LABELLED NON-ZERO PARTICLE FLOW FOR SMC-PHD FILTERING

Yang Liu[1], *Qinghua Hu*[2], *Yuexian Zou*[3], *Wenwu Wang*[1]

[1] Department of Electrical and Electronic Engineering, University of Surrey, UK
[2] School of Computer Science, Tianjin University, China
[3] School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, China
E-mail: yangliu@surrey.ac.uk; huqinghua@tju.edu.cn;
zouyx@pkusz.edu.cn; w.wang@surrey.ac.uk

ABSTRACT

The sequential Monte Carlo probability hypothesis density (SMC-PHD) filter assisted by particle flows (PF) has been shown to be promising for audio-visual multi-speaker tracking. A clustering step is often employed for calculating the particle flow, which leads to a substantial increase in the computational cost. To address this issue, we propose an alternative method based on the labelled non-zero particle flow (LPF) to adjust the particle states. Results obtained from the AV16.3 dataset show improved performance by the proposed method in terms of computational efficiency and tracking accuracy as compared with baseline AV-NPF-SMC-PHD methods.

Index Terms— Audio-visual Tracking, SMC-PHD Filter, Particle Flow

1. INTRODUCTION

Multi-speaker tracking based on audio-visual (AV) data in an enclosed space is an important task in several subject areas such as spatial audio and surveillance. Recently, sequential Monte Carlo PHD filter is proposed for tracking an unknown and variable number of speakers and AV-SMC-PHD filter is used to track speakers with audio-visual (AV) data. However, AV-SMC-PHD filter suffers from the weight degeneracy issue [1].

To solve this problem, particle flow filters [2, 3, 4, 5, 6, 7] have been used by migrating particles from the prior density to the posterior density. Zero diffusion particle flow (ZPF) and non-zero diffusion particle flow (NPF) have been used to improve the AV-SMC-PHD filter as AV-ZPF-SMC-PHD filter [8] and AV-NPF-SMC-PHD filter [9], respectively. As a result, the posterior density becomes more accurate and the chances required for particle resampling are decreased. However, since the label information about particles is unknown, the clustering is a necessary step for the SMC-PHD filter and particle flow and this, however, leads to a substantial increase in the computational cost.

For distinguishing the speakers, the labelled random finite set (RFS) is introduced to address target trajectories and their uniqueness, which is known as the delta-generalized labelled multi-Bernoulli (delta-GLMB) filter [10]. Although the weight degeneracy issue happens in the SMC implementation of the delta-GLMB, the ZPF can be used to improve the delta-GLMB at a cost of a higher computational complexity [11].

In this paper, we propose a labelled non-zero diffusion particle flow (LPF) SMC-PHD filter to address the weight degeneracy issue in the SMC-PHD filter. More specifically, label information of the particles is given by the born step and the particles are then predicted and updated by considering the label information. With label information, the covariance matrix of particles and speaker states can be accurately estimated. The partial derivatives of the flow are simplified. Numerical experiments show that the proposed AV-LPF-SMC-PHD filter significantly increases the acceptance rate of the AV-SMC-PHD filter with a lower computational cost than the baseline AV-NPF-SMC-PHD filter.

The remainder of the paper is organized as follows. Section 2 presents the problem and related work. We describe the proposed method in Section 3. The simulation results are presented in Section 4. Concluding remarks are provided in Section 5.

2. PROBLEM STATEMENT AND BACKGROUND

This section describes our problem formulation and the AV-NPF-SMC-PHD filter. We assume that the speaker dynamics and observations are described as:

$$\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k} = \mathbf{F}_{\tilde{\boldsymbol{m}}}\left(\{\tilde{\boldsymbol{m}}_{k-1}^j\}_{j=1}^{\tilde{N}_{k-1}}, \boldsymbol{\Upsilon}_k\right) \qquad (1)$$

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1), the BBC as part of the BBC Audio Research Partnership, the China Scholarship Council (CSC), and in part by the EPSRC grant EP/K014307/2.

