TRACKING SOUND SOURCES FOR OBJECT-BASED SPATIAL AUDIO IN 3D AUDIO-VISUAL PRODUCTION

Mohd Azri Mohd Izhar **Marco Volino** **Adrian Hilton**
Philip J. B. Jackson
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, GU3 7XH, UK
{m.mohdizhar, marco.volino, a.hilton, p.jackson}@surrey.ac.uk

ABSTRACT

In 3D audio-visual production, the positioning of sound sources relative to the visual display requires careful attention. This can be achieved with object-based audio which also allows the producer to maintain control over individual elements within the mix. However, each object's metadata is needed to define its position over time. Therefore, a 3D multiple-target tracking system relying on the fusion of audio and visual modalities using iterated-corrector probability hypothesis density (IPHD) filtering framework is proposed for tracking unknown and varying number of audio objects. Video data is analyzed to form a 3D reconstruction of a scene, and human figure detection is applied to the 2D frames of a camera-array. The steered response power of the acoustical signals at a microphone-array is used, with the phase transform (SRP-PHAT), to determine the dominant source position(s) at any time. Both position estimates from audio and visual processing are then fused using an IPHD filter and the reliabilities from each modality are utilized to provide more robust tracking performance. We apply the proposed 3D audio-visual IPHD (3DAV-IPHD) tracker for tracking audio objects in six degree-of-freedom (6DOF) virtual reality (VR) production and in our case, audio-visual measurements are captured using a compact audio-visual sensing platform consisting of 16-element microphone-array and 11-element light-field camera-array. Our experimental results show that audio information can successfully compensate the missed detections from visual-only tracking and hence, demonstrate the effectiveness of fusing audio and visual information using the proposed 3DAV-IPHD tracker. This gives an initial promising indication that such an approach may open the door to object-based VR production for live events.

1. INTRODUCTION

In immersive and interactive audio-visual content, there is very significant scope for spatial misalignment between the two main modalities. So, in productions that have both 3D video and spatial audio, the positioning of sound sources relative to the visual display requires careful attention. This may be achieved in the form of object-based audio, moreover allowing the producer to maintain control over individual elements within the mix and also, enabling the greatest flexibility in adapting a scene to the local reproduction system and the user's preferences [1, 2]. Each audio object contains metadata describing essential attributes and properties such as its position over time throughout the scene. As a result, tracking of sound sources needs to be performed for automated positional metadata extraction.

Audio tracking can be achieved from recordings collected via two or more microphones such as the binaural microphone or the microphone-array. The captured signals can then be processed using a sound source localization method to get source position estimates. Among the well-known approaches for sound source localization are the methods based on time delay estimation [3], binaural localization [4], steered-response power (SRP) [5, 6] and subspace techniques [7]. The SRP-based method is employed in our system due to its robustness against noisy and reverberant environments as compared to other methods. Despite that, the method still suffers from performance degradation in adverse acoustic environments involving multiple audio objects. To overcome this problem, extra information from the visual cue can be utilized as the method of visual tracking is typically reliable and accurate. Meanwhile, audio tracking is not restricted by the limitations in visual tracking, i.e., limited field of view, poor lighting and occlusions. Therefore, fusing both audio and visual information can provide more robust tracking performance as both modalities complement each other.

There have been several audio-visual trackers proposed including those that operate in the 3D space (e.g. [8–13]). In [11], a 3D single-speaker tracker was developed and, to fuse audio and visual cues, an adaptive particle filter was proposed based on the audio-visual reliabilities. A new fusion strategy based on particle filtering for multiple-speaker tracking was presented in [12]. However, it was developed under the assumption that the number of speakers is known and constant. In practice, however, this is not always the case as the speakers may appear and disappear in an unpredictable manner. As a result, a 3D multiple-speaker tracker using the probability hypothesis density (PHD) [14] filtering approach was proposed for tracking unknown and variable number of speakers [13]. Instead of fusing data from the two modalities in PHD filtering, only the visual-data was filtered and the audio part was fused later after PHD filtering to compensate for missed detections (gaps) in the visual tracking results. In that tracker, the contribution from the audio part was limited and it was

only considered when the speaker was visually detected.

Against this background, we introduce a 3D multiple-target tracking system relying on the fusion of audio and visual modalities using the PHD filtering framework. As opposed to the single-sensor PHD filter used in [13], we employ the multiple-sensor PHD filter explicitly the iterated-corrector PHD (IPHD) filter was chosen among other multiple-sensor PHD filter methods due to its good balance in performance and computational complexity [15]. The sequential Monte Carlo (SMC) implementation is considered in providing a practical solution for PHD filtering as it can perform well with non-Gaussian and non-linear scenarios [16]. The proposed 3D audio-visual (3DAV) IPHD filter termed as the 3DAV-IPHD filter can exploit the reliability from both audio and visual processing, i.e., the SRP and visual-detection confidence measure in computing the audio and visual likelihoods as well as audio and visual clutter intensities to improve the robustness and accuracy of the tracker. The overall system of our 3DAV-IPHD tracker consists of separate components for audio and visual processing, the 3DAV-IPHD filter and identification (ID) association.

