Multi-Modal Perception Attention Network with Self-Supervised Learning for Audio-Visual Speaker Tracking

Yidi Li[1], **Hong Liu**[1]*, **Hao Tang**[2]

[1]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, China
[2]Computer Vision Lab, ETH Zurich, Switzerland
{yidili, hongliu}@pku.edu.cn, hao.tang@vision.ee.ethz.ch

Abstract

Multi-modal fusion is proven to be an effective method to improve the accuracy and robustness of speaker tracking, especially in complex scenarios. However, how to combine the heterogeneous information and exploit the complementarity of multi-modal signals remains a challenging issue. In this paper, we propose a novel Multi-modal Perception Tracker (MPT) for speaker tracking using both audio and visual modalities. Specifically, a novel acoustic map based on spatial-temporal Global Coherence Field (stGCF) is first constructed for heterogeneous signal fusion, which employs a camera model to map audio cues to the localization space consistent with the visual cues. Then a multi-modal perception attention network is introduced to derive the perception weights that measure the reliability and effectiveness of intermittent audio and video streams disturbed by noise. Moreover, a unique cross-modal self-supervised learning method is presented to model the confidence of audio and visual observations by leveraging the complementarity and consistency between different modalities. Experimental results show that the proposed MPT achieves 98.6% and 78.3% tracking accuracy on the standard and occluded datasets, respectively, which demonstrates its robustness under adverse conditions and outperforms the current state-of-the-art methods.

Introduction

Speaker tracking is the foundation task for intelligent systems to implement behavior analysis and human-computer interaction. To enhance the accuracy of the tracker, multi-modal sensors are utilized to capture richer information (Kılıç and Wang 2017). Among them, auditory and visual sensors have received extensive attention from researchers as the main senses for human to understand the surrounding environment and interact with others. Similar to the process of human multi-modal perception, the advantage of integrating auditory and visual information is that they can provide necessary supplementary cues (Xuan et al. 2020). Compared with the single-modal case, the utilizing of the complementarity of audio-visual signals contributes to improving tracking accuracy and robustness, particularly when dealing with complicated situations such as target occlusion, limited view of cameras, illumination changes, and room reverberation (Katsaggelos, Bahaadini, and Molina 2015). Furthermore, multi-modal fusion shows distinct advantages when the information of one modality is missing, or neither modality is able to provide a reliable observation. As a result, it is critical to develop a multi-modal tracking method that is capable of fusing heterogeneous signals and dealing with intermittent noisy audio-visual data.

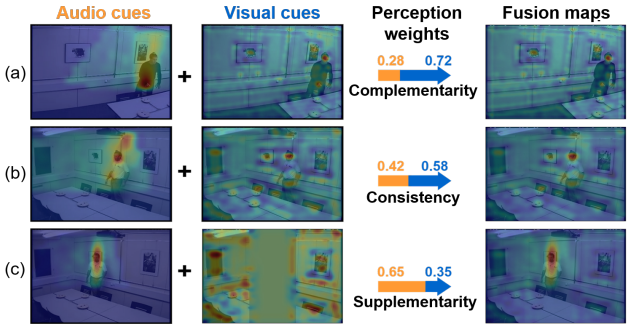

Figure 1: Keyframes of the working process of the proposed multi-modal perception attention network. (a)-(c) demonstrate the exploration of complementarity, consistency, and supplementarity between audio-visual signals, respectively.

Current speaker tracking methods are generally based on probabilistic generation models due to their ability to process multi-modal information. The representative method is Particle Filter (PF), which can recursively approximate the filtering distribution of tracking targets in nonlinear and non-Gaussian systems. Based on PF implementation, the Direction of Arrival (DOA) angle of the audio source is projected onto the image plane to reshape the typical Gaussian noise distribution of particles and increase the weights of particles near DOA line (Kılıç et al. 2015). A two-layered PF is proposed to implement feature fusion and decision fusion of audio-visual sources through the hierarchical structure (Liu, Li, and Yang 2019). Moreover, a face detector is employed to geometrically estimate the 3D position of the target to assist in the calculation of the acoustic map (Qian et al. 2021). However, these methods prefer to use the detection results of the single modality to assist the other modality to obtain more accurate observations, while neglecting to fully utilize the complementarity and redundancy of audio-visual information. In addition, most of the existing audio-visual trackers use generation algorithms (Ban et al. 2019; Schymura

*Corresponding Author: hongliu@pku.edu.cn
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

and Kolossa 2020; Qian et al. 2017), which are difficult to adapt to random and diverse changes of target appearance. Furthermore, the likelihood calculation based on the color histogram or Euclidean distance is susceptible to interference from observation noise, which limits the performance of the fusion likelihood.

To solve those limitations, we propose to adopt an attention mechanism to measure the confidence of multiple modalities, which determines the effectiveness of the fusion algorithm. The proposed idea is inspired by the human brain's perception mechanism for multi-modal sensory information, which integrates the data and optimizes the decision-making through two key steps: estimating the reliability of various sources and weighting the evidences based on the reliability (Zhang et al. 2016). Take the intuitive experience as an example: when determining a speaker's position in a noisy and bright environment, we mainly use eyes; conversely, in a quiet and dim situation, we rely on sounds. Based on this phenomenon, we propose a multi-modal perception attention network to simulate the human perception system that is capable of selectively capturing valuable event information from multiple modalities. Figure 1 depicts the working process of the proposed network, in which the first two rows show the complementarity and consistency of audio and video modalities. In the third row, the image frame is obscured by an artificial mask to show the supplementary effect of the auditory modality when the visual modality is unreliable. Different from existing end-to-end models, the specialized network focuses on perceiving the reliability of observations from different modalities. However, the perception process is usually abstract, making it difficult to manually label quantitative tags. Due to the natural correspondence between sound and vision, necessary supervision is provided for audio-visual learning (Hu et al. 2020) (Afouras et al. 2021). Therefore, we design a cross-modal self-supervised learning method, which exploits the complementarity and consistency of multi-modal data to generate weight labels of perception.

