INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Audio-Visual Multi-Speaker Tracking Based On the GLMB Framework

Shoufeng Lin⋆*, Xinyuan Qian*⋆

Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{elelins,eleqian}@nus.edu.sg

Abstract

Multi-speaker tracking using both audio and video modalities is a key task in human-robot interaction and video conferencing. The complementary nature of audio and video signals improves the tracking robustness against noise and outliers compared to the uni-modal approaches. However, the online tracking of multiple speakers via audio-video fusion, especially without the target number prior, is still an open challenge. In this paper, we propose a Generalized Labelled Multi-Bernoulli (GLMB)-based framework that jointly estimates the number of targets and their respective states online. Experimental results using the AV16.3 dataset demonstrate the effectiveness of the proposed method.

Index Terms: multi-speaker tracking, 3D, audio-visual fusion, GLMB filter

1. Introduction

Speaker tracking has attracted increasing attention in the past few decades due to its wide applications in the fields of video conferencing, surveillance and human-robot interaction [1, 2]. It is a temporal process of identifying and determining the kinematic state of each speaker (e.g., the position and the velocity) given noisy, incomplete or cluttered measurements of different modalities. To date, audio and video are the most popular modalities due to the low cost and convenient installation of microphones and cameras [3]. However, either modality faces respective challenges. The performance of the traditional visual trackers are degraded with target occlusions, limited camera's Field-of-View (FoV) and illumination changes [4, 5, 6], whereas the audio trackers are affected by intermittent voice activities, background noise, and strong room reverberation [7, 8, 9]. Therefore, an audio-visual tracker with the capability of exploiting the complementarity of both modalities is highly demanded, especially under challenging and rapid-varying scenarios [10, 11].

A number of audio-visual trackers have been proposed in the literature [11, 12, 13, 14, 15, 16], based on various Bayesian filters, viz. the Kalman Filter (KF) [12, 13], Particle Filter (PF) [11, 14, 15] and Probability Hypothesis Density (PHD) filter [16]. Among these trackers, the KF provides an efficient analytic solution that approximates the target posterior density with the first and second moments, i.e., mean and covariance. However, due to the assumptions of linear motion and a Gaussian distributed measurement noise model, KF is not the optimal choice for real-world applications. As an alternative to KF, PF is not limited by linear and Gaussian constraints, and hence obtained more popularity. It approximates the Probability Density Function (PDF) of the target using a group of weighted particles where the target is estimated to be located at the maximum likelihood expectation. However, most PF solutions assume a known and constant number of targets and often require a track management stage otherwise. This can be restrictive in practice. The PHD belongs to a family of Multiple Object Tracking (MOT) filters that propagate the multi-speaker PDF based on the Random Finite Set (RFS) theory. Compared to PF, it provides an analytic solution for tracking an unknown and varying number of targets. However, as the earliest RFS-based member, the PHD does not automatically track the identities of targets.

The GLMB filter is one of the most recent members in the RFS-based tracker family. It not only provides an elegant closed-form solution to the MOT [17, 18], but also jointly tracks the identity of each target. It has lately been successfully implemented in multi-speaker state-filtering using either audio features [19, 20] or video features [5]. However, to our best knowledge, the exploration of the GLMB in joint audio-visual tracking has not yet been found in the literature, despite its advantages. Therefore in this paper, we implement the GLMB framework for the joint audio-visual tracking, which we refer to as the Audio-Visual tracking with the Generalized Labelled Multi-Bernoulli (AV-GLMB). The resulting AV-GLMB produces a unique label for each tracked speaker without requiring the *a priori* knowledge of the number of speakers over time. Adaptations to the standard GLMB framework are necessary, which we will provide later in the paper.