The proposed 3DAV-IPHD tracker is applied to track sound sources in six degree-of-freedom (6DOF) virtual reality (VR) production. An audio-visual sensing platform with 16-element microphone-array and 11-element light-field camera-array was built for recording the VR content. Similar to [12], our audio-visual sensing platform is a compact platform and has benefits in terms of portability and easy setup. However, the main challenge in processing the signals captured from this compact sensing platform is depth estimation as targets are not surrounded by sensors. We rely on our camera-array processing to provide accurate depth estimation which is essential not just for 3D tracking but also for rendering the 6DOF VR videos. The novel contributions of this paper can be summarized as follows:

1. We conceive the 3DAV-IPHD tracker for tracking an unknown and varying number of multiple targets in a 3D space. In contrast to [13], our tracker can work with scenarios when one of the modalities is not available or when it has failed to detect a valid target. Furthermore, our tracker can be generalized to track any audio object and not specifically tailored to speaker-tracking unlike related works in [11–13].

2. We utilize the SRP and the confidence measure from audio and visual processing to improve the robustness and accuracy of the 3DAV-IPHD tracker.

3. We apply the proposed 3DAV-IPHD tracker to track sound sources in 6DOF VR production.

The remainder of this paper is organized as follows. We describe the proposed 3DAV-IPHD tracker in Section 2. Experimental setup and results are presented and discussed in Section 3. Our concluding remarks are provided in Section 4.

2. PROPOSED SYSTEM

The block diagram of our system to produce audio object streams is depicted in Fig. 1. The proposed system consists of three main processes and these processes are performed in stages beginning with positional metadata extraction, beamforming and finally, association of metadata with the audio signals. In order to extract the positional metadata, 3DAV-IPHD tracker is proposed utilizing audio-visual information captured using a camera-array and a microphone-array. The sound sources are visually detected from the multi-camera video frames to get 3D position estimates of sound sources together with the detection confidence measure. In parallel with visual detection, audio signals captured from the microphone-array are processed using an SRP-based method to yield another set of 3D position estimates with the associated SRP values. The 3D position estimates from both audio and visual processing are fused using the proposed 3DAV-IPHD tracker. On top of the fused position data, the confidence measures and SRP values can assist the tracker to produce more accurate and robust 3D position estimates. The proposed 3DAV-IPHD tracker is detailed in Section 2.3.

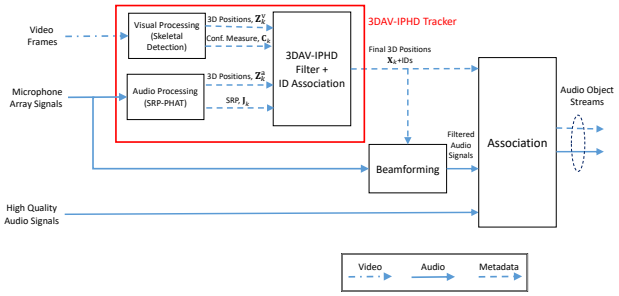

Figure 1. Block diagram of the proposed 3DAV-IPHD tracker and the overall system for audio objectification.

In the second stage, beamforming is invoked to filter and focus the audio signals from the microphone-array to specific directions based on the refined 3D position estimates. In the last stage, these filtered signals are served as reference signals in associating high quality audio signals (captured using close microphones) with the correct metadata based on the estimation from the 3DAV-IPHD tracker. The final output is given in the form of audio object streams which can then be used for further processing such as object-based production and rendering. In this paper, we focus only on the first stage which is sound source tracking.

2.1 Audio Processing

We employ the SRP method for acoustic source localization due to its robustness especially in a noisy and reverberant environment. SRP methods are based on the filter-and-sum beamforming technique that computes the received power when the microphone array is steered in the direction of a specific location.

Suppose there is a set of M microphones with locations in the (x, y, z)-coordinate system denoted as $\mathbf{m}_i \in \mathbb{R}^3$

where $i = 1, 2, ..., M$. For a given microphone pair of i and j, the generalized cross correlation (GCC) between time-discrete signals $s_i(t)$ and $s_j(t)$ can be written as

$$R_{ij}(\tau_{ij}(\mathbf{p})) = \int_{-\infty}^{\infty} S_i(\omega) S_j^*(\omega) \Psi_{ij}(\omega) e^{j\omega\tau_{ij}(\mathbf{p})} d\omega, \tag{1}$$

where $S_i(\omega)$ is the short time Fourier transform (STFT) of $s_i(t)$, $S_j^*(\omega)$ is the conjugate of the STFT of $s_j(t)$, Ψ_{ij} is the weighting function and $\tau_{ij}(\mathbf{p})$ is the time difference of arrival (TDOA) across the microphone pair i and j for a sound source located at $\mathbf{p} \in \mathbb{R}^3$. The TDOA can be calculated as

$$\tau_{ij}(\mathbf{p}) = \frac{\|\mathbf{p} - \mathbf{m}_i\| - \|\mathbf{p} - \mathbf{m}_j\|}{c}, \tag{2}$$

where c is the speed of sound.

