Audio-Visual Tracking of a Variable Number of Speakers with a Random Finite Set Approach

Volkan Kılıç*, Xionghu Zhong†, Mark Barnard*, Wenwu Wang* and Josef Kittler*
*CVSSP, Department of Electronic Engineering, University of Surrey, UK, GU2 7XH.
†CEMNET, School of Computer Engineering, Nanyang Technological University, Singapore, 639798.
*{v.kilic, mark.barnard, w.wang, j.kittler}@surrey.ac.uk, †xhzhong@ntu.edu.sg

***Abstract*—Speaker tracking in smart environments has attracted an increasing amount of attention in the past few years. Our recent studies show that fusing audio and visual modalities can provide improved robustness and accuracy in some challenging tracking scenarios such as occlusions (by the limited field of view of cameras or by other speakers), as compared with the tracking system based on individual modalities. In these previous works, however, the number of speakers is assumed to be known and remains fixed over the tracking process. In this paper, we focus on a more realistic and complex scenario where the number of speakers is unknown and variable with time. We extend the random finite set (RFS) theory for multi-modal data and devise a particle filter algorithm under the RFS framework for audio-visual (AV) tracking. The experiments on the AV16.3 dataset show the capability of our proposed algorithm for tracking both the number of speakers and the positions of the speakers in challenging scenarios such as occlusions.**

***Index Terms*—Audio-visual speaker tracking, random finite set.**

I. Introduction

The problem of tracking and localization of speakers in enclosed spaces has received much interest in the fields of computer vision and signal processing, driven by applications such as automatic camera steering in video conferencing and individual speaker discriminating in multi-speaker environments.

Speaker tracking may be achieved using either video or audio modalities. Video tracking [1], [2] is generally accurate when the targets are in the camera field of view, but it suffers from occlusions, changes in speaker appearance, illumination, and a limited camera view. Audio tracking [3], [4], on the other hand, is not restricted by these limitations, but could be affected by acoustic noise, room reverberations and the intermittency between utterance and silence. Fusing both audio and visual data has the potential to provide more robust tracking performance in the case that either modality is unavailable or both are corrupted, as demonstrated in our recent work [5], [6].

A comprehensive approach for tracking speakers with audio and video data is to use a state-space approach based on a Bayesian framework. Particle filter (PF) [7] is one of the widely employed algorithms which easily approaches the Bayesian optimal estimate with a sufficiently large number of particles [8]. In our previous works [5], [6], the PF is applied to multi speaker AV tracking under the assumption that the number of speakers is known and constant. In a practical tracking environment, however, the speakers to be captured by the audio-visual sensors may appear or disappear in a random manner. As a result, the number of speakers that can be observed from the audio-visual measurements may vary with time.

In this paper, we relax the above assumption and focus on the problem of tracking a variable number of speakers based on the AV data, where the variable number of speakers and their states are jointly estimated in a multi-speaker environment. A Bayesian tracking framework based on the random finite set (RFS) formulation [9], [10], [11] is used to deal with the unknown and variable number of speakers in audio-visual tracking. Our work is based on the PF implementation of the RFS in speaker tracking presented in [12] and [13]. Different from [12] and [13], however, the RFS approach is extended here to deal with both audio and visual measurements. We show in our experiments that, with this new extension, the proposed AV tracking system is able to track reliably a variable number of speakers in challenging scenarios such as occlusion.

The following section introduces the RFS formulation for multi-speaker tracking. The PF implementation of RFS is given in Section III, and experimental results are presented in Section IV, followed by the conclusion.

II. RFS State Model Formulation For Multi-Speaker Tracking

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

The state of each speaker is defined as $\mathbf{x} = \begin{bmatrix} x & \dot{x} & y & \dot{y} & s \end{bmatrix}^T$, where x and y are the horizontal and vertical positions of the rectangle centred around the face that we wish to track, $\dot{x}$ is the horizontal velocity, $\dot{y}$ is the vertical velocity and s is the scale of the rectangle centred around (x, y). The evolution of time dependent speaker state is

$$\mathbf{x}_{n,k} = \mathbf{F}\mathbf{x}_{n,k-1} + \mathbf{q}_{n,k} \tag{1}$$

where $\mathbf{q}_{n,k}$ is the zero-mean Gaussian noise with covariance $\mathbf{Q}$, $\mathbf{q}_{n,k} \sim \mathcal{N}(0, \mathbf{Q})$ at time frame $k = 1, ..., K$ and $\mathbf{F}$ is the linear motion model,

$$\mathbf{F}=\begin{bmatrix}1 & T & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & T & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\end{bmatrix} \quad \mathbf{Q}=\begin{bmatrix}\sigma_x^2 & 0 & 0 & 0 & 0\\ 0 & \sigma_{xv}^2 & 0 & 0 & 0\\ 0 & 0 & \sigma_y^2 & 0 & 0\\ 0 & 0 & 0 & \sigma_{yv}^2 & 0\\ 0 & 0 & 0 & 0 & \sigma_s^2\end{bmatrix}$$

where T is the period between two adjacent frames, and σ_x^2, σ_y^2, σ_{xv}^2, σ_{yv}^2 and σ_s^2 are the variances for the corresponding state component. In this work, the variances of the vertical components and the horizontal components are assumed to be the same, i.e., $\sigma_x^2=\sigma_y^2=\sigma^2$ and $\sigma_{xv}^2=\sigma_{yv}^2=\sigma_v^2$.

