3D Audio-Visual Speaker Tracking with A Novel Particle Filter

Yongheng Sun

Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School

Introduction

3D speaker tracking using co-located audio-visual sensors has received much attention recently, however, it is still challenging.

In this paper, a novel particle filter (PF) based method is proposed for 3D audio-visual speaker tracking. Compared with traditional PF based audio-visual speaker tracking method, our 3D audio-visual tracker has two main characteristics.
(1) In the prediction stage, we use audio-visual information at current frame to further adjust the direction of the particles after the particle state transition process.
(2) In the update stage, the particle likelihood is calculated by fusing both the visual distance and audio-visual direction information.

Experimental results show that the proposed tracker outperforms other methods and provides a favorable speaker tracking performance both in 3D space and on the image plane.

Methods

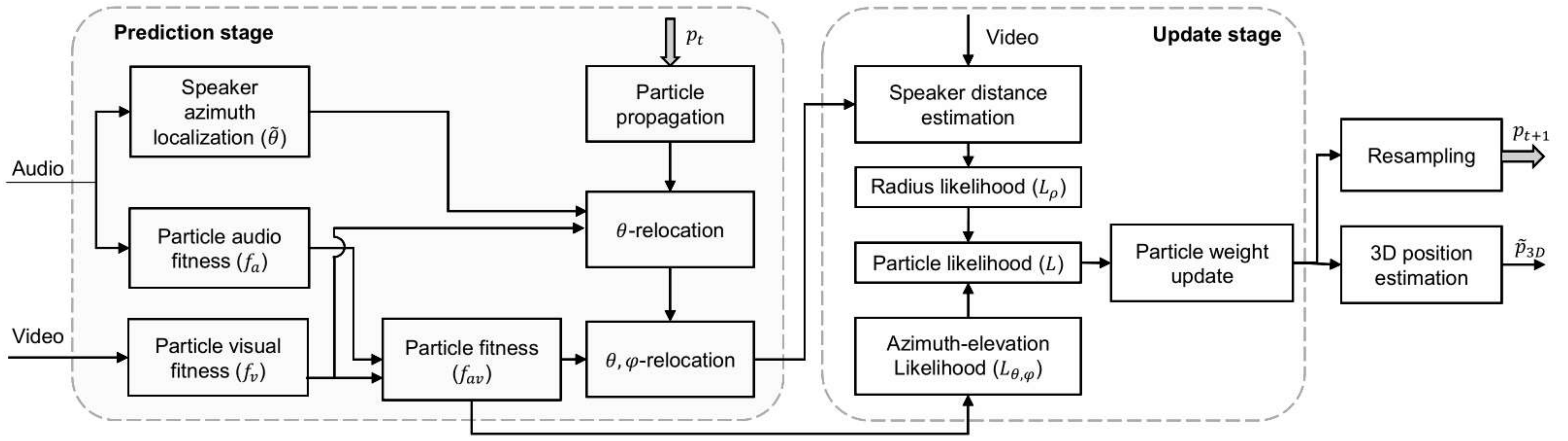
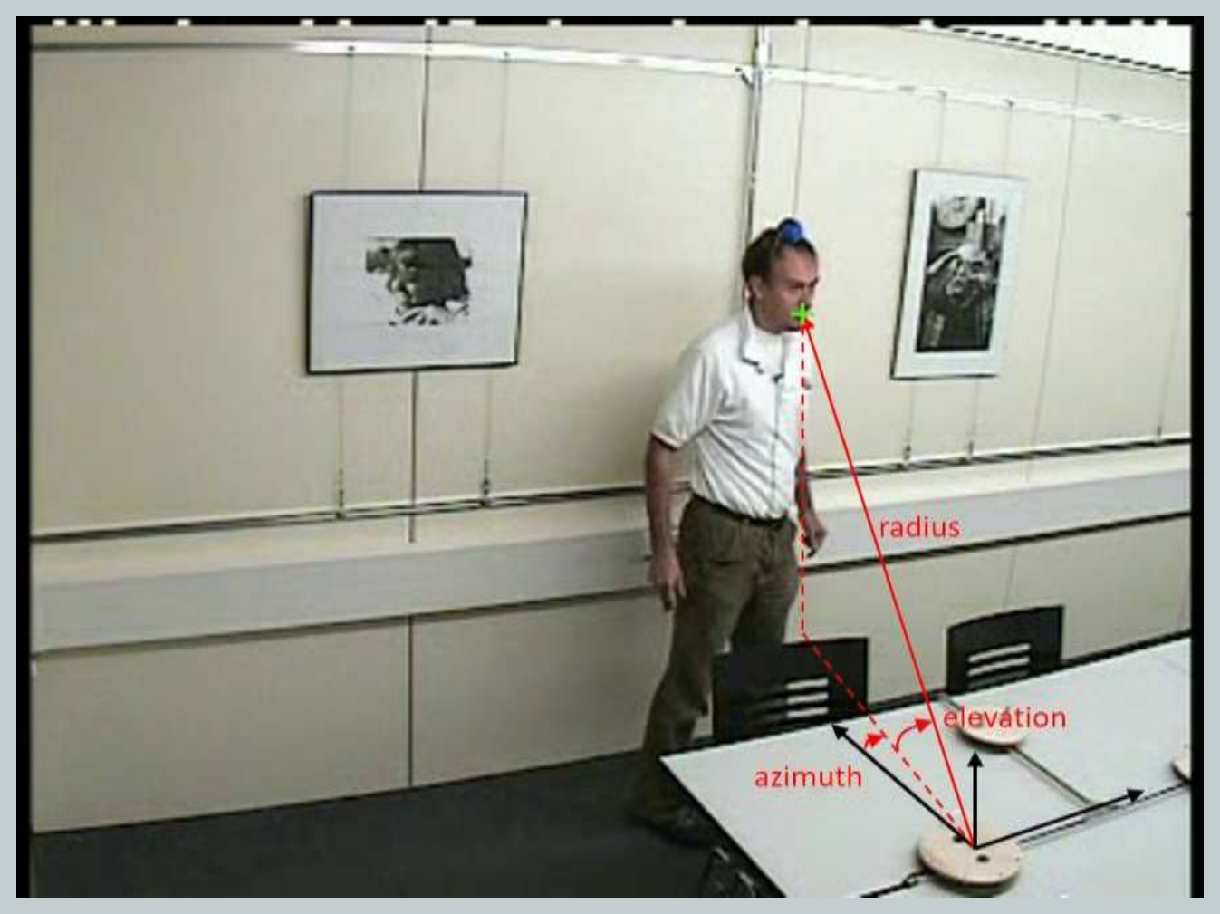
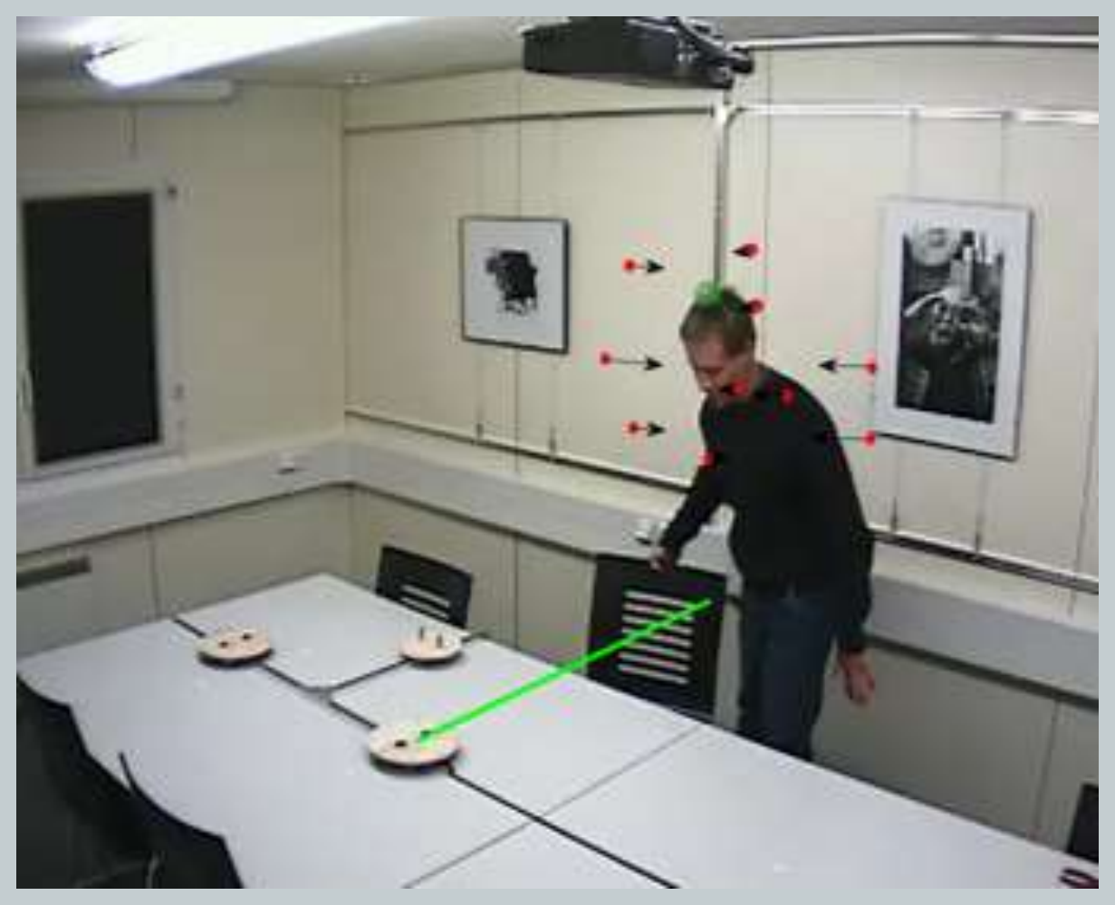
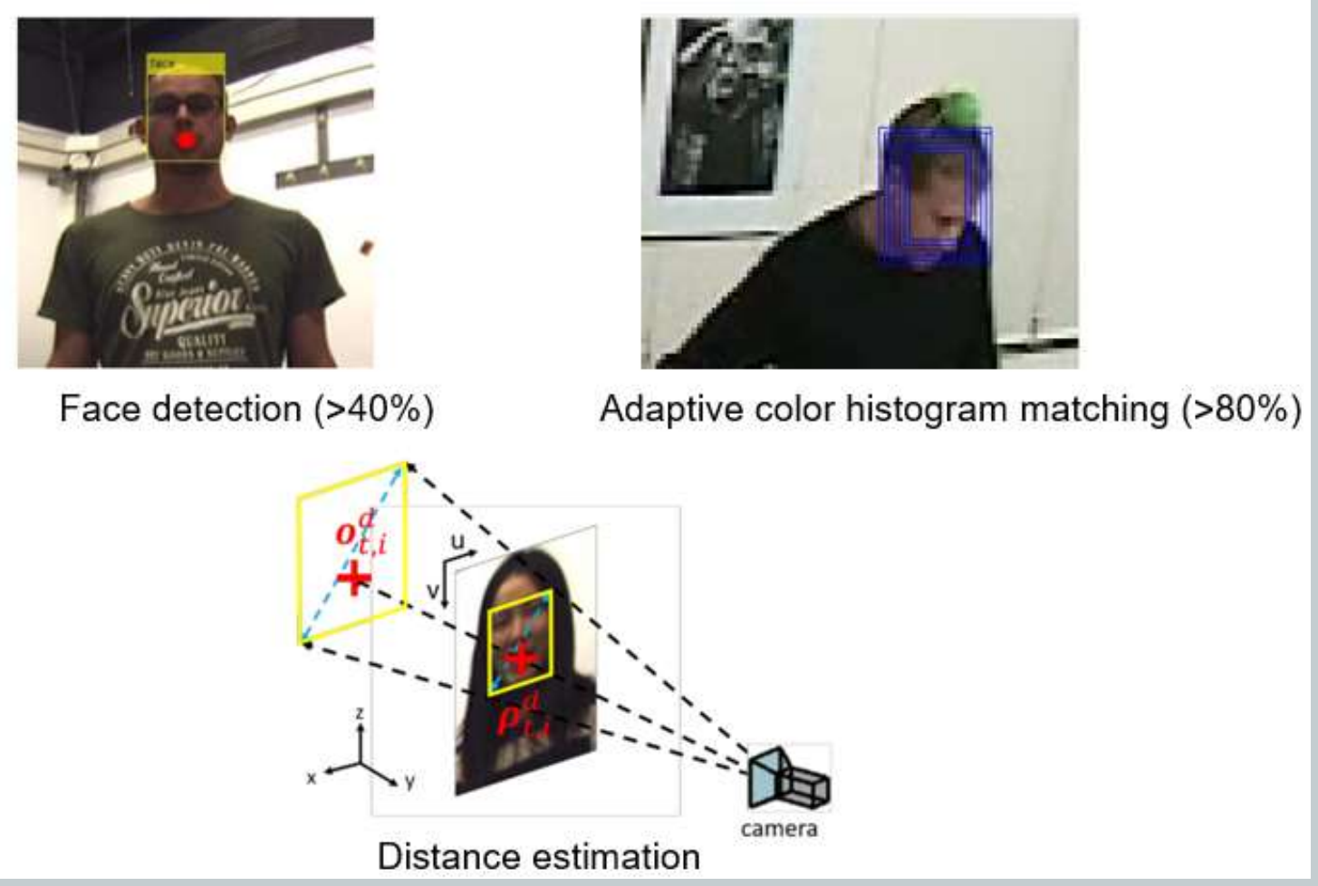