The SRP of a sound source located at the spatial position $\mathbf{p}$ can then be computed in terms of GCCs which is given by [5, 6]

$$J(\mathbf{p}) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} R_{ij}(\tau_{ij}(\mathbf{p})). \tag{3}$$

The weighting function in Eqn. (1) has a significant effect on the estimation accuracy. One of the common choices that can provide robustness against noise is to use the phase transform (PHAT) weighting function:

$$\Psi_{ij}(\omega) = \frac{1}{|S_i(\omega) S_j^*(\omega)|} \tag{4}$$

in the SRP computation and this method is known as SRP-PHAT. The source location can then be estimated as the one resulting maximum SRP value from a set of candidate source locations $\mathcal{G}$:

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p} \in \mathcal{G}} J(\mathbf{p}). \tag{5}$$

Despite the benefit of robustness against noise and reverberant environments, the SRP computation is based on intensive grid-search methods to find the global maximum and hence, it is computationally expensive for a real-time system. In [17], a global-maximum-finding algorithm called stochastic region contraction (SRC) was used in reducing the computational cost of the grid-search step in the SRP-PHAT computation. It was shown that their proposed SRP-PHAT-SRC method can reduce the computational cost by more than two orders of magnitudes with full or almost-full accuracy as compared to the conventional SRP-PHAT method. Therefore, the low-complexity SRP-PHAT-SRC method is used for acoustic source localization in our proposed system.

The extension of the SRP-PHAT method to multiple sound sources with unknown number of sources using the region-zeroing (RZ) approach was proposed in [18]. The RZ approach is considered in our work due to its performance superiority with less computational cost as compared to other similar methods. The procedure in estimating the position of multiple sound sources using the SRP-PHAT-SRC method and the RZ approach is described as follows:

- **Step-1** Evaluate SRP-PHAT values on a large set of R_{msrp} randomly selected points and keeping the highest N_{msrp} points.
- **Step-2** Apply the RZ method:
  1. Find the point with the highest SRP value. Store the position.
  2. Remove the point and all the neighboring points within l_{msrp}-meter radius in the space.
  3. Repeat Step-2(a) and Step-2(b) until no more point left in the space.
- **Step-3** Cluster all the stored points using Mahalanobis distance agglomerative clustering and the threshold for the minimum distance is set to d_{msrp} to result in P_{msrp} clusters.
- **Step-4** Apply SRC in each cluster. The point in each of the P_{msrp} clusters will be the candidate for the location estimates.
- **Step-5** From the P_{msrp} location estimates, keep only Q_{msrp} location estimates such that their SRP values are above the noise level J_{noise}. The Q_{msrp} represents the estimated number of sound sources and the associated locations are given as $\hat{\mathbf{p}}_i$ for $i = 1, 2, ..., Q_{\text{msrp}}$

2.2 Visual Processing

Here, our goal is to identify and track the sound source location from visual input for the purpose of ground truth comparison of audio only tracking and as a complimentary input to a multi-modal tracker. In the case of speaker-tracking, we assume that the sound originates from the mouth and use a human pose detector to locate body joints in the image. To this end, we employ OpenPose [19] [1] to detect the 2D skeletal keypoints of human subjects in the video streams. Due to the skeletal structure of OpenPose we use the nose joint and assume that there is a negligible offset between the nose and mouth when compared to the vertical resolution of the microphone array.

Given an input video stream, OpenPose provides 2D detections and confidences for each subject, however these are both spatially and temporally inconsistent. In order to accurately estimate and track the 3D location of the nose joint, 2D detection must be sorted such that they are spatially consistent between camera views and temporally consistent over time. We utilize a detection sorting algorithm as described by Malleson *et al.* [20] to effectively sort all 2D detections into a defined number of subjects, see Figure 2.

Sorted 2D detections are combined with confidences to remove the impact of potentially unreliable detections to estimate the 3D location via triangulation, as described in Eqn. (6):

$$\arg\min_{p} \sum_{i=1}^{N_C} ||\pi(i, p) - a_i||, \tag{6}$$

[1] Open-source implementation used in this work `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

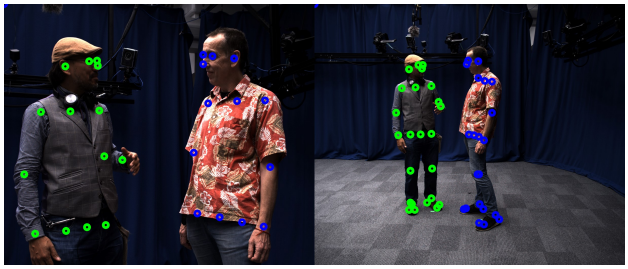

Figure 2. Captured image example with keypoints overlaid. Sorted 2D keypoints across camera views. Subject 1 (left,green) and Subject 2 (right,blue).

where $\pi(i,p)$ is the camera projection function that maps estimated 3D point p into the 2D cameras plane, a_i is the detected joint keypoint for the i^{th} camera of N_C cameras.

2.3 3D Audio-Visual PHD Filtering

In the previous two sections, the 3D positions of audio objects were estimated individually using either audio or visual processing. It is known that audio and visual cues complement each other and hence, instead of relying on these two modalities separately it is better to fuse them for providing more robust tracking in the case that either modality is unavailable or both are corrupted. Both positional data from audio and visual processing are fused using the proposed 3DAV-IPHD filter as shown in Fig. 1. On top of fusing the two modalities, PHD filtering can mitigate outliers and smooth the source trajectories from noisy 3D position estimates from audio and visual processing. The SMC implementation is considered in providing a practical solution for PHD filtering as it can perform well with non-Gaussian and non-linear scenarios [16]. Both the audio positional data and the visual positional data are fed as inputs to the PHD filter along with their reliabilities i.e., the SRP from audio processing and the confidence measure from visual processing.