Neural networks have been widely used in multi-modal fusion tasks, represented by Audio-Visual Speech Recognition (AVSR) (Baltrušaitis, Ahuja, and Morency 2018). However, except for preprocessing works such as target detection and feature extraction, neural network is rarely introduced to multi-modal tracking. This is because the positive samples in tracking task are simply random targets in the initial frame, resulting in a shortage of data to train a high-performing classifier. Therefore, using an attention network specifically to train the middle perception component provides a completely new approach to this problem. Another reason is that the heterogeneity of audio and video data makes it difficult to accomplish unity in the early stage of the network. Therefore, we propose the spatial-temporal Global Coherence Field (stGCF) map, which maps the audio cues to the image feature space through the projection operator of a camera model. To generate a fusion map, the integrated audio-visual cues are weighted by the perception weights estimated by the network. Finally, a PF-based tracker improved with the fusion map is employed to ensure smooth tracking of multi-modal observations.

All these components make up our Multi-modal Perception Tracker (MPT), and experimental results demonstrate that the proposed MPT achieves significantly better results than the current state-of-the-art methods.

In summary, the contributions of this paper are as follows:

- A novel tracking architecture, termed Multi-modal Perception Tracker (MPT), is proposed for the challenging audio-visual speaker tracking task. Moreover, we propose a new multi-modal perception attention network for the first time to estimate the confidence and availability of observations from multi-modal data.
- A novel acoustic map, termed stGCF map, is proposed, which utilizes a camera model to establish a mapping relationship between audio and visual localization space. Benefiting from the complementarity and consistency of audio-visual modalities, a new cross-modal self-supervised learning method is further introduced.
- Experimental results on the standard and occluded datasets demonstrate the superiority and robustness of the proposed methods, especially under noisy conditions.

Related Works

Sound Source Localization. As the preprocessing module of many applications, Sound Source Localization (SSL) has been extensively studied. Traditional microphone array-based acoustic sound source localization methods are based on Time Difference of Arrival (TDOA) (Cobos et al. 2020), steered beamforming (Chiariotti, Martarelli, and Castellini 2019), and high resolution spectral estimation (Yang et al. 2019). Among these, the Generalized Cross-Correlation (GCC) algorithm is a commonly used TDOA estimation method, which describes the similarity between signals received at two sensors. Its reduced computational intensity leads to shorter decision time and higher tracking efficiency. With the development of multi-modal technology, audio-visual learning is introduced to the SSL task. Aiming at the problem of sound source localization in visual scenes, a two-stream network structure with attention mechanism is designed (Senocak et al. 2019). The audio-visual category distribution matching method is developed to assist the selection and localization of the sounding object (Hu et al. 2020). Joint Deep Neural Networks (DNN) are proposed based on a probabilistic spatial audio model, including a visual DNN to localize candidate sound sources and an audio DNN to verify the localization of candidates (Masuyama et al. 2020). We improve the Global Coherence Field (GCF) method to extract audio features with both spatial and temporal cues under the guidance of visual information.

Audio-Visual Tracking. Commonly used methods are state-space approaches based on the Bayesian framework. Many works improve the PF architecture to integrate data streams from different modalities into a unified tracking framework. Among them, multi-modal observations are fused in a joint observation model, which is represented by improved likelihoods (Qian et al. 2019; Kılıç et al. 2015; Brutti and Lanz 2010). The tracking framework based on Extended Kalman Filter (EKF) realizes the fusion of an arbitrary number of multi-modal observations through dynamic

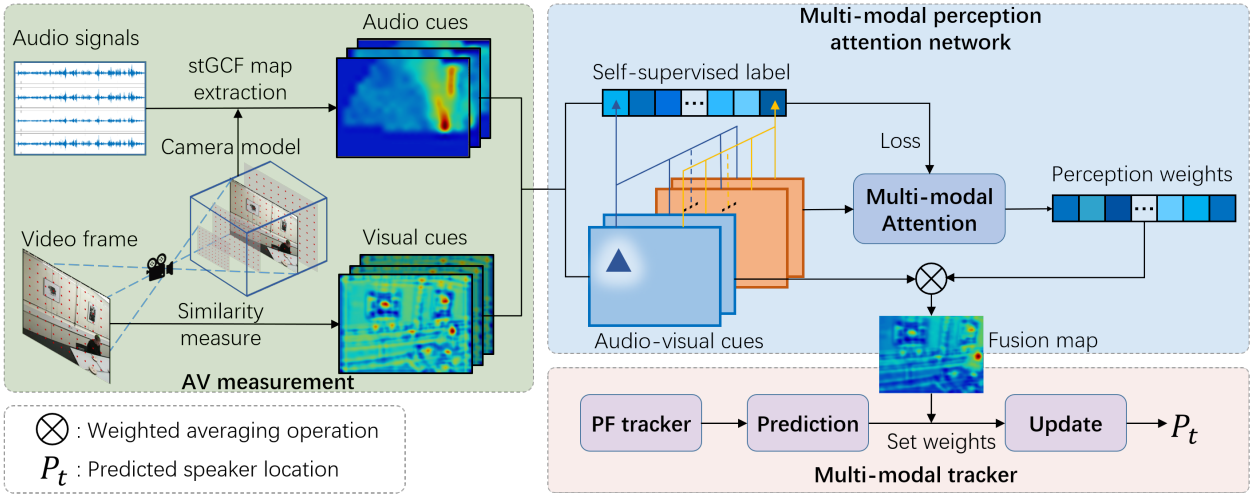

Figure 2: The framework of the proposed tracking architecture. The stGCF-based audio cues are mapped to the localization space consistent with the visual cues. The integrated audio-visual cues combined with perception weights evaluated by the multi-modal perception attention network generate a fusion map that guides update step of the PF-based multi-modal tracker.

