NON-ZERO DIFFUSION PARTICLE FLOW SMC-PHD FILTER FOR AUDIO-VISUAL MULTI-SPEAKER TRACKING

Yang Liu[1]*, Adrian Hilton*[1]*, Jonathon Chambers* [2]*, Yuxin Zhao*[3]*, Wenwu Wang*[1]

[1] Department of Electrical and Electronic Engineering, University of Surrey, UK
[2] School of Electrical and Electronic Engineering, Newcastle University, UK
[3] College of Automation, Harbin Engineering University, China
E-mail: yangliu@surrey.ac.uk; a.hilton@surrey.ac.uk; jonathon.chambers@newcastle.ac.uk; zhaoyuxin@hrbeu.edu.cn; w.wang@surrey.ac.uk

ABSTRACT

The sequential Monte Carlo probability hypothesis density (SMC-PHD) filter has been shown to be promising for audio-visual multi-speaker tracking. Recently, the zero diffusion particle flow (ZPF) has been used to mitigate the weight degeneracy problem in the SMC-PHD filter. However, this leads to a substantial increase in the computational cost due to the migration of particles from prior to posterior distribution with a partial differential equation. This paper proposes an alternative method based on the non-zero diffusion particle flow (NPF) to adjust the particle states by fitting the particle distribution with the posterior probability density using the non-zero diffusion. This property allows efficient computation of the migration of particles. Results from the AV16.3 dataset demonstrate that we can significantly mitigate the weight degeneracy problem with a smaller computational cost as compared with the ZPF based SMC-PHD filter.

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 various fields such as spatial audio. For tracking an unknown and variable number of speakers, the sequential Monte Carlo (SMC) PHD filter is a popular method and the AV-SMC-PHD filter has been recently introduced in [1]. However, the posterior density is estimated by the weighted particles. As a result, the AV-SMC-PHD filter suffers from the weight degeneracy issue [2].

To mitigate this problem, particle flow filters [3, 4, 5, 6, 7, 8] have been introduced to AV-SMC-PHD filtering recently by migrating particles from the prior distribution to the posterior distribution using the zero diffusion particle flow (ZPF) [9, 10]. As a result, the posterior density can be more accurately approximated, which leads to significant reduction in the number of times for particle re-sampling. However, due to the use of a partial differential equation defined on a zero diffusion matrix, the ZPF-SMC-PHD algorithm has doubled the computational load as compared with the baseline SMC-PHD algorithm.

In this paper, we propose to incorporate non-zero diffusion particle flow (NPF) [11] to mitigate the weight degeneracy problem of the AV-SMC-PHD filter. More specifically, we use the non-zero particle flow to adjust the particle states and weights before the update step of the AV-SMC-PHD filter. We use the diffusion matrix to cancel out partial derivatives of the flow in order to reduce the computational complexity. Numerical experiments show that the proposed AV-NPF-SMC-PHD filter significantly increases the acceptance rate and effective sample size (ESS) [12] of the AV-SMC-PHD filter with a lower computational cost than the AV-ZPF-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, the AV-SMC-PHD filter, and the non-zero diffusion particle flow. We assume that the speaker dynamics and observations are described as:

$$\widetilde{\boldsymbol{m}}_k = \boldsymbol{f}_{\widetilde{\boldsymbol{m}}}\left(\widetilde{\boldsymbol{m}}_{k-1}, \boldsymbol{\tau}_k\right), \tag{1}$$

$$\boldsymbol{z}_k = \boldsymbol{f}_{\boldsymbol{z}}\left(\widetilde{\boldsymbol{m}}_k, \varsigma_k\right) \tag{2}$$

where $\widetilde{\boldsymbol{m}}_k \in \mathbb{R}^M$ is the speaker state vector in time k, defined as $\widetilde{\boldsymbol{m}}_k = [x_k, y_k, \dot{x}_k, \dot{y}_k]^T$ which consists of positions

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.

