Xinyuan Qian, Alessio Brutti, Oswald Lanz, Maurizio Omologo, Andrea Cavallaro, **Multi-speaker tracking from an audio-visual sensing device**, IEEE TRANSACTIONS ON MULTIMEDIA, Volume: 21, Issue: 10, Oct. 2019, pp. 2576 – 2588, DOI: 10.1109/TMM.2019.2902489

The final published version is available online at: https://ieeexplore.ieee.org/document/8656587

When citing, please refer to the published version.

Multi-speaker tracking from an audio-visual sensing device

Xinyuan Qian, Alessio Brutti, Oswald Lanz, Maurizio Omologo, Andrea Cavallaro

***Abstract*—Compact multi-sensor platforms are portable and thus desirable for robotics and personal-assistance tasks. However, compared to physically distributed sensors, the size of these platforms makes person tracking more difficult. To address this challenge, we propose a novel 3D audio-visual people tracker that exploits visual observations (object detections) to guide the acoustic processing by constraining the acoustic likelihood on the horizontal plane defined by the predicted height of a speaker. This solution allows the tracker to estimate, with a small microphone array, the distance of a sound. Moreover, we apply a color-based visual likelihood on the image plane to compensate for misdetections. Finally, we use a 3D particle filter and greedy data association to combine visual observations, color-based and acoustic likelihoods to track the position of multiple simultaneous speakers. We compare the proposed multimodal 3D tracker against two state-of-the-art methods on the AV16.3 dataset and on a newly collected dataset with co-located sensors, which we make available to the research community. Experimental results show that our multimodal approach outperforms the other methods both in 3D and on the image plane.**

I. Introduction

Audio-visual person tracking is important for scene understanding, human-robot interaction, and speech enhancement. Exploiting the complementarity of multimodal signals by effectively fusing audio and video data helps improve accuracy and robustness [1–6]. In fact, combining information from multiple modalities is preferable to using each modality individually [3, 7–13]. For example, video analysis is challenging under clutter and varying lighting conditions, whereas sound sources may be intermittent or corrupted by background noise and reverberation. Tracking accuracy can improve when sound-source position estimates are combined with temporal visual observations on the image plane [12] or sound is used to help estimate people trajectories in unseen regions [9].

Visual trackers often use color histograms as features alongside object detections. In particular, face detectors are highly accurate [14–16], but they may fail under challenging poses, occlusions or low resolution. The signal from microphone pairs can generate Sound Source Localization (SSL) estimates [5, 10, 17–19], using, for example, the Generalized Cross Correlation (GCC) [20]. Among GCC methods, Generalized Cross Correlation with Phase Transform (GCC-PHAT) [20, 21] is preferable under reverberation or rapidly varying acoustic characteristics, when the spectral distribution of the noise cannot be estimated [22]. Recently, GCC and GCC-PHAT have also been applied in different fusion based tracking tasks, as described in [23, 24]. Combining observations from multiple (spatially distributed) microphone pairs in a 3D Global Coherence Field (GCF) acoustic map [25], also known as Steered Response Power PHAse Transform (SRP-PHAT) [26], performs well under noise and reverberation [22]. Audio-visual trackers combine these features to generate estimations on the image plane [1, 18, 27, 28], on a ground plane [5, 18, 29], or in 3D [4, 9, 10, 19, 30–35].

3D audio-visual trackers rely on stereo cameras [19, 32], depth cameras [35] or spatially distributed sensors [4, 9, 10, 30, 31, 33, 34]. The different views of spatially distributed sensors lead to a better coverage and estimates that can be obtained via triangulation. Moreover, there is a higher likelihood that at least one microphone pair captures the direct path of the target speech or at least one camera observes the target from a favourable view. Using a compact audio-visual sensing platform, instead, is challenging especially for depth estimation as targets are not surrounded by sensors (thus reducing the available spatial information). Moreover, the inter-microphone distance is small compared to the speaker-array distance [36]. For these reasons, triangulating the target position leads to noisy SSL estimates and therefore trackers for co-located sensors are usually constrained on the image plane [12, 18, 28, 37], except when using stereo vision [19].