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

$$\mathcal{X}_k=\{\mathbf{x}_{1,k},...,\mathbf{x}_{N_k,k}\} \tag{2}$$

where N_k=$|\mathcal{X}_k|$ is the number of speakers, with $|\cdot|$ representing the cardinality of the set. We assume that the maximum number of speakers at each time step is bounded by $N_{\max}$, i.e., $N_k \leq N_{\max}$. The complete multi-speaker dynamics at current step k can be addressed as

$$\mathcal{X}_k=\mathcal{B}_k(\mathbf{b}_k)\cup\mathcal{S}_k(\mathcal{X}_{k-1}) \tag{3}$$

where $\mathcal{S}_k(\mathcal{X}_{k-1})$ is the survived RFS of the states at time k from the previous speaker's finite set, $\mathcal{B}_k(\mathbf{b}_k)$ is the state vector of the speakers "born" at time step k. We assume that for the birth process at most one speaker is born at a time step and apply following hypotheses:

$$\mathcal{B}_k(\mathbf{b}_k)=\begin{cases}\emptyset, & \bar{\hbar}_{\text{birth}}\\ \{\mathbf{b}_k\}, & \hbar_{\text{birth}}\\ \emptyset, & |\mathcal{X}_{k-1}|=N_{\max}\end{cases} \tag{4}$$

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

$$\mathcal{S}_k(\mathcal{X}_{k-1})=\begin{cases}\mathcal{S}_k(\mathcal{X}_{k-1})\setminus\mathbf{x}_{n,k-1}, & \hbar_{\text{death}}^n\\ \bigcup_{n=1}^{|\mathcal{X}_{k-1}|}\{\mathbf{F}\mathbf{x}_{n,k-1}+\mathbf{q}_{n,k}\}, & \bar{\hbar}_{\text{death}}\end{cases} \tag{5}$$

where $\hbar_{\text{death}}^n$ and $\bar{\hbar}_{\text{death}}$ are, respectively, the death assumption for the nth speaker and the no-death hypothesis, and '$\setminus$' denotes the set subtraction. Here, we assume that each speaker has the same prior probability of disappearing P_{death}. In the case of the death process, the corresponding state is set as empty, and the other states will evolve according to the dynamic model (1).

In this study, we use video and audio modalities which give visual measurement set $\mathcal{Z}_k^{\text{vis}}$ and audio measurement set $\mathcal{Z}_k^{\text{aud}}$ under the assumption that they are independent. $\mathcal{Z}_k^{\text{vis}}$ is the normal distribution of color histogram measurements. $\mathcal{Z}_k^{\text{aud}}$ is also the normal distribution of the direction of arrival (DOA) angle recorded from microphone arrays.

Since the joint information from both the visual and audio measurements are employed, the complete measurement set at time k can also be addressed in an RFS, given as

$$\mathcal{Z}_k=\mathcal{Z}_k^{\text{vis}}\bigcup\mathcal{Z}_k^{\text{aud}} \tag{6}$$

The number of the measurements is the cardinality of the measurement set: $|\mathcal{Z}_k|=M_k^{\text{vis}}+M_k^{\text{aud}}\triangleq M_k$. The likelihood of the visual and audio measurement sets is explained in following sections.

A. Visual Tracking

The measurements observed from video are the color histogram q_k extracted from the video frames. In multi-speaker tracking, we have many color models of templates $\{r_1(u),r_2(u),...,r_t(u)\}$ which are used as references to compare their similarity with q_k in terms of the Bhattacharyya distance.

For the measurement model, each $\mathcal{Z}_k^{\text{vis}}=\{z_{1,k}^{\text{vis}},...,z_{M_k,k}^{\text{vis}}\}$ is modelled by

$$\mathcal{Z}_k^{\text{vis}}=\left\{\bigcup_{i=1,...,|\mathcal{X}_k|}\mathcal{D}_k(\mathbf{x}_{i,k})\right\}\cup\mathcal{C}_k^{\text{vis}} \tag{7}$$

where $\mathcal{C}_k^{\text{vis}}$ is the finite set of false measurements, and $\mathcal{D}_k(\mathbf{x}_{i,k})$ is the set of color measurements given by

$$\mathcal{D}_k(\mathbf{x}_{i,k})=\begin{cases}\emptyset, & \hbar_{\text{miss}}\\ \bigcup_{i=1}^{|\mathcal{X}_{k-1}|}D_{\mathbf{x}_i,k}, & \bar{\hbar}_{\text{miss}}\end{cases} \tag{8}$$

where $\hbar_{\text{miss}}$ and $\bar{\hbar}_{\text{miss}}$ are, respectively, the miss and detection hypotheses. The hypothesis $\hbar_{\text{miss}}$ happens with probability $P_{\text{miss}}^{\text{vis}}$. $D_{\mathbf{x}_i,k}$ is the Bhattacharyya distance:

$$D_{\mathbf{x}_i,k}=\min_j\left(\sqrt{1-\sum_{u=1}^{U}\sqrt{q_{\mathbf{x}_i,k}(u)r_j(u)}}\right) \tag{9}$$

where U is the number of histogram bins, and $r_j(u)$ is the Hue histogram of the reference image from the templates, and $q_{\mathbf{x}_i,k}(u)$ is the Hue histogram extracted from the rectangle centred on the position of the speaker.

