AUDIO INFORMED VISUAL SPEAKER TRACKING WITH SMC-PHD FILTER

Volkan Kılıç, Mark Barnard, Wenwu Wang, Adrian Hilton, and Josef Kittler

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
Emails:{v.kilic, mark.barnard, w.wang, a.hilton, j.kittler}@surrey.ac.uk

ABSTRACT

Sequential Monte Carlo probability hypothesis density (SMC-PHD) filter has received much interest in the field of nonlinear non-Gaussian visual tracking due to its ability to handle a variable number of speakers. The SMC-PHD filter employs surviving, spawned and born particles to model the state of the speakers and jointly estimates the variable number of speakers with their states. The born particles play a critical role in the detection of new speakers, which makes it necessary to propagate them in each frame. However, this increases the computational cost of the visual tracker. Here, we propose to use audio data to determine when to propagate the born particles and re-allocate the surviving and spawned particles. In our framework, we employ audio data as an aid to visual SMC-PHD (V-SMC-PHD) filter by using the direction of arrival (DOA) angles of the audio sources to reshape the distribution of the particles. Experimental results on the *AV*16.3 dataset with multi-speaker sequences show that our proposed audio-visual SMC-PHD (AV-SMC-PHD) filter improves the tracking performance in terms of estimation accuracy and computational efficiency.

***Index Terms*—** Audio-visual tracking, PHD filter, SMC implementation, multi-speaker tracking

1. INTRODUCTION

The problem of detection and tracking of multiple moving speakers in indoor environments using audio-visual (AV) modalities has attracted an increasing amount of attention in the last decade due to its potential applications such as automatic camera steering in video conferencing and individual speaker discrimination in multi-speaker environments. Several challenges are associated with AV tracking including estimation of the variable number of speakers and their states in various conditions like occlusion, limited view of cameras, illumination change and room reverberations.

A number of methods have been proposed to address these challenges. The PHD filtering [1] approach as a first order approximation of the random finite set (RFS) is a framework that has recently emerged and has been found to be promising for multi-speaker tracking. The sequential Monte Carlo (SMC) implementation [2] is introduced to obtain practical solutions of the PHD filter. We investigate and modify the standard SMC-PHD filtering algorithm aiming at improving its computational efficiency and estimation accuracy under challenging conditions as mentioned above. The SMC-PHD filter uses particles to model the surviving, spawned and born state of the speaker. The standard implementation of SMC-PHD filter [2] in visual tracking propagates the born particles every frame to detect the speaker present in the view, which is computationally expensive. Here, we propose to use the DOA information obtained from audio for the propagation of the particles. To reduce the computational complexity and improve the estimation accuracy, the propagation of born particles is decided based on DOA information and particles are re-located around the line drawn upon the DOA as illustrated in Figure 1. A similar approach has been used in [3], [4] and [5] under particle filter framework for a fixed number of speakers. Here, the SMC-PHD filter is used, and to the best of our knowledge, audio information has not been previously fused with visual information in a SMC-PHD filter as we do here.

(a) (b)

Fig. 1. Distribution of 50 particles for visual case in (a) and audio-visual case in (b).

To use the PHD filter, a set of random measurements is required to estimate both the position and the number of the targets. Random measurements can be obtained from sensors like GPS in application of SLAM (Simultaneous localization and mapping) [6] or microphone pairs in audio based speaker tracking [7]. In visual tracking, however, there are no sensors to generate random measurements except cameras. Therefore, pre-processing of visual data is needed to find measurements. There are two common ways. The first one is to use a detector as in [8] and [9] where background/foreground (B/F) detection algorithm is run on the frame and the centers of the foreground objects are used as the measurements set. Despite being computationally expensive, this method performs well in tracking of moving objects. However, if it is desired

to track particular parts like face, it does not work well. In that case, face detection algorithms are needed which have a higher computational cost than the B/F detection algorithms. The second method is to use color histogram templates [10] where the template is created for a single target. In this study, we extend the template for multi-target scenario. Color histograms of possible targets are stored in the template and during the tracking it is used as the measurements set. This method is computationally cheaper than the first method as the template is created only once at the beginning.

The rest of this paper is organized as follows: the next section introduces our proposed audio-visual SMC-PHD (AV-SMC-PHD) filtering algorithm. Section 3 shows experimental results performed on the *AV*16.3 dataset and compares the performance of the algorithms. Closing remarks are given in Section 4.