$$\{\mathring{\boldsymbol{z}}_k^o\}_{o=1}^{\mathring{N}_k} = \mathring{\mathbf{F}}_{\boldsymbol{z}} \left(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \mathring{\boldsymbol{\Psi}}_k \right) + \mathring{\boldsymbol{\epsilon}}_k \tag{2}$$

$$\{\breve{\boldsymbol{z}}_k^u\}_{u=1}^{\breve{N}_k} = \breve{\mathbf{F}}_{\boldsymbol{z}} \left(\{\tilde{\boldsymbol{m}}_k^j\}_{j=1}^{\tilde{N}_k}, \breve{\boldsymbol{\Psi}}_k \right) + \breve{\boldsymbol{\epsilon}}_k \tag{3}$$

where $\tilde{\boldsymbol{m}}_k^j \in \mathbb{R}^M$ is the speaker state vector at time k, $\tilde{}$ is used to distinguish the speaker state from the particle state used later, and $\tilde{N}_k$ is the number of speakers at time k. Let $\{\mathring{\boldsymbol{z}}_k^o\}_{o=1}^{\mathring{N}_k}$ and $\{\breve{\boldsymbol{z}}_k^u\}_{u=1}^{\breve{N}_k}$ denote the set of $\mathring{N}_k$ audio and $\breve{N}_k$ visual measurements at time k, respectively. In this paper, the state $\tilde{\boldsymbol{m}}_k^j = [x_k^j, y_k^j, \dot{x}_k^j, \dot{y}_k^j]^T$ consists of positions (x_k^j, y_k^j) and velocities $(\dot{x}_k^j, \dot{y}_k^j)$, while the measurement is a noisy version of the position. We define the system excitation as $\boldsymbol{\Upsilon}_k$. The measurement noise and clutter terms are denoted as $\mathring{\boldsymbol{\Psi}}_k$ and $\mathring{\boldsymbol{\epsilon}}_k$ for audio measurements, and $\breve{\boldsymbol{\Psi}}_k$ and $\breve{\boldsymbol{\epsilon}}_k$ for visual measurements, respectively. The transition model is denoted as $\mathbf{F}_{\tilde{\boldsymbol{m}}}$. The nonlinear measurement model for audio and visual information are denoted as $\mathring{\mathbf{F}}_{\boldsymbol{z}}$ and $\breve{\mathbf{F}}_{\boldsymbol{z}}$, respectively.

In [9], an AV-NPF-SMC-PHD filter is presented for audio-visual multi-speaker tracking. The audio information and visual information are applied in the prediction and update steps. The NPF is used to address the weight degeneracy issue. Direction of arrival (DOA) lines drawn from the microphone array to the speaker are applied for re-locating the existing particles [12]. In the prediction step, the particle set is $\{\boldsymbol{m}_{k-1}^i, \omega_{k-1}^i\}_{i=1}^{N_k}$, where N_k is the number of particles at time k, and $\boldsymbol{m}_{k-1}^i$ and ω_{k-1}^i are the state and weight of the i-th particle at time $k-1$. The particle weights are then calculated as:

$$\omega_{k|k-1}^i = \frac{\phi\left(\boldsymbol{m}_{k|k-1}^i | \boldsymbol{m}_{k-1}^i\right)\omega_{k-1}^i}{q_k\left(\boldsymbol{m}_{k|k-1}^i | \boldsymbol{m}_{k-1}^i, \boldsymbol{Z}_k\right)}, i = 1, ..., N_k \tag{4}$$

$$\omega_{k|k-1}^i = \frac{\gamma_k(\boldsymbol{m}_{k|k-1}^i)}{N_B p_k(\boldsymbol{m}_{k|k-1}^i | \boldsymbol{Z}_k)}, i = N_k + 1, ..., N_k + N_B \tag{5}$$

where ϕ and q_k are the state transition probability and the proposal distribution, respectively. If a new speaker appears, N_B particles are sampled from the new born importance function p_k and the PHD of the new born speaker γ_k.