(x_k, y_k) and velocities $(\dot{x}_k, \dot{y}_k)$. So $M = 4$. $\widetilde{\ }$ is used to distinguish the speaker state from the particle state used later. Let $\boldsymbol{Z}_k$ denote the set of observations in time k, $\boldsymbol{Z}_k = \{\boldsymbol{z}_k^1, \boldsymbol{z}_k^2, ..., \boldsymbol{z}_k^{R_k}\}$ where R_k is the number of observations at time k. The observation $\boldsymbol{z}_k^r \in \boldsymbol{Z}_k$ is a noisy version of the position (x_k, y_k), where r is the index of the observation. $\boldsymbol{\tau}_k$ and $\boldsymbol{\varsigma}_k$ are system excitation and observation noise terms, respectively. $\boldsymbol{f}_{\widetilde{\boldsymbol{m}}}$ is the transition model and $\boldsymbol{f}_{\boldsymbol{z}}$ is the nonlinear observation model.

2.1. AV-SMC-PHD filter

In [1], an AV-SMC-PHD filter is presented for audio-visual multi-speaker tracking. The audio information and visual information is applied in the prediction and update steps, respectively. Audio information, i.e. the DOA line which shows the approximate direction of the sound emanating from the speakers, is applied for re-locating the existing particles. The particle weights are then calculated as:

$$\omega_{k|k-1}^i = \begin{cases} \frac{\phi_{k|k-1}\left(\boldsymbol{m}_{k|k-1}^i | \boldsymbol{m}_{k-1}^i\right)\omega_{k-1}^i}{q_k\left(\boldsymbol{m}_{k|k-1}^i | \boldsymbol{m}_{k-1}^i, \boldsymbol{Z}_k\right)}, i = 1, ..., N \\ \frac{\gamma_k(\boldsymbol{m}_{k|k-1}^i)}{N_B p_k(\boldsymbol{m}_{k|k-1}^i | \boldsymbol{Z}_k)}, i = N+1, ..., N+N_B \end{cases} \tag{3}$$

where m_{k-1}^i and ω_{k-1}^i are the state and weight of the i-th particle at time $k-1$, respectively. N is the number of the surviving particles. $\phi_{k|k-1}$ is the analogue of the state transition probability and q_k is the proposal distribution. 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. The visual information is used to calculate the likelihood in the update step. The likelihood function is assumed to be Gaussian and is based on the color histogram:

$$h_k^u(\boldsymbol{m}_{k|k-1}^i) = \frac{1}{\breve{\sigma}_k\sqrt{2\pi}} \exp\left\{-\frac{D^u(\boldsymbol{m}_{k|k-1}^i)}{2\breve{\sigma}_k^2}\right\} \tag{4}$$

where $\breve{\sigma}_k^2$ is the variance of noise for the visual likelihood h_k^u and $D^u(\boldsymbol{m}_{k|k-1}^i, \boldsymbol{s}_k^u)$ are the Bhattacharyya distance between the u-th reference color histogram $\boldsymbol{s}_k^u$ and the color histogram for the candidate speaker at the state $\boldsymbol{m}_{k|k-1}^i$. We also define $h_k^{u,i} = h_k^u(\boldsymbol{m}_{k|k-1}^i)$ to be used later. Then, the weights are updated and the effective sample size (ESS) [12] is calculated. When the ESS is smaller than half of the total number of particles, re-sampling is performed for mitigating the weight degeneracy problem. The pseudo-code of the AV-SMC-PHD filter is presented in Algorithm 1, where U_k is the number of the reference histogram. More details can be found in [1].

2.2. Particle flow

In the particle flow filter, the posterior density is calculated based on a homotopy function [13]. The homotopy function can be defined to model the particle flow process [14],

$$\log(\acute{\psi}_k^i) = \log(g_k^i) + \lambda \log(h_k^i) - \log K_k^i \tag{5}$$

Algorithm 1 AV-SMC-PHD Filter

Input: $\{\boldsymbol{m}_{k-1}^i, \omega_{k-1}^i\}_{i=1}^{N_{k-1}}$, N_B, $\boldsymbol{Z}_k$ and DOA lines.

Output: $\{\widetilde{\boldsymbol{m}}_k^j, \widetilde{\omega}_k^j\}_{j=1}^{\widetilde{N}_k}$, and $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$.

Initialize: $\boldsymbol{\tau}_k$, N_B, k, q_k, $\phi_{k|k-1}$, p_k, γ_k, $\breve{\sigma}_k$ and $\{\boldsymbol{s}_k^u\}_{u=1}^{U_k}$.

