3D AUDIO-VISUAL SPEAKER TRACKING WITH A TWO-LAYER PARTICLE FILTER

Hong Liu, Yidi Li, Bing Yang

Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, China
{hongliu, yidili}@pku.edu.cn, bingyang@sz.pku.edu.cn

ABSTRACT

Audio-visual speaker tracking in 3D space is a challenging problem. Although the classical particle filter based methods have shown effectiveness in audio-visual speaker tracking, the performance degrades considerably when the measurements are disturbed by noise. To this end, a novel two-layer particle filter is proposed for 3D audio-visual speaker tracking. Firstly, two groups of particles, which are generated from the audio and video streams respectively, are propagated independently in the audio layer and visual layer. Then, the audio and visual likelihoods are combined in an adaptive sigmoid function, which can adjust particle weights according to the confidence of two modalities. Finally, an optimal particle set selected from two groups of particles is proposed to determine the speaker position and reset the particle positions in the next frame. Experiments on AV16.3 database show that our method outperforms the trackers using individual modalities and the existing approaches in the 3D space and on the image plane.

Index Terms— 3D speaker tracking, audio-visual fusion, particle filter, adaptive likelihood

1. INTRODUCTION

Speaker tracking using audio-visual information has attracted extensive attentions in the past decades due to its widespread applications in the fields of intelligent surveillance, human-robot interaction and smart space [1]. The conventional visual tracker suffers from occlusions, limited view of cameras and illumination variation [2–4], and the sound source tracker is affected by background noise, room reverberations and intermittency of speeches [5–7]. Therefore, a multi-modal fusion tracking method that can fully exploit the complementarity of audio and visual information is strongly demanded.

The most popular approach for audio-visual speaker tracking is particle filter (PF), which is applicable for nonlinear state space models to fuse multi-sensor data [8]. However, audio and visual measurements that are corrupted by noise tend to cause non-negligible errors to fusion algorithms [9]. Therefore, the main task is to effectively combine the different sensor streams and appropriately weight each modality in the complex dynamic environment. The audio, shape and spatial structure observations are fused in a joint observation model represented by multiplication of respective likelihoods [10]. The direction of arrival (DOA) estimations are projected to the image plane, and the visual tracker is combined with audio data by relocating the particles around the DOA line and re-calculating their weights according to their distance to the DOA line [9]. In [11], the normalized probabilities of the visual and acoustic observations are added by an adaptive weighting factor, which is adjusted dynamically according to an acoustic confidence measure based on a generalized cross correlation with phase transform (GCC-PHAT) approach.

Furthermore, most methods focus on tracking the speaker on the image plane [9, 12–14], rather than determining their positions in the real-world reference frame, because the calculation of accurate 3D coordinates requires more complex sensor configurations, error-prone parallax calculations and triangulation. The particle swarm optimization (PSO) is utilized in a stereo vision system, where the 3D projection is obtained by the calibrated stereo camera system and triangulation [15]. The approximate target position is mapped from the image plane to 3D word coordinates by assuming the width of shoulders and camera calibration information are known [16, 17]. However, the inaccurate radius estimation is the major limitation of this kind of method.

The main steps of PF include prediction (propagation), update (measurement) and estimation. Most current formulations for audio-visual tracking fuse data at the measurement level. In this paper, a novel two-layer PF (2-LPF) is constructed to fuse audio and video information at both measurement and decision level. In 2-LPF, two PFs are employed for audio and video streams and operated independently in the audio layer and visual layer. Firstly, two groups of particles are initialized with equal weights and propagated in respective coordinates. This hierarchical structure ensures the diversity of the particles. Secondly, audio and visual likelihoods are combined using an adaptive sigmoid function to update particle weights. In this way, 2-LPF can adaptively adjust particle weights according to the confidence of two modalities. Finally, an optimal particle set is selected to estimate the 3D position of the speaker and reset the positions of the particles.

This work is supported by National Natural Science Foundation of China (NSFC No.61673030, U1613209), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No. ZDSYS201703031405467).

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on December 07,2021 at 08:53:56 UTC from IEEE Xplore. Restrictions apply.