To address these limitations, we propose a novel Audio Visual 3D Tracker (AV3T), which uses multimodal signals from a compact audio-visual sensing platform composed of a small circular microphone array and a monocular camera. AV3T derives 3D visual observations by estimating the mouth positions of the targets from face detections, which also assist audio processing by reducing 3D localization uncertainties due to the small inter-microphone distance. This improvement is achieved by constraining the audio likelihood at the speaker-height plane, which is inferred from previous detections. By removing a degree of freedom in acoustic localization, the estimates are more accurate than the video estimates determined from the scaling factor on detected faces. Moreover, AV3T uses a color-based generative visual likelihood to compensate for misdetections within the Field-of-View (FoV) of the camera through color templates of the target that are updated from previously detected faces. Finally, after a greedy data association, a Particle Filter (PF) ensures a smooth tracking of the multimodal observations. In summary, the main novelties are (i) the conditional selective visual models, (ii) the cross-modal combination of the audio and video cues, (iii) the video-driven audio processing and (iv) a particle filter implementation for 3D audio-visual tracking with co-located sensor.

To the best of our knowledge, we are the first to perform 3D audio-visual tracking using a small co-located sensing platform with a single camera, and with multiple simultaneous

TABLE I

Audio-visual datasets, their sensor setup, annotations and content. KEY - # m: number of microphones; f_a : audio sampling frequency (kHz); CA: Circular Array; # c: number of cameras; fps: frame per second; col.: co-located platform; cal.: calibration information; VAD: voice activity detection; bbox: bounding box; # spk: maximum number of speakers present in the scene simultaneously; -: not applicable. The unit of the resolution column is pixels. Columns with gray shading indicate information that is required for our experiments.

	Sensors							Annotation				Content
	Audio			Video			col.	cal.	VAD	3D	Image	# spk
Dataset	# m	f_a	CA	# c	resolution	fps						
AVTRACK-1 [38]	4	44.1	-	1	640 × 480	25	C	-	C	-	active speaker(s), upper-body bbox	2
AVASM [39]	2	44.1	-	2	N/A	-	C	-	-	-	loudspeaker position	1
AVDIAR [12]	6	48	-	2	1920 × 1200	25	C	C	C	-	head, upper-body bbox	4
RAVEL [40]	4	48	-	2	1024 × 768	15	C	C	C	C	speaker bbox	5
CAVA [41]	2	44.1	-	2	1024 × 768	25	C	C	-	C	corner points	5
SPEVI [42]	2	44.1	-	1	360 × 288	25	C	-	-	-	face bbox	2
AMI [43]	14+	48	C	2+	720 × 576	25	-	-	C	C	head, face bbox, hand	N/A
CHIL [44]	88	44.1	-	5	1024 × 768	30	-	C	C	C	face bbox; head, eyes, nose position	5
AV16.3 [45]	16	16	C	3	360 × 288	25	-	C	C	C	mouth, head position	3
CAV3D (ours)	8	96	C	1	1024 × 768	15	C	C	C	C	mouth position	3

speakers who also move outside the FoV. Compared to our preliminary work ([33, 46]), this paper presents a new fusion strategy, the extension to multiple targets, a joint model of the likelihood functions through a repulsion mechanism, a multi-part color matching and an in-depth experimental analysis. We also contribute a new annotated audio-visual dataset with up to three simultaneous speakers recorded by a circular microphone array and a co-located camera.

II. Background

In this section, we discuss and compare audio-visual datasets and trackers. Depending on the relative position of the microphones and the camera(s), audio-visual datasets can be classified as co-located or spatially distributed.

