Deep Metric Learning-Assisted 3D Audio-Visual Speaker Tracking via Two-Layer Particle Filter

Yidi Li, Hong Liu, Bing Yang, Runwei Ding, and Yang Chen

1Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Shenzhen 518055, China
2School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China
3Yanka Kupala State University of Grodno, Grodno, Belarus

Correspondence should be addressed to Yidi Li; yidili@pku.edu.cn

Received 17 June 2020; Accepted 13 August 2020; Published 31 August 2020

1.Introduction

Audio-visual speaker tracking is a key technology of human-machine interaction, driven by applications such as intelligent surveillance, smart space, and multimedia systems. By analyzing the audio-visual data captured by multimodal sensor arrays, the positions of the speakers in the scene are continuously tracked, providing the underlying basis for subsequent action recognition and interaction. Compared with the conventional single-modal tracking, complementary information from audio and video streams is utilised to improve the tracking accuracy and robustness [1].

Current methods for speaker tracking are built on the probabilistic generation models due to their ability to process multimodal information. As the representative state-space approach based on Bayesian framework, Kalman filter (KF) [2], extended KF (EKF), and particle filter (PF) [3] are commonly used methods. Among them, PF can recursively approximate the filtering distribution of tracking targets by using dynamic models and random sampling. However, traditional PF assumes that the number of targets is a priori, which is not suitable for natural scene containing multiple speakers with random motion. Probability hypothesis density (PHD) filter [4] is introduced to solve the problem, which is another random method based on finite set statistics (FISST) theory. Different from the above Bayesian methods, the speaker number is estimated during the PHD-based tracking process, and therefore, the PHD filter is considered promising for multispeaker tracking. However, the PHD filter restricts the propagation of the multitarget posterior distribution to the first-order moment, resulting in the loss of high-order cardinal information, which leads to speaker number estimation errors in low signal-to-noise ratio situation [5]. PF is selected as the tracking framework in this paper since it easily approaches the Bayesian optimal estimates without being constrained by linear systems and Gaussian assumptions [6].

Many works try to improve the architecture of PF to integrate data streams from different modalities into a unified tracking framework. The direction of arrival (DOA) derived from the audio source is used to reshape the typical Gaussian noise distribution of the particles in
the propagation step of PF, and the weights of particles are recalculated according to their distance to the DOA results [7]. Tracking efficiency and accuracy usually depend on the number of particles and noise variance used in the state model and the propagation equation. Moreover, as an enhanced version of the PF, an adaptive algorithm is proposed to dynamically adjust the number of particles and noise variance by using audio-visual information [5].

The audio information obtained from the generalized cross correlation (GCC) algorithm and the video information extracted by the continuous adaptive mean shift (CAMshift) method are combined using a particle swarm optimization- (PSO-) based fusion technology [8]. The PSO algorithm can also be utilized to optimize the particle sampling in PF and improve the particle convergence to the active speaker region by incorporating an interaction mechanism [9].

To analyze and infer the dynamic system applied for speaker tracking, Bayesian theory provides an effective framework, which includes a state model and an observation model. Among them, the state model is used to describe the evolution of the state with time, and the observation model associates the observed information with the state of the speaker [10]. The prevailing fusion strategies in PF-based framework are performed by modifying the observation model to fuse the observations collected from multisensor devices [5]. Specifically, audio and visual likelihoods are constructed separately in the observation model to update the particle weights. A joint observation model is proposed in [11], which fuses audio, shape, and structure observations derived from audio and video in a multiplicative likelihood. The visual observation model [12] is derived by a face detector and reverts to a color-based generative model during misdetection. Furthermore, the visual observation is used to calculate the video-assisted global coherence field- (GCF-) based audio likelihood by limiting the acoustic map to the horizontal plane determined by the predicted height of the speaker [13]. Probabilities of the visual and acoustic observations are combined using 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 [14]. PF is combined to a pretrained convolutional neural network (CNN), which provides a generic target representation. The conventional color histogram-based appearance models cannot deal with sudden changes effectively, while a more stable observation model is presented by fusing deep features and manual features [15].