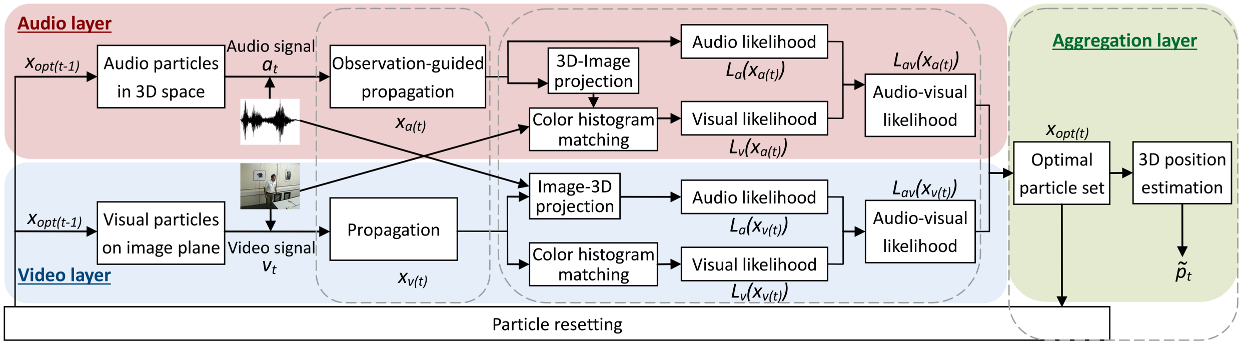

Fig. 1. Overall framework of the proposed 3D audio-visual speaker tracking method based on a two-layer particle filter.

2. PROPOSED 3D AUDIO-VISUAL TRACKER

The dynamic state space model of the proposed method consists of a state transition model $P(x_t|x_{t-1})$ and a multi-modal measurement model $P(a_t, v_t|x_t)$, where x_t is the state variable at time t, and (a_t, v_t) indicates the current observations of the audio signal and the video frame respectively. The expected posterior probability distribution $P(x_t|a_t, v_t)$ is approximated using two groups of particles, $\mathbf{x}_a$ and $\mathbf{x}_v$, corresponding to the audio and visual modalities.

Fig.1 depicts the framework of the proposed method. Audio and visual particles are propagated and weighted in audio layer and visual layer respectively, and then integrated in the aggregation layer. 2-LPF is employed to track the speaker using a three-step process of propagation (Sec.2.1), update (Sec.2.2), estimation and resetting (Sec.2.3).

2.1. Propagation of audio and visual particles

The proposed 2-LPF is equipped with two groups of particles with associated weights, which are denoted as $\{\mathbf{x}_a^i, \omega_a^i\}$, $\{\mathbf{x}_v^j, \omega_v^j\}$ with $i, j \in \{1, ..., N_s\}$ respectively, where N_s is the number of audio and visual particles. The audio particles are propagated in 3D world coordinates and modeled in spherical coordinates, while the visual particles are propagated on the image plane and modeled in rectangular coordinates. The state vectors of two groups of particles are defined as:

$$\mathbf{x}_{a(t)} = [\alpha_t, \beta_t, \gamma_t, \dot{\alpha}_t, \dot{\beta}_t, \dot{\gamma}_t]^T, \mathbf{x}_{v(t)} = [h_t, l_t, \dot{h}_t, \dot{l}_t]^T, \quad (1)$$

where α_t, β_t, γ_t are the azimuth, elevation and radius of the audio particle at time t, $\dot{\alpha}_t$, $\dot{\beta}_t$, $\dot{\gamma}_t$ are corresponding velocities, h_t and l_t are horizontal and vertical coordinates of the visual particle, and $\dot{h}_t$, $\dot{l}_t$ are corresponding velocities.

In the propagation step, two groups of particles are propagated in the audio layer and the visual layer independently in respective spaces by the dynamic model:

$$\mathbf{x}_{a(t)} = G_a \mathbf{x}_{a(t-1)} + A_{(t)} + q_{a(t)}, \quad (2)$$

$$\mathbf{x}_{v(t)} = G_v \mathbf{x}_{v(t-1)} + q_{v(t)}, \quad (3)$$

where G is a linear motion model and $q_{a(t)}$, $q_{v(t)}$ are zero-mean Gaussian-distributed noises with covariance Q_a, Q_v.