2. AUDIO-VISUAL TRACKING WITH RFS

This section describes our problem formulation for multi-speaker tracking based on the RFS framework, provides a brief review on using color information in visual tracking, and then introduces the proposed tracking algorithm.

2.1. RFS State Model Formulation for Multi-Speaker Tracking

In a single speaker tracking system, the state of a speaker is defined as $\mathbf{x} = \begin{bmatrix} x & \dot{x} & y & \dot{y} & s \end{bmatrix}^T$, where x and y are, respectively, the horizontal and vertical positions of the rectangle centred around the face that we wish to track, $\dot{x}$ and $\dot{y}$ are, respectively, the horizontal and vertical velocity, and s is the scale of the rectangle centred around (x, y). The constant velocity model is employed for the evolution of time dependent state [5] ;

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

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

Here, the state of the speaker is defined as a vector with five dimensions. Measurements can also be defined as a vector with a different dimension depending on the application. The state and measurement of a single speaker system evolve in time with their dimensions fixed which is not the case for multi-speaker tracking since the number of speakers and measurements may change. Therefore, the dimensions of the multi-speaker state and measurements also evolve in time.

Multi-speaker state and measurement are represented by finite collections $\mathcal{X}_k$ and $\mathcal{Z}_k$,

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

$$\mathcal{Z}_k = \{\mathbf{z}_{1,k}, ..., \mathbf{z}_{M_k,k}\} \tag{3}$$

where $\mathcal{Z}_k$ consists of M_k observations which may be corrupted by noise due to clutter, and N_k is the number of speakers in the view.

In a single speaker Bayesian tracking, uncertainty is introduced by modelling $\mathbf{x}_k$ and $\mathbf{z}_k$ as random vectors. Similarly, uncertainty in multi-speaker tracking is introduced by modelling the $\mathcal{X}_k$ and $\mathcal{Z}_k$ as RFSs

$$\mathcal{X}_k = \mathcal{S}_k\left(\mathcal{X}_{k-1}\right) \cup \mathcal{B}_k(\mathcal{X}_{k-1}) \cup \Gamma_k \tag{4}$$

$$\mathcal{Z}_k = \Theta_k(\mathcal{X}_k) \cup \mathcal{C}_k \tag{5}$$

where $\mathcal{S}_k(\mathcal{X}_{k-1})$ denotes the RFS of surviving speakers, $\mathcal{B}_k(\mathcal{X}_{k-1})$ is the RFS of speakers spawned from the previous set of speakers $\mathcal{X}_{k-1}$ and Γ_k is the RFS of new speakers that appear spontaneously at time k [2]. $\Theta_k(\mathcal{X}_k)$ denotes the RFS of the measurements generated by the speakers $\mathcal{X}_k$ and $\mathcal{C}_k$ is the RFS of clutter or false alarms. Besides, the dynamics in the state evolution $\mathcal{X}_k$ are described by the multi-speaker transition density $f_{k|k-1}(\mathcal{X}_k|\mathcal{X}_{k-1})$, while the randomness in the observations are described by the multi-speaker likelihood $g_k(\mathcal{Z}_k|\mathcal{X}_k)$. Then, the RFS formulation can be employed in the optimal multi-speaker Bayes filter by propagating the multi-speaker posterior density. However, it is computationally intractable since multiple integrals are involved in the recursion of the multi-speaker posterior. To reduce the computational complexity, the PHD filter is proposed which propagates the first-order moment of the multi-speaker posterior instead of the posterior itself [1].

The PHD filter is defined as the intensity $v_{k|k}$ whose integral on any region of the state space gives the expected number of speakers. The local maxima of the PHD function indicate the highest local concentration of the expected number of speakers, which also identify the likely positions of the speakers. The PHD filter has two iterative steps: prediction and update. The prediction step of the PHD is defined as

$$\begin{aligned} v_{k|k-1}\left(\mathbf{x}_k\right) &= \gamma_k\left(\mathbf{x}_k\right) \\ &\quad + \int \phi_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) v_{k-1|k-1}\left(\mathbf{x}_{k-1}\right) d\mathbf{x}_{k-1} \end{aligned} \tag{6}$$

where $\gamma_k\left(\mathbf{x}_k\right)$ denotes the intensity function of the new speaker birth RFS Γ_k, and $\phi_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ is the analog of the single-speaker state transition probability