AVTRACK-1 [38], AVASM [39], AVDIAR [12], RAVEL [40] and CAVA [41] were captured with *co-located audio-visual sensors* mounted on a dummy head, recording speakers talking in turns and, occasionally, simultaneously. AVTRACK-1, AVASM, and AVDIAR have image-plane annotations only. In AVTRACK-1 [38], speakers move slowly inside the FoV, close to the platform and mostly facing the camera. In AVASM [39], the sound source is a stationary loudspeaker that emits white noise or speech from different positions. AVDIAR [12] includes a multi-party dialog with speakers moving while turning their heads towards other participants, rather than facing the platform. RAVEL [40], was designed for human-robot interaction tasks and therefore the movements of the speakers are limited and very close to the platform. CAVA [41] uses one microphone pair and therefore supports only the study of limited audio processing functionalities, such as azimuth estimation. The scenario considered here mimics natural head movements of an active perceiver that also pans or moves around, and joins different small groups of people chatting. SPEVI [42] uses Stereo Audio and Cycloptic Vision (STAC) sensors [18], which consist of two microphones mounted on a 95-cm long bar with a camera in the middle. Audio direction information can be mapped onto the image plane through the audio-visual sensor's geometric relationship, without calibration information. However, the size of the platform limits the range of its possible applications.

CHIL [44], AMI [43], and AV16.3 [45] were captured with *spatially distributed audio-visual sensors*. CHIL [44] recorded meetings and seminars in different rooms with four corner-cameras and one ceiling-camera, and a variety of acoustic sensors, including three 4-element table-top microphones, three 4-element T-shaped arrays and a 64-element linear array. Annotations include the centroid of the head, the position of nose and eyes, the face bounding box on the image plane, as well as the position in 3D. AMI [43] was collected in three meeting rooms, each equipped with cameras at the corners or on the ceiling, and with an 8-element circular array and a circular or linear compact array. A close-up camera was also used for each participant. Image-plane annotations are available as well as the location of people when seated. AV16.3 [45] is commonly used for audio-visual person tracking [1, 10, 33] and was recorded in one AMI meeting room, with multiple simultaneous speakers captured by three cameras on the walls and two circular microphone arrays on a table. Annotations include head bounding boxes on the image plane and mouth positions in 3D, as well as voice activity detection labels.

Tab. I compares audio-visual datasets in terms of sensor types and configuration, annotation and content.

Audio-visual trackers operate on the image plane, on a plane parallel to the ground, or in 3D. Acoustic information, such as Direction of Arrival (DoA) estimates, may be derived independently to assist color-based trackers by re-weighting visual posteriors *on the image plane* [1, 48, 49]. However, due to the non-stationarity of speech signals, this approach is subject to audio-estimate inaccuracies. These inaccuracies can be dealt with a Kalman Filter (KF) to validate the measurements with a Gaussian reliability window based on audio-visual correspondence [18]. However, reverberation and the absence of a direct acoustic path may introduce errors in the DoA projection on the image plane. As an alternative, an audio-visual alignment method can be trained to map onto the image binaural spectral features extracted from a microphone pair [12, 28]. Audio features are then combined with a multi-person visual tracker [15], where a semi-supervised Gaussian mixture model assigns observations to targets. Similarly, the color descriptor of bounding boxes can be combined with

TABLE II

Audio-visual trackers, their processing and fusion methods. KEY – Ref: reference; co.: co-located platform; mic.: microphone array information; cam.: camera information; loc.: localization; trk.: tracker; MP: Microphone Pair (length in *cm*); CA: Circular microphone Array (diameter in *cm*); LA: Linear microphone Array (length in *cm*); TA: T-shaped microphone Array (length in *cm*); #s: number of cameras on the sensing platform; #w: number of cameras on the wall; DoA: Direction of Arrival; TDoA: Time Difference of Arrival; GCC-PHAT: Generalized Cross Correlation with Phase Transform; SSM: Sam Sparse Mean [47]; RTF: Relative Transfer Function; ILD: Interaural Level Difference; IPD: Interaural Phase Difference; GCF: Global Coherence Field; v-GCF: video-assisted Global Coherence Field;H : color histogram;S color spatiogram; MSM: Multi-body Shape Model; SIFT: Scale-Invariant Feature Transform; KF: Kalman Filter; EKF: Extended Kalman Filter; PF: Particle Filter; MOT: Multiple Object Tracking; ILDA: Incremental Linear Discriminant Analysis; CAMShift: Continuously Adaptive Mean Shift; PHD: Probability Hypothesis Density filter; PSO: Particle Swarm Optimization; GM: Graphical Model; N/A: information Not Available; -: not applicable.