2. Problem Formulation

As shown in Fig.1, given the synchronized audio signals $\mathbf{a}_{1:t}$, video images $\mathbf{v}_{1:t}$ and the sensor calibration information ζ_t, we aim to find the 3D location $\mathbf{o}_{t,i}$ of each speaker $i \in I_t$ at time t, where I_t is the set of speakers at t. In existing works [11, 13, 12, 14, 15], I_t is often assumed known *a priori* and constant. However, in what follows, we show that the proposed method does not require such prior knowledge. The system consists of two stages, i.e., the AV measurements, and the AV-GLMB filter.

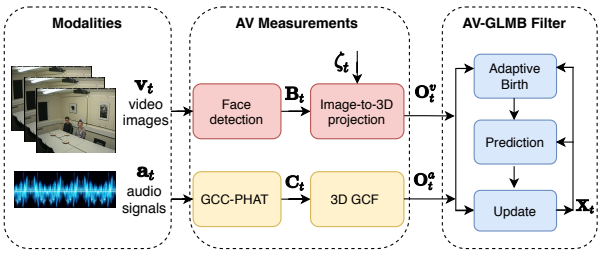

Figure 1: *Block diagram of the proposed AV-GLMB tracker. (Notations: $\mathbf{v}_t$: video images; $\mathbf{a}_t$ audio signals; $\mathbf{B}_t$: the set of face detection bounding boxes at time t ; $\mathbf{C}_t$: the set of GCC-PHAT; $\mathbf{O}_t^a$, $\mathbf{O}_t^v$: the set of audio and video measurements respectively; $\mathbf{X}_t$: the set of target 3D positional and identity state estimates.)*

⋆ The first two authors contributed equally to this work.

http://dx.doi.org/10.21437/Interspeech.2020-1969

3. Audio-visual Measurements

We first extract the AV measurements from the audio signals and video images (see Fig.1). In this paper, all audio and visual measurements are in the 3D space.

3.1. Audio Measurement

Speaker localization methods can be mainly divided into three classes: the Time Difference of Arrival (TDoA)-based method, the energy ratio-based method and the learning-based method [9]. We use Generalized Cross Correlation with Phase Transform (GCC-PHAT) to estimate TDoA due to its higher robustness to environmental noise and room reverberation [21] and the result is derived as:

$$C_t(\tau_m(\mathbf{o})) = \int_{-\infty}^{+\infty} \frac{S_{m1}(t,f)S^*_{m2}(t,f)}{|S_{m1}(t,f)|\,|S^*_{m2}(t,f)|} e^{j2\pi f \tau_m(\mathbf{o})} df \tag{1}$$

where f indicates the frequency, S_{m1} and S_{m2} are the Short-Time-Fourier-Transform (STFT) computed at the m-th pair of microphones, i.e., $m1$ and $m2$ ($m \in [1, M]$, M is the number of microphone pairs), $\tau_m(\mathbf{o})$ denotes the TDoA between microphone pair m from a sound source at a generic 3D position $\mathbf{o}$, and $*$ is the complex conjugate operator.

The location estimate comes from the peak of the 3D Global Coherence Field (GCF), i.e.,

$$\mathbf{o}^a_t = \underset{\mathbf{o}}{\operatorname{argmax}} \frac{1}{M} \sum_{m=1}^{M} C_t(\tau_m(\mathbf{o})) \tag{2}$$

3.2. Visual Measurement

A MXNet algorithm [22] is used for face detection on the image plane where the target mouth position is geometrically estimated from the detection bounding box $\mathbf{b}_{t,i}$. Since there is no explicit solution to derive the target 3D location from a monocular camera, given the sensor calibration information ζ_t, by using the pin-hole camera model [23] and assuming the target face width and height in 3D (i.e., W and H), we can derive the 3D video measurement as:

$$\mathbf{o}^v_{t,i} = \Psi(\mathbf{b}_{t,i}|\zeta_t, W, H) \tag{3}$$

where Ψ is the image-to-3D projection, and $\mathbf{o}^v_{t,i} \in \mathbf{O}^{\mathbf{v}}_{\mathbf{t}}$. More details can be found in [15].