Prediction stage
Update stage
Audio
Video
Speaker azimuth localization ($\bar{\theta}$)
Particle audio fitness (f_a)
Particle visual fitness (f_v)
p_t
Particle propagation
θ-relocation
Particle fitness (f_{av})
θ,φ-relocation
Video
Speaker distance estimation
Radius likelihood (L_ρ)
Particle likelihood (L)
Azimuth-elevation Likelihood ($L_{\theta,\varphi}$)
Particle weight update
Resampling
p_{t+1}
3D position estimation
$\tilde{p}_{3D}$

Fig. 1. Overall framework of the proposed 3D audio-visual speaker tracker.







Experiments

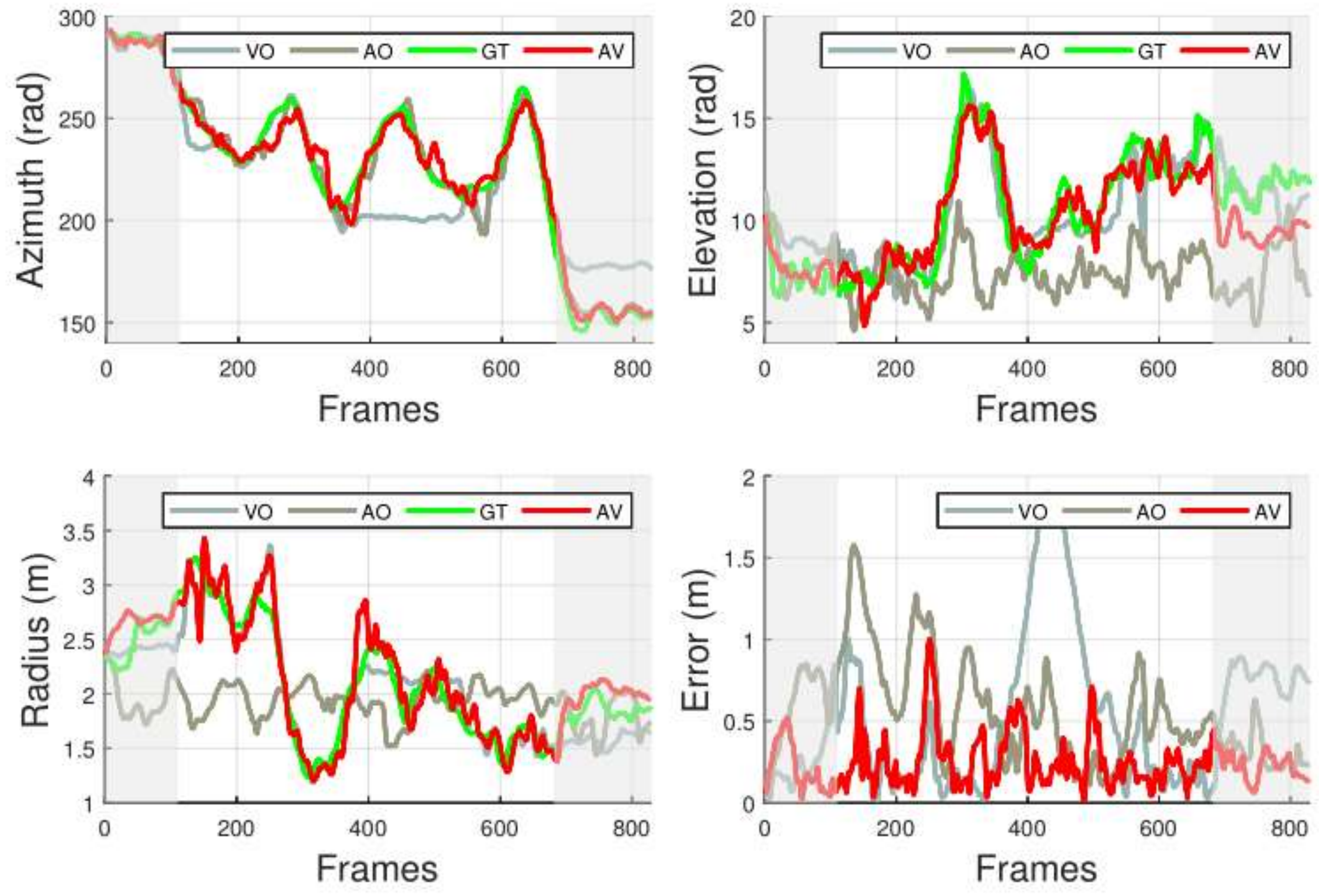


Fig. 6. 3D speaker tracking results on CAV3D $seq08$.