Ref	**Space**	**Sensor**			**Audio Processing**			**Video Processing**			**Fusion**	
		co.	**mic.**	**cam.**	**loc.**	**features**	**trk.**	**detection**	**features**	**trk.**	**level**	**method**
[1]	Image	-	1 CA (20)	1w	DoA	SSM	-	-	H_{HSV}	-	hybrid	PF
[48]	Image	-	2 CA (20)	1w	DoA	SSM	-	-	H_{HSV}	-	hybrid	PHD
[49]	Image	-	1 CA (20)	1w	DoA	SSM	-	-	H_{HSV}	Mean-shift	hybrid	PHD
[18]	Image	C	1 MP (95)	1s	DoA	GCC-PHAT	KF	change	H_{RGB}	-	late	PF
[12]	Image	C	1 MP (12), 1 LA (22.6), 2 CA (20)	1s	TDoA	RTF	-	upper body	$H_{N/A}$	MOT tracker [15] (tracklet + ILDA)	hybrid	GM
[28]	Image	C	1 MP (12)	1s	DoA	ILD, IPD	-	N/A	$H_{N/A}$	MOT tracker [15] (tracklet + ILDA)	hybrid	GM
[37]	Image	C	1 MP (12)	2s	-	ILD, IPD	-	person	H_{RGB}	-	hybrid	GM+KF
[5]	Ground	-	4 MP (N/A)	1w	TDoA	GCC-PHAT	EKF	-	$H_{N/A}$	Mean-shift [50], KF	late	KF
[17]	Ground	-	5 MP (95)	5s	DoA	GCC	-	person	-	-	late	KF
[19]	3D	C	1 MP (47)	2s	DoA	GCC-PHAT	-	-	H_{HSV}	CAMShift	late	PSO
[30]	3D	-	3 TA (40)	4w	-	GCC-PHAT	-	foreground, face upper body	-	PF	late	PF
[4]	3D	-	7 TA (40)	4w	-	GCC-PHAT	-	-	H_{RGB}, MSM	-	hybrid	PF
[9]	3D	-	1 LA (126), 3 TA (40)	5w	-	GCC-PHAT	PF	face	H_{RGB}	CAMShift, KF	late	PF
[10]	3D	-	2 CA (30)	1-2w	TDoA	GCC	-	face	H_{RGB}	head tracker [51]	late	PF
[31]	3D	-	N/A	N/A	-	GCF	-	-	H_{RGB}, MSM	PF	hybrid	PF
[34]	3D	-	1 CA (20)	3w	DoA	SSM	-	face	H_{Hue}, SIFT	-	hybrid	PF
[33]	3D	-	1 CA (20)	1w	3D	GCF	-	upper body	H_{RGB}	-	late	PF
ours	3D	C	1 CA (20)	1s	-	v-GCF	-	face	S_{HSV}	-	hybrid	PF

binaural features for multi-speaker tracking using individual KFs [37]. Spatially distributed microphone pairs and STAC sensors can be used for tracking on a *plane parallel to the ground*, without explicit height estimation [5, 17]. Individual KFs can be used on each modality prior to fusion by PF [5]. Moreover, DoA estimates can be used to estimate trajectories in regions unobserved by cameras [17]. Spatially distributed sensors facilitate tracking in *3D*. Signals from multiple microphone pairs and cameras can be processed independently and then fused in a PF [10]. DoA estimates are projected to the multi-camera views to initialize a 3D visual tracker [34]. Other strategies rely on the existence of a subset of sensors providing a direct audio-visual observation of the objects [4, 9, 30, 31]. Finally, a compact platform with a microphone pair and a stereo camera can be used to combine the multi-modal features with confidence measurements using Particle Swarm Optimization [19]. However, this approach needs objects to be always inside the overlapping FoV of the stereo pair.