The purpose of this work is to adopt deep metric learning to optimize the observation model in the PF tracker. The method can effectively describe the similarity between samples by learning a distance metric. By designing the network structure and constructing a distance-based cost function, the similarity between the particle diffusion area and the matching template can be obtained. The likelihood function based on the network output can better define the particle weights and reflect the confidence of observations from different modalities. This work is based on a two-layer PF framework, which achieves audio-visual fusion through a hierarchical structure including an audio layer and a visual layer [16].

In the propagation step, two groups of particles from the audio and video streams are diffused through the audio-guided motion model. In the update step, the similarity between the particle diffusion area and the template is obtained through a pretrained Siamese network to calculate particle weight. In the estimation step, an optimal particle set is constructed to determine the speaker position. Finally, the target template is updated by a long short-term mechanism.

The main work of the proposed 3D audio-visual speaker tracker is discussed in following sections: Section 2 describes the methodology of the tracker in detail, including the motion model, the observation model, the ensemble method, and the template update method; Section 3 presents the experimental results and detailed analysis; finally, Section 4 concludes this work.

2. Methodology

For the state-space model in the tracking task, the recursive filtering approach is commonly used to realize the dynamic system estimation based on the Bayesian theory. In PF, the state of the speaker is estimated according to the posterior probability distribution \( p(x_t | a_t, v_t) \), which is approximated by a set of random particles with associated weights, where \( x_t \) is the state vector at time \( t \) and \( (a_t, v_t) \) indicates the current observations from the audio signals and the video frames. Assuming that the target state transition is a first-order Markov process, the required posterior probability distribution is formulated in terms of likelihood function \( L(a_t, v_t | x_t) \) and state transition model \( p(x_t | x_{t-1}) \):

\[
p(x_t | a_t, v_t) \propto L(a_t, v_t | x_t) p(x_t | x_{t-1}) p(x_{t-1} | a_{t-1}, v_{t-1}).
\]  

(1)

Sampling importance resampling (SIR) algorithm is applied to implement the recursive Bayesian filter by Monte Carlo (MC) simulations [10]. The importance density is chosen to be the prior density \( p(x_t | x_{t-1}) \). As the particles are drawn from the proposal importance density, \( x^i_t \sim p(x_t | x^i_{t-1}) \), the weights are given by \( w^i_t \propto L(a_t, v_t | x^i_t) \).

Figure 1 depicts the framework of the proposed audio-visual speaker tracker. Audio and visual particles, \( x_a \) and \( x_v \), are propagated, respectively, in audio layer and visual layer driven by the audio-guided motion model (Section 2.1). Through the pretrained Siamese network, the likelihoods are obtained to weigh the particles (Section 2.2), and the speaker position is estimated using a set of optimal particles (Section 2.3). Finally, a long short-term mechanism is applied to update templates of the target (Section 2.4).

2.1. Audio-Guided Motion Model. In audio processing, the DOAs of input audio signals are adopted as position observations to assist particle diffusion. To obtain the DOA observations, the two-step sam-sparse-mean (SSM) approach is employed for sound source localization [17]. The
The first step is to perform sector-based detection and localization. The space around the microphone array is dispersed into multiple sectors with corresponding SSIM activity value. The existence of the activity source in the sector is determined by comparing the activity value with the adaptive threshold. The second step is to conduct a point-based search in each active sector. The parametric approach is utilized for localization, which uses the cost function SRP-PHAT to optimize the spatial position parameters.

The proposed audio-visual speaker tracker is equipped with a set of audio particles \( \{ x_a^k, j = 1, \ldots, N_a \} \) and a set of visual particles \( \{ x_v^k, k = 1, \ldots, N_v \} \), where \( x_a^j \) and \( x_v^k \) are support points with associated weights \( \{ \omega_a^j, j = 1, \ldots, N_a \} \) and \( \{ \omega_v^k, k = 1, \ldots, N_v \} \) and \( N_a \) and \( N_v \) is the number of particles. The audio particle that propagated in 3D world is modeled in spherical coordinates, which is represented as the state vector:

\[
x_a(t) = \begin{bmatrix} \alpha_t, \beta_t, \gamma_t, \alpha_t', \beta_t', \gamma_t' \end{bmatrix}^T,
\]

where \((\alpha_t, \beta_t, \gamma_t)\) indicates azimuth, elevation, and radius of the audio particle at time \( t \) and \((\alpha_t', \beta_t', \gamma_t')\) indicates corresponding velocities. Unlike the above definition, the visual particle is propagated on the image plane and modeled in rectangular coordinates. The state vector of visual particle is defined as follows:

\[
x_v(t) = \begin{bmatrix} h_t, l_t, \hat{h}_t, \hat{l}_t \end{bmatrix}^T,
\]

where \( h_t \) and \( l_t \) are horizontal and vertical coordinates of the image frame and \((\hat{h}_t, \hat{l}_t)\) represents velocities in the corresponding direction.