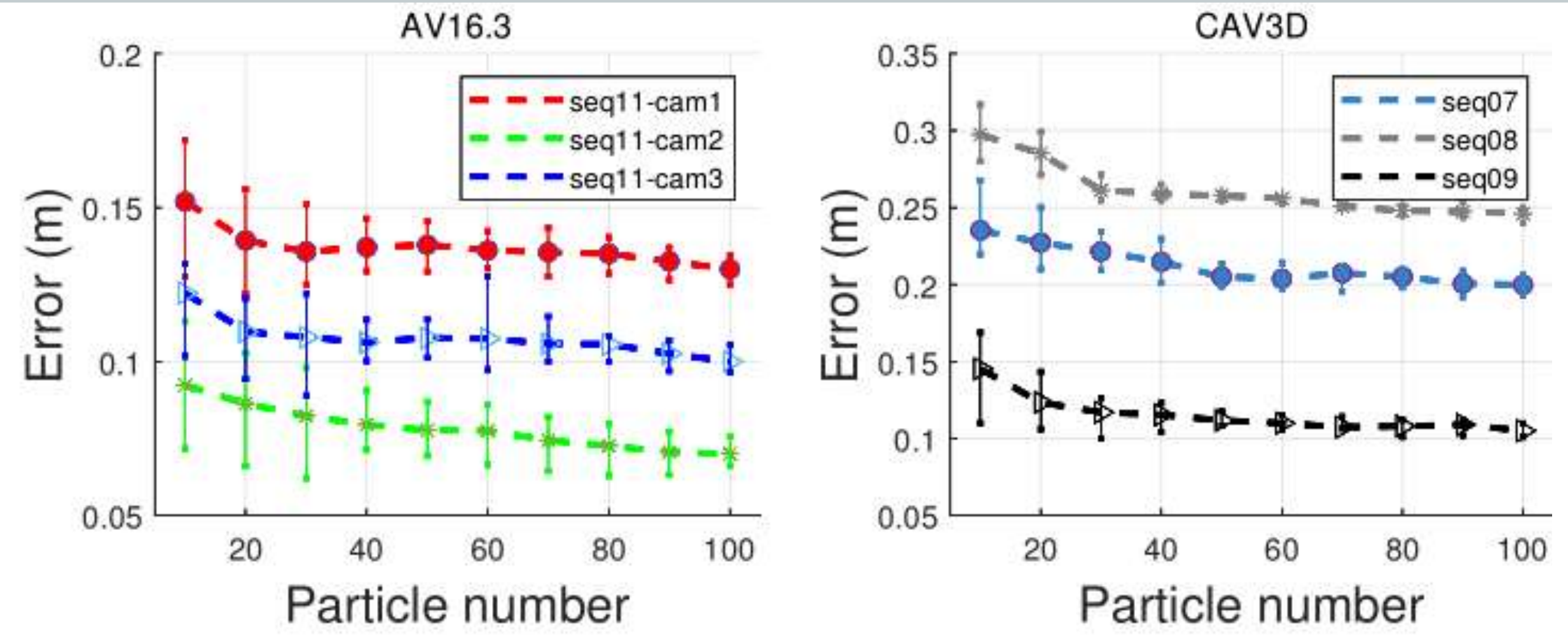


Fig. 5. Tracking accuracy under different numbers of particles.

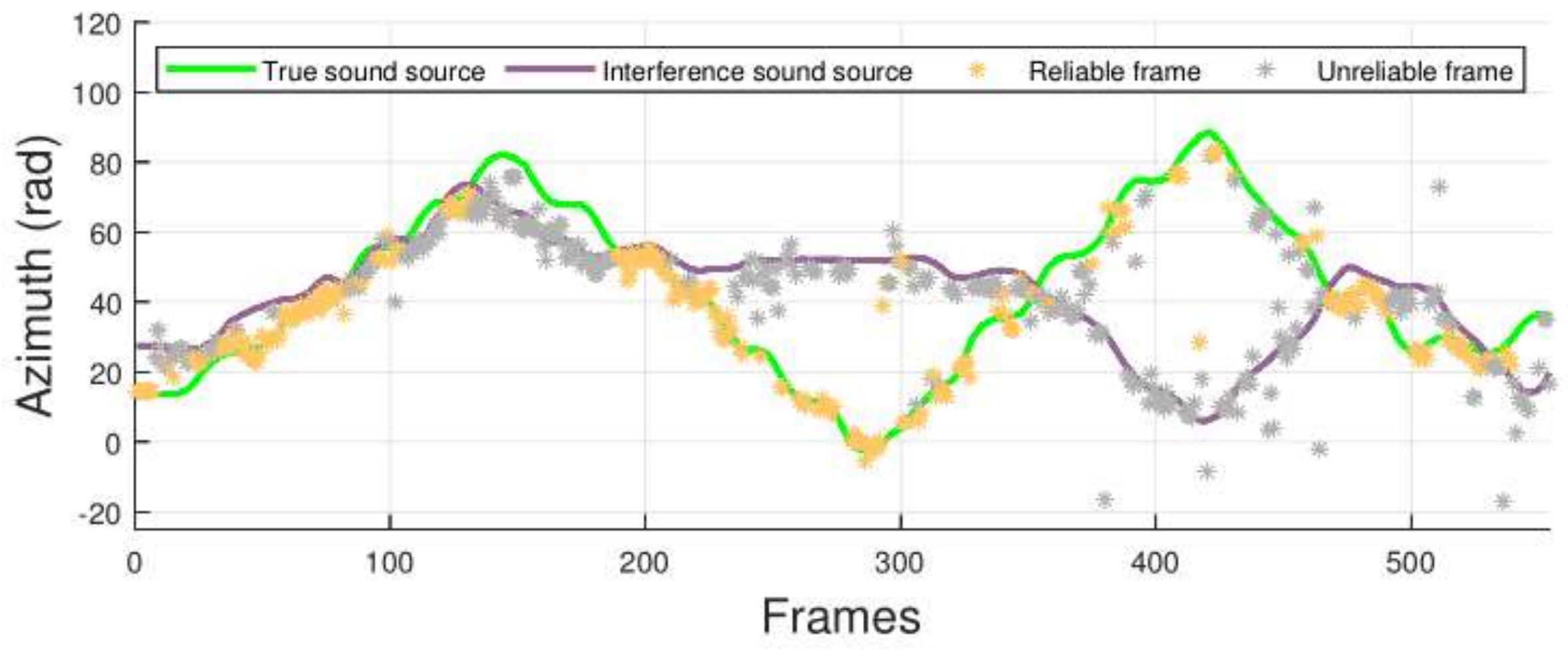


Fig. 4. Video assisted audio reliability measurement.

TABLE I
3D SPEAKER TRACKING RESULTS WITH OR WITHOUT PARTICLE RELOCATION. "√" REPRESENTS WITH THIS OPERATION, AND "-" REPRESENTS WITHOUT THIS OPERATION.