Fusion is a key component of multimodal trackers [52]. With *late fusion*, decisions are first obtained from individual modalities and then combined, thus making the final result sensitive to errors in each individual modality [5, 9, 10, 17–19, 30, 33] . *Early fusion* would integrate features immediately after their extraction [53], for example by concatenation. However, to the best of our knowledge, no audio-visual tracker uses early fusion, mainly due to their different working spaces [2].

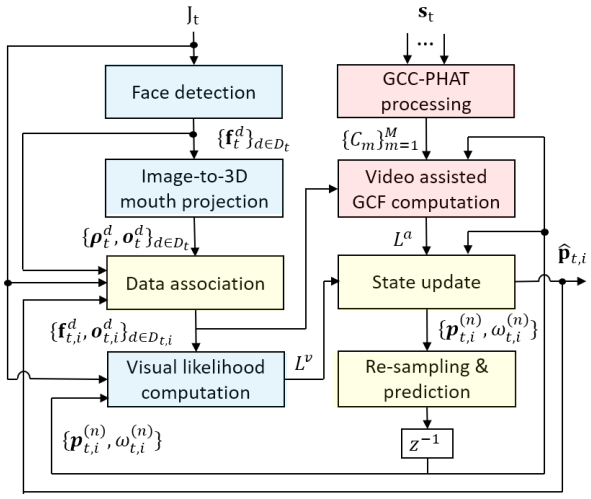


Fig. 1. Block diagram of AV3T, the proposed audio-visual 3D tracker. The blue blocks represent the computation of the visual likelihood (Sec. III-A). The red blocks represent the computation of the audio likelihood (Sec. III-B). The yellow blocks represent the audio-visual tracking (Sec. IV).

Finally, with *hybrid fusion* modalities interact with each other before the final fusion stage [1, 4, 12, 28, 31, 34, 37, 48, 49]. Tab. II compares audio-visual trackers in terms of tracking space, sensor types, audio-visual processing, and fusion strategy.

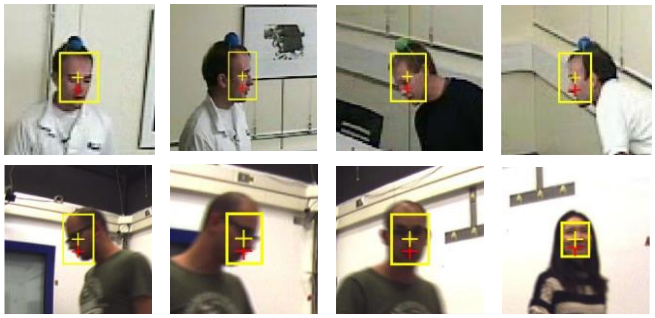
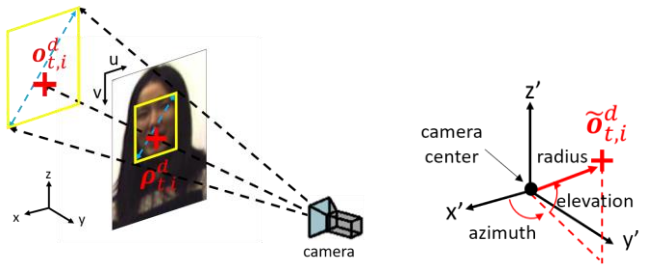


Fig. 2. + sample mouth position estimates from face detections under varying poses; + center of the bounding box.