Run:

Propagate surviving particles $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_{k-1}}$.

if DOA lines exists **then**

Concentrate $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_{k-1}}$ around the DOA lines.

if new speaker **then**

Sample N_B born particles by $\boldsymbol{m}_{k|k-1}^i \sim p_k(\cdot|\boldsymbol{Z}_k)$.

end if

end if

Calculate the particle weight $\omega_{k|k-1}^i$ by Eq. (3).

Combine all the particles:$\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_k} = \{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_{k-1}} \cup \{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=N_{k-1}+1}^{N_{k-1}+N_B}$.

(Optional) Update $\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_k}$ by the particle flow.

Update the weights of the particles $\{\omega_{k|k-1}^i\}_{i=1}^{N_k}$ to obtain $\{\omega_k^i\}_{i=1}^{N_k}$ and calculate $\widetilde{N}_k = \sum_{i=1}^{N_k} \omega_k^i$.

Set $\{\boldsymbol{m}_k^i\}_{i=1}^{N_k}$ as $\{\boldsymbol{m}_{k|k-1}^i\}_{i=1}^{N_k}$.

Cluster $\{\widetilde{\boldsymbol{m}}_k^j, \widetilde{\omega}_k^j\}_{j=1}^{\widetilde{N}_k}$ by k-means.

if ESS $< N_k/2$ **then**

(Optional) Re-sample $\{\boldsymbol{m}_k^i, \omega_k^i\}_{i=1}^{N_k}$.

end if

where K_k^i is the normalization term and g_k^i is the prior density. λ is the step size parameter taking values from $[0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ and $N_\lambda\triangle\lambda = 1$. Eq. (5) represents the motion of the particles from the prior to the posterior density [15]. With λ varied to 1, $\acute{\psi}_k^i(.)$ is translated into the normalized posterior density ψ_k^i [4]. The particle flow obeys the Ito stochastic differential equation [14]:

$$\triangle\boldsymbol{m}_{k|k-1}^i = \boldsymbol{f}_k^i(\boldsymbol{m}_{k|k-1}^i, \lambda)\triangle\lambda + \upsilon_k^i \boldsymbol{w}_k^i \tag{6}$$

where $\boldsymbol{f}_k^i \in \mathbb{R}^M$ is the particle flow vector which moves the particle $\boldsymbol{m}_{k|k-1}^i$ with the distance $\triangle\boldsymbol{m}_{k|k-1}^i$ for the time period $\triangle\lambda$. $\boldsymbol{w}_k^i \in \mathbb{R}^M$ is the Wiener process with the diffusion coefficient υ_k^i. Based on the Fokker-Planck equation [16], $\boldsymbol{f}_k^i$ is calculated by the partial differential equation:

$$\begin{aligned} \boldsymbol{\nabla}\log(h_k^i) = &-\boldsymbol{\nabla}^2\log(\acute{\psi}_k^i)\boldsymbol{f}_k^i - \boldsymbol{\nabla}\boldsymbol{f}_k^i\boldsymbol{\nabla}\log(\acute{\psi}_k^i) \\ &- \boldsymbol{\nabla}(\mathrm{div}(\boldsymbol{f}_k^i)) + \frac{1}{2\acute{\psi}_k^i}\boldsymbol{\nabla}(\mathrm{div}(\boldsymbol{Q}_k^i\boldsymbol{\nabla}\acute{\psi}_k^i)) \end{aligned} \tag{7}$$

where $\boldsymbol{\nabla}$ is the spatial vector differentiation operator $\frac{\partial}{\partial\boldsymbol{m}_{k|k-1}^i}$. div(.) is the divergence operator [17], e.g. $\mathrm{div}(\boldsymbol{f}_k^i) = \sum_{l=1}^M \frac{\partial\boldsymbol{f}_k^i}{\partial\boldsymbol{\varepsilon}^l}$ where $\boldsymbol{\varepsilon}^l \in \mathbb{R}^M$ is the l-th basis vector. $\boldsymbol{Q}_k^i$ is the diffusion matrix. In ZPF, $\boldsymbol{Q}_k^i = 0$. Due to the involvement of the differentiation $\boldsymbol{\nabla}\boldsymbol{f}_k^i$ and the divergence

$\text{div}(\boldsymbol{f}_k^i)$, the complexity for calculating $\boldsymbol{f}_k^i$ is high. In NPF [11], for simplifying the solution of Eq. (7), $\boldsymbol{Q}_k^i$ is set as $\boldsymbol{\nabla}\log(\acute{\psi}_k^i)\boldsymbol{\nabla}\boldsymbol{f}_k^i + \boldsymbol{\nabla}(\text{div}(\boldsymbol{f}_k^i))$ and Eq. (7) becomes:

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

where

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

where $\boldsymbol{P}_{k|k-1}^i$ is the covariance matrix of $\boldsymbol{m}_{k|k-1}^i$. Eqs. (7-8) show that the computational complexity is reduced by the diffusion term $\boldsymbol{Q}_k^i$. The derivation of Eq. (8) can be found in [11].