$$\begin{aligned} \phi_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) &= p_{S,k}\left(\mathbf{x}_{k-1}\right) f_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) \\ &\quad + \beta_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) \end{aligned} \tag{7}$$

where $p_{S,k}\left(\mathbf{x}_{k-1}\right)$ is the survival probability for the speakers still existing and $f_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ is the single-speaker state transition density. The intensity function of RFS $\mathcal{B}_k(\mathcal{X}_{k-1})$ is denoted by $\beta_{k|k-1}\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ for the speaker spawned at time k with previous state $\mathbf{x}_{k-1}$. The PHD update is defined as

$$\begin{aligned} v_{k|k}\left(\mathbf{x}_k\right) &= \left(1 - p_{D,k}\right)\left(\mathbf{x}_k\right) v_{k|k-1}\left(\mathbf{x}_k\right) \\ &\quad + \sum_{\mathbf{z}_k \in \mathcal{Z}_k} \frac{p_{D,k}\left(\mathbf{x}_k\right) g_k\left(\mathbf{z}_k|\mathbf{x}_k\right) v_{k|k-1}\left(\mathbf{x}_k\right)}{\kappa_k\left(\mathbf{z}_k\right) + \int p_{D,k}\left(\mathbf{x}_k\right) g_k\left(\mathbf{z}_k|\mathbf{x}_k\right) v_{k|k-1}\left(\mathbf{x}_k\right)} \end{aligned} \tag{8}$$

where $p_{D,k}(\mathbf{x}_k)$ is detection probability and $g_k\left(\mathbf{z}_k|\mathbf{x}_k\right)$ is the single-speaker likelihood defining the probability that $\mathbf{z}_k$ is

generated by a speaker state $\mathbf{x}_k$. The intensity of clutter RFS $\mathcal{C}_k$ is defined as $\kappa_k(\mathbf{z}_k)$ which is $\kappa_k(\mathbf{z}_k) = \lambda u(\mathbf{z}_k)$, where λ is the average number of Poisson clutter points per scan and $u(\mathbf{z}_k)$ is the probability distribution of each clutter point.

With the multiple integrals in the PHD prediction (6) and update (8) steps, there are no closed-form solutions in general. To obtain a numerical solution for the integrals in PHD recursion, SMC method has been proposed which approximates the PHD with a set of random samples (particles) [2].

Suppose that at time step $k-1$, the PHD $v_{k-1|k-1}(\mathbf{x}_{k-1})$ is approximated by $\left\{w_{k-1}^{(i)}, \mathbf{x}_{k-1}^{(i)}\right\}_{i=1}^{L_{k-1}}$ of L_{k-1} particles and their corresponding weight as

$$v_{k-1|k-1}(\mathbf{x}_{k-1}) \approx \sum_{i=1}^{L_{k-1}} w_{k-1}^{(i)} \delta\left(\mathbf{x}_{k-1} - \mathbf{x}_{k-1}^{(i)}\right) \tag{9}$$

Prediction of the PHD $v_{k|k-1}(\mathbf{x}_k)$ is obtained with weighted particles $\left\{\tilde{w}_{k|k-1}^{(i)}, \tilde{\mathbf{x}}_k^{(i)}\right\}_{i=1}^{L_{k-1}+J_k}$ (The quantities with tilde are discussed later.)

$$v_{k|k-1}(\mathbf{x}_k) \approx \sum_{i=1}^{L_{k-1}+J_k} \tilde{w}_{k|k-1}^{(i)} \delta\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_k^{(i)}\right) \tag{10}$$

where J_k new particles arise from birth process. By substituting (9) into (6) and then applying importance sampling, we get predicted weights $\tilde{w}_{k|k-1}^{(i)}$;

$$\tilde{w}_{k|k-1}^{(i)} = \begin{cases} \frac{\phi_{k|k-1}\left(\tilde{\mathbf{x}}_k^{(i)}, \mathbf{x}_{k-1}^{(i)}\right) w_{k-1}^{(i)}}{q_k\left(\tilde{\mathbf{x}}_k^{(i)} | \mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_k\right)}, & i = 1, ..., L_{k-1} \\ \frac{\gamma_k\left(\tilde{\mathbf{x}}_k^{(i)}\right)}{J_k p_k\left(\tilde{\mathbf{x}}_k^{(i)} | \mathcal{Z}_k\right)}, & i = L_{k-1}+1, ..., L_{k-1}+J_k \end{cases} \tag{11}$$

Practically, L_{k-1} particles are first drawn from importance sampling $q_k\left(\tilde{\mathbf{x}}_k^{(i)} | \mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_k\right)$ to propagate the particles from time step $k-1$, then J_k particles from the new born importance function $p_k\left(\tilde{\mathbf{x}}_k^{(i)} | \mathcal{Z}_k\right)$ are drawn to model the state of new speakers appearing in the scene.