weight flow (Schymura and Kolossa 2020). Probability Hypothesis Density (PHD) filter is introduced for tracking an unknown and variable number of speakers with the theory of Random Finite Sets (RFSs). The analytical solution is derived by introducing a Sequential Monte Carlo (SMC) implementation (Liu et al. 2019). By analyzing the task as a generative audio-visual association model formulated as a latent-variable temporal graphical model, a variational inference model is proposed to approximate the joint distribution (Ban et al. 2019). An end-to-end trained audio-visual object tracking network based on Single Shot Multibox Detector (SSD) is proposed, where visual and audio inputs are fused by an add merge layer (Wilson and Lin 2020). Deep learning methods are less utilized in the audio-visual tracking task, leading to further research prospects.

Attention-Based Models. Recently, the attention mechanism has been widely used in several tasks (Duan et al. 2021b; Tang et al. 2021; Yang et al. 2021; Liu et al. 2021; Duan et al. 2021a; Tang et al. 2019; Xu et al. 2018). In visual object tracking, the Siamese network-based tracker is further developed by designing various attention mechanisms (Wang et al. 2018; Yu et al. 2020). Based on the MDNet architecture, two modules of spatial attention and channel attention are employed to increase the discriminative properties of tracking (Zeng, Wang, and Lu 2019). In audio-visual analysis, a cross-modal attention framework for exploring the potential hidden correlations of same-modal and cross-modal signals is proposed for audio-visual event localization (Xuan et al. 2020). For video emotion recognition, (Zhao et al. 2020) integrates spatial, channel and temporal attention into visual CNN, and temporal attention into audio CNN. In audio-visual speech separation, the attention mechanism is used to help the model measure the differences and similarities between the visual representations of different speakers (Li and Qian 2020). To the best of our knowledge, attention has not been studied on the audio-visual speaker tracking task. In this paper, a self-supervised multi-modal perception attention network is introduced to investigate the perceptive ability of different modalities on the tracking scene.

Proposed Method

In this work, we propose a novel tracking architecture with a multi-modal perception attention network for audio-visual speaker tracking. Figure 2 shows the overall framework of the proposed MPT, which consists of four main modules: audio-visual (AV) measurements, multi-modal perception attention network, cross-modal self-supervised learning, and PF-based multi-modal tracker.

Audio-Visual Measurements

Through audio-visual measurements, the corresponding cues are extracted from audio signals and video frames. To integrate multi-modal cues in the same state space, we map the audio cues to the same localization plane as visual cues.

Audio Measurement. The acoustic map that highlights the source positions can be accomplished through a coherence measure based on cross-power spectrum phase, such as the Generalized Cross Correlation with Phase Transfrom (GCC-PHAT) (Omologo and Svaizer 1997). On this basis, we introduce the stGCF method to extract the audio cues. Define $\mathbf{r}_{ik}^{PHAT}(t,\tau)$ as the GCC-PHAT derived from microphone pair (i,k). It shows a prominent peak where the delay τ is equal to the actual TDOA. Let $\tau_{ik}(p)$ denote the theoretical time delay of a generic point p relative to the microphone pair (i,k). For the set Ω of M microphone pairs, the GCF value is defined as the average of the GCC-PHAT for each microphone pair belonging to Ω:

$$\mathbf{r}_{\Omega}^{GCF}(t,p) = \frac{1}{M}\sum_{(i,k)\in\Omega}\mathbf{r}_{ik}^{PHAT}(t,\tau_{ik}(p)). \quad (1)$$

Given a spatial grid with potential sound source positions, the GCF value represents the probability of the existence of a sound source at each position. To construct the spatial grid,

a pinhole camera model is utilized to project the 2D points on the image plane into a series of 3D points with different depths in 3D world coordinates, where the depth refers to the vertical distance from the 3D point to the camera's optical center. Assuming that a set D with d depths is represented as $D = \{D_k, k = 1, ..., d\}$, given depth D_k, the image-to-3D projection process is formulated as:

$$p^{3d}_{ijk} = \Phi(p^{2d}_{ij}; D_k), \tag{2}$$

where Φ is the projection operator, i and j are the index of vertical and horizontal coordinate of the point, $i = 1, ..., h$ and $j = 1, ..., w$. A 2D sampling point set, $\mathbf{P}^{2d} = \{p^{2d}_{11}, ..., p^{2d}_{hw}\}$, is constructed by sampling on the image plane. Through Eq. (2), $\mathbf{P}^{2d}$ is projected to multiple planes with different depths, $\mathbf{P}^{2d} \xrightarrow{\Phi} \{\mathbf{P}^{3d}_k, k = 1, ..., d\}$, where $\mathbf{P}^{3d}_k$ is the sample set on the plane with the depth D_k. The GCF map derived from $\mathbf{P}^{3d}_k$ is formulated as:

$$\mathbf{R}^{GCF}_{\Omega}(t, \mathbf{P}^{3d}_k) = \begin{bmatrix} \mathbf{r}(p_{11k}) & \dots & \mathbf{r}(p_{1wk}) \\ \vdots & \ddots & \vdots \\ \mathbf{r}(p_{h1k}) & \dots & \mathbf{r}(p_{hwk}) \end{bmatrix}_{h \times w}, \tag{3}$$

where $\mathbf{r}(p_{..k})$ is short for $\mathbf{r}^{GCF}_{\Omega}(t, p^{3d}_{..k})$. Assuming that the peak of GCF map is at the k_{max}-th depth, the spatial GCF (sGCF) map at time t is defined as:

$$\mathbf{R}^{sGCF}_{\Omega}(t, \mathbf{P}^{3d}) = \mathbf{R}^{GCF}_{\Omega}(t, \mathbf{P}^{3d}_{k_{max}}). \tag{4}$$

Due to the intermittent nature of speech and the continuity of the speaker's movement, the speech signals over a period provide references for audio cues at the current moment. Considering the signal in the time interval $[t - m_1, t]$, the m_2 frames with largest peak values of sGCF maps are selected among $m_1 + 1$ frames. The stGCF map at time t is defined as:

$$\mathbf{R}^{stGCF}_{\Omega}(t, \mathbf{P}^{3d}) = \{\mathbf{R}^{sGCF}_{\Omega}(t', \mathbf{P}^{3d}) | t' \in \mathbf{T}\}, \tag{5}$$

where $\mathbf{T}$ denotes the time set of the m_2 frames.