In the propagation step, due to the inaccuracy of elevation and radius estimation, the relatively accurate azimuth estimates are used to optimize the motion model based on the Langevin process [18]. In the azimuth direction of audio particles, the Langevin motion model is expressed as

\[
\dot{\alpha}_t = \dot{\alpha}_{t-1} e^{-\lambda_a \Delta T} + \Gamma_a \dot{\alpha}_t \sqrt{1 - e^{-2\lambda_a \Delta T}},
\]

\[
\dot{\alpha}'_t = \dot{\alpha}_{t-1} + \Delta T \dot{\alpha}_t,
\]

where \( \lambda_a \) and \( \xi_a \) are designed parameters, \( \Gamma_a \) is the zero-mean Gaussian-distributed noise, and \( \Delta T \) is the time interval between two consecutive frames. In addition, the position of particles is further modified according to the DOA estimation results:

\[
\alpha_t = \alpha_t' + \rho_a (\hat{\theta}_t - \alpha_t'),
\]

where \( \rho_a \) is a correction factor and \( \hat{\theta}_t \) is the DOA (azimuth) estimated by the microphone measurements. In the elevation and radius direction, the particles are only diffused through the Langevin model without additional adjustment.

For the motion model of visual particles, additional operations related to coordinate transformation are required. A pinhole camera model is employed to project the point \((\tilde{b}_{t-1}, \tilde{c}_{t-1}, \tilde{y}_{t-1})\) located in 3D world coordinates to the image plane:

\[
[\omega \tilde{h}_t, \omega \tilde{l}_t, \omega]^T = M [\tilde{b}_t, \tilde{b}_{t-1}, \tilde{y}_{t-1}, 1]^T,
\]

where \( \tilde{b}_{t-1} \) and \( \tilde{y}_{t-1} \) are elevation and radius from the tracking result of the previous frame and \( M \) is normalization coefficient and projection matrix. \((\tilde{h}_t, \tilde{l}_t)\) is the projected point on the image plane, and its direction relative to the microphone array center can be calculated:

\[
\tilde{d}_{t}^{\text{img}} = \arctan \frac{\tilde{h}_t - h}{\tilde{l}_t - l},
\]

where \((h, l)\) indicates the coordinates of the array center. The visual particle is firstly propagated through the Langevin model, whose coordinate is denoted as \((\hat{h}_t, \hat{l}_t)\). The audio-guided motion model and coordinate transformation for visual particles are expressed as

\[
\hat{\theta}_t = \arctan \frac{\hat{h}_t - h}{\hat{l}_t - l},
\]

\[
\hat{\alpha}_t = \hat{\alpha}_{t-1} + \rho_v (\tilde{d}_{t}^{\text{img}} - \hat{\alpha}_t'),
\]

\[
[\hat{h}_t, \hat{l}_t]^T = [\cos \hat{\theta}_t, \sin \hat{\theta}_t]^T \sqrt{(\hat{h}_t - h)^2 + (\hat{l}_t - l)^2} + [h, l]^T,
\]

where \((\hat{h}_t, \hat{l}_t)\) is the modified particle coordinate. The DOA information is projected onto the image plane through the pinhole camera model. In this way, the particles converge to the sound source direction by moving toward the projection point.
2.2. Deep Metric Learning-Assisted Observation Model.