$\mathcal{C}_k^{\text{vis}}$ is the finite set of false color measurements. For the false measurements, we assume that each $c_k^{\text{vis}}\in\mathcal{C}_k^{\text{vis}}$ follows a Beta distribution. As for the false color measurement pdf, it is shown that [12]

$$c^{\text{vis}}(\{z_{1,k}^{\text{vis}},...,z_{m,k}^{\text{vis}}\})=P_{|\mathcal{Z}_k^{\text{vis}}|}(m)\left(m!\prod_{i=1}^{m}\kappa^{\text{vis}}(z_{i,k}^{\text{vis}})\right) \tag{10}$$

where $P_{|\mathcal{Z}_k^{\text{vis}}|}(m)$=$P[|\mathcal{Z}_k^{\text{vis}}|=m]$ is the probability of false measurements and $\kappa^{\text{vis}}(z^{\text{vis}})$ is a Beta distribution. The number false measurements $|\mathcal{C}_k^{\text{vis}}|$ is assumed to follow a Poisson distribution with an average rate of λ_c^{vis}, $P_{|\mathcal{Z}_k^{\text{vis}}|}(m)$= $e^{-\lambda_c^{\text{vis}}}\frac{(\lambda_c^{\text{vis}})^m}{m!}$. Therefore, equation (10) can be expressed as

$$c(\mathcal{Z}_k^{\text{vis}})=e^{-\lambda_c^{\text{vis}}}\prod_{z_k^{\text{vis}}\in\mathcal{Z}_k^{\text{vis}}}\lambda_c^{\text{vis}}\kappa^{\text{vis}}(z_k^{\text{vis}}) \tag{11}$$

Assuming that noise on the visual likelihood function is Gaussian, then the likelihood function of the measured color histogram can be written as [14]:

$$g(z_k^{\text{vis}}|\mathbf{x}_k) \propto \mathcal{N}(z_k^{\text{vis}} : 0, \sigma_{\text{vis}}^2) = \frac{1}{\sigma_{\text{vis}}\sqrt{2\pi}} \exp\left\{-\frac{\mathcal{D}_k(\mathbf{x}_{i,k})^2}{2\sigma_{\text{vis}}^2}\right\} \quad (12)$$

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

B. Audio Detection and Tracking

The previous section described the visual measurement set and calculation of the likelihood function. Here, we discuss the estimation of the DOAs and the enhancement of the estimates using a smoothing process based on the Auto-Regressive (AR) model.

As in our previous work [5], [6], the two-step method proposed in [15] is used to estimate DOAs. The first step is the sector based combined detection and localization where the space around a circular microphone array is divided into a number of sectors. Using the SAM-SPARSE-MEAN approach [16], an "activeness" measure for each sector is taken at each time frame. This measure of activeness is then compared to a threshold to decide whether there is an active source in that sector. In the second step a point based search is applied in each of the sectors labelled as having at least one active source. A parametric approach [15] is used for localization and the location parameters are optimized with respect to a cost function such as SRP-PHAT [17]. To improve the estimate of the azimuth we apply a pth order AR model,

$$\theta_{n,k} = \sum_{i=1}^{p} \varphi_i \theta_{n,k-i} + \varepsilon_k \quad (13)$$

where $\theta_{n,k}$ is the DOA (azimuth) angle (in degrees) of the nth speaker, φ_i are the parameters of the model and ε_k is white noise. In this work, the AR order $p = 3$ is found to be adequate.

Then, DOA measurement model for each $\mathcal{Z}_k^{\text{aud}} = \{z_{1,k}^{\text{auds}}, ..., z_{M_k,k}^{\text{aud}}\}$ takes the form

$$\mathcal{Z}_k^{\text{aud}} = \left\{ \bigcup_{i=1,...,|\mathcal{X}_k|} \mathcal{E}_k(\mathbf{x}_{i,k}) \right\} \cup \mathcal{C}_k^{\text{aud}} \quad (14)$$

where $\mathcal{C}_k^{\text{aud}}$ is the finite set of false measurements and $\mathcal{E}_k(\mathbf{x}_{i,k})$ is the difference between DOA angle, $\theta_{n,k}$, and $\psi_{(\mathbf{x}_i,k)}$ which is the speaker position in terms of angle with respect to the microphone

$$\mathcal{E}_k(\mathbf{x}_{i,k}) = \begin{cases} \emptyset, & \hbar_{\text{miss}} \\ \bigcup_{i=1}^{|\mathcal{X}_{k-1}|} \left\{ \min_n \left(\psi_{(\mathbf{x}_i,k)} - \theta_{n,k} \right) \right\}, & \bar{\hbar}_{\text{miss}} \end{cases} \quad (15)$$

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

$\mathcal{C}_k^{\text{aud}}$ is the finite set of false DOA measurements and we assume that each $c_k^{\text{aud}} \in \mathcal{C}_k^{\text{aud}}$ follows a uniform distribution for the false measurements. The false DOA measurement pdf can be shown as

$$c^{\text{aud}}(\{z_{1,k}^{\text{aud}}, ..., z_{m,k}^{\text{aud}}\}) = P_{|\mathcal{Z}_k^{\text{aud}}|}(m) \left(m! \prod_{i=1}^{m} \kappa^{\text{aud}}(z_{i,k}^{\text{aud}}) \right) \quad (16)$$