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

This section presents an improved version of the AV particle flow SMC-PHD filter based on non-zero diffusion. As shown in Eq. (8), NPF can be calculated according to a likelihood function. First, the audio-visual likelihood vector $\boldsymbol{h}_k$ is obtained by audio likelihood $\mathring{\boldsymbol{h}}_k^i \in \mathbb{R}^{O_k}$ and visual likelihood $\breve{\boldsymbol{h}}_k^i \in \mathbb{R}^{N_k}$. We assume the audio likelihood is $\mathring{\boldsymbol{h}}_k^i = \{\mathring{h}_k^{o,i}\}_{o=1}^{O_k}$ where O_k is the number of DOA lines at time k as in [10]. $\mathring{h}_k^{o,i}$ is calculated by the distance $\boldsymbol{d}_k^{o,i}$ from the i-th particle to the o-th DOA line.

$$\mathring{h}_k^{o,i} = \frac{1}{\mathring{\sigma}_k\sqrt{2\pi}} \exp\left\{-\frac{\left\| \boldsymbol{d}_k^{o,i}\right\|_2^2}{2\mathring{\sigma}_k^2}\right\} \tag{10}$$

where $\mathring{\sigma}_k^2$ is the variance of the audio likelihood. The visual likelihood $\breve{h}_k^{u,i} \in \breve{\boldsymbol{h}}_k^i$ is calculated based on the color histogram for the image patch taken at the particle $\boldsymbol{m}_{k|k-1}^i$ and the u-th reference histogram in Eq. (4). Then the audio-visual likelihood function h_k^i is obtained as:

$$h_k^i = \frac{\mathring{\boldsymbol{h}}_k^{i\,T}\mathring{\boldsymbol{\omega}}_k + \breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k}{\|\mathring{\boldsymbol{\omega}}_k\|_1 + \|\breve{\boldsymbol{\omega}}_k\|_1} \tag{11}$$

where $\mathring{\boldsymbol{\omega}}_k \in \mathbb{R}^{O_k}$ and $\breve{\boldsymbol{\omega}}_k \in \mathbb{R}^{U_k}$ are the weight sets for the audio and visual likelihood, respectively. $\|\cdot\|_1$ is the l_1 norm. The first and second derivative of the likelihood function are then calculated respectively as:

$$\boldsymbol{\nabla}\log h_k^i = \frac{\boldsymbol{\nabla}(\mathring{\boldsymbol{h}}_k^{i\,T}\mathring{\boldsymbol{\omega}}_k) + \boldsymbol{\nabla}(\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k)}{(\|\mathring{\boldsymbol{\omega}}_k\|_1 + \|\breve{\boldsymbol{\omega}}_k\|_1)h_k^i} \tag{12}$$

$$\begin{aligned}\boldsymbol{\nabla}^2 \log h_k^i =& \frac{\boldsymbol{\nabla}^2(\mathring{\boldsymbol{h}}_k^{i\,T}\mathring{\boldsymbol{\omega}}_k) + \boldsymbol{\nabla}^2(\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k)}{(\|\mathring{\boldsymbol{\omega}}_k\|_1 + \|\breve{\boldsymbol{\omega}}_k\|_1)h_k^i} \\ &- (\boldsymbol{\nabla}\log h_k^i)(\boldsymbol{\nabla}\log h_k^i)^T\end{aligned} \tag{13}$$

where

$$\boldsymbol{\nabla}(\mathring{\boldsymbol{h}}_k^{i\,T}\mathring{\boldsymbol{\omega}}_k) = -\sum_{o=1}^{O_k}\frac{1}{\mathring{\sigma}_k^2}\boldsymbol{d}_k^{o,i}\mathring{h}_k^{o,i}\mathring{\omega}_k^o \tag{14}$$