In the update step, the particle state is updated by the NPF,

$$\boldsymbol{m}_k^i \Leftarrow \boldsymbol{m}_k^i + \triangle\boldsymbol{m}_k^i \lambda \tag{6}$$

where

$$\triangle\boldsymbol{m}_k^i = \boldsymbol{f}_k^i(\boldsymbol{m}_k^i, \lambda)\triangle\lambda + \upsilon_k^i \boldsymbol{w}_k^i \tag{7}$$

where $\boldsymbol{f}_k^i \in \mathbb{R}^M$ is the particle flow vector and $\boldsymbol{w}_k^i \in \mathbb{R}^M$ is the Wiener process with the diffusion coefficient υ_k^i. It moves the particle $\boldsymbol{m}_{k|k-1}^i$ with the distance $\triangle\boldsymbol{m}_{k|k-1}^i$ for the time period $\triangle\lambda$. Based on the Fokker-Planck equation [13], the non-zero particle flow $\boldsymbol{f}_k^i$ is calculated by the partial differential equation:

$$\boldsymbol{f}_k^i = -[\boldsymbol{\nabla}^2 \log \psi_k^i]^{-1}(\boldsymbol{\nabla} \log h_k^i) \tag{8}$$

where

$$\boldsymbol{\nabla}^2 \log \psi_k^i \approx -(\boldsymbol{P}_{k|k-1}^i)^{-1} + \lambda \boldsymbol{\nabla}^2 \log h_k^i \tag{9}$$

where $\boldsymbol{P}_{k|k-1}^i$ is the covariance matrix of $\boldsymbol{m}_{k|k-1}^i$. The derivation of Eq. (8) can be found in [14]. Then the audio-visual likelihood function h_k^i is obtained as:

$$h_k^i = \frac{\mathring{\boldsymbol{h}}_k^{i\,T} \mathring{\boldsymbol{\omega}}_k + \breve{\boldsymbol{h}}_k^{i\,T} \breve{\boldsymbol{\omega}}_k}{\|\mathring{\boldsymbol{\omega}}_k\|_1 + \|\breve{\boldsymbol{\omega}}_k\|_1} \tag{10}$$

where $\mathring{\boldsymbol{\omega}}_k$ and $\breve{\boldsymbol{\omega}}_k$ are the weight sets for the audio and visual likelihood, respectively, and $\|\cdot\|_1$ denotes the L_1 norm. The first and second derivative of the likelihood function can be found in [9]. Then the weights of particles are calculated as

$$\omega_k^i = \left[1 - p_{D,k}^i + \sum_{\boldsymbol{z}_k^r \in \boldsymbol{Z}_k} \frac{p_{D,k}^i h_k^{i,r}}{\kappa_k(\boldsymbol{z}_k^r) + G_k^r}\right] \omega_{k|k-1}^i \tag{11}$$

where

$$G_k^r = \sum_{i=1}^{N_k} p_{D,k}^i h_k^{i,r} \omega_{k|k-1}^i \tag{12}$$

in which $\kappa_k(\boldsymbol{z}_k^r)$ denotes the clutter intensity of the r-th measurement $\boldsymbol{z}_k^r$ at time k, $p_{D,k}^i$ is the detection probability at time k, and $h_k^{i,r}$ is the likelihood of the i-th particle for the r-th measurement. The measurement $\boldsymbol{z}_k^r$ is calculated by $\mathring{\boldsymbol{z}}_k^o$ and $\breve{\boldsymbol{z}}_k^u$ [9]. The number of speakers is estimated as the sum of the weights. The states and weights of the speakers $\{\tilde{\boldsymbol{m}}_k^j, \tilde{\omega}_k^j\}_{j=1}^{\tilde{N}_k}$ can be calculated using a clustering step e.g. the k-means clustering method [15]. Finally, resampling is performed when the effective sample size (ESS) [16] is smaller than half number of particles. More detail of the AV-NPF-SMC-PHD filter can be found in [9].