The observation model is constructed to measure the candidate samples determined by the particles. Deep metric learning method using Siamese network provides a solution for similarity metric tasks, which convert the tracking problem into a similarity problem in the feature space of the known target and search area [19]. The framework of the adopted Siamese network is shown in Figure 2. It is equipped with two subnetworks with the same network structure and identical shared parameters. Each branch consists of three convolutional layers with 15 kernels, followed by a $2 \times 2$ MaxPooling layer. The filter size is $5 \times 5$, $3 \times 3$, $3 \times 3$. The input of the network is an image pair $(I^p, I^t)$ with a label $Y \in \{0, 1\}$, where $Y$ indicates whether the image pair represents the same speaker. Each image is fed to a branch of CNN. Through the network, the feature mapping function, $G_W(I)$ with parameter $W$, is trained to map the input image pair to the target feature space. In this space, a distance-based metric function, $L_{total}(W) = \sum_{n=1}^{N} \mathcal{L}(W, (I^p, I^t, Y)^n)$, is used to measure the similarity of two images. The loss function proposed in [19] is adopted, which is defined as follows:

$$
\mathcal{L}(W, (I^p, I^t, Y)^n) = (1 - Y) \mathcal{L}_{G}(E_W^n) + Y \mathcal{L}_{F}(E_W^n)
$$

$$
= (1 - Y) \frac{2}{Q} (E_W^n)^2 + Y \cdot 2Qe^{-(2/Q)E_W^n},
$$

where $\mathcal{L}(W, (I^p, I^t, Y)^n)$ with parameter $(I^p, I^t, Y)^n$ is trained to map the input image pair to the target feature space.

$$
\mathcal{L}(W, (I^p, I^t, Y)^n) = (1 - Y) \mathcal{L}_{G}(E_W^n) + Y \mathcal{L}_{F}(E_W^n)
$$

$$
= (1 - Y) \frac{2}{Q} (E_W^n)^2 + Y \cdot 2Qe^{-(2/Q)E_W^n},
$$

where $E_W^n = E_W^n(I^p, I^t)^n$, $n$ is the index of image pairs, $N$ is the number of training sample pairs, and $\mathcal{L}_{G}$ and $\mathcal{L}_{F}$ indicate the loss of positive and negative image pairs. Constant $Q$ is set to the upper bound of $E_W^n$.

The rectangular boxes around the particles are cropped as candidate samples. The audio particles in the 3D world coordinates are projected to the image plane to obtain their rectangular boxes. Candidate samples are fed into the network, and the outputs indicate the similarity of the sample to the template, which is used to calculate the likelihood:

$$
L(n \mid x_i) = \exp\left(\frac{-\kappa_v E_W(n, x_i)^2}{2}\right),
$$

where $\kappa_v$ is a designed parameter, $E_W(n, x_i)$ is the normalized output, and $x_i$ is the bounding box of the particle. Observation noise is assumed to follow a Gaussian distribution. To prevent tracking failures due to occlusion, deformation, and speaker walking out of the camera view, audio likelihood is added to modify the particle weights. Using the current DOA estimation result, the audio likelihood is defined as follows:

$$
L_a(a_i \mid x_i) = \exp\left(\frac{-\kappa_a |\hat{\theta}_i - \theta_i|^2}{2}\right),
$$

where $\kappa_a$ is a designed parameter. Reverse process of the pinhole camera model is used to reconstruct the 3D coordinates of the visual particle to obtain its azimuth. This process requires a prior parameter that is derived from the radius of the audio particle closest to the visual particle. The weight of particles is calculated as follows:

$$
\omega_i = \left\{ \begin{array}{ll}
L(n \mid x_i), \\
\rho_w L(n \mid x_i) + (1 - \rho_w)L_a(a_i \mid x_i),
\end{array} \right.
$$

where $\rho_w$ is a user-defined threshold. When the likelihoods of all particles are less than the threshold, it indicates that visual observations are unreliable; therefore, audio observations are added to improve the particle weights.

2.3. Ensemble Method with Optimal Particle Set.

Ensemble method is used to integrate the decisions of audio and visual particles to estimate the position. By comparing the weights of all particles, the $N_s$ particles with the largest weight are defined as the optimal particles and used to form the optimal particle set:

$$
x_{opt} = \left\{ x_{a, opt}, x_{v, opt} \right\},
$$

where $x_{a, opt}$ and $x_{v, opt}$ represent the set of optimal audio particles and visual particles, respectively. The speaker position is estimated by the optimal particle set:

$$
\text{max}_{j} L(x_{a, opt}^j) \geq \xi_1 \text{ or } \text{max}_{k} L(x_{v, opt}^k) \geq \xi_1,
$$

where $a_{opt}$ denotes the normalized weight of the optimal particle. Finally, the optimal particle set is utilised to reset the audio and visual particles at the next frame, which ensures the effectiveness and diversity of the particles.