The update step of the PHD recursion is obtained by updating the weight of the predicted particles when the likelihood $g_k\left(\mathbf{z}_k | \tilde{\mathbf{x}}_k^{(i)}\right)$ is available. Then $v_{k|k-1}(\mathbf{x}_k)$ is substituted into (8) and the predicted weights $\left\{\tilde{w}_{k|k-1}^{(i)}\right\}_{i=1}^{L_{k-1}+J_k}$ are updated according to

$$\tilde{w}_k^{(i)} = \left[p_M\left(\tilde{\mathbf{x}}_k^{(i)}\right) + \sum_{\mathbf{z}_k \in \mathcal{Z}_k} \frac{p_D\left(\tilde{\mathbf{x}}_k^{(i)}\right) g_k\left(\mathbf{z}_k | \tilde{\mathbf{x}}_k^{(i)}\right)}{\kappa_k(\mathbf{z}_k) + C_k(\mathbf{z}_k)}\right] \tilde{w}_{k|k-1}^{(i)} \tag{12}$$

where

$$C_k(\mathbf{z}_k) = \sum_{j=1}^{L_{k-1}+J_k} p_D\left(\tilde{\mathbf{x}}_k^{(j)}\right) g_k\left(\mathbf{z}_k | \tilde{\mathbf{x}}_k^{(i)}\right) w_{k|k-1}^{(j)} \tag{13}$$

Note that J_k new particles, sampled for the born speakers at each iteration, are added to the old ones $L_k = L_{k-1} + J_k$ which causes the number of particles to grow over time and makes the PHD filter inefficient. In addition, the low weight particles need to be removed and the particles with high weights should be duplicated in order to concentrate the particles on the zones around the targets. To this extent, a resampling step is performed after the update step. L_k particles are resampled from $\left\{\tilde{w}_k^{(i)} / \hat{N}_{k|k}, \tilde{\mathbf{x}}_k^{(i)}\right\}_{i=1}^{L_{k-1}+J_k}$ where $\hat{N}_{k|k}$ is the total mass and $\hat{N}_{k|k} = \sum_{i=1}^{L_{k-1}+J_k} \tilde{w}_k^{(i)}$. L_k is estimated by $L_k = \rho \hat{N}_{k|k}$ where ρ is the constant number of particles per speaker. Therefore the complexity of the SMC-PHD filter increases *linearly* with the number of speakers. After the resampling step, the new weights of set $\left\{w_k^{(i)}, \mathbf{x}_k^{(i)}\right\}_{i=1}^{L_k}$ are normalized to preserve the total mass.

2.2. Color Likelihood Model

The color information of the state $\mathbf{x}_k$ is represented using a color histogram. Let the speaker candidate be defined with the rectangle centred around the location (x_k, y_k) on the frame. This rectangle is converted to color histogram $\hat{q}(\mathbf{x}_k)$ in order to calculate its similarity with the reference speaker models. In multi-speaker tracking, we have many color models of speakers $\left\{r_1^{(u)}, r_2^{(u)}, ..., r_M^{(u)}\right\}$ where u is the index of histogram bins. The color similarities between speaker candidate and reference models are calculated in terms of the Bhattacharyya distance.

$$D_m(\mathbf{x}_k) = \sqrt{1 - \sum_{u=1}^{U} \sqrt{\hat{q}^{(u)}(\mathbf{x}_k) r_m^{(u)}}} \tag{14}$$

Assuming that noise on the color likelihood function is Gaussian, then the likelihood function of each measured color histogram can be written as [11]:

$$\begin{aligned} g_m(\mathbf{z}_k | \mathbf{x}_k) &\propto \mathcal{N}(\mathbf{z}_k; 0, \sigma_c^2) \\ &= \frac{1}{\sigma_c \sqrt{2\pi}} \exp\left\{-\frac{D_m(\mathbf{x}_k)^2}{2\sigma_c^2}\right\} \end{aligned} \tag{15}$$

where σ_c^2 is the variance of noise for the color likelihood.