3. AUDIO-VISUAL LABELLED NON-ZERO DIFFUSION PARTICLE FLOW SMC-PHD FILTER

In the AV-NPF-SMC-PHD filter, the speaker states and the covariance matrix of the particles are estimated by a clustering step, which directly affects the tracking performance. In this section, an improved version of the AV-NPF-SMC-PHD filter is proposed with the label information. At time k, the label sets for the audio and visual particles are given as $\{\mathring{l}_k^i\}_{i=1}^{N_k}$ and $\{\breve{l}_k^i\}_{i=1}^{N_k}$, respectively, where $\mathring{l}_k^i \in \{0, ..., \mathring{N}_k\}$ and $\breve{l}_k^i \in \{0, ..., \breve{N}_k\}$ with $\mathring{N}_k$ and $\breve{N}_k$ being the number of audio and visual measurements, respectively. When $\mathring{l}_k^i = 0$ or $\breve{l}_k^i = 0$, the i-th particle is not detected by audio or visual

Algorithm 1 AV-LPF-SMC-PHD Filter

```
Input: $\{\boldsymbol{m}^i_{k-1}, \omega^i_{k-1}\}_{i=1}^{N_{k-1}}$, $N_B$, $k$, $\{\mathring{\boldsymbol{z}}^o_k\}_{o=1}^{\mathring{N}_k}$ and $\{\breve{\boldsymbol{z}}^u_k\}_{u=1}^{\breve{N}_k}$.
Output: $\{\tilde{\boldsymbol{m}}^j_k, \tilde{\omega}^j_k\}_{j=1}^{\tilde{N}_k}$, and $\{\boldsymbol{m}^i_k, \omega^i_k\}_{i=1}^{N_k}$.
    Initialize: $\boldsymbol{\Upsilon}_k$, $N_B$, $q_k$, $\phi_{k|k-1}$, $p_k$, $\gamma_k$, $\breve{\boldsymbol{\Psi}}$ and $\mathring{\boldsymbol{\Psi}}$.
    Run:
    for Each visual measurement $\breve{\boldsymbol{z}}^u_k$ do
        for Each audio measurement $\mathring{\boldsymbol{z}}^o_k$ do
            Create $\frac{N_B}{\mathring{N}_k+\breve{N}_k}$ particles as Eq. (13)
            Calculate particle weight $\omega^i_{k|k-1}$ as Eq. (5).
        end for
    end for
    Propagate surviving particles $\{\boldsymbol{m}^i_{k|k-1}\}_{i=1}^{N_{k-1}}$ as Eq. (1).
    Calculate particle weights $\omega^i_{k|k-1}$ as Eq. (4).
    Set particle label $\mathring{l}^i_k$ and $\breve{l}^i_k$ as Eq. (15, 16)
    Set $\{\boldsymbol{m}^i_k\}_{i=1}^{N_k}$ as $\{\boldsymbol{m}^i_{k|k-1}\}_{i=1}^{N_{k-1}}$.
    for $i \in [1, ..., N_k]$ do
        Calculate the audio-visual likelihood $h^i_k$ by Eq. (21).
        for $\lambda \in [0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ do
            Evaluate flow $\boldsymbol{f}^i_k$ by Eq. (8).
            Update $\triangle\boldsymbol{m}^i_k$ by Eq. (7) and $\boldsymbol{m}^i_k \Leftarrow \boldsymbol{m}^i_k + \triangle\boldsymbol{m}^i_k\lambda$.
        end for
    end for
    Combine all the particles:$\{\boldsymbol{m}^i_k, \omega^i_{k|k-1}\}_{i=1}^{N_k} \Leftarrow$
    $\{\boldsymbol{m}^i_k, \omega^i_{k|k-1}\}_{i=1}^{N_k} \cup \{\boldsymbol{m}^i_{k|k-1}, \omega^i_{k|k-1}\}_{i=N_k+1}^{N_k+N_B}$.
    Calculate the particle weight $\omega^i_k$ by Eq. (22).
    Set $j = 0$.
    for Each audio measurement $\mathring{\boldsymbol{z}}^o_k$ do
        for Each visual measurement $\breve{\boldsymbol{z}}^u_k$ do
            Calculate the speaker weight $\tilde{\omega}^j_k$ by Eq. (22).
            if $\tilde{\omega}^j_k \geq 0.5$ then
                Calculate the speaker state $\tilde{m}^j_k$ by Eq. (24).
                $j \Leftarrow j + 1$.
            end if
        end for
    end for
    if ESS $< N_k/2$ then
        (Optional) Re-sample $\{\boldsymbol{m}^i_k, \omega^i_k\}_{i=1}^{N_k}$.
    end if
```

information. For the audio and visual measurements $\mathring{\boldsymbol{z}}^o_k$ and $\breve{\boldsymbol{z}}^u_k$, the new born particle state is shown as

$$\boldsymbol{m}^i_{k|k-1} \propto \mathcal{N}(\mathring{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\mathring{\boldsymbol{z}}^o_k, \mathring{\boldsymbol{\Psi}})\mathcal{N}(\breve{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\breve{\boldsymbol{z}}^u_k, \breve{\boldsymbol{\Psi}}) \tag{13}$$