4. AV-GLMB Tracker

As shown in Fig.1, the proposed AV-GLMB filter consists of the prediction (with the adaptive birth model) and update recursions. It is fed with the AV measurements, and produces target state estimates.

4.1. Labeled RFS

First we define the AV-GLMB random finite set (RFS) $\mathbf{X} \triangleq \{(\mathrm{x}_i, \ell_i) \mid i \in \mathbb{N}\}$ as a **labeled** RFS with state space $\mathbb{X}$ and label space $\mathbb{L}$, ($\mathrm{x}_i \in \mathbb{X}$ and $\ell_i \in \mathbb{L}$), where the labels are unique, i.e., $\ell_i \neq \ell_{i'}$, $\forall i \neq i'$. Its probability density in the δ-GLMB form is given as [17]:

$$\pi(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \omega^{(I,\xi)} \delta_I(\mathcal{L}(\mathbf{X})) \left[p^{(\xi)}\right]^{\mathbf{X}} \tag{4}$$

where $\omega^{(I,\xi)}$ is the probability of the hypothesis (I, ξ), I is a set of labels, ξ represents a history of association map between targets and measurements. $\mathcal{F}(\mathbb{L})$ denotes the class of the finite subsets of $\mathbb{L}$ and Ξ is the discrete space of the association map. $p^{(\xi)}$ is the probability distribution of a target state, The generalized delta function $\delta_I(\mathcal{L}(\mathbf{X}))$ indicates whether the set of labels in $\mathbf{X}$ matches that of I, i.e.,

$$\delta_Y(X) \triangleq \begin{cases} 1, \text{ if } X = Y \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

$\Delta(\mathbf{X}) \triangleq \delta_{|\mathbf{X}|}(|\mathcal{L}(\mathbf{X})|)$ is called the *distinct label indicator*. The δ-GLMB is completely characterized by the set of parameters $\{(\omega^{(I,\xi)}, p^{(\xi)}) : (I, \xi) \in \mathcal{F}(\mathbb{L}) \times \Xi\}$.

The AV-GLMB recursion also consists of the multi-object "update" step based on Bayes inference and the Chapman-Kolmogorov [24] "prediction" step based on the state transition models.

4.2. AV-GLMB Recursion: Update

If the current RFS prediction density is a δ-GLMB of the form (4), using the current AV measurements $Z_t \triangleq \{\{\mathbf{o}^a_t\} \times \{\mathbf{o}^v_{t,i}\}\}$, (where $\mathbf{o}^a_t, \mathbf{o}^v_{t,i}$ denote the concurrent 3D location estimates from (2) and (3), respectively), the posterior density is a δ-GLMB [18], i.e.,

$$\pi(\mathbf{X}|Z) = \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \sum_{\theta\in\Theta(I)} \omega^{(I,\xi,\theta)}(Z_t)\delta_I(\mathcal{L}(\mathbf{X})) \left[p^{(\xi,\theta)}(\cdot|Z_t)\right]^{\mathbf{X}} \tag{6}$$

where $\Theta(I)$ denotes the subset of current track-measurement association maps with domain I, and standard derivations of $\omega^{(I,\xi,\theta)}(Z_t)$ and $p^{(\xi,\theta)}(\mathrm{x}, \ell|Z_t)$ are provided in [18]. (For denotation simplicity, we drop the subscript t hereafter.)

The detection probability for a state is $p_D(\mathrm{x}, \ell) \equiv 0.95$ empirically in this paper. Here $\mathrm{g}(z_{\theta(\ell)}|\mathrm{x}, \ell)$ denotes the AV-GLMB likelihood for the measurement $z_{\theta(\ell)} \in Z$ being generated by state (x, ℓ). Standard implementations of the GLMB filter use a single modality, which we extend to the multi-modality scenario here. The proposed AV likelihood function is:

$$\mathrm{g}(z_{\theta(\ell)}|\mathrm{x}, \ell) \triangleq \mathrm{g}_a(\mathbf{o}^a|\mathrm{x}, \ell)\ \mathrm{g}_v(\mathbf{o}^v_{\theta(\ell)}|\mathrm{x}, \ell) \tag{7}$$