2.4. Template Update with Long Short-Term Mechanism.

The traditional method performs speaker tracking according to the template provided in the first frame, which is not updated in subsequent frames to avoid contamination of target features. However, in real scenes, nonrigid targets such as speakers have various deformations, thus showing large differences in appearance. Therefore, template update method is used to adapt to changes in speaker appearance and prevent drift problem in tracking.
An indicator is proposed to measure the tracking confidence and selectively update the template. The confidence of the tracking result $\hat{p}_t$ is defined as follows:

$$R(\hat{p}_t) = \frac{L\left(\hat{p}_t\right) - L_{\min}}{(1/2N_s)\left(\sum_{j=1}^{N_s}L(x_{a(t)}^j) + \sum_{k=1}^{N_s}L(x_{v(t)}^k)\right)},$$  

(15)

where $L_{\min}$ is the smallest likelihood among all particles and $L(\hat{p}_t)$ is the likelihood of $\hat{p}_t$, which is calculated by equation (12). Threshold $\xi_2$ is set. When $R(\hat{p}_t) \geq \xi_2$, the tracking success rate is considered high enough to update the template, and the area around $\hat{p}_t$ is cropped as a short-term template $\hat{I}_p$. In addition, the past appearance of the speaker is essential for tracking. It is inevitable that noise will be merged into template through successive updating. Therefore, the target image defined in the first frame by user is continuously adopted as a long-term template $I_p^0$. The samples are matched with $I_p^0$ and $I_p$, respectively, and the similarities are measured by the Siamese network. The modified likelihood is defined as follows:

$$E_W(x'_t) = \rho_aE_W(I_p^0, x'_t) + (1 - \rho_a)E_W(I_p, x'_t),$$  

(16)

$$L(v_t | x'_t) = \exp(-\kappa_vE_W(x'_t)^2),$$  

(17)

where $\rho_a$ is a designed weighting factor, by which the long-term template and the short-term template are combined. Equation (17) is substituted into equation (12) to calculate the particle weight.

### 3. Experiments and Discussion

#### 3.1. Experimental Setting

The proposed tracker is evaluated on the AV16.3 corpus [20], which is a commonly used dataset for audio-visual speaker tracking captured by spatially distributed audio-visual sensors. The corpus is collected in a conference room with three cameras on the wall and two microphone arrays on the table. Audio signals are recorded at 16 kHz using two 10 cm radius, 8-microphone uniform circular arrays. Video sequences are recorded at 25 Hz captured by 3 monocular cameras. Each frame is a color image of 360×288 pixels. The camera calibration information provided by the dataset is used for coordinate conversion in the pinhole camera model. The trackers are evaluated in single speaker case with three camera views on seq08, 11, and 12, where the speaker is moving and speaking at the same time with some challenging poses, such as outside of the camera view, not facing the cameras or fast motion. In each experiment, we use data streams from a camera and a microphone array.

The number of audio particles, visual particles, and optimal particles is set to 50, respectively. The audio-guided motion model is set with $(\lambda_v, \zeta_v) = (5, 0.25)$ and $(\lambda_a, \zeta_a) = (4, 5)$, which depends on the particle velocity at different coordinates and $(\rho_a, \rho_v) = (0.3, 0.25)$. Parameters of likelihood functions are set as $(\kappa_v, \kappa_a) = (15, 5)$. Thresholds and weighting factors are set as $(\xi_1, \xi_2) = (0.45, 1.4)$ and $(\rho_a, \rho_v) = (0.5, 0.32)$. The Siamese network is trained with image pairs extracted from videos with continuous speaker location annotations. Images within radius $r$ of the annotation center are considered as positive samples and otherwise as negative samples. Two positive samples in the interval $F$ frames are randomly matched into positive image pairs with label $Y = 0$. Random matches of positive and negative samples are denoted as negative sample pairs with label $Y = 1$. Randomly chosen coordinates of azimuth (rad), elevation (rad), and radius (m) on seq11c2. The AO method uses the sound source localization result as the observation value in the PF algorithm. It performs well in azimuth localization, but it is difficult to accurately track in the other two directions. VO tracker achieves effective tracking on the image plane. Assuming that the speaker height is 1.7 m, and the tracking results on the image plane can be projected on board.
into the 3D space. The results show that, in the case of using a monocular camera, insufficient depth information makes it difficult to perform 3D tracking. The tracking results obtained using the AV method show that tracking in the azimuth direction is superior to the other two directions, which is due to the guidance of the azimuth estimation in

Figure 3: (a–c) 3D tracking results of VO, AO, and AV trackers on azimuth (rad), elevation (rad), and radius (m) on seq11c2, where green line represents ground truth. (d) MAE in 3D (m). (e, f) 3D trajectories on seq08c1 and seq11c2, where green lines indicate ground truth trajectories.