	θ -relocation	θ,φ -relocation	3D MAE (m)	2D MAE (pixel)	TLR (3D)	Percentage of resample (%)
AV16.3	√	√	**0.10**	**4.13**	**4.48**	**6.61**
	√	-	0.29	7.60	30.76	20.10
	-	√	0.31	7.99	32.36	7.80
	-	-	0.46	9.80	49.66	25.80
CAV3D	√	√	**0.21**	**12.0**	**20.7**	**18.42**
	√	-	0.23	18.9	25.4	27.01
	-	√	0.59	26.5	53.8	20.40
	-	-	0.95	40.5	85.1	31.41

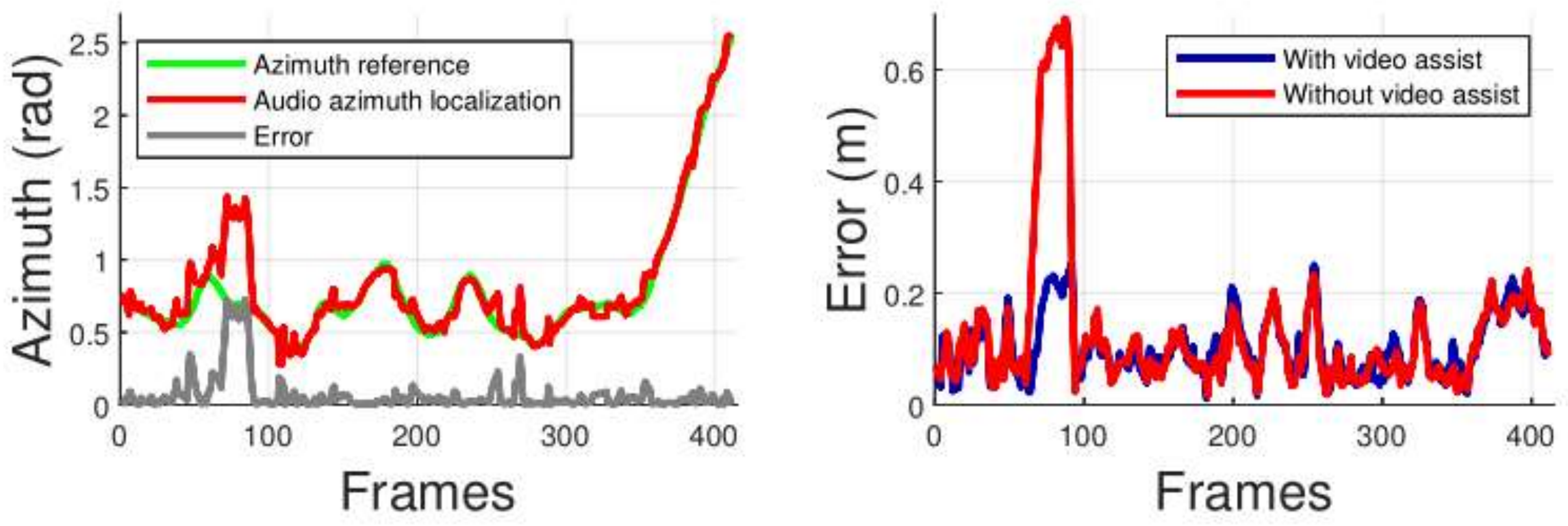


Fig. 2. Audio azimuth localization and 3D tracking error on $seq11$-$cam2$-$mic1$.

Conclusions

(1) This paper presents a novel particle filter based method for 3D audio-visual speaker tracking using a co-located monocular camera and microphone array.
(2) The audio azimuth relocation and audio-visual azimuth-elevation relocation are successively performed, aiming to make the particles more concentrated around the speaker direction
(3) Face detection combined with the proposed adaptive color histogram matching method can provide continuous speaker distance information when the speaker is in the visual FoV.

Authors:

Hong Liu, hongliu@pku.edu.cn, Peking University

Yongheng Sun, 1801213394@pku.edu.cn, Peking University

Yidi Li, yidili@pku.edu.cn, Peking University

Bing Yang, bingyang@sz.pku.edu.cn