security. International Society for Optics and Photonics; 2004

[96] Bar-Shalom Y, Li X-R. Multitarget-Multisensor Tracking: Principles and Techniques. Storrs, CT: University of Connecticut; 1995

[97] Chakravorty R, Challa S. Multitarget tracking algorithm-Joint IPDA and Gaussian mixture PHD filter. In: FUSION'09. 12th International Conference on Information Fusion. IEEE; 2009

[98] Jaward M, et al. Multiple object tracking using particle filters. In: 2006 IEEE Aerospace Conference. IEEE; 2006

[99] Kim K, Davis LS. Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In: European Conference on Computer Vision. Springer; 2006

[100] Wang Y, Jing Z, Hu S. Data association for PHD filter based on MHT. In: 2008 11th International Conference on Information Fusion. IEEE; 2008

[101] Arulampalam MS, et al. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing. 2002;**50(2)**:174–188

[102] Gordon N, Doucet A, Freitas J. Sequential Monte Carlo Methods in Practice. Springer-Verlag; 2001

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

[104] Feng P, et al. Adaptive retrodiction particle PHD filter for multiple human tracking. IEEE Signal Processing Letters. 2016;**23(11)**:1592–1596

[105] Yu W, et al. Multi-scale mean shift tracking. IET Computer Vision. 2014;**9(1)**:110–123

[106] Anand S, et al. Semi-supervised kernel mean shift clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;**36(6)**:1201–1215

[107] Yuan X, Li SZ. Half quadratic analysis for mean shift: With extension to a sequential data mode-seeking method. In: 2007 IEEE 11th International Conference on Computer Vision. IEEE; 2007

[108] Tao W, Jin H, Zhang Y. Color image segmentation based on mean shift and normalized cuts. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2007;**37 (5)**:1382–1389

[109] Comaniciu D, Ramesh V. Mean shift and optimal prediction for efficient object tracking. In: Proceedings of the 2000 International Conference on Image Processing. IEEE; 2000

[110] Ukrainitz Y, Sarel B. Mean shift theory and applications. Internet: http://www.wisdom.weizmann.ac.il/~vision/courses/2004_2/files/mean_shift/mean_shift.ppt [Oct. 24, 2014].

[111] Carletta J, et al. The AMI meeting corpus: A pre-announcement. In: International Workshop on Machine Learning for Multimodal Interaction. Springer; 2005

[112] Mostefa D, et al. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. Language Resources and Evaluation. 2007;**41(3–4)**:389–407

[113] M. Taj, School of Electron. Eng. and Comput. Sci., Queen Mary Univ. of London, London, U.K., Surveillance performance evaluation initiative (SPEVI) audiovisual people dataset, 2007 [Online]. Available: http://www.eecs.qmul.ac.uk/~andrea/spevi.html, accessed Aug. 24, 2014.

[114] Lathoud G, Odobez J-M, Gatica-Perez D. AV16. 3: an audio-visual corpus for speaker localization and tracking. In: International Workshop on Machine Learning for Multimodal Interaction. Springer; 2004

[115] *S3A*. 2017. Available from: http://cvssp.org/data/s3a/

[116] Campos T, Liu Q, Barnard M. S3A Speaker Tracking with Kinect2. 2017. Available from: http://epubs.surrey.ac.uk/807708/

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

[118] Li Y, Huang C, Nevatia R. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: CVPR 2009. IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009

[119] Wu B, Nevatia R. Tracking of multiple, partially occluded humans based on static body part detection. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE; 2006

[120] Ristic B, et al. A metric for performance evaluation of multi-target tracking algorithms. IEEE Transactions on Signal Processing. 2011;**59(7)**:3452–3457

[121] Schuhmacher D, Vo B-T, Vo B-N. A consistent metric for performance evaluation of multi-object filters. IEEE Transactions on Signal Processing. 2008;**56(8)**:3447–3457

[122] Barnard M, et al. Robust multi-speaker tracking via dictionary learning and identity modeling. IEEE Transactions on Multimedia. 2014;**16(3)**:864–880

[123] Pham NT, Huang W, Ong SH. Tracking multiple objects using probability hypothesis density filter and color measurements. In: IEEE International Conference on Multimedia and Expo. IEEE; 2007