At the k-th frame, suppose there are m_k^{a} and m_k^{v} targets detected from audio and visual processing, respectively. The audio observation set can be denoted as $\mathbf{Z}_k^{\mathrm{a}} = \{\mathbf{z}_1^{\mathrm{a}}, ..., \mathbf{z}_{m_k^{\mathrm{a}}}^{\mathrm{a}}\}$ and the visual observation set can be denoted as $\mathbf{Z}_k^{\mathrm{v}} = \{\mathbf{z}_1^{\mathrm{v}}, ..., \mathbf{z}_{m_k^{\mathrm{v}}}^{\mathrm{v}}\}$, where $\mathbf{z}$ is a 3D position vector $\mathbf{z} = [x, y, z]^\top$. From the noisy observation sequences $\mathbf{Z}_1^{\mathrm{a}}, \mathbf{Z}_2^{\mathrm{a}}, ..., \mathbf{Z}_k^{\mathrm{a}}$ and $\mathbf{Z}_1^{\mathrm{v}}, \mathbf{Z}_2^{\mathrm{v}}, ..., \mathbf{Z}_k^{\mathrm{v}}$, the real target status $\mathbf{X}_k = \{\mathbf{x}_1, ..., \mathbf{x}_{n_k}\}$ can be estimated using PHD filtering where n_k is the number of detected targets and $\mathbf{x}$ represents the (x, y, z)-position and also the associated $(\dot{x}, \dot{y}, \dot{z})$-velocity of a target, i.e. $\mathbf{x} = [x, y, z, \dot{x}, \dot{y}, \dot{z}]^\top$. The SRP from audio processing is denoted as $\mathbf{J}_k = \{J_k(\mathbf{z}_1^{\mathrm{a}}), ..., J_k(\mathbf{z}_{m_k^{\mathrm{a}}}^{\mathrm{a}})\}$ whereas the confidence measure from visual processing is denoted as $\mathbf{C}_k = \{C_k(\mathbf{z}_1^{\mathrm{v}}), ..., C_k(\mathbf{z}_{m_k^{\mathrm{v}}}^{\mathrm{v}})\}$.

There are four main steps in SMC-PHD filtering namely the prediction, update, resampling and state estimation. The joint observation set is denoted as $\mathbf{Z}_k = \mathbf{Z}_k^{\mathrm{a}} \cup \mathbf{Z}_k^{\mathrm{v}}$ and correspondingly, the cardinality of this set is represented by $m_k = m_k^{\mathrm{a}} + m_k^{\mathrm{v}}$. The number of new-born particles per target and the new-born target intensity are set to be M_b and ν, respectively. In the prediction step, m_k groups of M_b Gaussian-distributed particles are formed and each group centered at each of the targets in $\mathbf{Z}_k$. All the new-born particles have equal weight of $\frac{\nu}{m_k M_b}$. Meanwhile, there are N_{k-1} surviving particles from the previous frame with particle state of $\mathbf{x}_{k-1}^{(n)}$ for $n = 1, 2, ..., N_{k-1}$. The evolution of a particle state $\mathbf{x}_{k-1}^{(n)}$ to the new state $\mathbf{x}_{k|k-1}^{(n)}$ is given by the following motion model:

$$\mathbf{x}_{k|k-1}^{(n)} = \mathbf{F}\mathbf{x}_{k-1}^{(n)} + \mathbf{q}_k^{(n)}, \tag{7}$$

where $\mathbf{F}$ is a 6×6 prediction matrix following the first-order linear motion model and $\mathbf{q}_k^{(n)}$ is the zero-mean independent and identically distributed Gaussian noise with a pre-defined covariance matrix $\mathbf{Q}$. The particle weight is updated as $w_{k|k-1}^{(n)} = P_S w_{k-1}^{(n)}$ where P_S is the probability of target survivability.

Based on the IPHD method [15], the update step needs to be performed two times iteratively taking into account the two sensors in our system namely the microphone-array (audio sensor) and the camera-array (visual sensor). In [21, 22], it was shown that the order of the sensor updates can result in different tracking performance and it was suggested to start the iterated update using the observation set from the sensor having the lower probability of target detection. Generally, the probability of detection of the visual sensor P_D^{v} is higher than the probability of detection of the audio sensor P_D^{a} and hence, the weight of the particle $w_{k|k-1}^{(n)}$ is updated based on the audio sensor first as

$$\widetilde{w}_{k|k}^{(n)} = \left[(1 - P_D^{\mathrm{a}}) + \sum_{\mathbf{z} \in \mathbf{Z}_k^{\mathrm{a}}} \frac{P_D^{\mathrm{a}} g^{\mathrm{a}}(\mathbf{z}|\mathbf{x}_{k|k-1}^{(n)})}{\mathcal{L}^{\mathrm{a}}(\mathbf{z})}\right] w_{k|k-1}^{(n)}, \tag{8}$$

where

$$\mathcal{L}^{\mathrm{a}}(\mathbf{z}) = \kappa^{\mathrm{a}}(\mathbf{z}) + \sum_{j=1}^{N_{k-1}+m_k M_b} P_D^{\mathrm{a}} g^{\mathrm{a}}(\mathbf{z}|\mathbf{x}_{k|k-1}^{(j)}) w_{k|k-1}^{(j)}, \tag{9}$$