Then we set the audio and visual label of the i-th particle as o and u. As the states of the born particles are estimated by the measurements, particle flow is not applied to them and their weights are estimated by Eq. (5). For surviving particles, the particle state $\boldsymbol{m}^i_{k|k-1}$ is calculated by the transition function.

$$\{\boldsymbol{m}^i_{k|k-1}\}_{i=1}^{N_k} = \mathbf{F}_{\boldsymbol{m}}\left(\{\boldsymbol{m}^i_{k-1}\}_{i=1}^{N_{k-1}}, \boldsymbol{\Upsilon}_k\right) \tag{14}$$

where $\mathbf{F}_{\boldsymbol{m}} = \mathbf{F}_{\tilde{\boldsymbol{m}}}$. The labels of the surviving particles are given as:

$$\mathring{l}^i_k = \delta_{r(1)}(\mathring{P}_D)\arg\max_o(r(1)\mathcal{N}(\mathring{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\mathring{\boldsymbol{z}}^o_k, \mathring{\boldsymbol{\Psi}})) \tag{15}$$

$$\breve{l}^i_k = \delta_{r(1)}(\breve{P}_D)\arg\max_u(r(1)\mathcal{N}(\breve{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\breve{\boldsymbol{z}}^u_k, \breve{\boldsymbol{\Psi}})) \tag{16}$$

where

$$\delta_X(Y) = \begin{cases} 1, X < Y \\ 0, X \geq Y \end{cases} \tag{17}$$

where $r(1)$ is a random value from 0 to 1. The covariance matrix of the i-th particle is calculated as

$$\boldsymbol{P}^i_{k|k-1} = \frac{\sum_{\alpha=1}^{N_k} s_{\mathring{l}^i_k}(\mathring{l}^\alpha_k)s_{\breve{l}^i_k}(\breve{l}^\alpha_k)[\omega^i_{k|k-1}\boldsymbol{e}(\boldsymbol{m}^i_{k|k-1})\boldsymbol{e}(\boldsymbol{m}^i_{k|k-1})^T]}{\sum_{\alpha=1}^{N_k} s_{\mathring{l}^i_k}(\mathring{l}^\alpha_k)s_{\breve{l}^i_k}(\breve{l}^\alpha_k)\omega^i_{k|k-1}} \tag{18}$$

where

$$\boldsymbol{e}(\boldsymbol{m}^i_{k|k-1}) = \boldsymbol{m}^i_{k|k-1} - \frac{\sum_{\alpha=1}^{N_k} s_{\mathring{l}^i_k}(\mathring{l}^\alpha_k)s_{\breve{l}^i_k}(\breve{l}^\alpha_k)\left(\omega^\alpha_{k|k-1}\boldsymbol{m}^\alpha_{k|k-1}\right)}{\sum_{\alpha=1}^{N_k} s_{\mathring{l}^i_k}(\mathring{l}^\alpha_k)s_{\breve{l}^i_k}(\breve{l}^\alpha_k)\omega^\alpha_{k|k-1}} \tag{19}$$