(a) (b)

Fig. 3. (a) Image-to-3D mouth projection. $\boldsymbol{\rho}^d_{t,i}$: d-th estimated mouth position of target i at time t on the image plane; $\mathbf{o}^d_{t,i}$: d-th projected mouth estimate in 3D. Yellow: face bounding box on the image plane and its projection in 3D; cyan dashed line: diagonal of the bounding box (used for distance estimation); red cross: estimated mouth position; (x, y, z): 3D world coordinates; (u, v): image coordinates; (b) camera's spherical coordinates. (x', y', z'): shifted world coordinates at the camera center; $\tilde{\mathbf{o}}^d_{t,i}$: the counterpart of $\mathbf{o}^d_{t,i}$ in the spherical coordinates of the camera.

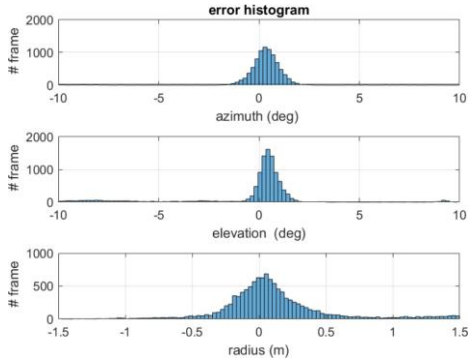


Fig. 4. Sample distribution of the 3D mouth-estimate errors in the image-to-3D projection process.

III. Audio-visual Processing

We aim to track the mouth position, $\mathbf{p}_{t,i}$, of each target $i \in I$ over time, t, in 3D world coordinates, given audio signals, $\mathbf{s}_t$, captured by a small K-element circular microphone array, and video frames, J_t, captured by a monocular camera. We use a probabilistic tracking framework, with $\mathbf{p}_{t,i}$ estimated as an expectation, given the past observations:

$$\mathbf{p}_{t,i} \approx \hat{\mathbf{p}}_{t,i} = \mathsf{E}[\mathbf{p} | J_{0:t}, \mathbf{s}_{0:t}], \tag{1}$$

where $\mathbf{p}$ is a generic 3D point.

Fig. 1 shows the block diagram of the proposed tracker. We assume the audio-visual signals to be synchronized, the sensors calibrated, and the number of targets $|I|$ known and constant ($|\cdot|$ indicates the cardinality of a set).

A. Visual likelihood

Let $\mathbf{f}^d_{t,i} = (u, v, w, h)^\top$ be the bounding box of the d-th detected face of target i at time t, where $d \in D_{t,i}$ and $D_{t,i}$ is the set of face detections associated to target i; (u, v) is the position of the top left corner of the box on the image plane, and (w, h) is its width and height ($^\top$ denotes transpose). Given $\mathbf{f}^d_{t,i}$, we geometrically extract the mouth position as:

$$\boldsymbol{\rho}^d_{t,i} = [\mathbf{I}_{2\times 2}, \Lambda]\mathbf{f}^d_{t,i}, \tag{2}$$

where $\Lambda = diag(0.5, 0.75)$ is a diagonal matrix and $\mathbf{I}_{2\times 2}$ is a 2-dimensional identity matrix. Fig. 2 shows sample mouth position estimates from face detections under different poses.

We then derive the 3D mouth position estimate, $\mathbf{o}^d_{t,i}$, with the pinhole camera model:

$$\mathbf{o}^d_{t,i} = \Psi(\boldsymbol{\rho}^d_{t,i}; w, h, W, H), \tag{3}$$

where W and H are the expected width and height of the face bounding box in 3D and Ψ is the image-to-3D projection [54]. To estimate the scaling factor, we use the length of the diagonal, $\sqrt{W^2 + H^2}$, which is less sensitive to changes in face orientation than the width of the bounding box (Fig. 3(a)).

The main uncertainties of 3D mouth position estimation from a monocular camera are in the range (distance), due to the inaccuracy in the hypothesized sizes of a face (W, H). We model these uncertainties, especially distinguishing the range estimates from azimuth and elevation, by designing the visual likelihood in spherical coordinates, originated at the geometrical center of the camera (Fig. 3(b)).