Figure 4: (a) MAE on image plane (pixels). (b) Tracking results on frame samples from seq11c1 produced by comparison method (green) and our method (red).
the proposed motion model. The errors in elevation and radius are mainly caused by the speaker walking out of the camera field of view and large movements. Figure 3(d) shows the 3D MAE of above three methods, reflecting the superiority of the tracking performance with audio-visual fusion.

The effectiveness of the proposed deep metric learning-assisted observation model is evaluated by comparing with a two-layer particle filter (2-LPF), which uses the color histogram matching method to measure the similarity between the rectangle around the particle and the reference template. The HSV color model is extracted to calculate the Bhattacharyya distance, thereby defining likelihood and particle weights in the same form as equation (10). Partial 3D trajectories of two trackers on seq08c1 and seq11c2 are shown in Figures 3(e) and 3(f), where the tracking results of the proposed method (yellow) are closer to the ground truth trajectories (green). Compared with traditional color features, the features extracted by the network are more distinguishable. Siamese network can better handle the similarity measurement of input image pairs, based on which more accurate particle weights are applied. The trajectory errors shown in figures are mainly due to the large error in the image 3D reconstruction process when the speaker, microphone, and camera are in the same vertical plane.

In order to investigate the effect of the proposed template update method on tracking performance, another comparative experiment is conducted. In the comparison method, the fixed template is the target image defined in the first frame without updating during the tracking process. Figure 4(a) shows the MAE on image plane (pixels) of two methods on seq11c1. The first peak of the error curve is caused by the speaker leaving the screen for a while and then reentering the scene. After frame 460, the target scale and appearance changes obviously due to the speaker moving close to the camera. As shown in the frame samples in Figure 4(b), the tracker equipped with template update mechanism achieves stable tracking (red), while the rectangular box produced by the comparison method (green) deviates from the target.

Finally, the tracking accuracy of the proposed tracker and the existing audio-visual trackers [5, 12, 21] are tested on nine single-person sequences captured by three cameras. Table 1 lists the MAE on the image plane (pixels) and 3D (m), respectively. The proposed tracker achieves outstanding performance both on image plane and in 3D space.

### Table 1: MAE in 3D (m) and on the image plane (pixels) on seq08, seq11, and seq12 over cameras 1, 2, and 3.

| Sequence | Camera | MAE (m) | MAE (pixels) |
|----------|--------|---------|--------------|
| Seq 08    | [21]   | 0.15    | 0.12         |
|           | [12]   | 0.08    | 0.08         |
|           | Ours   | 10.75   | 4.31         |
|           | [5]    | 3.24    |              |
|           | [12]   | 0.17    |              |
| Seq 11    | 2      | 0.24    | 0.11         |
|           | 0.07   | 3.33    | 4.66         |
|           | 0.07   | 9.85    | 5.34         |
|           | 0.34   | 3.28    |              |
|           | 0.23   | 14.66   | 8.15         |
|           | 0.13   | 6.04    |              |
|           | 0.32   | 0.11    |              |
|           | 0.33   | 14.01   | 7.48         |
|           | 0.07   | 5.13    |              |
|           | 0.12   | 13.96   | 6.64         |
|           | 0.18   | 4.06    |              |
|           | 0.41   | 12.49   | 6.86         |
|           | 0.18   | 4.15    |              |
|           | 0.51   | 10.81   | 10.67        |
|           | 0.10   | 5.19    |              |
|           | 0.47   | 10.96   | 7.14         |
|           | 0.20   | 5.58    |              |
| Average   | 0.32   | 0.17    | 0.11         |
|           | 11.75  | 7.09    | 4.42         |

### 4. Conclusions

This paper presents a deep metric learning-assisted 3D audio-visual speaker tracker, which integrates a designed Siamese network into the two-layer PF framework. In the proposed observation model, the similarity measures of the template and the particle diffusion areas are calculated to update weights of audio particles and visual particles. The template that adapts to the speaker appearance change is obtained through a long short-term update mechanism, which prevents drift during tracking. Audio information and video information are fused through the audio-guided motion model, conditional weighting formula, and the optimal particle set. The proposed algorithm is evaluated on single speaker sequences and achieves substantial performance improvement compared to the trackers using individual modalities and the comparison of audio-visual methods. Future work will focus on acoustic feature extraction and multimodal confidence evaluation methods.