2.3. Proposed AV-SMC-PHD Filtering

The way we introduce the DOA data into SMC-PHD filter is based on [3] and [5] where the efficiency of the particles is improved with DOA information under a particle filter framework. The DOA is used to draw a line, named as DOA line,

from the center of microphone array to a point in the image frame estimated by the projection of DOA to $2D$ image plane. Then, all particles are re-allocated around the DOA line. Details on projecting $3D$ DOA information to $2D$ can be found in [3] and [5].

Here, DOA data is not used in the same way for all particles as in [3] and [5], instead the contribution of DOA information is varied by the type of the particles. In the RFS, multi speaker state is defined in equation (4) as the union of surviving, spawned and born particles in the SMC-PHD filter. In our proposed algorithm, the born particles are generated only when the detection of a new speaker occurs. The born particles are uniformly distributed around the DOA line as illustrated in Figure 1 (b). The surviving and spawned particles are also concentrated around the DOA line if DOA data exists at that time. Although for a short silence, the missing DOA data is completed by interpolation, DOA data will be lost in the case where the speaker stops talking for a long time. In that situation, our proposed algorithm continues tracking without the DOA information. Reallocating the particles around the DOA line is likely to increase the possibility of speaker detection since the DOA indicates the approximate direction of the sound emanating from the speaker.

The state vector used for surviving and spawned particles is defined as $\tilde{\mathbf{x}}_{s,k}$ for time k since DOA information is used for surviving and spawned particles in the same way. In addition, born particles are defined as $\tilde{\mathbf{x}}_{b,k}$. In each iteration, surviving particles from the previous iteration and the particles spawned from them are distributed by a dynamic model given in equation (1). More details about the generation of surviving, spawned and born particles can be found in [7] and [2]. If the DOA exists in current time, the DOA line is drawn [3] and perpendicular Euclidean distances $\mathbf{d}_k = \begin{bmatrix} d_k^{(1)} & \dots & d_k^{(L_{k-1})} \end{bmatrix}$ of the particles to the DOA line are calculated. In the case that multiple DOA lines exist, the closest DOA line is chosen for the particles as long as the distance to the DOA line is under a pre-determined threshold value. It is required since one speaker may keep quiet while another speaker is still talking which causes the particles belonging to silent speaker to converge to the DOA line that belongs to another speaker. Then, the movement distances $\hat{\mathbf{d}}_k$ of the particles are calculated as follows [5]:

$$\hat{\mathbf{d}}_k = \frac{\mathbf{d}_k}{\|\mathbf{d}_k\|_1} \odot \mathbf{d}_k \tag{16}$$

where $\odot$ is the element-wise product and $\|.\|_1$ is the ℓ_1 norm. $\hat{\mathbf{d}}_k$ is used to relocate the particles $\tilde{\mathbf{x}}_{s,k}$ to around the DOA line:

$$\tilde{\mathbf{x}}_{s,k} = \tilde{\mathbf{x}}_{s,k} \oplus \mathbf{h}_k \hat{\mathbf{d}}_k \tag{17}$$

where $\oplus$ is the element-wise addition and $\mathbf{h}_k = \begin{bmatrix} cos(\theta_k) & 0 & sin(\theta_k) & 0 & 0 \end{bmatrix}^T$. It is multiplied by $\mathbf{h}_k$ to update only the position (x, y) of the particle state vector $\begin{bmatrix} x & \dot{x} & y & \dot{y} & s \end{bmatrix}^T$ in order to provide the perpendicular movement to the DOA line. After that, new speaker case is checked using the DOA information. If the number of DOA lines is greater than the number of estimated speakers in $k-1$, it means a new speaker appears in the scene. Then J_k born particles $\tilde{\mathbf{x}}_{b,k}$ are generated and distributed uniformly around the new DOA line. All particles are combined under $\tilde{\mathbf{x}}_k$ and the prediction step is employed to calculate the weights of particles $\tilde{w}_{k|k-1}$. After the estimation of color likelihood using equation (15), the update step is performed to calculate $\tilde{w}_k$. The total mass, which gives the number of estimated speakers, is calculated by summing the weights of the particles. After the particles are resampled, then positions of the estimated speakers are estimated using the clustering algorithm. Lastly, the identity of the speakers is detected by measuring the similarity between the color histogram of the estimated speakers and the reference speakers models.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the SMC-PHD algorithm on the *AV*16.3 dataset for AV tracking. First, the experimental setup and the performance metric for tracking error analysis are described, and then the comparative results between V-SMC-PHD and AV-SMC-PHD are discussed.