$\kappa^{\mathrm{a}}(\mathbf{z})$ is the audio clutter intensity and $g^{\mathrm{a}}(\mathbf{z}|\mathbf{x})$ is the audio likelihood function. Both $\kappa^{\mathrm{a}}(\mathbf{z})$ and $g^{\mathrm{a}}(\mathbf{z}|\mathbf{x})$ can be computed using the normalized SRP $\bar{J}_k(\mathbf{z}) = \frac{J_k(\mathbf{z})}{\max\{\max(\mathbf{J}_j)\}_{j=1}^K}$ as

$$\kappa^{\mathrm{a}}(\mathbf{z}) = \lambda^{\mathrm{a}} \frac{1 - \bar{J}_k(\mathbf{z})}{\sum_{\mathbf{z} \in \mathbf{Z}_k} \left(1 - \bar{J}_k(\mathbf{z})\right)} \tag{10}$$

and

$$g^{\mathrm{a}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\|\mathbf{z} - \mathbf{x}\|_2 | 0, \left(\frac{\alpha^{\mathrm{a}}}{\bar{J}_k(\mathbf{z})}\right)^2\right), \tag{11}$$

where $\|.\|_2$ is the l_2 norm, λ^{a} is the average number of audio clutter points and α^{a} is the user-defined constant indicating uncertainty of reliabilities in audio measurement. In the second iteration of the update, the weight $\widetilde{w}_{k|k}^{(n)}$ is

updated as

$$w_{k|k}^{(n)} = \left[(1 - P_D^{\mathrm{v}}) + \sum_{\mathbf{z} \in \mathbf{Z}_k^{\mathrm{v}}} \frac{P_D^{\mathrm{v}} g^{\mathrm{v}}(\mathbf{z}|\mathbf{x}_{k|k-1}^{(n)})}{\mathcal{L}^{\mathrm{v}}(\mathbf{z})} \right] \widetilde{w}_{k|k}^{(n)}, \tag{12}$$

where

$$\mathcal{L}^{\mathrm{v}}(\mathbf{z}) = \kappa^{\mathrm{v}}(\mathbf{z}) + \sum_{j=1}^{N_{k-1}+m_k M_b} P_D^{\mathrm{v}} g^{\mathrm{v}}(\mathbf{z}|\mathbf{x}_{k|k-1}^{(j)}) \widetilde{w}_{k|k}^{(j)}, \tag{13}$$

$\kappa^{\mathrm{v}}(\mathbf{z})$ is the visual clutter intensity and $g^{\mathrm{v}}(\mathbf{z}|\mathbf{x})$ is the visual likelihood function. The computation of $\kappa^{\mathrm{v}}(\mathbf{z})$ and $g^{\mathrm{v}}(\mathbf{z}|\mathbf{x})$ are similar to the audio counterpart as given in Eqn. (10) and Eqn. (11), respectively. Instead of using $\bar{J}_k(\mathbf{z})$, λ^{a} and α^{a}, we use the confidence measure $C_k(\mathbf{z})$, λ^{v} and α^{v}, respectively in the visual update.

After the update step, the estimated number of targets can be computed by the sum of weights of all N_{k-1} surviving particles and $m_k \cdot M_b$ new-born particles. The number of particles of the current frame N_k can then be determined by multiplying the estimated number of targets with $(M_p + M_b)$ where M_p is the number of particles per surviving target and after that, rounding the result to the nearest integer. In the resampling step, the particles are resampled from $\{(\mathbf{x}_{k|k}^{(n)}, w_{k|k}^{(n)})\}_{n=1}^{N_{k-1}+m_k M_b}$ to $\{(\mathbf{x}_k^{(n)}, w_k^{(n)})\}_{n=1}^{N_k}$. Particles are selected based on their assigned weights where particles with high weights are duplicated while particles with low weights are discarded. After that, the particles are grouped using the K-means clustering algorithm. If the accumulated weight is greater than a pre-defined threshold ξ, then we consider there exists a target in that cluster with the estimated position given by the centroid of the cluster. The pseudo code of the 3DAV-IPHD filter is shown in Algorithm 1. After PHD filtering, an ID is assigned to every final 3D position estimate using the ID association scheme proposed in [23].