Here the likelihood function for the visual modality is:

$$\mathrm{g}_v(\mathbf{o}^v_{\theta(\ell)}|\mathrm{x}, \ell) \sim \mathcal{N}(\cdot; \phi_v(\mathrm{x}), \Sigma^2_v) \tag{8}$$

where $\phi_v(\mathrm{x})$ denotes a one-to-one mapping from target state space to the 3D visual measurement space in spherical coordinates originated at the camera center, and $\Sigma_v = \mathrm{diag}(2^\circ, 2^\circ, 0.4m)$ represents the measurement accuracy. The likelihood function for the audio modality is:

$$\mathrm{g}_a(\mathbf{o}^a|\mathrm{x}, \ell) = \frac{1}{M} \sum_{m=1}^{M} C_t(\tau_m(\phi_a(\mathrm{x}, \mathbf{o}^v))) \tag{9}$$

Different from (2), here $\phi_a(\mathrm{x}, \mathbf{o}^v)$ maps x to the height of $\mathbf{o}^v$.

4.3. AV-GLMB Recursion: Prediction

If the current RFS filtering density from its previous update step is a δ-GLMB of the form (4), the prediction density to the next time is a δ-GLMB given as [18]:

$$\pi_+(\mathbf{X}_+) = \Delta(\mathbf{X}_+) \sum_{(I_+,\xi)\in\mathcal{F}(\mathbb{L}_+)\times\Xi} \omega^{(I_+,\xi)}_+ \delta_{I_+}(\mathcal{L}(\mathbf{X}_+)) \left[p^{(\xi)}_+\right]^{\mathbf{X}_+} \tag{10}$$

where standard derivations of $\omega_{+}^{(I_{+},\xi)}$ and $p_{+}^{(\xi)}(\mathrm{x},\ell)$ can be found in [18].

The standard implementation of GLMB assumes known target birth distributions, which can be restrictive. Here we use an adaptive birth model as detailed in [25]. The probability density of the new-born labeled multi-Bernoulli RFS is:

$$\pi_B(\mathbf{X}_+) = \Delta(\mathbf{X}_+) w_B(\mathcal{L}(\mathbf{X}_+)) \left[p_B\right]^{\mathbf{X}_+} \tag{11}$$

where

$$w_B(I) = \prod_{i\in\mathbb{B}} \left(1 - r_B^{(i)}\right) \prod_{\ell\in I} \frac{1_{\mathbb{B}}(\ell) r_B^{(\ell)}}{1 - r_B^{(\ell)}} \tag{12}$$

Meanwhile, the new-born likelihood for each measurement $z \in Z$ can be formulated as:

$$r_U(z) = 1 - \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \sum_{\theta\in\Theta(I)} 1_{z_\theta}(z)\omega^{(I,\xi,\theta)} \tag{13}$$

where $\omega^{(I,\xi,\theta)}$ is given in (10), and the inclusion function here indicates if the measurement z has been assigned to a target by any of the updated hypotheses. It can be seen from (13) that, a measurement that has been used in all hypotheses cannot initiate a new-born target ($r_U(z) = 0$), while for measurements that have not been assigned to any of the targets, the new-born likelihood is 1.

In (12), the existence probability of the adaptive birth at the next time depends on its new-born likelihood obtained from current time, i.e.,

$$r_B(z) = \min\left(r_{B_{\max}},\ \lambda_B \frac{r_U(z)}{\sum_{\zeta\in Z} r_U(\zeta)}\right) \tag{14}$$

where λ_B is the expected number of target births at the next time, and $r_{B_{\max}} \in [0,1]$ is the maximum existence probability of a new-born target to ensure that the resulting $r_B(z)$ does not exceed 1 when λ_B is too large. Here we choose $\lambda_B = 0.3$ and $r_{B_{\max}} = 0.15$ empirically.