$$s_X(Y) = \begin{cases} 1, X = Y \\ 0, X \neq Y \end{cases} \tag{20}$$

The likelihood of the i-th particle is defined as

$$h^i_k = \begin{cases} \mathcal{N}(\mathring{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\mathring{\boldsymbol{z}}^{\mathring{l}^i_k}_k, \mathring{\boldsymbol{\Psi}})\mathcal{N}(\breve{\boldsymbol{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\breve{\boldsymbol{z}}^{\breve{l}^i_k}_k, \breve{\boldsymbol{\Psi}}) \\ \text{,if } \mathring{l}^i_k > 0 \text{ and } \breve{l}^i_k > 0 \\ \mathcal{N}(\mathring{\mathbf{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\mathring{\boldsymbol{z}}^{\mathring{l}^i_k}_k, \mathring{\boldsymbol{\Psi}})\text{,if } \mathring{l}^i_k > 0 \text{ and } \breve{l}^i_k = 0 \\ \mathcal{N}(\breve{\boldsymbol{F}}_{\boldsymbol{z}}(\boldsymbol{m}^i_{k|k-1})|\breve{\boldsymbol{z}}^{\breve{l}^i_k}_k, \breve{\boldsymbol{\Psi}})\text{,if } \mathring{l}^i_k = 0 \text{ and } \breve{l}^i_k > 0 \\ 0\text{,if } \mathring{l}^i_k = 0 \text{ and } \breve{l}^i_k = 0 \end{cases} \tag{21}$$

By particles flow Eq. (6-8), the particle state is updated and then the weight of the particle is calculated as:

$$\omega^i_k = \begin{cases} (1-\mathring{p}^i_{D,k})(1-\breve{p}^i_{D,k}), \mathring{l}^i_k = 0 \text{ and } \breve{l}^i_k = 0 \\ \frac{(\mathring{p}^i_{D,k})^{\delta_0(\mathring{l}^i_k)}(\breve{p}^i_{D,k})^{\delta_0(\breve{l}^i_k)}h^i_k\phi\left(\boldsymbol{m}^i_k|\boldsymbol{m}^i_{k-1}\right)\omega^i_{k|k-1}}{(\mathring{\kappa}_k+\breve{\kappa}_k+G^i_k)\phi\left(\boldsymbol{m}^i_{k|k-1}|\boldsymbol{m}^i_{k-1}\right)|\det(\boldsymbol{I}+\Delta\lambda\nabla f)|^{-1}}, \text{others} \end{cases} \tag{22}$$

where

$$G^i_k = \sum_{\alpha=1}^{N_k} s_{\mathring{l}^\alpha_k}(\mathring{l}^i_k)s_{\breve{l}^\alpha_k}(\breve{l}^i_k)(\mathring{p}^i_{D,k})^{\delta_0(\mathring{l}^\alpha_k)}(\breve{p}^i_{D,k})^{\delta_0(\breve{l}^\alpha_k)}h^\alpha_k\omega^\alpha_{k|k-1} \tag{23}$$

where $\mathring{p}^i_{D,k}$ and $\breve{p}^i_{D,k}$ are the detection possibility for audio and visual measurements, respectively, det denotes the determinant, and $\mathring{\kappa}_k$ and $\breve{\kappa}_k$ are the clutter densities for audio and visual measurements, respectively. For $o \in [0, ..., \mathring{N}_k]$ and $u \in [0, ..., \breve{N}_k]$, the estimated state for speakers is given by

$$\tilde{\boldsymbol{m}}^j_k = \frac{\sum_{i=1}^{N_k} s_{\mathring{l}^i_k}(o)s_{\breve{l}^i_k}(u)\left(\omega^i_{k|k-1}\boldsymbol{m}^i_{k|k-1}\right)}{\tilde{\omega}^j_k} \tag{24}$$

Table 1. The OSPA for the AV-LPF-SMC-PHD, AV-NPF-SMC-PHD, AV-ZPF-SMC-PHD, AV-PF-PF, AV-PF-GLMB filters, which are denoted in short as LPF, NPF, ZPF, PPF, GLMB, respectively.