where $P_{|\mathcal{Z}_k^{\text{aud}}|}(m)$=$P[|\mathcal{Z}_k^{\text{aud}}| = m]$ is the probability of the false measurements and $\kappa^{\text{aud}}(z^{\text{aud}})$ is a uniform density with an interval $[-\theta_{\max}, \theta_{\max}]$. The number false measurements $|\mathcal{C}_k^{\text{aud}}|$ is assumed to follow a Poisson distribution with an average rate of λ_c^{aud}, $P_{|\mathcal{Z}_k^{\text{aud}}|}(m)$= $e^{-\lambda_c^{\text{aud}}} \frac{(\lambda_c^{\text{aud}})^m}{m!}$ and (16) can be expressed as

$$c(\mathcal{Z}_k^{\text{aud}}) = e^{-\lambda_c^{\text{aud}}} \prod_{z_k^{\text{aud}} \in \mathcal{Z}_k^{\text{aud}}} \lambda_c^{\text{aud}} \kappa^{\text{aud}}(z_k^{\text{aud}}) \quad (17)$$

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

$$g(z_k^{\text{aud}}|\mathbf{x}_k) \propto \mathcal{N}(z_k^{\text{aud}} : 0, \sigma_{\text{aud}}^2) = \frac{1}{\sigma_{\text{aud}}\sqrt{2\pi}} \exp\left\{-\frac{\mathcal{E}_k(\mathbf{x}_{i,k})^2}{2\sigma_{\text{aud}}^2}\right\} \quad (18)$$

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

III. Particle Filter Implementation of RFS

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

We can define pdfs for $\mathcal{X}_k$ and $\mathcal{Z}_k$ using the RFS model given in the previous section. RFS state transition density is denoted by

$$f(\mathcal{X}_k|\mathcal{X}_{k-1}) \quad (19)$$

and RFS likelihood function is denoted by

$$g(\mathcal{Z}_k|\mathcal{X}_k) \quad (20)$$

To derive these pdfs, some mathematical concepts are required which are beyond the scope of this paper. Detailed descriptions of the RFS pdf concepts can be found in [11] and [18]. Based on RFS pdf concepts, the derivation of (19) and (20) are given in [12].

The Bayesian recursive estimation of the posterior distribution of the RFS state $p(\mathcal{X}_k|\mathcal{Z}_k)$ can be written as

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}) = \int_{\mathcal{F}} f(\mathcal{X}_k|\mathcal{X}_{k-1}) p(\mathcal{X}_{k-1}|\mathcal{Z}_{1:k-1}) \mu(d\mathcal{X}_{k-1}); \quad (21)$$

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}) \propto p(\mathcal{Z}_{1:k}|\mathcal{X}_k) p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}), \quad (22)$$

where $p(\mathcal{X}_k|\mathcal{X}_{k-1})$ characterizes the birth, death and survival processes of the state dynamics. The subscript $\mathcal{F}$ is the collection of all finite subsets of the state space, and $\mu(d\mathcal{X}_{k-1})$ is a measure on $\mathcal{F}$. Considering that the visual measurement

set $\mathcal{Z}_k^{\text{vis}}$ and the audio measurement set $\mathcal{Z}_k^{\text{aud}}$ are independent, the likelihood for the joint measurements can be written as

$$p(\mathcal{Z}_k|\mathcal{X}_k) = p(\mathcal{Z}_k^{\text{vis}}, \mathcal{Z}_k^{\text{aud}}|\mathcal{X}_k) = p(\mathcal{Z}_k^{\text{vis}}|\mathcal{X}_k)p(\mathcal{Z}_k^{\text{aud}}|\mathcal{X}_k) \quad (23)$$

where $p(\mathcal{Z}_k^{\text{vis}}|\mathcal{X}_k)$ and $p(\mathcal{Z}_k^{\text{aud}}|\mathcal{X}_k)$ are the likelihood for the visual measurements and the audio measurements respectively. Since states have nonlinear relationship with the measurements, closed-form solution for the PDF of the source state is not available. In this paper, a particle filtering approach is employed to approximate the PDFs. Assume that we have particles $\mathcal{X}_{k-1}^{(\ell)}$ for $\ell = 1, \ldots, L$ at the previous time step $k-1$, and the corresponding importance weight $w_{k-1}^{(\ell)}$. The particles at the current time step k are generated according to

$$\mathcal{X}_k^{(\ell)} \sim f(\mathcal{X}_k^{(\ell)}|\mathcal{X}_{k-1}^{(\ell)}). \quad (24)$$

The particles are weighted by

$$w_k^{(\ell)} = w_{k-1}^{(\ell)} g(\mathcal{Z}_k|\mathcal{X}_k^{(\ell)}). \quad (25)$$

After resampling, the posterior distribution is thus approximated by

$$p(\mathcal{X}_k^{(\ell)}|\mathcal{Z}_k) \approx \sum_{\ell=1}^{L} \tilde{w}_k^{(\ell)} \delta_{\mathcal{X}_k^{(\ell)}}(\mathcal{X}_k), \quad (26)$$

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

The proposed tracking algorithm, called RFS-PF algorithm, is presented in Algorithm 1. This algorithm describes how to use RFS-PF for visual or audio-visual tracking. In visual tracking, the color likelihood function is calculated for $N_{\max} = 2$ using equations (27), (28) and (29), respectively for no speaker, one speaker and two speakers. The audio likelihood function is calculated when it exists, again using equations (27) - (29) and is fused with the color likelihood function according to (23).