In particular, the PF in the audio layer can directly guide audio particles to source direction by current azimuth observations. In Eq.(2), $A_{(t)} = [\hat{\alpha}_t - \tilde{\alpha}_{t-1}, 0, 0, 0, 0, 0]^T$, where $\hat{\alpha}_t$ indicates the observed azimuth at time t, $\tilde{\alpha}_{t-1}$ is the azimuth of estimated speaker position at time $t-1$.

2.2. Weight update with audio-visual likelihood

1) Audio likelihood: A two-step sam-sparse-mean (SSM) approach is employed for multi-source detection and localization, since SSM appears to be superior to classical detection features such as energy or SNR [18].

In general, the elevation and radius estimations of most microphone array based sound source localization technologies are inaccurate. Therefore, the comparatively precise azimuth estimated value is used to calculate the audio likelihood. Let $\hat{\alpha}_t$ represent the current azimuth estimated by SSM approach, the audio likelihood is calculated as:

$$L_a(a_t|x_t) = exp(-\lambda_a |\hat{\alpha}_t - \alpha_t|^2), \quad (4)$$

where λ_a is a designed parameter, and α_t is the azimuth of the particle located in 3D world coordinates.

2) Visual likelihood: The color histogram matching method is used to measure the similarity between the rectangle around the particle and the reference template defined at initialization step. The HSV color model is extracted to calculate Bhattacharyya distance, since HSV is observed to be more robust to illumination variation [9]. Assuming that the visual observation noise obeys the Gaussian distribution, then the visual likelihood is calculated as:

$$L_v(v_t|x_t) = exp(-\lambda_v D_{HSV}^2), \quad (5)$$

where λ_v is a designed parameter and D_{HSV} is the Bhattacharyya distance. The scale of the rectangle is estimated by a scale space filter implemented in fDSST [19].

3) Audio-visual likelihood: From the Bayesian perspective, the tracking problem is to calculate the confidence of each particle recursively. In order to measure the reliability of particles more accurately, a particular sigmoid function is designed to combine audio likelihood L_a with visual likelihood L_v to update the particle weights:

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on December 07,2021 at 08:53:56 UTC from IEEE Xplore. Restrictions apply.

$$L_{av}(a_t, v_t|x_t) = f(L_a, L_v) = \frac{\kappa_0 exp(\eta_a L_a + \eta_v L_v)}{1 + \kappa_0 exp(\eta_a L_a + \eta_v L_v)}, \quad (6)$$

$$\omega_{a(t)} = L_{av}(\mathbf{x}_{a(t)}) = f(L_a(a_t|\mathbf{x}_{a(t)}), L_v(v_t|\mathbf{x}_{a(t)})), \quad (7)$$

$$\omega_{v(t)} = L_{av}(\mathbf{x}_{v(t)}) = f(L_a(a_t|\mathbf{x}_{v(t)}), L_v(v_t|\mathbf{x}_{v(t)})), \quad (8)$$

where κ_0 is the initial value of the function, η_a and η_v determine the intercept and slope of the sigmoid function and indicate the reliability of two modalities. They are adjusted adaptively according to confidence measures of the tracking result $\tilde{p}_{t-1}$ at previous frame:

$$\eta_{a(t)} = \varsigma_a * L_a(\tilde{p}_{t-1}) / [\frac{1}{2N_s}(\sum_{i=1}^{N_s} L_a(\mathbf{x}^i_{a(t-1)}) + \sum_{j=1}^{N_s} L_a(\mathbf{x}^j_{v(t-1)}))], \quad (9)$$

$$\eta_{v(t)} = \varsigma_v * L_v(\tilde{p}_{t-1}) / [\frac{1}{2N_s}(\sum_{i=1}^{N_s} L_v(\mathbf{x}^i_{a(t-1)}) + \sum_{j=1}^{N_s} L_v(\mathbf{x}^j_{v(t-1)}))], \quad (10)$$

where ς_a and ς_v are user-defined parameters used to balance the measurement results.

4) Calculation of likelihood: In order to calculate the Bhattacharyya distance of the audio particle in $L_v(v_t|\mathbf{x}_{a(t)})$, the particle $\mathbf{x}_a$ located in 3D world coordinates is projected to the particle $\mathbf{x}_a^{img}$ located on the image plane using a pinhole camera model:

$$[\varpi h_t, \varpi l_t, \varpi]^T = M[\alpha_t, \beta_t, \gamma_t, 1]^T, \quad (11)$$

where (h_t, l_t) is the coordinate in pixels, ϖ and M is normalization coefficient and projection matrix.