3.1. Setup and Performance Metric

The SMC-PHD was tested using the *AV*16.3 corpus developed by the IDIAP Research Institute [12]. 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 only two multi-speaker sequences because of the space constraints. The first one is Sequence 24 (camera #1) where two moving speakers are walking back and forth, crossing the field of view twice and occluding each other. Sequence 45 (camera #3) is the second sequence where three moving speakers are occluding each other for many times. In these two sequences, the speakers are speaking continuously and the number of speakers is changing up to 3. 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 SMC-PHD are set similar to [9], [10] as: $p_D = 0.98$, $p_S = 0.99$, $\lambda = 0.26$ and $\sigma_c = 0.1$. The uniform density u is $(360 \times 280)^{-1}$ and the number of particles per speaker is $\rho = 50$. As a performance metric, the well-known OSPA-T (Optimal Subpattern Assignment for Tracks) metric [13] is chosen which is a mathematically consistent metric for the evaluation of multi-speaker tracking systems.

The OSPA-T is based on OSPA metric [14] and extends it for tracking management evaluation. The 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.

3.2. Results and Discussions

To test the computational efficiency of the proposed and baseline algorithms, we ran experiments on Sequence 24 and 45 where the number of particles per speaker changes from 25 to 500 in Intel core $i7$ 2.2 GHz processor with 8 GB memory under Windows 7 operating system. Experiments are repeated 10 times and average time costs are given in Figure 2.

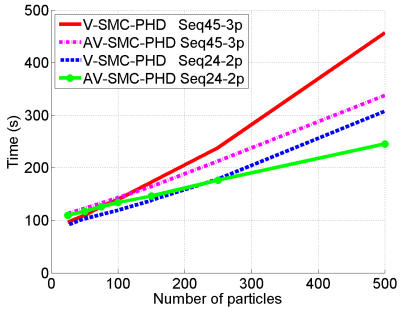


Fig. 2. Computational cost

The computational cost increases with the number of particles and the cost of the V-SMC-PHD filter is larger than that of the AV-SMC-PHD filter. Time cost for Sequence 45 is expected to be higher than Sequence 24 since the maximum number of speakers is three in Sequence 45 while it is two in Sequence 24. It is true that using audio information brings some computational cost to our proposed tracking system. However, Figure 2 shows this cost is negligible and makes the proposed algorithm computationally more efficient than the baseline algorithm. The following experiments are about the estimation accuracy of the algorithms. Some frames from Sequence 24 are shown in Figure 3. The first row shows the results of V-SMC-PHD filter, and the second row for AV-SMC-PHD filter.

In the first column, only one speaker is detected by the V-SMC-PHD filter while both speakers are detected by our proposed AV-SMC-PHD filter. After occlusion in the second column, our proposed AV-SMC-PHD filter tracks the speakers more accurately. There is no DOA information in the third column, but still our proposed AV-SMC-PHD filter manages to track both speakers while the V-SMC-PHD filter lost one speaker. Figure 4 shows the estimation of the number of speakers.

Here, the number of active speakers is changing from 2 to 0 and our proposed AV-SMC-PHD filter shows better performance than the V-SMC-PHD filter. For visualization, downsampling is performed to the plots.

The same experiments are conducted on Sequence 45 and frames are given in Figure 5. Here, the three speakers occlude

Fig. 3. Sequence 24 (camera #1): Two speakers with occlusions. The first row shows the results of the V-SMC-PHD filter and the second row shows tracking results of our proposed AV-SMC-PHD filter.

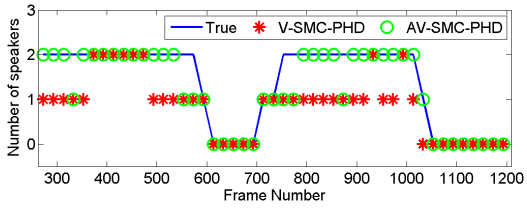


Fig. 4. Number of speakers estimation for Sequence 24.

each other many times and the AV-SMC-PHD filter is able to detect and follow all speakers even after occlusions.

Fig. 5. Sequence 45 (camera #3): Three speakers with occlusions. V-SMC-PHD performance is shown in the first row, and AV-SMC-PHD shows better performance in the second.