Seq (Cam)	LPF	NPF	ZPF	PPF	GLMB
24 (1)	10.64	12.32	12.99	12.18	10.66
24 (2)	11.22	13.20	13.82	13.12	11.98
24 (3)	10.98	13.23	14.01	13.02	11.62
25 (1)	13.09	15.96	16.80	14.90	13.43
25 (2)	13.93	15.29	15.88	13.08	12.94
25 (3)	13.62	16.29	17.56	14.98	14.45
30 (1)	13.44	15.76	17.15	15.29	13.55
30 (2)	11.66	13.41	14.22	13.86	11.68
30 (3)	13.06	15.93	17.63	15.61	16.38
45 (1)	15.53	17.65	19.33	24.50	18.14
45 (2)	15.62	18.60	20.85	22.26	20.35
45 (3)	16.77	19.50	21.35	24.34	20.36
Avg. OSPA	**13.29**	**15.60**	**16.80**	**16.43**	**14.63**

where

$$\tilde{\omega}_k^j = \sum_{i=1}^{N_k} s_{\mathring{l}_k^i}(o) s_{\breve{l}_k^i}(u) \omega_{k|k-1}^i \tag{25}$$

After calculating the target weight $\tilde{\omega}_k^j$, the target which has a weight lower than a threshold ξ $(0 < \xi < 1)$ is considered as noise and ignored. When the noise level of the measurements is high, ξ should be set as a low value. In our experiment, we set ξ as 0.5. The pseudo-code of the AV-LPF-SMC-PHD filter is presented in Algorithm 1.

4. EXPERIMENTAL RESULTS

In this section, the proposed algorithm is compared to the other particle flow SMC filters, which include AV-PF-PF, AV particle flow superpositional GLMB filter (AV-PF-GLMB) [11], AV-ZPF-SMC-PHD [8] and AV-NPF-SMC-PHD algorithms [9] using the AV16.3 dataset. Zero diffusion particle flow has been used for improving the tracking accuracy of PF-PF [17, 18] and GLMB filter [19, 11]. In this paper, they are used as AV-PF-PF and AV-PF-GLMB filter for the AV data. For all the filters, the same measurements are applied. The color histograms are used as the visual measurements while the DOA lines are used as the audio measurements. More detail is given in [9]. The experiments are run 10 times in Matlab on Windows 7 with Intel i7. The AV16.3 dataset, in which multiple speakers keep speaking and walking, is applied to test the performance of all the filters. The speakers are recorded by three video cameras at 25 Hz and two circular eight-element microphone arrays at 16 kHz. The pixels of the image frame are 288x360. The audio and video streams are synchronized.

Table 2. Experimental results for the AV-LPF-SMC-PHD, AV-NPF-SMC-PHD, AV-ZPF-SMC-PHD, AV-PF-PF and AV-PF-GLMB filters respectively, which are denoted in short as LPF, NPF, ZPF, PPF and GLMB, in terms of ESS, resampling times and the running times for sequence 45 (camera 1).

Filter	ESS	Resampling	Time(s)
LPF	81.4	37	114.7
NPF	82.1	36	163.8
ZPF	77.8	58	268.0
PPF	70.5	68	215.9
GLMB	75.7	63	1325.8

The Optimal Sub-pattern Assignment (OSPA) for trackers [20], which gives a combined score for the estimation performance in the number of sources and their positions, is used to evaluate the tracking accuracy. Apart from that, ESS [16] is used to show the level of the weight degeneracy problem [21, 7]. The parameters are set as: $N_B = 50$, $\mathring{p}_{D,k}^i = 0.98$, $\breve{p}_{D,k}^i = 0.98$, $\mathring{\kappa}_k = 0.0035$ and $\breve{\kappa}_k = 0.0035$. Other parameters of the PHD filter and particle flow filters are set as in [12, 8, 9]. The order parameter in the OSPA metric is 2. The number of particles per speaker is 50 and the particles are spread randomly in the tracking area.

Table 1 reports the average OSPA over 10 random tests. The first column, e.g. 24 (1), shows the sequence and camera number. The OSPA of AV-LPF-PHD filter is only 13.29 which is the lowest among all the compared methods. With the contribution of the label information, 15% reduction in tracking error has been achieved as compared with the AV-NPF-PHD filter. In addition, AV-LPF-SMC-PHD filter also improves the estimation accuracy by 21% and 19% over the AV-ZPF-SMC-PHD and AV-PF-PF respectively.

Due to the space limitation, Table 2 only shows the average ESS for sequence 45 (camera 1). The running time of the AV-LPF-SMC-PHD filter is only 114.7s, the lowest among the compared methods. Compared to the AV-NPF-SMC-PHD filter, the running time is decreased 30% and the ESS is similar i.e. about 82. This is expected as the clustering step is removed and the covariance matrix is estimated without using the extra Kalman filter.