To obtain the azimuth information of the video particles, the 3D coordinates are reconstructed by the inverse process of pinhole camera model in Eq.(11) with a prior parameter of radius $\breve{\gamma}_t$,

$$\mathbf{x}_v^{3D} = [\alpha_t, \beta_t, \breve{\gamma}_t]^T, \quad (12)$$

where $\breve{\gamma}_t$ is determined by the radius of the nearest audio particle around the video particle.

In addition, to solve the tracking error caused by the missing visual observation when the speaker is outside the field of view, a designed threshold is utilized:

$$\omega^j_{v(t)} = \begin{cases} L_{av}(\mathbf{x}^j_{v(t)}), \max_j L_v(\mathbf{x}^j_{v(t)}) \geq \xi \\ 0, \qquad \max_j L_v(\mathbf{x}^j_{v(t)}) < \xi. \end{cases} \quad (13)$$

The weights of visual particles are reset to zero when the maximum of the visual likelihoods of all video particles at time t are below a threshold ξ.

2.3. Estimation and resetting using optimal particle set

In the estimation step, the optimal particle set, $\mathbf{x}_{opt}$, is constructed by comparing the fusion weights of all the particles. The N_s particles with the largest weight are selected to form the optimal particle set:

$$\mathbf{x}_{opt} = \{\mathbf{x}_a^{i_{best}}, \mathbf{x}_v^{j_{best}}\}, \quad (14)$$

where $\mathbf{x}_a^{i_{best}}$ and $\mathbf{x}_v^{j_{best}}$ are optimal audio and visual particles in the set of best N_s particles, i_{best} and j_{best} are respective indexes. $\mathbf{x}_{opt}$ is then used to estimate the position of the speaker:

$$\tilde{p_t} = \sum_{k=1}^{N_s} \omega^k_{opt(t)} \mathbf{x}^k_{opt(t)}, \quad (15)$$

where $\tilde{p_t}$ is the final estimated target position and the weights are normalized to ensure that $\sum_{k=1}^{N_s} \omega^k_{opt(t)} = 1$, and k is the index of optimal particles.

In the resetting step, the initial states of audio and visual particles at the next frame are reset as the optimal particle set and a resampling procedure is utilized to eliminate particles with relatively small weights and to prevent degeneracy phenomenon. Lastly, two groups of particles are returned to the prorogation step and continue recursively.

3. EXPERIMENTS AND DISCUSSIONS

3.1. Experiment setting

The proposed tracker is tested on the AV16.3 corpus [20], which provides the synchronized audio-visual data with continuous 3D speaker location annotations. The audio data is recorded by a 10 cm-radius, 8-microphone uniform circular array at the sampling rate of 16 kHz. The video is captured by 3 monocular cameras at 25 Hz and each frame is a colour image of 360×288 pixels. The camera calibration information provided by the database is used in our pinhole camera model for coordinate transformation. The experiments are tested on $seq08, 11, 12$ and we only use the data captured by one camera and one microphone array. The sequences contain a single speaker with some challenging poses, such as outside the field of the view, not facing the cameras or fast motion. The number of audio particles, video particles and optimal particles is set to 50. Covariance matrix Q_a, Q_v are diagonal matrixes with $\sigma_a^2 = 0.05$, $\sigma_v^2 = 5$. The number of bins used for the Hue histogram is 8. The parameters λ_a, λ_v and κ_0 in Eq.(4-6) are chosen as 15, 5 and 0.1. The threshold ξ in Eq.(13) is set to 0.15. The average mean absolute error (MAE) in 3D (m) and on the image plane (pixels) for 10 runs are considered to evaluate the precision of the trackers.

3.2. Results and analysis

The proposed 3D audio-visual tracker (AV) is compared against the PF trackers that use individual modalities only, namely audio-only (AO) and video-only (VO). Fig.2 shows the 3D tracking results for $seq11cam3$ of azimuth (rad), elevation (rad) and radius (m). With the aid of sound source localization technology, AO has a high accuracy in azimuth estimation, but lower precision of elevation and radius estimations. VO method shows better performances on the image plane, but cannot locate accurate 3D coordinates using a monocular camera. AV method can make up for the above

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on December 07,2021 at 08:53:56 UTC from IEEE Xplore. Restrictions apply.