Algorithm 1: 3DAV-IPHD filter algorithm

Input: $\{\mathbf{Z}_i^{\mathrm{a}}\}_{i=1}^K, \{\mathbf{S}_i\}_{i=1}^K, \{\mathbf{Z}_i^{\mathrm{v}}\}_{i=1}^K, \{\mathbf{C}_i\}_{i=1}^K$
Output: Target state $\mathbf{X}$
$k = 1$
while $k <= K$ **do**
  % Prediction step
  foreach $n = \{1, 2, ..., N_{k-1}\}$ **do**
    Propoagate surviving particles using Eqn. (7)
    $w_{k|k-1}^{(n)} = P_S w_{k-1}^{(n)}$
  end
  foreach $\mathbf{z} \in \mathbf{Z}_k$ **do**
    Draw M_b new particles
    $\mathbf{x}_{k|k-1}^{(n)} \sim \mathcal{N}(\cdot|\mathbf{z}, \Sigma)$ with equal weight of $w_{k|k-1}^{(n)} = \frac{\nu}{m_k M_b}$
  end
  % Update step
  foreach $n = \{1, 2, ..., N_{k-1} + m_k M_b\}$ **do**
    Update the particle weight using Eqn. (8)
    Reupdate the particle weight using Eqn. (12)
  end
  % Resampling
  Resample $\{(\mathbf{x}_{k|k}^{(n)}, w_{k|k}^{(n)})\}_{n=1}^{N_{k-1}+m_k M_b}$ to obtain $\{(\mathbf{x}_k^{(n)}, w_k^{(n)})\}_{n=1}^{N_k}$
  % State estimation
  Cluster particles for the final state $\mathbf{X}_k$ estimation
  $k = k + 1$
end

3. EXPERIMENTAL EVALUATION

In this section, we first describe the experimental setup and performance metrics before presenting the analysis and comparison of the results.

3.1 Setup and Performance Metric

The multi-sensing platform consists of a 16-element microphone-array and an 11-element light-field camera-array, as illustrated in Fig. 3 and was built for 6DOF VR capturing. The cameras are located across three rows (3,5,3) with cameras positioned equidistant from one another excluding the central camera that is located in the center of the array. The distance between the microphones is log-spaced in the horizontal-direction in order to cover a broad frequency range. The microphone-array arrangement (in Fig. 3) can provide high azimuthal resolution and lower elevation resolution considering that in nature, we are more sensitive to the sound changes in the horizontal-direction than the vertical-direction [24]. This integrated microphone-array and camera-array rig was mounted on a static tripod and used in our audio-visual studio recordings alongside more conventional sound recording by spot microphones and a first-order ambisonic room microphone.

The recordings were made of short scenes representing three genres which are the drama, sport and music. In this paper, we focus on the drama scene which consists of two subjects in a conversation standing side-by-side with considerable movement from where they stand as depicted in the image example in Fig. 2. The audio and video were recorded at 48 kHz and 25 Hz, respectively. They were synchronized before being processed in our system and the coordinate system's origin was aligned to the center camera. The duration of the sequence was 20s with a total of 500 video frames.

In our audio processing, the positions of the sound sources were estimated using all 16 acoustical signals from the microphone-array. The parameters for the SRP-PHAT method were set as follows. The whole sequence was divided into several frames with frame duration of 0.2s. In each frame, we applied short-time Fourier transform (STFT) to the signals with the window size of 512 samples and using the Hanning window function. In the SRP computation, 100,000 random (x, y, z) points were evaluated over horizontal $x \in [-2, 2]$, vertical $y \in [-1, 1]$ and depth

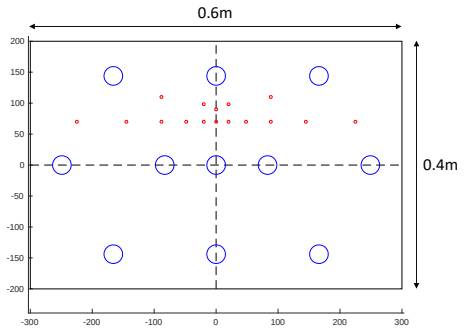

Figure 3. The schematic diagram of the AV-sensing platform for capturing 6DOF VR content. The small red circles indicate the positions of the 16 microphones while the blue cirles indicate the positions of the 11 light-field cameras.

$z \in [0, 5]$ meters. From all these points, $N_{\text{msrp}} = 4500$ best points were selected for region-zeroing and clustering with $l_{\text{msrp}} = 0.5$m and $d_{\text{msrp}} = 1.2$m. The noise power level J_{noise} was set to 0.2. The PHD filter had the same step size as the video frame rate which was 25 Hz (0.04s) and, since the audio frame duration was 0.2s, any missing audio data was completed by interpolation in the case of short silence (in this case, less than 1.5s). The parameters for the PHD filter were set as: $P_D^{\text{v}} = 0.7, P_D^{\text{a}} = 0.45, P_S = 0.98, \Sigma = 0.005, \kappa^{\text{v}} = \kappa^{\text{a}} = 0.02, \alpha^{\text{v}} = 0.05, \alpha^{\text{a}} = 0.1, M_p = 200$ and $M_b = 100$. The above parameters were empirically set based on our measured data.

In order to evaluate the proposed audio-visual tracker, the ground truth was established by using visual detection and tracking from two separately-positioned camera arrays and by manually annotating the voice activity throughout the sequence. The established ground truth represents the actual positions of sound sources in the sequence. The evaluation was done based on the spherical coordinate system (in azimuth, elevation and radius as illustrated in Fig. 4) rather than the Cartesian coordinate system, to reflect real listening experience. In accordance with human spatial auditory perception [24, 25], a detection was considered to be valid, i.e., true positive (tp) if the estimated position was within the tolerance value of $\pm 5^\circ$ in azimuth, $\pm 10^\circ$ in elevation and ± 0.7m in radius from the ground truth, and otherwise it was labeled as an outlier, i.e., false positive (fp). In the case of missed detection including due to inaccurate detection, the frame was labeled as a false negative (fn). Two evaluation metrics were considered namely recall and precision, where recall $= \frac{\#\text{tp}}{\#\text{tp}+\#\text{fn}}$ and precision $= \frac{\#\text{tp}}{\#\text{tp}+\#\text{fp}}$. Since the two speakers did not talk at the same time but engaged throughout the sequence, there can only be one sound source in each frame and $\#\text{tp}+\#\text{fp}$ is equal to the total frame number. Since there were more visually-valid targets than the audibly-valid targets from the ground truth, the additional visually-valid targets were excluded in the evaluation with the ground truth.