The number of speakers $\bar{N}_k$ at time k is approximated by

$$\bar{N}_k \approx \sum_{\ell=1}^{L} w_k^{(\ell)} |\mathcal{X}_k^{(\ell)}| \quad (30)$$

$\bar{N}_k$ is the floating number and since the number of speaker should be an integer, rounding operation is applied $\hat{N}_k = \lceil \bar{N}_k \rfloor$. After the K-means algorithm is performed to cluster all the RFS particles, the centroids of these clusters $\{\hat{\mathbf{x}}_{N,k}\}_{N=1}^{\hat{N}_k}$ are taken as the final state estimates. Lastly, the final states of the speakers are used to detect their identities by comparing an image patch centred on the estimated position with the reference image from the templates using the Bhattacharyya distance defined in (9).

Algorithm 1: RFS-PF algorithm for multi-speaker tracking

Initialization: for $\ell = 1, ..., L$, draw particles $\mathbf{x}_0^{(\ell)} \sim \mathcal{N}(.|\mathbf{x}_0, \mathbf{Q})$
Set initial weights $\tilde{w}_0^{(\ell)} = \frac{1}{L}$
while $k \leftarrow 1$ *to* K **do**
  Set $\mathcal{X}_k^{(\ell)} = \emptyset$; Source death, survival and birth:
  - draw a random number $u_d \sim \mathcal{U}[0,1)$; $u_b \sim \mathcal{U}[0,1)$
  if $u_d > P_{death}$ **then**
    - compute $\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{q}_k$
    - set $\mathcal{X}_k^{(\ell)} = \mathcal{X}_k^{(\ell)} \cup \{\mathbf{x}_k\}$
  end
  if $u_b > P_{birth}$ *and* $|\mathcal{X}_{k-1}^{(\ell)}| < N_{max}$ **then**
    - draw an initial state $\mathbf{b}_k$
    - set $\mathcal{X}_k^{(\ell)} = \mathcal{X}_k^{(\ell)} \cup \{\mathbf{b}_k\}$
  end
  for $\ell \leftarrow 1$ *to* L **do**
    - compute the color likelihood $g(\mathcal{Z}_k^{\text{vis}}|\mathcal{X}_k^{(\ell)})$ using related equations (27) - (29)
    if θ_k *exists* **then**
      - compute the audio likelihood $g(\mathcal{Z}_k^{\text{vis}}|\mathcal{X}_k^{(\ell)})$ using related equations (27) - (29)
      - compute the likelihood for the joint measurements:
      $g(\mathcal{Z}_k|\mathcal{X}_k^{(\ell)}) = g(\mathcal{Z}_k^{\text{vis}}|\mathcal{X}_k^{(\ell)})g(\mathcal{Z}_k^{\text{aud}}|\mathcal{X}_k^{(\ell)})$
    else
      $g(\mathcal{Z}_k|\mathcal{X}_k^{(\ell)}) = g(\mathcal{Z}_k^{\text{vis}}|\mathcal{X}_k^{(\ell)})$
    end
    - compute the importance weight:
    $w_k^{(\ell)} = \tilde{w}_{k-1}^{(\ell)} g(\mathcal{Z}_k|\mathcal{X}_k^{(\ell)})$
  end
  - normalize the weight $\tilde{w}_k^{(\ell)} = \frac{w_k^{(\ell)}}{\sum_{\ell=1}^{L} w_k^{(\ell)}}$
  - resample the particles
  - output the estimates using the K-means approach
  - detect speaker id
end

IV. Experimental Evaluations

In this section, evaluations of the RFS-PF algorithm on the $AV16.3$ dataset for visual and AV tracking are presented with performance comparison. First, the experimental setup and the performance metric for tracking error analysis are described, and then comparative results between visual RFS-PF and AV RFS-PF are discussed.

A. Setup and Performance Metric

The RFS-PF was tested using the $AV16.3$ corpus developed by the IDIAP Research Institute [19]. The corpus consists of subjects moving and speaking at the same time whilst being recorded by three calibrated video cameras and two circular eight-element microphone arrays. The audio was recorded at 16 kHz and video was recorded at 25 Hz. They were synchronized before being used in our system. Each video frame is a colour image of 288x360 pixels.

In the sequences, the speakers wear a ball for annotation but in our application this ball is never used. In this paper, we used two multi-speaker sequences. The first one is Sequence 30 (camera #2) where two moving speakers are walking back

$$g(\mathcal{Z}_k|\emptyset) = e^{-\lambda_c}\left(\lambda_c\kappa(z_k)\right)^{|\mathcal{Z}_k|} \quad (27)$$

$$g(\mathcal{Z}_k|\{\mathbf{x}_k\}) = g(\mathcal{Z}_k|\emptyset)\left(P_{\text{miss}} + (1-P_{\text{miss}})\sum_{z_k\in\mathcal{Z}_k}\left(\frac{1}{\lambda_c\kappa(z_k)}\right)g(z_k|\mathbf{x}_k)\right) \quad (28)$$

$$g(\mathcal{Z}_k|\{\mathbf{x}_{1,k},\mathbf{x}_{2,k}\}) = g(\mathcal{Z}_k|\emptyset)\Bigg\{\prod_{i=1,2}\left(P_{\text{miss}} + (1-P_{\text{miss}})\sum_{z_k\in\mathcal{Z}_k}\left(\frac{1}{\lambda_c\kappa(z_k)}\right)g(z_k|\mathbf{x}_{i,k})\right)$$
$$-(1-P_{\text{miss}})^2\sum_{z_k\in\mathcal{Z}_k}\left(\frac{1}{\lambda_c\kappa(z_k)}\right)^2 g(z_k|\mathbf{x}_{1,k})g(z_k|\mathbf{x}_{2,k})\Bigg\} \quad (29)$$

and forth once, one behind the other at a constant distance and both are speaking continuously. Sequence 25 (camera #3) is the second sequence where two moving speakers are walking back and forth twice, each speaker is starting from the opposite side and occluding the other once and the two speakers are talking most of the time.