Since both the audio and visual measurements fall in the 3D space, and we are interested in the kinetic states, the same state transition function is used for each dimension. The surviving rate for each state is $p_S(\mathrm{x},\ell) \equiv 0.65$ in this paper. We choose the Langevin model [7, 26, 27], which is also a first-order Markov process, i.e.,

$$\mathrm{f}(\mathrm{x}_{+}^{(d)}|\mathrm{x}^{(d)},\ell) = \begin{bmatrix} 1 & t_\Delta \\ 0 & e^{-\beta_\mathrm{x}\cdot t_\Delta} \end{bmatrix} \cdot \mathrm{x}^{(d)} + w_\mathrm{x} \begin{bmatrix} 0 \\ \sqrt{1 - e^{-2\beta_\mathrm{x}\cdot t_\Delta}} \end{bmatrix} \tag{15}$$

where $\mathrm{x}^{(d)} = [\mathbf{o}_\mathrm{x}^{(d)}, \dot{\mathbf{o}}_\mathrm{x}^{(d)}]^T$, $\dot{\mathbf{o}}^{(d)}$ is the velocity of speaker at $\mathbf{o}_\mathrm{x}$, and $d = 1,2,3$ is the dimension index. $t_\Delta = 0.04$s is the time step, $w_\mathrm{x} \sim \mathcal{N}(\cdot;0,\sigma_\mathrm{x}^2)$ follows the normal distribution. Model parameters $\beta_\mathrm{x} = 0.5\mathrm{s}^{-1}$ and $\sigma_\mathrm{x} = 0.2$m/s are respectively the rate constant and the steady-state root-mean-square velocity for the random motions of speakers.

5. Numerical Results

5.1. Implementation Details

The proposed AV-GLMB tracker is tested on the AV16.3 dataset [28], which provides the audio signals captured by two 8-element circular microphone arrays (20cm diameter), the synchronized image sequences recorded by three standard RGB cameras, the sensor calibration information and the 3D target location annotations. The audio sampling frequency is 16kHz while the image frame rate is 25Hz. The recording room is of size $8.2\times3.6\times2.4\mathrm{m}^3$ with the approximate reverberation time $T_{60} = 0.5$s. The participants wander around, cross each other, move in and out the camera's FoV and mostly speak concurrently. For each experiment, we only use the first circular array and the individual cameras. The same parameter settings as in [15] are used unless otherwise specified in the paper. The Sequential Monte Carlo (SMC) implementation for the GLMB is used, with 100 particles for each track. The experimental results are averaged over 10 runs.

5.2. Performance Metrics

In this work, we use two metrics for performance evaluation of the proposed AV-GLMB, i.e., the Mean Absolute Error (MAE) and the Optimal Sub-Pattern Assignment (OSPA), defined as follows.

$$\epsilon_{\mathrm{MAE}} = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{|I_t|} \sum_{i\in I_t} ||\mathbf{o}_{t,i} - \hat{\mathbf{o}}_{t,i}||_2 \tag{16}$$

where $||\cdot||_2$ calculates the Cartesian distance between the ground truth and the target state estimate, and $|I_t|$ is the ground truth number of speakers. For a sequence of audio-visual data, the MAE gives an overall performance score. To have a closer view of the performance over time, the OSPA can be used.