Visual Measurement. The tracking task aims to localize an arbitrary target selected in the first frame of the video, which makes it impossible to collect data in advance to train a specific detector for tracking. Therefore, the general deep metric learning method is introduced to train the model at the initial offline stage, which considers the tracking problem as the similarity measurement between a known target and the search area. A pre-trained Siamese network (Bertinetto et al. 2016) is employed in this module, which uses cross-correlation as the metric function completed by the convolution operation. The output response maps are equipped as visual cues, which can be formulated as:

$$\mathbf{S}(I_t) = \{f(I_t, I^{ref}) | I^{ref} \in \mathbf{I}\}, \tag{6}$$

where I_t is the current video frame, I^{ref} is the reference template which is the user-defined tracking target in the first frame, and $\mathbf{I}$ is the set of the reference templates with different scales. $f(\cdot)$ denotes the metric function that outputs a representative score map. The $\mathbf{S}(I_t)$ reflects the probability of the tracking target at each position in the search image, which is consistent with the meaning of the stGCF maps referring to the audio cues.

Multi-Modal Perception Attention Network

Given the extracted audio and visual cues, the multi-modal perception attention network (see Figure 2) generates a confidence score map as a speaker location representation. The brain's attention mechanism is able to selectively improve the transmission of information that attracts human attention, weighing the specific information that is more critical to the current task goal from abundant information. Inspired by this signal processing mechanism, a neural attention mechanism is exploited in this module to learn to measure the plausibility of multiple modalities.

To integrate the audio and visual cues, the stGCF maps $\mathbf{R}^{stGCF}_{\Omega}$ and visual response maps $\mathbf{S}(I_t)$ are normalized and reshaped into 3D matrix form, expressed as:

$$\begin{aligned} \mathbf{R} &= [\mathbf{R}_1, ..., \mathbf{R}_{D^a}] \in \mathbb{R}^{U \times D^a}, \\ \mathbf{S} &= [\mathbf{S}_1, ..., \mathbf{S}_{D^v}] \in \mathbb{R}^{U \times D^v}, \end{aligned} \tag{7}$$

where U denotes the size of each input video frame, $U = H \times W$. D^a is the dimension of the audio cues, which depends on m_2 referring to temporal cues, and D^v is dimension of the video cues, which is determined by the number of I^{ref}. The fused audio-visual cues, $\mathbf{V} = [\mathbf{R}_1, ..., \mathbf{R}_{D^a}, \mathbf{S}_1, ..., \mathbf{S}_{D^v}]$, are processed through a base network, which draws on the architecture of the channel attention module (Woo et al. 2018), where the channel corresponds to the observation extracted from the audio or visual modality. For each channel $i \in \{1, ..., D^a + D^v\}$, the attention mechanism G_{att} generates a positive score α_i to measure the reliability of the observation on the i-th channel. The processing is formulated as:

$$G_{att}(\mathbf{V}) = [\alpha_1, ..., \alpha_{D^a + D^v}] \in \mathbb{R}^{1 \times (D^a + D^v)}, \tag{8}$$

where the score α_i, termed the perceptual weight, reflects the confidence level of the multi-modal cues measured according to the previous section. The α_i is higher in reliable observations and turns to lower in ambiguous observations disturbed by background noise, room reverberation, visual occlusion, confusing background, etc. This gets benefits from the statistical features learned by the network from observation maps. Through this, the network exhibits the perceived ability to multi-modal observations, which describes the working interpretability of the proposed network.

Cross-Modal Self-Supervised Learning

The sensing capability accomplished by the network is an abstract process, which makes it impossible to label data artificially for essential supervision. To this end, a new cross-modal self-supervised learning strategy is proposed to train the network. The self-supervision includes a temporal factor and a spatial factor, which consider the temporal continuity of moving targets and the positional consistency in multi-modal observations, respectively. For the i-th channel, assuming that point $p^{max}_{t,i}$ is the position of the peak of the feature map at time t, the corresponding spatial factor of the observation on channel i is defined as the averaging operators in and across the multiple modalities. The cross-modal

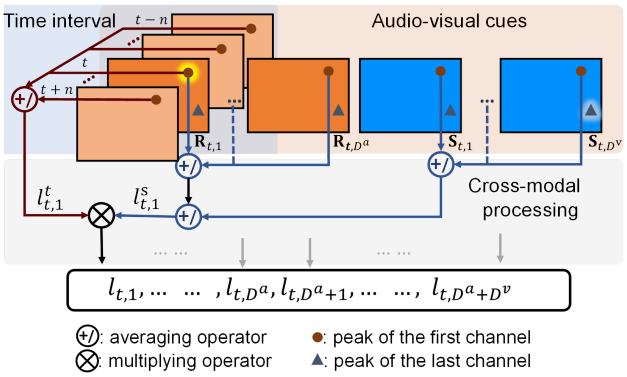

Figure 3: Illustration of the self-supervised labels generated across modalities in a time interval.

spatial factor is formulated as:

$$l^s_{t,i} = \frac{1}{2}[\frac{1}{D^a}\sum_{j=1}^{D^a}\mathbf{R}_{t,j}(p^{max}_{t,i}) + \frac{1}{D^v}\sum_{k=1}^{D^v}\mathbf{S}_{t,k}(p^{max}_{t,i})], \quad (9)$$

where $\mathbf{S}_{t,k}(p)$ denotes the normalized visual response value at position p. Note that $\mathbf{R}_{t,j}(p)$ is the normalized sGCF value at position p, where j is the depth index.