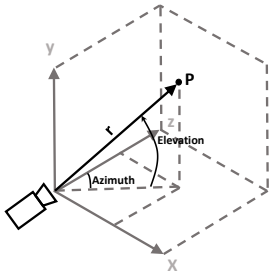

Figure 4. The transformation from the Cartesian coordinate system to the spherical coordinate system.

3.2 Results and Discussion

Fig. 5 depicts the ground truth and the results from audio and visual processing before PHD filtering. Two tracks from visual processing correspond to the positions of the two subjects throughout the sequence even when they did not speak. The recalls for audio and visual processing are 0.814 and 0.906, respectively whereas the precisions for audio and visual processing were 0.883 and 1.000, respectively. In both metrics, visual processing outperforms audio processing which is expected based on the observed results in Fig. 5. There was no outlier in visual processing as all detections were within the tolerable range from the ground truth. However, there were some detections missing, as can be seen from frame number 124 until 158 (speaker 1) and also from frame number 407 until 422 (speaker 2). These missed detection occurred when the subjects turned their head away from the cameras. Meanwhile, it is worth noting that audio processing has better accuracy in the azimuth-direction than other directions and this is due to the arrangement of our microphone-array that has more microphones in the horizontal-direction than the vertical-direction (11 horizontal positions, 4 vertical positions).

The result after PHD filtering and ID association with detection threshold $\xi = 0.25$ is shown in Fig. 6. Despite the small number of outliers introduced which labeled as ID=3 and ID=4, we found that there is no missed frame in both IDs that are associated with our two subjects (ID=1 and ID=2). The audio information successfully compensated the missed detection from visual detection, and hence demonstrates the effectiveness of the proposed filter in fusing audio and visual information. We benchmarked our 3DAV-IPHD tracker with the visual-only 3DV-IPHD tracker and the results are summarized in Tab. 1 for $\xi = 0.25$ and $\xi = 0.5$. In both ξ cases, it can be observed that the 3DAV-IPHD tracker has better recall and precision performance over the visual-only 3DV-IPHD tracker. By lowering the ξ value, the number of valid detections increased which also may include outliers as can be observed from $\xi = 0.5$ to $\xi = 0.25$. However, the number of outliers is not particularly an issue in our application as the outliers will automatically be removed in the later stage when associating the metadata with the audio signals. Explicitly, the reference signal from the outliers will not result in a good match with the high-quality audio signal and will not

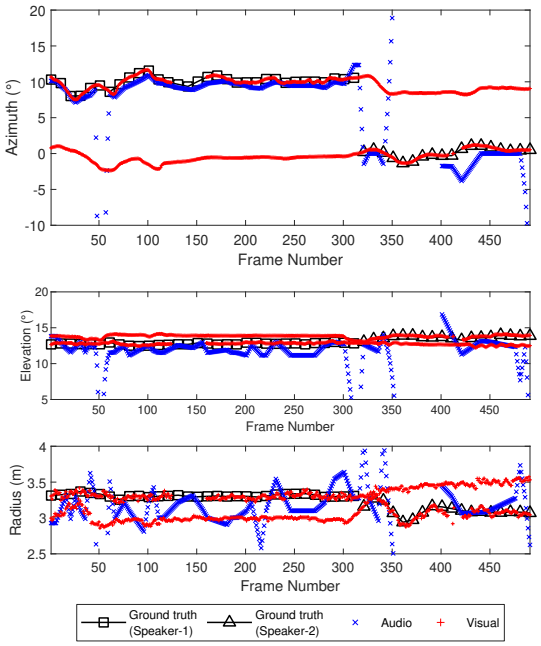

Figure 5. Estimated 3D positions from audio and visual processing. The data points on the ground truth curves were downsampled for better visualisation.

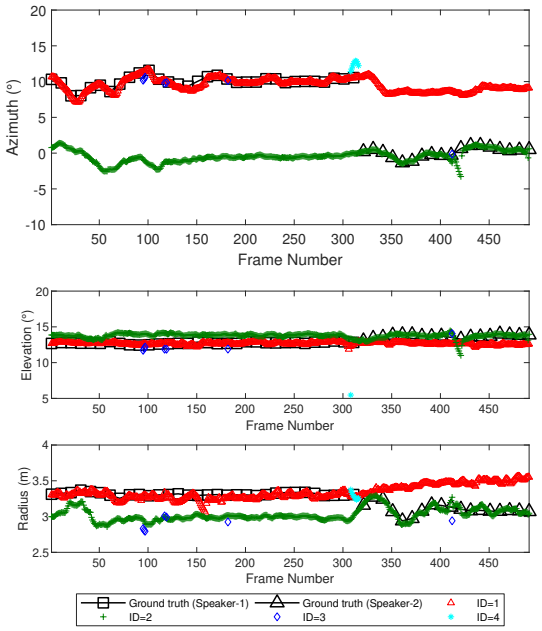

Figure 6. Estimated 3D positions from the proposed 3DAV-IPHD tracker. The data points on the ground truth curves were downsampled for better visualisation.

be considered in the association as long as there is a valid detection in that particular frame. Therefore, the priority is always recall over precision.