In Sequence 30 and 25, the number of speakers is changing between 0 to 2. Two speakers occlude each other in Sequence 25 which makes it more challenging than Sequence 30. Therefore, with these two sequences, we are able to evaluate the proposed algorithm on the following two challenging tracking scenarios: a variable number of speakers and speaker occlusion. The parameters for the RFS-PF are set as: $P_{\text{birth}} = 0.2$, $P_{\text{death}} = 0.01$, $P_{\text{miss}}^{\text{vis}} = 0.02$, $P_{\text{death}}^{\text{aud}} = 0.25$, λ_c^{vis} and λ_c^{aud} are set to 3, $\theta_{\max} = \pi/2$ and $L = 500$.

An optimal subpattern assignment (OSPA) metric [20] is used to evaluate the performance of our multi-speaker tracking algorithm. OSPA employs a penalty value to transfer the cardinality error into the state error and is able to present the performance on source number estimation as well as speaker position estimation. Assume that $\hat{\mathcal{X}}_k = \{\hat{\mathbf{x}}_{1,k},...,\hat{\mathbf{x}}_{\hat{N}_k,k}\}$ is an estimation of the ground truth state set $\mathcal{X}_k = \{\mathbf{x}_{1,k},...,\mathbf{x}_{N_k,k}\}$ and $\Pi_{\hat{N}_k,N_k}$ is the set of maps $\pi : 1,...,\hat{N}_k \to 1,...,N_k$. Here the state cardinality estimation $\hat{N}_k$ may not be same as the ground truth N_k. Then, the OSPA error metric for $\hat{N}_k \le N_k$ is given as [20]

$$e_{\text{OSPA}}(\hat{\mathcal{X}}_k,\mathcal{X}_k) = \min_{\pi\in\Pi_{\hat{N}_k,N_k}}\sqrt[p]{\frac{1}{N_k}\left(\sum_{i=1}^{\hat{N}_k} d^{(c)}(\hat{\mathbf{x}}_{i,k},\mathbf{x}_{\pi_i,k})^p + c^p(N_k-\hat{N}_k)\right)} \quad (31)$$

If $N_k < \hat{N}_k$, then $e_{\text{OSPA}}(\hat{\mathcal{X}}_k,\mathcal{X}_k) = e_{\text{OSPA}}(\mathcal{X}_k,\hat{\mathcal{X}}_k)$. The function $d^{(c)}(\cdot)$ is defined as $\min(c,d(\cdot))$. In our case, the cut-off parameter $c = 25$, the OSPA metric order parameter $p = 1$.

B. Results and Discussion

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

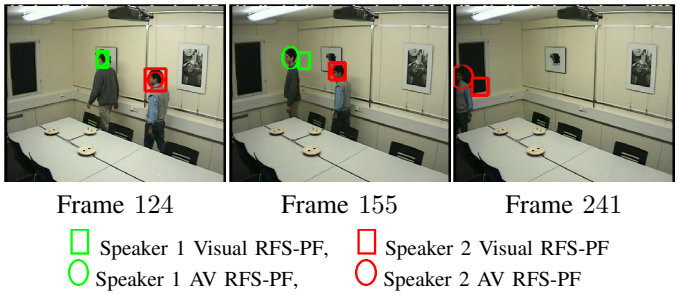


Fig. 1. Tracking in Sequence 30 (camera #2).

At the beginning, both the visual and AV tracker track the speakers well, but when the speakers go to the corner of the room, the visual tracker starts to drift away because of the illumination effects. Figure 2 shows the estimation of the number of active speakers.

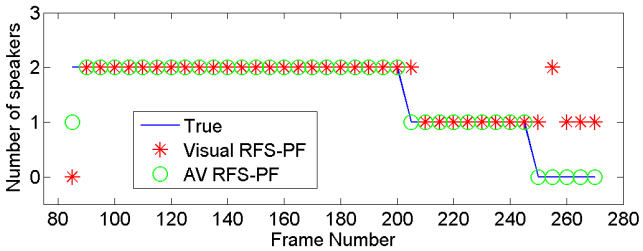
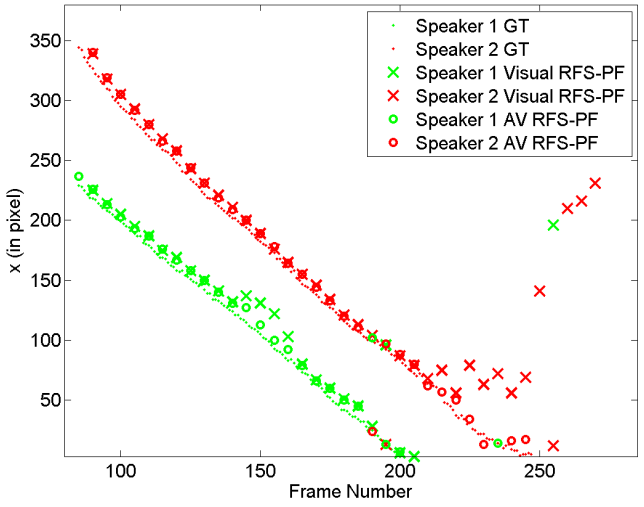