### Data Availability

The dataset AV16.3 corpus used in this study is an open source dataset provided by the Idiap Research Institute. The dataset is publicly available on the website https://www.idiap.ch/dataset/av16-3/. We cited this dataset in Section 3.1 of the article with a corresponding reference in the reference list as [20].

### Conflicts of Interest

The authors declare that there are no conflicts of interest.

### Authors’ Contributions

Hong Liu and Yidi Li conceived and designed the study. Yidi Li and Bing Yang developed the theory and performed the experiments. Runwei Ding and Yang Chen verified the analytical methods and supervised the findings of this work. Yidi Li wrote the manuscript with support from Bing Yang and Yang Chen. Hong Liu and Runwei Ding reviewed and edited the manuscript. All authors discussed the results and contributed to the final manuscript.

### Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61673030 and U1613209) and National Natural Science Foundation of Shenzhen (No. JCYJ20190808182209321).

### References

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

[2] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, “Kalman filters for audio-video source localization,” in *Proceedings of the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 118–121, New Palz, NY, USA, October 2005.

[3] M. Heuer, A. Al-Hamadi, B. Michaelis, and A. Wendemuth, “Multi-modal fusion with particle filter for speaker localization and tracking,” in *Proceedings of the 2011 International Conference on Multimedia Technology*, pp. 6450–6453, Hangzhou, China, July 2011.

[4] Y. Liu, V. Kilic, J. Guan, and W. Wang, “Audio-visual particle flow SMC-PHD filtering for multi-speaker tracking,” *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 934–948, 2020.

[5] 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.

[6] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[7] V. Kilic, M. Barnard, W. Wang, and J. Kittler, “Audio constrained particle filter based visual tracking,” in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3627–3631, Vancouver, Canada, May 2013.

[8] F. Keyrouz, U. Kirchmaier, and K. Diepold, “Three dimensional object tracking based on audiovisual fusion using particle swarm optimization,” in *Proceedings of the 2008 International Conference on Information Fusion*, pp. 1–5, Cologne, Germany, June 2008.

[9] K. Wu, V. G. Reju, A. W. H. Khong, and S. T. Goh, “Swarm intelligence based particle filter for alternating talker localization and tracking using microphone arrays,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1384–1397, 2017.

[10] 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.

[11] D. Gatica-Perez, G. Lathoud, J.-M. 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.

[12] 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 *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3071–3075, Calgary, Canada, April 2018.

[13] 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, no. 10, pp. 2576–2588, 2019.

[14] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, “A joint particle filter for audio-visual speaker tracking,” in *Proceedings of the 2005 International Conference on Multimodal Interfaces*, pp. 61–68, Trento, Italy, October 2005.

[15] 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.

[16] H. Liu, Y. Li, and B. Yang, “3D audio-visual speaker tracking with a two-layer particle filter,” in *Proceedings of the 2019 IEEE International Conference on Image Processing*, pp. 1955–1959, Taipei, Taiwan, September 2019.

[17] G. Lathoud and M. Magimai-Doss, “A sector-based, frequency-domain approach to detection and localization of multiple speakers,” in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 265–268, Philadelphia, PA, USA, March 2005.

[18] H. Liu, H. Lan, B. Yang, and C. Pang, “Multiple concurrent sound source tracking based on observation-guided adaptive particle filter,” in *Proceedings of the 2018 Annual Conference of the International Speech Communication Association*, pp. 826–830, Hyderabad, India, September 2018.

[19] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 539–546, San Diego, CA, USA, June 2005.

[20] 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*, pp. 182–195, Springer, Berlin, Germany, 2004.

[21] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, “3D audiovisual speaker tracking with an adaptive particle filter,” in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2896–2900, New Orleans, LA, USA, March 2017.