System	ξ	Recall	Precision
Audio (without PHD)	-	0.814	0.883
Visual (without PHD)	-	0.906	1.000
Visual-only 3DV-IPHD	0.25	0.910	1.000
	0.50	0.906	1.000
3DAV-IPHD	0.25	1.000	0.986
	0.50	0.910	1.000

Table 1. Performance comparison between the proposed 3DAV-IPHD tracker and the visual-only 3DV-IPHD tracker relying on different threshold values $\xi = 0.25$ and $\xi = 0.5$. The recall and precision from audio and visual processing before PHD filtering are included for reference purposes.

4. CONCLUSION

In this paper, we have proposed the 3DAV-IPHD tracker that can track an unknown and varying number of audio objects using audio-visual measurements. Separate processing is first performed on each measurement where the audio measurement is processed using the SRP-PHAT-SRC method in getting the position estimates of the sound sources while the visual measurement is processed for visual detection and tracking of the sound sources. Both sets of 3D position estimates are then fused using the IPHD filter and the reliabilities from each modality are utilized to provide more robust tracking performance. We have applied the proposed 3DAV-IPHD tracker for tracking audio objects in 6DOF VR production. In our case, audio-visual measurements have been captured using a compact audio-visual sensing platform that consists of 16-element microphone-array and 11-element light-field camera-array. Experimental results have shown that audio information can successfully compensate the missed detections from visual-only tracking as recall went from 91% to 100% and hence, demonstrating the effectiveness of the proposed 3DAV-IPHD tracker in fusing audio and visual information.

As future work, the proposed tracker may be applied to more challenging scenarios such as involving more movements of different types of audio objects, and also scenarios where some of the objects are located outside the cameras' field of view. Benchmarking may be done with other state-of-the-art methods.

5. ACKNOWLEDGMENT

This research was supported by the InnovateUK project Polymersive: Immersive Video Production Tools for Studio and Live Events (105168). The authors would like to thank Qingju Liu for useful discussions.

6. REFERENCES

[1] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proceedings of the IEEE*, vol. 101, pp. 1920–1938, Sep. 2013.

[2] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. J. Hughes, D. Menzies, M. F. S. Gálvez,

Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, "An audio-visual system for object-based audio: From recording to listening," *IEEE Transactions on Multimedia*, vol. 20, pp. 1919–1931, Aug 2018.

[3] G. C. Carter, "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, pp. 236–255, Feb 1987.

[4] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 68–77, Jan 2010.

[5] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberent Environments using Microphone Arrays*. Ph.d. thesis, Brown University, Providence, RI, USA, 2000.

[6] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust Localization in Reverberant Rooms*, pp. 157–180. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, March 1986.

[8] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio–visual active speaker tracking in cluttered indoors environments *," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, pp. 7–15, Feb 2009.

[9] D. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Adv. Signal Process.*, vol. 2002, pp. 1154–1164, Dec 2002.

[10] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Transactions on Multimedia*, vol. 16, pp. 864–880, April 2014.

[11] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3D audio-visual speaker tracking with an adaptive particle filter," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2896–2900, March 2017.

[12] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio–visual sensing device," *IEEE Transactions on Multimedia*, vol. 21, pp. 2576–2588, Oct 2019.

[13] Q. Liu, W. Wang, T. de Campos, P. J. B. Jackson, and A. Hilton, "Multiple speaker tracking in spatial audio via PHD filtering and depth-audio fusion," *IEEE Transactions on Multimedia*, vol. 20, pp. 1767–1780, July 2018.

[14] R. P. S. Mahler, "Multitarget bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, pp. 1152–1178, Oct 2003.

[15] R. Mahler, "Approximate multisensor CPHD and PHD filters," in *2010 13th International Conference on Information Fusion*, pp. 1–8, July 2010.

[16] B. 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, pp. 1224–1245, Oct 2005.

[17] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 1, pp. I–121–I–124, April 2007.

[18] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 125–128, March 2010.

[19] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[20] C. Malleson, J. Collomosse, and A. Hilton, "Real-time multi-person motion capture from multi-view video and IMUs," *International Journal of Computer Vision*, 2019.

[21] L. Liu, H. Ji, and Z. Fan, "Improved iterated-corrector PHD with gaussian mixture implementation," *Signal Processing*, vol. 114, pp. 89 – 99, 2015.

[22] S. Nagappa and D. E. Clark, "On the ordering of the sensors in the iterated-corrector probability hypothesis density (PHD) filter," in *Signal Processing, Sensor Fusion, and Target Recognition XX* (I. Kadar, ed.), vol. 8050, pp. 275 – 280, International Society for Optics and Photonics, SPIE, 2011.

[23] Q. Liu, T. E. de Campos, W. Wang, and A. Hilton, "Identity association using PHD filters in multiple head tracking with depth sensors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1506–1510, March 2016.

[24] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.

[25] P. Zahorik, D. Brungart, and A. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, pp. 409–420, 05 2005.