Algorithm 2 Non-zero exact particle flow of the AV-ZPF-SMC-PHD Filter

Input: $\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_k}$, $\{\boldsymbol{m}_{k-1}^i, \omega_{k-1}^i\}_{i=1}^{N_{k-1}}$ and $\{\boldsymbol{s}_k^u\}_{u=1}^{U_k}$.
Output: $\{\boldsymbol{m}_{k|k-1}^i, \omega_{k|k-1}^i\}_{i=1}^{N_k}$
  Initialize: $\mathring{\sigma}_k$, $\breve{\sigma}_k$, $\boldsymbol{I}^4$, $\boldsymbol{P}_{k|k-1}^i$, $\mathring{\boldsymbol{\omega}}_k$, υ_k^i, $\{\varepsilon^l\}_{l=1}^M$ and $\breve{\boldsymbol{\omega}}_k$.
  Run:
  for $i \in [1, ..., N_k]$ **do**
    Calculate the visual likelihood $\breve{\boldsymbol{h}}_k^i$ by Eq. (4).
    Calculate the audio likelihood $\mathring{\boldsymbol{h}}_k^i$ by Eq. (10).
    Calculate the audio-visual likelihood h_k^i by Eq. (11).
    Calculate $\boldsymbol{\nabla}\log h_k^i$ and $\boldsymbol{\nabla}^2\log h_k^i$ by Eqs. (12)-(17).
    for $\lambda \in [0, \triangle\lambda, 2\triangle\lambda, \cdots, N_\lambda\triangle\lambda]$ **do**
      Evaluate flow $\boldsymbol{f}_k^i$ by Eq. (8).
      Update $\triangle\boldsymbol{m}_{k|k-1}^i$ by Eq. (6) and $\boldsymbol{m}_{k|k-1}^i = \boldsymbol{m}_{k|k-1}^i + \triangle\boldsymbol{m}_{k|k-1}^i\lambda$.
    end for
    Re-calculat the particle weights by Eq. (3).
  end for

$$\boldsymbol{\nabla}^2(\mathring{\boldsymbol{h}}_k^{i\,T}\mathring{\boldsymbol{\omega}}_k) = -\frac{1}{\mathring{\sigma}_k^2}\sum_{o=1}^{O_k}(\boldsymbol{\nabla}(\mathring{\boldsymbol{h}}_k^{i\,T}\mathring{\boldsymbol{\omega}}_k)\boldsymbol{d}_k^{o,i^T} + \mathring{h}_k^{o,i}\mathring{\omega}_k^o\boldsymbol{I}^4) \tag{15}$$

where $\boldsymbol{I}^4$ is a 4×4 identity matrix. $\boldsymbol{d}_k^{o,i}$ is the distance from the o-th DOA line to the i-th particle. By the finite-difference method [18], the first derivative of $\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k$ as an M-dimensional vector is calculated. The l-th element of $\boldsymbol{\nabla}(\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k)$ is defined as:

$$(\boldsymbol{\nabla}(\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k))^l = \frac{1}{||\boldsymbol{\varepsilon}^l||_1}\sum_{u=1}^{U_k}(\breve{h}_k^u(\boldsymbol{m}_{k|k-1}^i + \boldsymbol{\varepsilon}^l) - \breve{h}_k^{u,i}) \tag{16}$$

where $\boldsymbol{\varepsilon}^l$ is used as the unit step for approximating the derivatives. The second derivative of $\breve{\boldsymbol{h}}_k^i\breve{\boldsymbol{\omega}}_k$ has a similar form. As $\boldsymbol{\nabla}^2(\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k) \in \mathbb{R}^{M*M}$, the l-th row vector is calculated as:

$$\begin{aligned}(\boldsymbol{\nabla}^2(\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k))^l =& \frac{1}{||\boldsymbol{\varepsilon}^l||_1}(\boldsymbol{\nabla}((\breve{\boldsymbol{h}}_k(\boldsymbol{m}_{k|k-1}^i + \boldsymbol{\varepsilon}^l))^T\breve{\boldsymbol{\omega}}_k) \\ &- \boldsymbol{\nabla}(\breve{\boldsymbol{h}}_k^{i\,T}\breve{\boldsymbol{\omega}}_k))\end{aligned} \tag{17}$$

The pseudo-code of the particle flow of the AV-NPF-SMC-PHD filter is presented in Algorithm 2 and is shown as the optional step before the update step in Algorithm 1.