The temporal factor is derived by performing the above averaging operation on a time interval centered on time t. The temporal factor and the self-supervised label are expressed as:

$$\begin{aligned} l^t_{t,i} &= \frac{1}{2n+1}\sum_{q=t-n}^{t+n}\mathbf{V}_{q,i}(p^{max}_{t,i}), \\ l_{t,i} &= l^s_{t,i} \times l^t_{t,i}, \end{aligned} \quad (10)$$

where $\mathbf{V}$ denotes the audio map or visual map. As shown in Figure 3, the self-supervised label integrates the evaluations from different modalities in a time interval. When the target drifts on one observation, according to the complementarity between the modalities and the continuity of target motion, the lower value is provided by the other channel with more accurate observation. In addition, when the peaks of all observations are located in the same area, the value will increase accordingly. The general L2 loss is used to evaluate the generated labels and the attention measures.

Multi-Modal Tracker

The attention network introduced above supports multi-modal tracking through an improved PF algorithm. The attention measure output by the network is used to weight the audio-visual cues $\mathbf{V}$. Compared with the traditional additive likelihood and multiplicative likelihood, the weighting method based on the attention mechanism is essentially closer to the human sensory selective attention mechanism. Fusion map obtained after weighted average is expressed as:

$$\mathbf{Z} = \frac{1}{D^a + D^v}\sum_{i=1}^{D^a+D^v}\alpha_i\mathbf{V}_i. \quad (11)$$

The perception attention values of different modalities are fused in the map and used to weight particles in the update step of the PF. After diffusion, the value of the fusion map at the particle position is set as the new particle weight. Moreover, in order to utilize the global information of the fusion map, we simply improve the resampling step as well. At the beginning of each iteration, a group of the particles is reset to the peak position of the fusion map. Through the correction of the peak value, the tracking drift problem caused by the observation noise of some frames is avoided. The method is outstanding when the observation is severely disturbed by the environment noise.

Experiments and Discussions

Datasets. In this section, the proposed tracker is evaluated on the AV16.3 corpus (Lathoud, Odobez, and Gatica-Perez 2004), which provides true 3D mouth location derived from calibrated cameras and 2D measurements on the various images for systematic assessment. The audio data is recorded at the sampling rate of 16 kHz by two circular eight-element microphone arrays placed 0.8m apart on the table. The images are captured by 3 monocular color cameras installed in 3 corners of the room at 25Hz with size $H \times W = 288 \times 360$. The experiments are tested on seq08, 11, and 12, where a single participant wandered around, moved quickly, and spoke intermittently. Each set of experiments uses signals from two microphone arrays and an individual camera.

Implementation Details. Visual cues are generated by a pretrained Siamese network (Bertinetto et al. 2016) based on AlexNet backbone. Reference image set $\mathbf{I}$ contains two target rectangles with scales of 1 and 1.25, which are defined by users in the first frame. For audio measurement, the number of 2D sampling points in the horizontal and vertical directions on the image plane are $w = 20$ and $h = 16$. A $0.8m$ high table is placed in a $(3.6 \times 8.2 \times 2.4)m$ room. Therefore, the sampling points located outside the room range and below the desktop are removed, which is in accord with the real situation and avoids the ambiguity caused by the symmetry of the circular microphone. The depths number of projected 3D points is set to $d = 6$. The speech signal is enframed to $40ms$ by a Hamming window with a frame shift of $1/2$. The parameters to calculate stGCF are set to $M = 120$, $m_1 = 15$, $m_2 = 5$. Backbone of the attention network is MobileNetv3-large (Howard et al. 2019). The network is trained on single speaker sequences seq01, 02, 03, which contain more than 4500 samples. The parameters to generate self-supervised label are set to $D^a = 5$, $D^v = 2$, $n = 6$. All models are trained for 20 epochs with batch size 16 and learning rate 0.01. Our method and comparison methods are based on Sampling Importance Resampling (SIR)-PF for tracking. The number of particles is set to 100. Our source codes are available at https://github.com/liyidi/MPT.

Evaluation Metrics. Mean Absolute Error (MAE) and the Accuracy (ACC) is used to evaluate performance of tracking methods. MAE calculates the Euclidean distance in pixel between the estimated position and the ground truth (GT), divided by the number of frames. ACC measures the percentage of correct estimates, whose error distance in pixel does not exceed 1/2 of the diagonal of the bounding-box of GT.

Sequences		Uni-modal		Multi-modal				Uni-modal+Occ		Multi-modal+Occ				Occ rate
Seq	Cam	AO	VO	AV-A	AV3D	2LPF	MPT(ours)	VO		2LPF		MPT(ours)		
		MAE↓		MAE↓				MAE↓	ACC↑	MAE↓	ACC↑	MAE↓	ACC↑	
08	1	32.87	21.41	10.75	4.31	3.32	3.67	103.46	26.35	94.45	42.97	11.54	88.79	29.88
	2	18.76	16.58	7.33	4.66	3.06	3.58	181.82	19.14	75.41	62.42	7.88	92.75	37.74
	3	27.01	15.73	9.85	5.34	3.47	3.43	141.76	21.53	68.54	50.51	14.3	81.52	54.62
11	1	28.27	14.69	14.66	8.15	6.15	6.77	30.12	81.06	26.35	82.91	13.57	88.93	15.66
	2	24.16	16.42	14.01	7.48	5.58	4.55	116.87	19.45	111.47	27.31	26.06	65.65	70.17
	3	25.66	21.54	13.96	6.64	3.86	3.84	86.60	42.95	49.97	50.00	21.98	77.40	66.32
12	1	40.67	17.83	12.49	6.86	4.11	4.67	93.07	39.88	122.72	16.87	17.43	77.71	32.55
	2	24.26	19.03	10.81	10.67	5.39	4.84	145.54	23.12	104.3	31.15	23.96	65.75	74.50
	3	34.02	22.29	11.86	9.71	5.65	3.78	157.37	21.78	144.25	25.48	21.97	66.57	78.35
Average		28.40	18.39	11.74	7.09	4.51	**4.34**	117.40	32.80	88.60	43.29	**17.63**	**78.34**	51.08