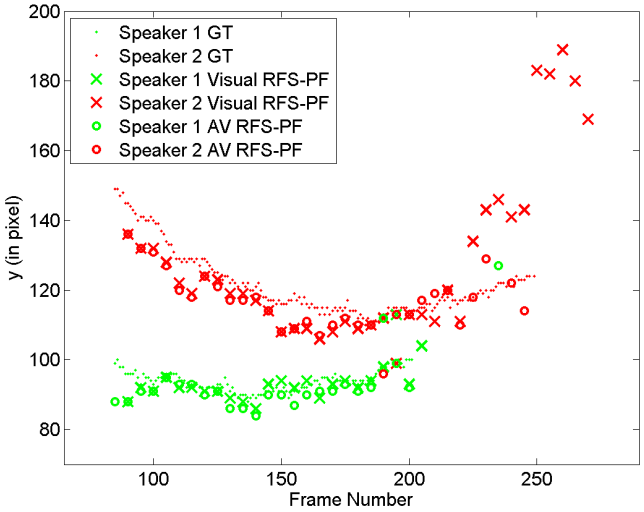
Fig. 2. Estimation of the number of active speakers for Sequence 30.

Here, the number of active speakers is changing from 2 to 0 and the AV tracker shows better performance than the visual tracker. For clearer presentation, downsampling is performed to the plots. Position estimates of the trackers are given in Figure 3 where GT is the abbreviation for ground truth positions. It can be observed that the visual tracker starts to deviate from the ground truth trajectory in the last few frames.

The tracking results of the proposed algorithm for Sequence 25 are demonstrated in Figure 4. Here, the two speakers occlude each other and the AV tracker is able to follow the second speaker after occlusion earlier the visual tracker. The number of active speakers estimated for Sequence 25 is given



(a)



(b)

Fig. 3. Position estimates of the Visual and AV trackers for Sequence 30.

in Figure 5. It can be observed that the performance of the visual tracker is not good as the AV tracker.

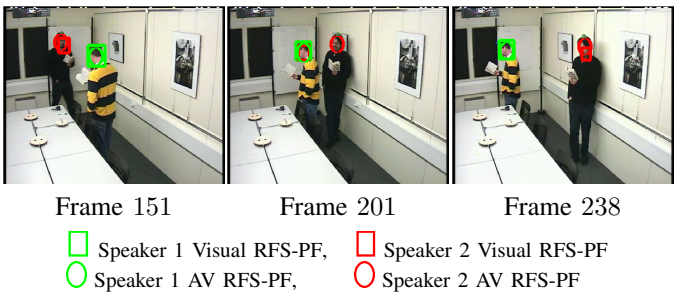


Fig. 4. Tracking in Sequence 25 (camera #3).

The position estimates for x- and y- trajectories are given in Figure 4-(a) and 4-(b), respectively. The AV tracker trajectories follow the ground truth trajectories closer than the visual tracker. To see the performance difference between the

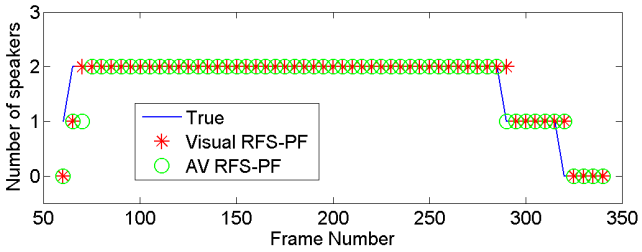
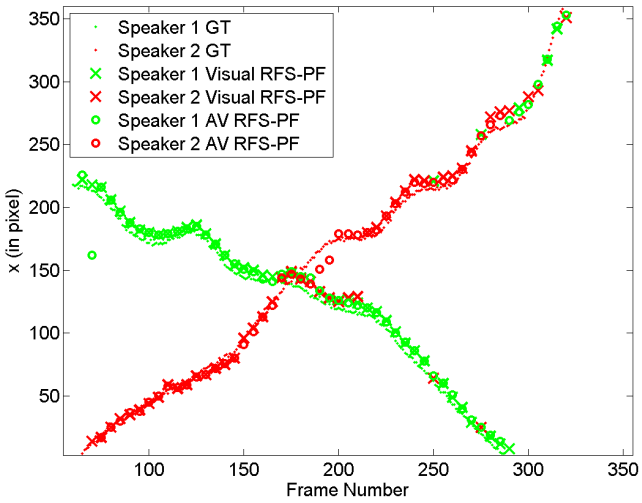
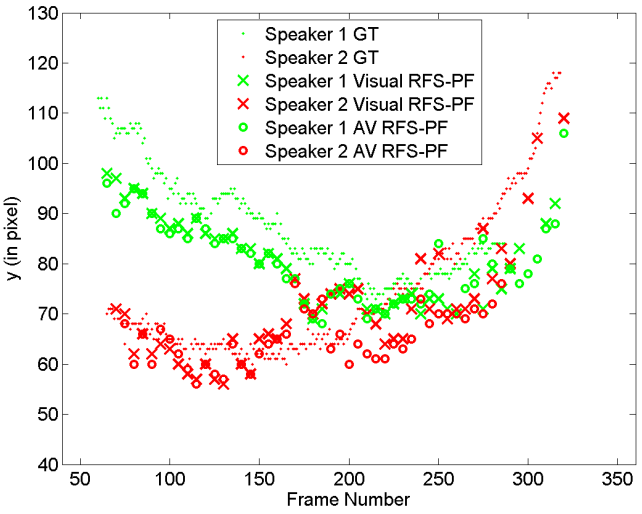


Fig. 5. Estimation of the number of active speakers for Sequence 25.

trackers, the OSPA errors are plotted for Sequence 30 and Sequence 25 in Figure 6-(a) and (b), respectively. To get more reliable results, the experiments are repeated 10 times and the average error is plotted. It is clearly seen that adding audio information to the visual tracker leads to an increase in performance.