The OSPA metric $\varepsilon_\rho^{(c)}$ of two finite sets $R = \{r_1,...,r_{n_R}\}$ and $S = \{s_1,,...,s_{n_S}\}$, (integers $n_R \le n_S$) is defined as follows [29].

$$\varepsilon_\rho^{(c)}(R,S) \triangleq \Big(\frac{1}{n_S}\big(\min_{\pi\in\Pi_{n_S}} \sum_{i=1}^{n_R} d^{(c)}(r_i, s_{\pi(i)})^\rho + c^\rho(n_S - n_R)\big)\Big)^{\frac{1}{\rho}} \tag{17}$$

where the order and cut-off parameters are $\rho \ge 1$ and $c > 0$ respectively, $d^{(c)}(r,s) \triangleq \min(c, \|r-s\|_2)$, and Π_{n_S} denotes the set of permutations on $\{1,2,...n_S\}$, $n_S \in \mathbb{N}$. The distance $\varepsilon_\rho^{(c)}(R,S)$ stands for a ρ-th order per-target error. If $n_R > n_S$, $\varepsilon_\rho^{(c)}(R,S) = \varepsilon_\rho^{(c)}(S,R)$. Note that when there is no cardinality error, i.e., $n_S = n_R$, the OSPA is proportional to the MAE when $\rho = 1$, i.e., $\varepsilon_1^{(c)} = \epsilon_{\mathrm{MAE}}$.

5.3. Tracking Results

Fig. 2 shows the tracking results from the proposed AV-GLMB method. It can be seen that the most of the AV measurements are close to the ground truth. Overall, the video measurements seem more accurate than the audio measurements, especially for the heights of speaker's mouth (see the bottom panel). However, video measurements still have deviations (e.g., , at the x-dimension around time frame 30) and miss-detections (e.g., , video measurements at time frames 54, 55 and 56). Moreover, it is not clear which speaker each measurement belongs to. Thus in the proposed tracking method, the audio measurements are mapped to the heights of the previous estimate in the likelihood function (see (9)). At the other two dimensions, the audio measurements can help improve the estimation accuracy, together with the video measurements (see (7)). Moreover, the proposed method produces filtered estimates of the 3D locations of speakers, with a unique label attached to each estimate (marked in different colors). Even though we do not assume prior knowledge of the number of speakers, the proposed AV-GLMB still correctly identifies two speakers.

Fig. 3 shows the resulting errors in the OSPA metric. We choose $\rho = 1$ and $c = 1$m in this paper. The top panel shows the

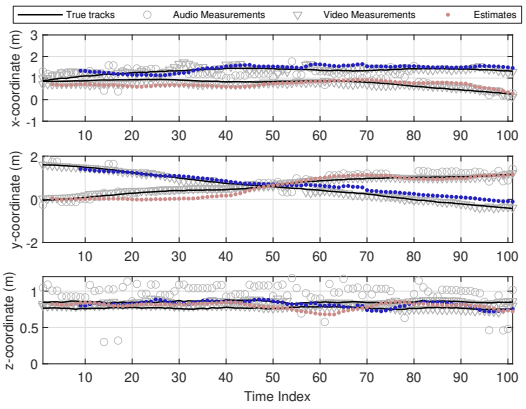

Figure 2: *Tracking results of the proposed AV-GLMB, using AV16.3 sequence 25, camera 1 and microphone array 1 data.*

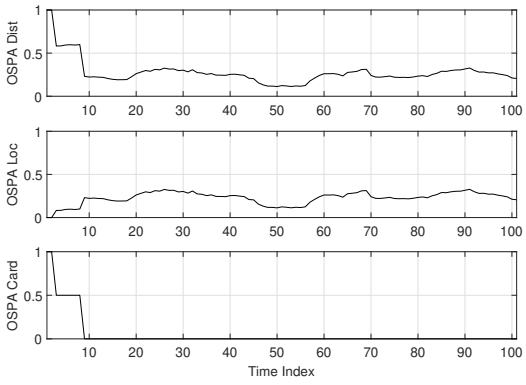

Figure 3: *OSPA errors* ($\downarrow$) *from the proposed AV-GLMB, using AV16.3 sequence 25, camera 1 and microphone array 1 data.*