Table 1: Comparison results with uni-modal methods and the state-of-the-art audio-visual methods on the original dataset and the occluded dataset. Occ rate is the percentage of frames in which the speaker is occluded by the mask. MAE is in pixel, ACC is in %. The proposed method achieves robust tracking in the presence of occlusion. (Occ: occluded sequences, AO: audio-only, VO: visual-only, AV-A: (Kılıç et al. 2015), AV3D: (Qian et al. 2017), 2LPF: (Liu, Li, and Yang 2019), MPT: ours)

AM	AN	TR	Org		Occ	
			MAE↓	ACC↑	MAE↓	ACC↑
GCF	-	-	80.15	45.23	80.15	45.23
stGCF	-	-	28.40	63.58	28.40	63.58
stGCF	AvgAtt	-	22.63	74.16	33.48	60.59
stGCF	MPAtt	-	12.56	89.50	24.57	72.17
stGCF	AvgAtt	IPF	17.33	78.22	26.88	67.54
stGCF	**MPAtt**	**IPF**	**4.34**	**98.55**	**17.63**	**78.34**

Table 2: Influence of each innovative component in MPT, compared with the general GCF feature and average attention. (AM: audio measurement, AN: attention network, TR: tracker, AvgAtt: average attention, MPAtt: multi-modal perception attention, IPF: improved PF, Org: original dataset)

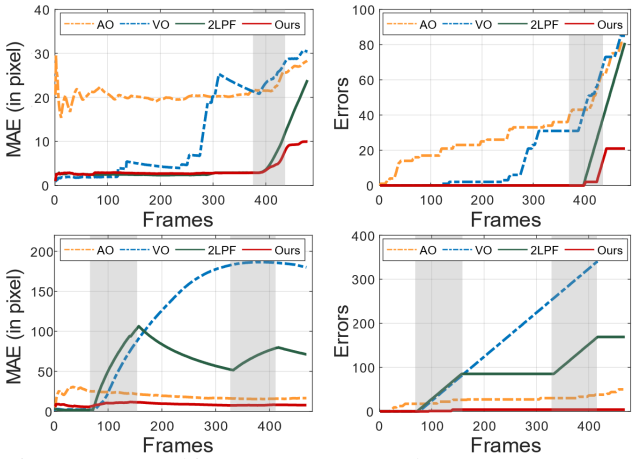

Figure 4: Tracking accuracy on the Seq11cam1 and Seq08cam2. (Errors: numbers of miss tracking).

Comparison Results. The proposed MPT is compared with the uni-modal method and the state-of-the-art audio-visual methods, which are based on the PF architecture. The AO and VO methods are implemented based on the audio cues and visual cues proposed in the previous section. Furthermore, in order to verify the robustness of the tracker under interference conditions, we conducted comparative experiments on the occluded data. The occlusion area is artificially covered in the middle of the image (1/3 of the frame), which is used to simulate the situation where the field of view is limited or the camera viewfinder is obscured. In the sequences, the speaker walks behind the occluded area and then appears on the screen again. For better evaluation, we count the proportion of frames in each sequence where the target was occluded by the mask.

Comparison results are listed in Table 1. Firstly, the combination of audio and visual modalities shows great benefits for speaker tracking. On the standard dataset, the MAE of the proposed MPT is 4.34 in pixel, which is superior to the state-of-the-art. 2LPF has achieved accurate estimation by employing additional particle filters in audio and visual space, respectively. However, the calculation of fusion likelihood in 2LPF depends on the stable observations, which leads to a rapid decline when visual observation is unavailable. In contrast, MPT achieves a better tracking accuracy of 78.34% on sequences with an average occlusion rate of 51.08%. Figure 4 shows the MAE and error numbers of two typical sequences, where the shaded box represents the frames in which the target is occluded. VO and 2LPF are severely affected by occlusion, which can be seen from the significant rise of curves in the shaded area. Our MPT is also affected by occlusion, but the impact is relatively minor.

Ablation Study and Analysis. Ablation studies are conducted to evaluate the effectiveness of the main innovative components of our method in Table 2, including audio measurement, attention network, and PF-based tracker. The general GCF (Brutti, Omologo, and Svaizer 2006) calculates the plausibility of the existence of active sound sources at specific coordinates in all possible source positions in a given room. Without the guidance of prior information, it is difficult to derive accurate coordinates within limited calculations. In the case of the stGCF method, the search range is reduced to multiple depth planes in 3D space using the projection relationship to achieve sound source localization on the image plane, which has never been studied before. How-

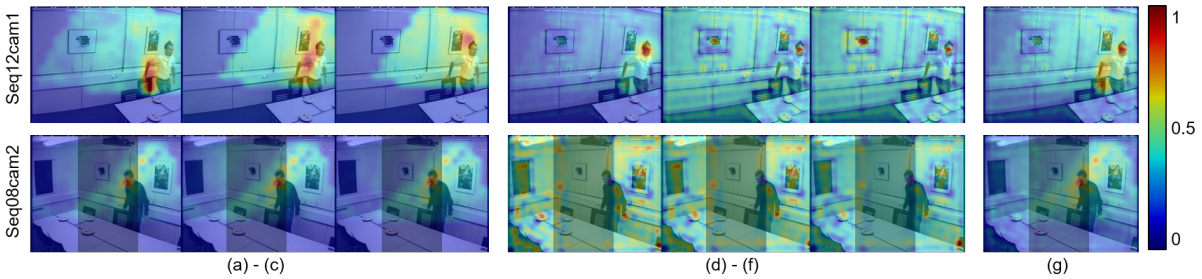

Figure 5: Visualization of the complementarity of multiple modalities. (a)-(c) are audio cues, (d)-(f) are visual cues, (g) is the fusion map generated according to perception weights. The two rows respectively represent the scene where the audio signal and the video observation are disturbed. The shadow in the middle of the image represents the invisible part.