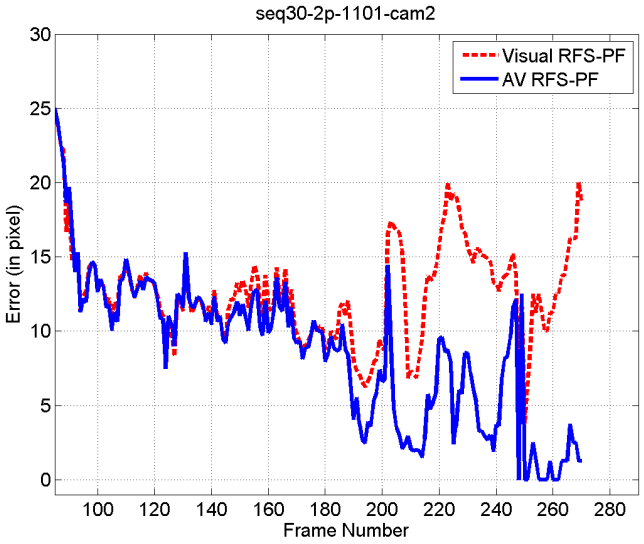


(a)

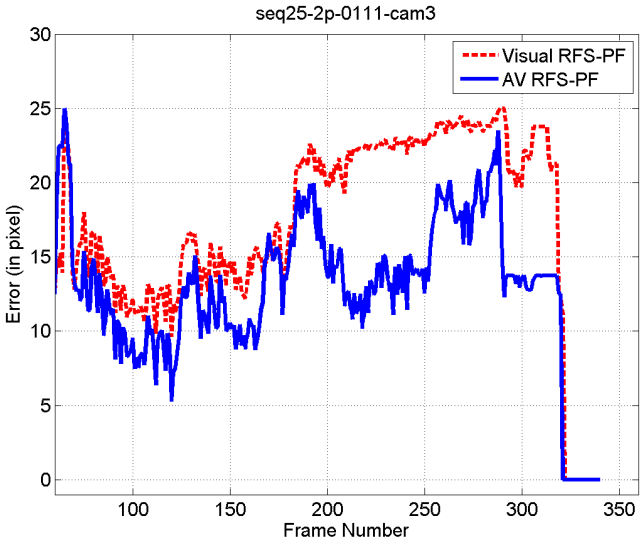


(b)

Fig. 6. Position estimates of the Visual and AV trackers for Sequence 25.



(a)



(b)

Fig. 7. Performance comparison in terms of the OSPA error.

V. Conclusion

In this study, we have proposed a random finite set approach for tracking a variable number of speakers in a smart room environment using audio-visual measurements. The proposed RFS-PF algorithm has been evaluated on two different sequences from the $AV16.3$ dataset. Experimental results demonstrated that the proposed technique can reliably estimate both the number of speakers in the tracking environment and the positions of the speakers for up to three speakers within a challenging tracking scenario such as occlusions.

Acknowledgment

This research was supported by the Engineering and Physical Sciences Research Council of the UK (grant no. EP/H050000/1).

References

[1] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[2] M. M. Trivedi, H. S. Kohsia, and I. Mikic, "Dynamic context capture and distributed video arrays for intelligent spaces," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 35, no. 1, pp. 145–163, 2005.

[3] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[4] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1409–1415, 2012.

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

[6] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Adaptive particle filtering approach to audio-visual tracking," in *Proceedings of the 21st European Signal Processing Conference, Marrakech, Morocco, 9-13 September*, 2013.

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

[8] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[9] B.-N. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 2, pp. ii–357.

[10] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[11] R. P. S. Mahler, "Statistical multisource-multitarget information fusion," *Artech House*, 2007.

[12] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[13] X. Zhong and A. B. Premkumar, "Particle filtering approaches for multiple acoustic source detection and 2-d direction of arrival estimation using a single acoustic vector sensor," *IEEE Transactions on Signal Processing*, vol. 60, pp. 4719–4733, Sept. 2012.

[14] J. Czyz, B. Ristic, and B. Macq, "A color-based particle filter for joint detection and tracking of multiple objects," in *International Conference on Pattern Recognition*, 2005.

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

[16] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Mar. 2005, vol. 3, pp. iii/265 – iii/268.

[17] J. DiBiase, "A high-accuracy, low-latency technique for talker localisation in reverberant environments," in *Ph.D. dissertation, Brown University, Providence, RI, USA*, 2000.

[18] B.-N. Vo, S. Singh, and A. Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1224–1245, 2005.

[19] G. Lathoud, J. M. Odobez, and D. Gatica-Perez, "AV16.3: an audio-visual corpus for speaker localization and tracking," in *Proceedings of the 2004 Machine Learning in Medical Imaging Workshop, S. Bengio and H. Bourlard Eds.* 2005, pp. 182–195, Springer Verlag.

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