overall OSPA distance considering both the estimation inaccuracy and the cardinality error, while the respective components are shown in the middle and bottom panels. Note that most of the cardinality estimates are accurate, except that at frame 0 and 1, the AV-GLMB takes two frames of measurements to initialize the first track, and a few more for the second track. For a cardinality error of 1 while the ground truth is 2, the corresponding OSPA error is $\frac{1}{2}c = 0.5$m. The overall OSPA distance is mostly less than 0.3m for the two tracks. Most of state-of-the-art (SoA) methods assume *a priori* knowledge of the number of speakers, and initialize tracks using ground truth. We do assume such *a priori* knowledge. Thus for reasonable comparison, we do not include the errors due to cardinality estimate in calculating MAE for our proposed method, and the result is equal to the averaged OSPA distance (i.e., $\bar{\varepsilon}_1^{(c)}$). Since there are two speakers in AV16.3 MOT sequences, cardinality estimates that are not equal to 2 are counted as an error. The percentage of all estimated with cardinality errors can be used to compute the track loss rate (TLR) and compared with the SoA methods.

The MAE and TLR comparisons between the proposed AV-GLMB and other SoA methods [11, 14, 16, 15], and the results are shown in Table 1. Methods marked with * use either the target number prior or the target ground truth initial locations, or both, such as [11, 14, 15]. We evaluate the performance in both the 2D image plane and the 3D space. We can see that

Table 1: *Tracking errors of the proposed AV-GLMB and the other SoA methods. The MOT sequences of the AV16.3 dataset are used (Key - $-$: information not available; methods marked with * use the target number prior or ground truth initials).*

		Image ($pixel$)			3D (m)		
seq	cam	[14]*	[16]	**prop.**	[15]*	[11]*	**prop.**
18	1	14.3	-	15.7	.13	.31	.22
	2	11.7	-	10.9	.14	.28	.26
	3	15.8	-	6.3	.39	.33	.23
19	1	11.9	-	15.3	.13	.32	.25
	2	9.6	-	11.6	.16	.27	.25
	3	12.1	-	5.4	.27	.29	.18
24	1	10.0	14.0	16.5	.12	.28	.30
	2	8.9	15.0	10.6	.55	.29	.33
	3	10.0	14.1	7.0	.11	.41	.29
25	1	14.8	15.7	17.7	.16	.31	.24
	2	7.7	13.9	10.8	.11	.21	.33
	3	8.9	17.1	10.7	.17	.21	.25
30	1	13.8	16.7	14.8	.19	.55	.26
	2	8.9	16.9	10.4	.18	.26	.38
	3	10.3	19.3	15.7	.28	.38	.28
average		11.3	15.8	**12.0**	.21	.31	**.27**
TLR (%)		-	-	-	15.8	37.7	**12.0**

the performance of the proposed AV-GLMB is comparable to other SoA methods. Explicitly, the implementation of [11] here uses known number of targets and visual observation ground truth, and both [14] and [15] use ground truth initial locations and known number of targets, which give significant advantages to their tracking accuracy but can be quite restrictive in practice. [16] uses the SMC-PHD tracker and detects speaker identity by measuring the color histogram of speakers. Its resulting averaged tracking error on the image plane is 15.8 pixels from data available in the literature. The proposed AV-GLMB does not assume *a priori* knowledge of the speaker number or the ground truth initial locations, which poses significant challenges to the estimation accuracy. However, as shown in the table, the averaged 3D tracking error of the AV-GLMB is 0.27m, which is comparable to existing SoA methods, i.e., 0.21m in [15] and 0.31m in [11]. The TLR of the other two 3D audio-visual trackers are respectively 15.8% in [15] and 37.7% in [11]. The proposed AV-GLMB has 12.0% loss rate, which is encouraging.

6. Conclusions

In this paper, we propose a novel 3D audio-visual multi-speaker tracking framework which exploits the complementarity of the audio and visual modalities. Different from existing state-of-the-art methods, the proposed AV-GLMB provides a closed-form solution for multi-speaker tracking without relying on ground truth locations to initialize, or assuming *a priori* knowledge of the number of speakers. This contributes to the first successful implementation of the GLMB filter in the joint audio-visual multi-speaker online tracking. Experimental results have demonstrated the benefits of the proposed method.