The number of speakers estimated for Sequence 45 is given in Figure 6. It can be observed that the performance of the V-SMC-PHD filter is not as good as the AV-SMC-PHD filter.

To see the performance difference between the filters, the OSPA-T errors are plotted for Sequence 24 and Sequence 45 in Figure 7-(a) and (b), respectively. To get more reliable results, the experiments are repeated 10 times and the average error is plotted. In Figure 7-(a), average OSPA-T error is 27.12 for V-SMC-PHD and 17.71 for AV-SMC-PHD. It means that AV-SMC-PHD offers 34.71% improvements over V-SMC-PHD in Sequence 24. The average OSPA-T errors for Sequence 45 in Figure 7-(b) are 39.09 and 28.43 for V-SMC-PHD and AV-SMC-PHD, respectively. Again, AV-

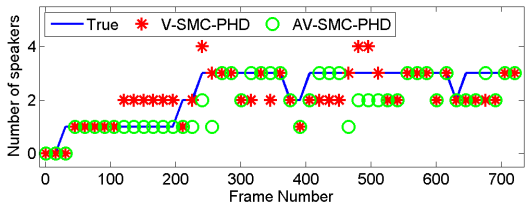
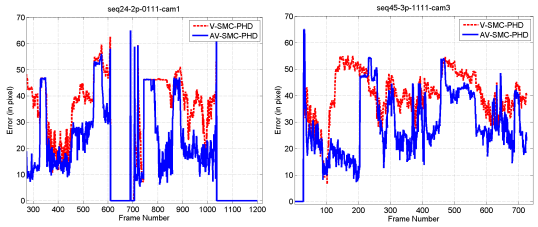


Fig. 6. Number of speakers estimation for Sequence 45.

SMC-PHD filter performs better with a 27.27% improvement. It is clearly seen that adding audio information to the visual tracker leads to an increase in performance.



(a) (b)

Fig. 7. Performance comparison in terms of OSPA-T error.

4. CONCLUSION

In this study, we have proposed a SMC-PHD approach for tracking a variable number of speakers in a smart room environment using audio-visual measurements. Efficient distribution of the born particles based on the DOA information reduces both the computational complexity and the estimation error. The proposed AV-SMC-PHD algorithm has been evaluated on two different sequences from the *AV*16.3 dataset. Experimental results demonstrated that the proposed technique can reliably estimate both the number of speakers and the positions of the speakers with significant improvement in a challenging tracking scenario such as occlusions.

5. ACKNOWLEDGEMENT

This work was supported by the Engineering and Physical Sciences Research Council of the U.K. under Grants EP/K014307/1 and EP/L000539/1.

6. REFERENCES

[1] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[2] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multitarget filtering with random finite sets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 4, pp. 1224–1245, Oct. 2005.

[3] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio constrained particle filter based visual tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 3627–3631.

[4] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Adaptive particle filtering approach to audio-visual tracking," in *Proc. IEEE 21st Eur. Signal Process. Conf.*, Sep. 2013, pp. 1–5.

[5] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 186–200, Feb. 2015.

[6] J. Mullane, B.-N. Vo, M. D. Adams, and B.-T. Vo, "A random-finite-set approach to Bayesian SLAM," *IEEE Trans. Robot.*, vol. 27, no. 2, pp. 268–282, Jan. 2011.

[7] 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 Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.

[8] E. Maggio, M. Taj, and A. Cavallaro, "Efficient multitarget visual tracking using random finite sets," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1016–1027, Aug. 2008.

[9] X. Zhou, Y. F. Li, and B. He, "Entropy distribution and coverage rate-based birth intensity estimation in GM-PHD filter for multi-target visual tracking," *Signal Process.*, vol. 94, pp. 650–660, 2014.

[10] J. Wu, S. Hu, and Y. Wang, "Adaptive multifeature visual tracking in a probability-hypothesis-density filtering framework," *Signal Process.*, vol. 93, no. 11, pp. 2915–2926, Nov. 2013.

[11] J. Czyz, B. Ristic, and B. Macq, "A color-based particle filter for joint detection and tracking of multiple objects," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 217–220.

[12] G. Lathoud, J. M. Odobez, and D. Gatica-Perez, "AV16.3: an audio-visual corpus for speaker localization and tracking," in *Proc. 2004 Mach. Learn. Med. Imag. Workshop*, 2005, pp. 182–195.

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

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