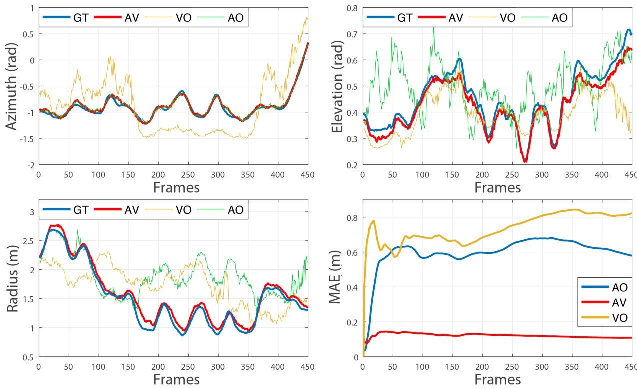

Fig. 2. 3D tracking results for AV, AO and VO at azimuth (rad), elevation (rad) and radius (m) direction on $seq11c3$ and 3D MAE (m).

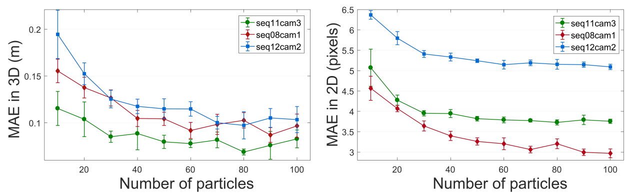

Fig. 3. MAE of proposed method in 3D (m) and on the image plane (pixels) versus number of particles.

Table 1. MAE of proposed 2-LPF using different likelihoods on three sequences. (AL: additive likelihood, ML: multiplicative likelihood, F-SL: sigmoid likelihood with fixed parameters, A-SL: sigmoid likelihood with adaptive parameters, 3D: MAE in 3D (m), 2D: MAE on the image plane (pixels).)

Method		seq11cam3		seq12cam2		seq08cam1	
Likelihood	Setup	3D	2D	3D	2D	3D	2D
AL	ϵ=0.5	0.106	5.843	0.172	9.137	0.134	4.274
ML	-	0.097	5.204	0.156	7.218	0.128	3.931
F-SL	(η_a, η_v)=(2,8)	0.094	5.031	0.135	5.813	0.121	3.835
A-SL	$(\varsigma_a, \varsigma_v)$=(2,1)	0.082	**3.684**	0.128	**5.047**	0.114	3.616
A-SL	$(\varsigma_a, \varsigma_v)$=(4,2)	**0.074**	3.859	**0.114**	5.392	**0.103**	**3.324**

Table 2. MAE of audio-visual tracking in 3D (m) and on the image plane (pixels) on $seq08$, $seq11$, and $seq12$ of AV16.3, over camera 1, 2, 3.

Sequence		3D MAE(m)			2D MAE(pixels)		
seq	cam	[16]	[17]	2-LPF	[9]	[17]	2-LPF
08	1	0.15	0.12	0.10	10.75	4.31	3.32
	2	0.24	0.11	0.08	7.33	4.66	3.06
	3	0.20	0.09	0.06	9.85	5.34	3.47
11	1	0.31	0.33	0.26	14.66	8.15	6.15
	2	0.29	0.14	0.08	14.01	7.48	5.58
	3	0.26	0.12	0.07	13.96	6.64	3.86
12	1	0.41	0.26	0.20	12.49	6.86	4.11
	2	0.51	0.17	0.11	10.81	10.67	5.39
	3	0.47	0.20	0.12	11.86	9.71	5.65
Average		0.32	0.17	**0.12**	11.75	7.09	**4.51**

deficiencies and significantly improve the tracking performances in three directions. The bottom right graph in Fig.2 shows that the 3D error of the proposed method is far less than AO and VO trackers.

Table 1 shows the MAE of proposed 2-LPF using additive likelihood (AL), multiplicative likelihood (ML), proposed sigmoid likelihood with fixed parameters (F-SL) and proposed sigmoid likelihood with adaptive parameters (A-SL). In ML, L_{av} is proportional to the product of the individual likelihoods: $L_{av} = L_a * L_v$. In AL, L_{av} is calculated by combining audio likelihood with visual likelihood using a weighting factor: $L_{av} = \epsilon L_a + (1-\epsilon)L_v$. After extensive experiments, η_a and η_v of F-SL in Eq.(6) are determined as 2 and 8. ς_a and ς_v of A-SL in Eq.(9-10) are set to $(1, 2)$ and $(2, 4)$. A-SL performs better on three sequences, because the nonlinear sigmoid function has stronger representational capacity to express particle weights and can better balance the reliability of the two modalities with adaptive parameters.