Acknowledgements: This research work is supported by the Neuromorphic Computing project, Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), and Human-Robot Interaction Phase 1 (Grant No. 192 25 00054) from the National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme.

7. References

[1] X. Qian, A. Xompero, A. Cavallaro, A. Brutti, O. Lanz, and M. Omologo, "3D mouth tracking from a compact microphone array co-located with a camera," in *IEEE Int. Conf. on Audio, Speech and Signal Processing*, Calgary, Canada, Apr 2018, pp. 3071–3075.

[2] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational Bayesian inference for audio-visual tracking of multiple speakers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Nov 2019.

[3] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proc. of the IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct 2010.

[4] D. Fox, "Adapting the sample size in particle filters through KLD-sampling," *The Int. Journal of Robotics Research*, vol. 22, no. 12, pp. 985–1003, Dec 2003.

[5] D. Y. Kim, B.-N. Vo, B.-T. Vo, and M. Jeon, "A labeled random finite set online multi-object tracker for video data," *Pattern Recognition*, vol. 90, pp. 377–389, 2019.

[6] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomput.*, vol. 74, no. 18, pp. 3823–3831, Nov 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. on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[8] S. Lin, "Reverberation-robust localization of speakers using distinct speech onsets and multichannel cross correlations," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 26, no. 11, pp. 2098–2111, 2018.

[9] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: a systematic review," *ACM Computing Surveys*, vol. 48, no. 4, p. 52, 2016.

[10] A. Jaimes and N. Sebe, "Multimodal human computer interaction: A survey," in *Int. Workshop on Human-Computer Interaction*, Las Vegas, Nevada, USA, Jul 2005, pp. 1–15.

[11] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1154–1164, Nov 2002.

[12] M. Taj and A. Cavallaro, "Audio-assisted trajectory estimation in non-overlapping multi-camera networks," in *IEEE Int. Conf. on Audio, Speech and Signal Processing*, Taipei, Taiwan, Apr 2009.

[13] E. D'Arca, N. M. Robertson, and J. Hopgood, "Person tracking via audio and video fusion," in *Data Fusion & Target Tracking Conf.: Algorithms & Applications*, London,UK, May 2012.

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

[15] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Trans. on Multimedia*, vol. 21, no. 10, pp. 2576–2588, Oct 2019.

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

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

[18] B.-N. Vo, B.-T. Vo, and D. Phung, "Labeled random finite sets and the bayes multi-target tracking filter," *IEEE Trans. on Signal Processing*, vol. 62, no. 24, pp. 6554–6567, 2014.

[19] S. Lin, "Jointly tracking and separating speech sources using multiple features and the generalized labeled multi-bernoulli framework," in *IEEE Int. Conf. on Audio, Speech and Signal Processing*, Calgary, Canada, April 2018, pp. 3211–3215.

[20] ——, "Robust pitch estimation and tracking for speakers based on subband encoding and the generalized labeled multi-bernoulli filter," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 27, no. 4, pp. 827–841, 2019.

[21] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[22] X. Wu, R. He, and Z. Sun, "A lightened CNN for deep face representation with noisy labels," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, Nov 2018.

[23] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[24] C. W. Gardiner *et al.*, *Handbook of stochastic methods*. Springer Berlin, 1985, vol. 3.

[25] S. Lin, B. T. Vo, and S. E. Nordholm, "Measurement driven birth model for the generalized labeled multi-bernoulli filter," in *2016 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2016, pp. 94–99.

[26] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE Int. Conf. on Audio, Speech and Signal Processing*, vol. 5. IEEE, 2001, pp. 3021–3024.

[27] D. B. Ward, E. Lehmann, R. C. Williamson *et al.*, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[28] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16. 3: an audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*. Martigny, Switzerland: Springer, Jun 2004, pp. 182–195.

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