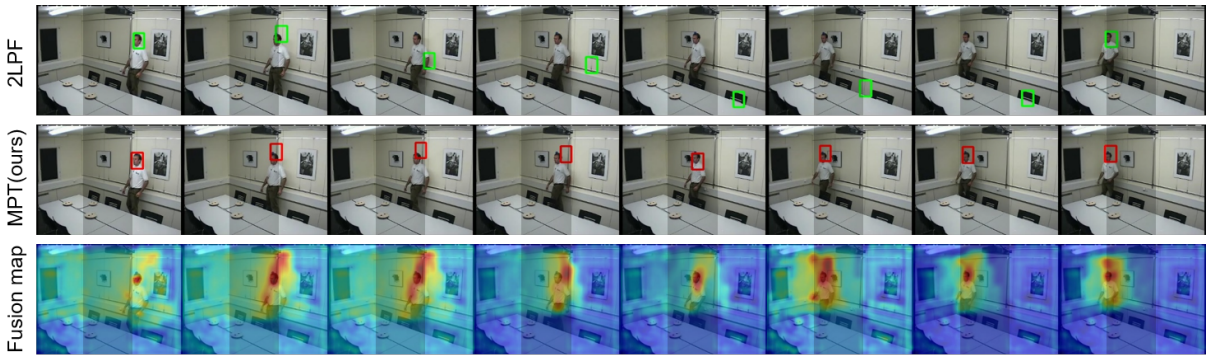

Figure 6: Comparison of tracking results on the occluded sequence. Green rectangles are from the contrast method 2LPF, and red rectangles come from the proposed MPT method. The bottom row shows the fusion maps of corresponding frames.

ever, the stGCF method is affected by the geometric configuration of the camera and the microphone array, especially when the speaker is located on the line connecting the camera and the array. In addition, due to the directionality of the sound signal, the peak usually appears in a large highlighted area in the stGCF map, which provides an ambiguous search result. Nevertheless, the results we calculated using two microphone arrays are better than traditional methods, with the MAE decreasing from 80.15 to 28.40. Note that the result is not changed by visual occlusion.

In the last four sets of the ablation study, visual cues are added to evaluate the contribution of the attention mechanism. The enhancements made by AvgAtt shows the strength of audio-visual fusion, even if it works as a set of weights with the same value. By comparison, our MPAtt achieves higher performance gains. Compared with AvgAtt, the accuracy of MPAtt has increased by 21% and 19%, respectively, on the original dataset and the occlusion dataset. In addition, an improved PF is employed for tracking, which smoothes the trajectory through the time series model. The improved resampling method using the global maximum of the fusion map avoids the particles being restricted to the local optimum due to the target missing of individual frames.

Visualization Analysis. In this section, the audio-visual cues and fusion maps are generated as the heat map to visualize the sub-process of the proposed method, which demonstrates the interpretability of the perception attention network. As illustrated in Figure 5, in the sample in the first row, the speech is disturbed by the noise emitted by the chair, and in the sample in the second row, the face of the speaker is completely obscured. Nevertheless, the correct area in the fusion map is highlighted. This indicates that benefit from the network's ability to perceive the state of each modality, the model can learn the corresponding perception weights by using the complementarity across the audio-visual cues. Figure 6 shows the robustness of our tracker, which can achieve continuous tracking when the field of view is limited. Since the auditory sense is not interfered by the visual distraction, audio cues hold dominance over such difficult samples. When the speaker walks to the occluded area, the tracker can roughly estimate the speaker's position, which is beneficial to re-track when the target is visible again.

Conclusions

In this paper, we propose a novel multi-modal perception tracker for the challenging audio-visual speaker tracking task. We also propose a new multi-modal perception attention network and a new acoustic map extraction method. The proposed tracker utilizes the complementarity and consistency of multiple modalities to learn the availability and reliability of observations between various modalities in a self-supervised fashion. Extensive experiments demonstrate that the proposed tracker is superior over the current state-of-the-art counterparts, especially showing sufficient robustness under adverse conditions. Lastly, the intermediate process is visualized to demonstrate the interpretability of the proposed tracker network.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No.62073004), and the Science and Technology Plan of Shenzhen (No.JCYJ20200109140410340).

References

Afouras, T.; Asano, Y. M.; Fagan, F.; Vedaldi, A.; and Metze, F. 2021. Self-supervised object detection from audio-visual correspondence. *arXiv:2104.06401*.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2): 423–443.

Ban, Y.; Alameda-Pineda, X.; Girin, L.; and Horaud, R. 2019. Variational bayesian inference for audio-visual tracking of multiple speakers. *IEEE TPAMI*, 43(5): 1761–1776.

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *ECCV*, 850–865.

Brutti, A.; and Lanz, O. 2010. A joint particle filter to track the position and head orientation of people using audio visual cues. In *European Signal Processing Conference*, 974–978.

Brutti, A.; Omologo, M.; and Svaizer, P. 2006. Speaker localization based on oriented global coherence field. In *ICSLP*, 2606–2609.