5. CONCLUSION

We have presented a novel AV-LPF-SMC-PHD filter for audio-visual multi-speaker tracking using label information. The proposed algorithm has been tested on the AV16.3 dataset and compared with other particle flow methods and PHD filters. The experimental results show that the proposed filter offers a higher tracking accuracy than the baseline method with a lower computational cost. The proposed filter also gives a similar ESS to that by the baseline AV-NPF-SMC-PHD filter.

6. REFERENCES

[1] A. Beskos, D. Crisan, A. Jasra, and N. Whiteley, "Error bounds and normalizing constants for sequential Monte Carlo in high dimensions," *arXiv preprint arXiv:1112.1544*, 2011.

[2] F. Daum and J. Huang, "Particle flow for nonlinear filters," *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, vol. 769704, pp. 5920–5923, Apr. 2011.

[3] ——, "Particle flow for nonlinear filters with log-homotopy," *Proc. SPIE Conf. Signal Processing Sensor Fusion, Target Recognition*, vol. 6969, pp. 696 918–1 – 696 918–12, 2008.

[4] ——, "Renormalization group flow and other ideas inspired by physics for nonlinear filters, Bayesian decisions, and transport," *Proc. SPIE Defense and Security*, pp. 90 910I–1 – 90 910I–14, 2014.

[5] ——, "Small curvature particle flow for nonlinear filters," *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, pp. 8393 – 8393–11, 2012.

[6] J. Heng, A. Doucet, and Y. Pokern, "Gibbs flow for approximate transport with applications to Bayesian computation," *arXiv preprint arXiv:1509.08787*, 2015.

[7] P. Bunch and S. Godsill, "Approximations of the optimal importance density using Gaussian particle flow importance sampling," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 748–762, 2016.

[8] Y. Liu, W. Wang, J. Chambers, V. Kilic, and A. Hilton, "Particle flow SMC-PHD filter for audio-visual multi-speaker tracking," *Proc. IEEE Intl Conf. Latent Variable Analysis and Signal Separation*, pp. 344–353, Mar. 2017.

[9] Y. Liu, A. Hilton, J. Chambers, Y. Zhao, and W. Wang, "Non-zero diffusion particle flow SMC-PHD filter for audio-visual multi-speaker tracking," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[10] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Trans. Signal Processing*, vol. 61, no. 13, pp. 3460–3475, 2013.

[11] A.-A. Saucan, Y. Li, and M. J. Coates, "Particle flow SMC delta-GLMB filter," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[12] V. Kilic, M. Barnard, W. Wang, A. Hilton, and J. Kittler, "Mean-shift and sparse sampling based SMC-PHD filtering for audio informed visual speaker tracking," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2417–2431, 2016.

[13] L. P. Kadanoff, *Statistical Physics: Statics, Dynamics and Renormalization*. World Scientific Publishing Co Inc, 2000.

[14] F. Daum and J. Huang, "Particle flow with non-zero diffusion for nonlinear filters," *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, vol. 04, pp. 87 450P–87 450P–13, 2013.

[15] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," *Proc. the Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[16] A. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and Bayesian missing data problems," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 278–288, 1994.

[17] Y. Li and M. Coates, "Particle filtering with invertible particle flow," *IEEE Trans. Signal Processing*, vol. 65, no. 15, pp. 4102–4116, 2016.

[18] ——, "Fast particle flow particle filters via clustering," *Proc. IEEE Int. Conf. Information Fusion (FUSION)*, pp. 2022–2027, 2016.

[19] A.-A. Saucan, Y. Li, and M. Coates, "Particle flow superpositional GLMB filter," *Proc. SPIE Conf. Signal Processing, Sensor Fusion, and Target Recognition*, vol. 10200, p. 102000F, 2017.

[20] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. Signal Processing*, vol. 59, no. 7, pp. 3452–3457, 2011.

[21] Y. Li, L. Zhao, and M. Coates, "Particle flow for particle filtering," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3979–3983, 2016.