To evaluate the convergence of the proposed algorithm, experiments are performed by changing the number of particles from 10 to 100 on three sequences. It can be observed from Fig.3 that the MAE decreases as the number of particles increases, and remain stable when $N_s > 50$. A large number of particles lead to extra computation cost in the implementation of the PF, while the proposed method performs well with the limited number of particles, which effectively reduces the computation time and ensures the real-time performance of the tracker.

Table 2 depicts the comparison results in 3D (m) and on the image plane (pixels) between the proposed 2-LPF tracker and the particle filter based audio-visual trackers in [16], [17], and [9]. The average MAEs of $0.12m$ and 4.51 pixels are obtained by the proposed 2-LPF tracker, which outperform the comparison methods in all the cases. In the framework of 2-LPF, each layer can be operated as a particle filter, and more specially, information from two modalities is fused in the update and resetting step. It is clearly seen that the two-layer structure leads to an increase in tracking performance.

4. CONCLUSIONS

This paper presents a novel two-layer particle filter for 3D speaker tracking using audio-visual information. The layered structure increases particle diversity and implement feature fusion and decision fusion in the update and resetting steps. The adaptive sigmoid likelihood can better balance the reliability of two modalities with its strong representational capacity. Therefore, 2-LPF with proposed fusion likelihood outperforms contrast methods. In particular, an optimal particle set is constructed to reset particle positions, which ensures the effectiveness of the particles. The proposed tracker is tested on the single speaker sequences, including challenging situations such as outside the field of the view, not facing the cameras or fast motion. Experimental results show that the proposed method can track the speaker in 3D space as well as on the image plane with high accuracy, and outperforms the trackers using individual modalities and existing approaches.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on December 07,2021 at 08:53:56 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

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

[2] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, 2015.

[3] N. Song, K. Li, and W. Chen, "Robust visual tracking via adaptive structure-enhanced particle filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1578–1582.

[4] X. Qian, L. Han, Y. Wang, and M. Ding, "Deep learning assisted robust visual tracking with adaptive particle filtering," *Signal Processing Image Communication*, vol. 60, pp. 183–192, 2017.

[5] C. Pang, H. Liu, J. Zhang, and X. Li, "Binaural sound localization based on reverberation weighting and generalized parametric mapping," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1618–1632, 2017.

[6] H. Liu, H. Lan, B. Yang, and C. Pang, "Multiple concurrent sound source tracking based on observation-guided adaptive particle filter," in *INTERSPEECH*, 2018, pp. 826–830.

[7] Y. Ban, X. Li, X. Alameda-Pineda, L. Girin, and R. Horaud, "Accounting for room acoustics in audio-visual multi-speaker tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 6553–6557.

[8] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

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

[10] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 601–616, 2007.

[11] Kai Nickel, Tobias Gehrig, Rainer Stiefelhagen, and John McDonough, "A joint particle filter for audio-visual speaker tracking," in *International Conference on Multimodal Interfaces*, 2005, pp. 61–68.

[12] V. Kilic, M. Barnard, W. Wang, and J. Kittler, "Audio constrained particle filter based visual tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3627–3631.

[13] D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *IEEE International Conference on Image Processing*, 2003, vol. 3, pp. 25–28.

[14] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *IEEE International Conference on Computer Vision Workshops*, 2017, pp. 446–454.

[15] F. Keyrouz, U. Kirchmaier, and K. Diepold, "Three dimensional object tracking based on audiovisual fusion using particle swarm optimization," in *International Conference on Information Fusion*, 2008, pp. 1–5.

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

[17] 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 International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 3071–3075.

[18] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, vol. 3, pp. 265–268.

[19] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.

[20] J. M. Odobez G. Lathoud and D. Gatica-perez, "Av16.3: an audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*, 2004, pp. 182–195.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on December 07,2021 at 08:53:56 UTC from IEEE Xplore. Restrictions apply.