4. EXPERIMENTAL RESULTS

In this section, the proposed algorithm is compared to the baselines, which include AV-PF-PF, AV-GPF-PHD, AV-ZPF-SMC-PHD [10] and AV-SMC-PHD algorithms [1] using the AV16.3 dataset. Particle flow has been used for improving the tracking accuracy for the particle filter (PF-PF) [19, 20, 21] and Gaussian mixture PHD (GPF-PHD) filter [22]. In this

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

Seq (Cam)	NPF	ZPF	PPF	GPF	SMC
24 (1)	12.32	12.99	12.18	13.92	17.71
24 (2)	13.20	13.82	13.12	14.58	19.83
24 (3)	13.23	14.01	13.02	15.09	18.94
25 (1)	15.96	16.80	14.90	18.20	19.13
25 (2)	15.29	15.88	13.08	15.14	18.47
25 (3)	16.29	17.56	14.98	17.95	21.61
30 (1)	15.76	17.15	15.29	18.50	25.22
30 (2)	13.41	14.22	13.86	15.19	19.37
30 (3)	15.93	17.63	15.61	18.51	25.31
45 (1)	17.65	19.33	24.50	23.12	29.46
45 (2)	18.60	20.85	22.26	21.71	29.47
45 (3)	19.50	21.35	24.34	21.96	28.43
Avg. OSPA	**15.60**	**16.80**	**16.43**	**17.82**	**22.75**

paper, they are used as AV-PF-PF and AV-GPF-PHD filter for the AV data. The observations are the same as in the proposed AV-NPF-SMC-PHD filter. The experiments are run in Matlab on Windows 7 with Intel i7.

The AV16.3 dataset consists of sequences where multiple speakers keep speaking and walking. Those actions are recorded by three calibrated video cameras at 25 Hz and two circular eight-element microphone arrays at 16 kHz. Each image frame has 288x360 pixels. Before running the tests, the audio and video streams are synchronized. The Optimal Subpattern Assignment (OSPA) for trackers [23], 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 [12] is used to show the accuracy of the resulting approximation of the posterior density [19, 8]. The parameters are set as: $\boldsymbol{\tau}_k = [0, 0, 0, 0]^T$, $N_B = 50$, $q_k = 1$, $p_k = 1$, $\mathring{\sigma}_k = 2$, $\breve{\sigma}_k = 2$, $\upsilon_k^i = 1$, $\varepsilon = 5$ and $\dot{\varepsilon} = 5$. Each weight of $\mathring{\boldsymbol{\omega}}_k$ and $\breve{\boldsymbol{\omega}}_k$ is set as 0.5. Other parameters of the PHD filter and particle flow filters are set as in [1] and [10]. The OSPA metric order parameter 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. With the contribution of the non-zero particle flow, 31% reduction in tracking error has been achieved as compared with AV-SMC-PHD filter. In addition, AV-NPF-SMC-PHD filter also improves the estimation accuracy by 7%, 5%, 12% over the AV-ZPF-SMC-PHD, AV-PF-PF and AV-GPF-PHD filters, respectively.

Due to the space limitation, Table 2 only shows average ESS for the sequence 45 (camera 1). Compared to the AV-SMC-PHD filter, ESS is increased 103.7% by the NPF. Apart

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

Filter	ESS	Resampling	Time(s)
NPF	82.1	36	163.8
ZPF	77.8	58	268.0
PPF	70.5	68	215.9
GPF	63.6	90	386.3
SMC	40.3	354	124.3

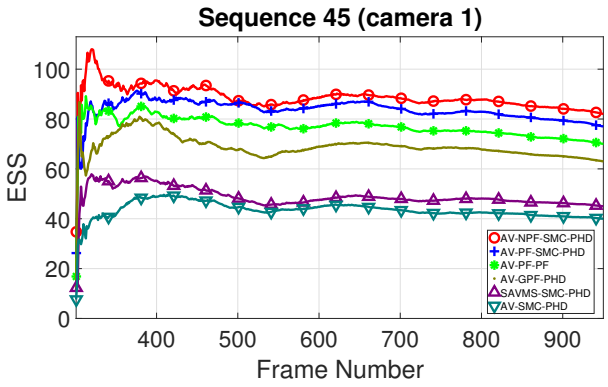

Fig. 1. The ESS of the AV-NPF-SMC-PHD, AV-ZPF-SMC-PHD, AV-PF-PF, AV-GPF-PHD and AV-SMC-PHD filters in the frames 300-960 for the sequence 45 with camera 1.