Chiariotti, P.; Martarelli, M.; and Castellini, P. 2019. Acoustic beamforming for noise source localization-reviews, methodology and applications. *Mechanical Systems and Signal Processing*, 120: 422–448.

Cobos, M.; Antonacci, F.; Comanducci, L.; and Sarti, A. 2020. Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach. *IEEE/ACM TASLP*, 28: 1270–1281.

Duan, B.; Tang, H.; Wang, W.; Zong, Z.; Yang, G.; and Yan, Y. 2021a. Audio-Visual Event Localization via Recursive Fusion by Joint Co-Attention. In *WACV*, 4013–4022.

Duan, B.; Wang, W.; Tang, H.; Latapie, H.; and Yan, Y. 2021b. Cascade attention guided residue learning gan for cross-modal translation. In *ICPR*, 1336–1343.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *ICCV*, 1314–1324.

Hu, D.; Qian, R.; Jiang, M.; Tan, X.; Wen, S.; Ding, E.; Lin, W.; and Dou, D. 2020. Discriminative sounding objects localization via self-supervised audiovisual matching. *NeurIPS*, 33.

Katsaggelos, A. K.; Bahaadini, S.; and Molina, R. 2015. Audiovisual fusion: Challenges and new approaches. *Proceedings of the IEEE*, 103(9): 1635–1653.

Kılıç, V.; Barnard, M.; Wang, W.; and Kittler, J. 2015. Audio assisted robust visual tracking with adaptive particle filtering. *IEEE TMM*, 17(2): 186–200.

Kılıç, V.; and Wang, W., eds. 2017. *Audio-visual speaker tracking*. IntechOpen.

Lathoud, G.; Odobez, J.-M.; and Gatica-Perez, D. 2004. AV16. 3: An audio-visual corpus for speaker localization and tracking. In *International Workshop on MLMI*, 182–195.

Li, C.; and Qian, Y. 2020. Deep audio-visual speech separation with attention mechanism. In *ICASSP*, 7314–7318.

Liu, G.; Tang, H.; Latapie, H. M.; Corso, J. J.; and Yan, Y. 2021. Cross-view exocentric to egocentric video synthesis. In *ACM MM*, 974–982.

Liu, H.; Li, Y.; and Yang, B. 2019. 3D audio-visual speaker tracking with a two-layer particle filter. In *ICIP*, 1955–1959.

Liu, Y.; Kılıç, V.; Guan, J.; and Wang, W. 2019. Audio-visual particle flow smc-phd filtering for multi-speaker tracking. *IEEE TMM*, 22(4): 934–948.

Masuyama, Y.; Bando, Y.; Yatabe, K.; Sasaki, Y.; Onishi, M.; and Oikawa, Y. 2020. Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling. In *IROS*, 4848–4854.

Omologo, M.; and Svaizer, P. 1997. Use of the crosspower-spectrum phase in acoustic event location. *IEEE TSAP*, 5(3): 288–292.

Qian, X.; Brutti, A.; Lanz, O.; Omologo, M.; and Cavallaro, A. 2019. Multi-speaker tracking from an audio-visual sensing device. *IEEE TMM*, 21(10): 2576–2588.

Qian, X.; Brutti, A.; Lanz, O.; Omologo, M.; and Cavallaro, A. 2021. Audio-visual tracking of concurrent speakers. *IEEE TMM*.

Qian, X.; Brutti, A.; Omologo, M.; and Cavallaro, A. 2017. 3D audio-visual speaker tracking with an adaptive particle filter. In *ICASSP*, 2896–2900.

Schymura, C.; and Kolossa, D. 2020. Audiovisual speaker tracking using nonlinear dynamical systems with dynamic stream weights. *IEEE/ACM TASLP*, 28: 1065–1078.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and Kweon, I. S. 2019. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE TPAMI*.

Tang, H.; Liu, H.; Xu, D.; Torr, P. H.; and Sebe, N. 2021. Attention-gan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE TNNLS*.

Tang, H.; Xu, D.; Sebe, N.; Wang, Y.; Corso, J. J.; and Yan, Y. 2019. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2417–2426.

Wang, Q.; Teng, Z.; Xing, J.; Gao, J.; Hu, W.; and Maybank, S. 2018. Learning attentions: Residual attentional siamese network for high performance online visual tracking. In *CVPR*, 4854–4863.

Wilson, J.; and Lin, M. C. 2020. AVOT: Audio-visual object tracking of multiple objects for robotics. In *ICRA*, 10045–10051.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *ECCV*, 3–19.

Xu, D.; Wang, W.; Tang, H.; Liu, H.; Sebe, N.; and Ricci, E. 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, 3917–3925.

Xuan, H.; Zhang, Z.; Chen, S.; Yang, J.; and Yan, Y. 2020. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI*, 279–286.

Yang, B.; Liu, H.; Pang, C.; and Li, X. 2019. Multiple sound source counting and localization based on tf-wise spatial spectrum clustering. *IEEE/ACM TASLP*, 27(8): 1241–1255.

Yang, G.; Tang, H.; Ding, M.; Sebe, N.; and Ricci, E. 2021. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 16269–16279.

Yu, Y.; Xiong, Y.; Huang, W.; and Scott, M. R. 2020. Deformable siamese attention networks for visual object tracking. In *CVPR*, 6728–6737.

Zeng, Y.; Wang, H.; and Lu, T. 2019. Learning spatial-channel attention for visual tracking. In *ICCC*, 277–282.

Zhang, W.-H.; Chen, A.; Rasch, M. J.; and Wu, S. 2016. Decentralized multisensory information integration in neural systems. *Journal of Neuroscience*, 36(2): 532–547.

Zhao, S.; Ma, Y.; Gu, Y.; Yang, J.; Xing, T.; Xu, P.; Hu, R.; Chai, H.; and Keutzer, K. 2020. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *AAAI*, 303–311.