from that, the re-sampling times are decreased for 89.8%, which means the weight degeneracy problem has been significantly mitigated. Except for the baseline method, AV-NPF-SMC-PHD filter has the lowest computational cost. The cost of the AV-NPF-SMC-PHD filter is only 61% of the AV-ZPF-SMC-PHD filter. The average ESS among all tested algorithms is also illustrated in Fig. 1. The frames 300-960 are shown, since in the remaining frames only one (or no speaker) appears in the scene. The AV-SMC-PHD filter has the lowest ESS (about 40.3), which is lower than $N_k/2$ (75) and shows a strong degeneracy problem. The AV-NPF-SMC-PHD filter maintains a particle cloud with the highest ESS (about 82), which is higher than 75 with a lower computational cost.

5. CONCLUSION

We have presented a novel AV-NPF-SMC-PHD filter for audio-visual multi-speaker tracking by smoothly migrating the particles. The proposed algorithm has been tested on the AV16.3 dataset. The experimental results show that the proposed filter offers a higher tracking accuracy and ESS than the baseline method and a lower computational cost than the zero diffusion particle flow PHD filter.

6. REFERENCES

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

[2] 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.

[3] F. Daum and J. Huang, "Particle flow for nonlinear filters," *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 5920–5923, 2011.

[4] ——, "Particle flow for nonlinear filters with log-homotopy," in *SPIE Defense and Security Symposium*. International Society for Optics and Photonics, 2008, pp. 696 918–6 969 180.

[5] ——, "Renormalization group flow and other ideas inspired by physics for nonlinear filters, Bayesian decisions, and transport," in *SPIE Defense and Security Symposium*. International Society for Optics and Photonics, 2014, pp. 90 910I–90 910I–14.

[6] J. H. Fred Daum, "Small curvature particle flow for nonlinear filters," *Proc. SPIE Conference*, pp. 8393 – 8393–11, 2012.

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

[8] 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.

[9] Y. Liu, W. Wang, and Y. Zhao, "Particle flow for sequential monte carlo implementation of probability hypothesis density," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[10] Y. Liu, W. Wang, J. Chambers, V. Kılıç, and A. Hilton, "Particle flow SMC-PHD filter for audio-visual multi-speaker tracking," in *Proc. IEEE Intl. Conf. on Latent Variable Analysis and Signal Separation*, 2017.

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

[12] 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.

[13] F. Daum and J. Huang, "A plethora of open problems in particle flow research for nonlinear filters, Bayesian decisions, Bayesian learning, and transport," in *Proc. SPIE Symposium on Signal Processing, Sensor/Information Fusion, and Target Recognition*, 2016, pp. 98 420I–98 420I.

[14] F. Daum, J. Huang, and A. Noushin, "Exact particle flow for nonlinear filters," in *SPIE Defense, Security, and Sensing*. International society for optics and photonics, 2010, pp. 769 704–1–769 704–19.

[15] A. Doucet, N. de Freitas, N. Gordon, and A. Smith, *Sequential Monte Carlo Methods in Practice*. Springer Science & Business Media, 2001.

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

[17] C. H. Edwards, *Advanced Calculus of Several Variables*. Courier Corporation, 2012.

[18] T. H. Meyer, M. Eriksson, and R. C. Maggio, "Gradient estimation from irregularly spaced data sets," *Mathematical Geology*, vol. 33, no. 6, pp. 693–717, 2001.

[19] Y. Li, L. Zhao, and M. Coates, "Particle flow for particle filtering," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2016, pp. 3979–3983.

[20] Y. Li and M. Coates, "Fast particle flow particle filters via clustering," in *Proc. IEEE Intl. Conf. on Information Fusion*, 2016, pp. 2022–2027.

[21] ——, "Particle filtering with invertible particle flow," *IEEE Transactions on Signal Processing*, vol. 65, no. 15, pp. 4102–4116, 2016.

[22] L. Zhao, J. Wang, Y. Li, and M. J. Coates, "Gaussian particle flow implementation of PHD filter," in *SPIE Defense, Security, and Sensing*, 2016, pp. 98 420D – 98 420D–10.

[23] 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.