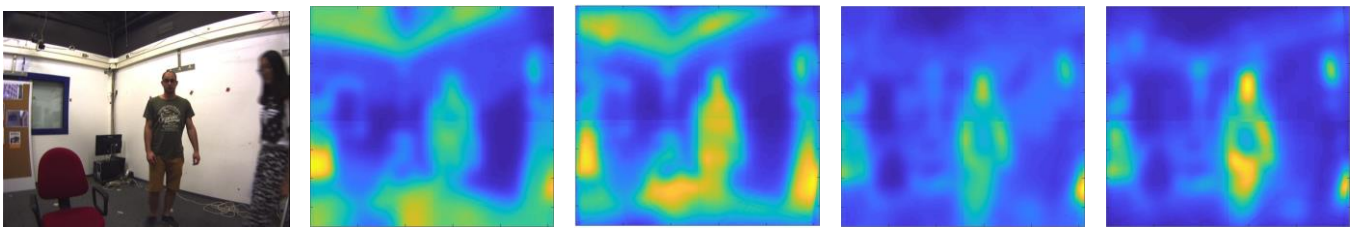
Let $\tilde{\mathbf{o}}^d_{t,i}$ be the estimated mouth position $\mathbf{o}^d_{t,i}$ in the camera's spherical coordinates. Assuming a Gaussian distribution on the accuracy of the 3D estimate (see Fig. 4), we evaluate the likelihood at $\mathbf{p}$ as:

$$L^v_{\text{det}}(J_t \mid \mathbf{p}) = \sum_{d \in D_{t,i}} \exp\left(-\left(\tilde{\mathbf{o}}^d_{t,i} - \tilde{\mathbf{p}}\right)^\top \Sigma_v^{-1} \left(\tilde{\mathbf{o}}^d_{t,i} - \tilde{\mathbf{p}}\right)\right), \tag{4}$$

where $\tilde{\mathbf{p}}$ is the equivalent of $\mathbf{p}$ in the camera's spherical coordinates and Σ_v is a diagonal matrix whose elements represent different estimation accuracies. Note that we constrain in $[0, \pi]$ the absolute difference between any two angles.

When a face detection is unavailable, we revert to a color-based generative model to find the most likely target position on the image. To this end, we use a color spatiogram [55], a histogram augmented with spatial means and covariances for each histogram bin, which provides a more discriminative target description. To better separate the target from the background (see Fig. 5) we use the HSV color space [1, 32].

To extract the spatiogram to evaluate the likelihood of $\mathbf{p}$, we need to define the image region corresponding to a face in an hypothesized 3D location $\mathbf{p}$. To this end, we create in $\mathbf{p}$



(a) test image (b) histogram (RGB) (c) spatiogram (RGB) (d) histogram (HSV) (e) spatiogram (HSV)

Fig. 5. Comparison of color descriptors. (a) Sample image. (b-e) Corresponding histogram and spatiogram in RGB and HSV. Yellow (blue) indicates a higher (lower) probability of target (mouth) presence at the corresponding pixel position.

a 3D hyper-rectangle $\mathbf{b}(\mathbf{p}; W, H)$ that is perpendicular to the ground (assuming an upstanding pose) and oriented towards the camera (to indicate a profile view in that pose). This 3D hyper-rectangle at $\mathbf{p}$ is then projected onto the image plane to obtain the rectangular bounding box of the person's face in $\mathbf{p}$:

$$\mathbf{v} = \Phi[\mathbf{b}(\mathbf{p}; W, H)], \tag{5}$$

where Φ indicates the 3D-to-image projection [54]. Finally, we compare the color feature surrounded by the bounding box $\mathbf{v}$ and a reference image of the target, updated from the last associated face detection $\mathbf{f}^{d}_{t',i}$ whose 3D mouth estimate $\mathbf{o}^{d}_{t',i}$ is closer to the averaged target position estimate $\hat{\mathbf{p}}_{t'|\Delta t,i}$ during the time interval $[t'-\Delta t, t'-1]$. The similarities of two spatiograms are measured using [56], which is derived from the Bhattacharyya coefficient:

$$L^{v}_{\mathrm{HSV}}(J_t \mid \mathbf{p}) = \sum_{b=1}^{B} \sqrt{r^{b}_{\mathbf{v}} r^{b}_{\mathbf{f}}} \left| 8\pi |\Sigma^{b}_{\mathbf{v}} \Sigma^{b}_{\mathbf{f}}|^{\frac{1}{4}} \mathcal{N}(\mu^{b}_{\mathbf{v}} | \mu^{b}_{\mathbf{f}}, 2(\Sigma^{b}_{\mathbf{v}} + \Sigma^{b}_{\mathbf{f}})) \right|, \tag{6}$$

where $r^{b}_{\mathbf{v}}$ indicates the b^{th} bin of the spatiogram computed at the image region, surrounded by $\mathbf{v}$ and $b = 1, \ldots, B$. $\mu^{b}_{\mathbf{v}}$ and $\Sigma^{b}_{\mathbf{v}}$ are spatial mean and covariance of the image pixels surrounded by $\mathbf{v}$ and belonging to the b^{th} bin. Analogous definitions apply to $r^{b}_{\mathbf{f}}$, $\mu^{b}_{\mathbf{f}}$ and $\Sigma^{b}_{\mathbf{f}}$ where the dependency of $\mathbf{f}$ on d, t and i is dropped for simplicity.

Let us define a target i as *visible* when it is inside the FoV and unoccluded by any other tracked targets, $\tilde{i}$. When there is no detection ($D_{t,i} = \varnothing$) and the target is *not visible*, the likelihood follows a uniform distribution, $\mathcal{U}$. If $J^{0.9}$ is a rectangular crop corresponding to the central 90% region of the image, the first condition (inside the FoV) can be expressed as:

$$\mathbf{p}'_{t|\Delta t,i} \in J^{0.9}, \tag{7}$$

where $\mathbf{p}'_{t|\Delta t,i}$ is the averaged target estimate on the image plane during the time interval $[t-\Delta t, t-1]$. The second condition (unoccluded) is that the distance between the position estimate of target i and any other target $\tilde{i}$ on the image plane is farther than half-diagonal-size of the last face detection. Otherwise, the target closer to the camera is considered unoccluded.

Finally, we define the visual likelihood as:

$$L^{v}(J_t \mid \mathbf{p}) = \begin{cases} L^{v}_{\mathrm{det}}(J_t \mid \mathbf{p}) & \text{if } D_{t,i} \neq \varnothing \\ L^{v}_{\mathrm{HSV}}(J_t \mid \mathbf{p}) & \text{else if } \eta_{t,i} = 1, \\ \mathcal{U}(\mathbf{p}) & \text{otherwise.} \end{cases} \tag{8}$$

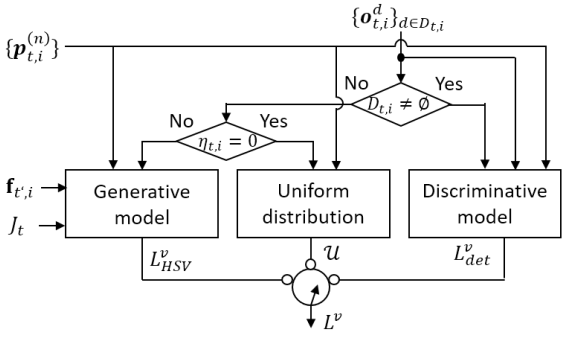


Fig. 6. Details of the visual likelihood computation block in Fig. 1.

where $\eta_{t,i}$ is a flag set to 1 when target i is *visible* at time t.

The process of the overall visual likelihood computation is shown in Fig. 6.

B. Audio likelihood

Acoustic source localization can be accomplished by combining information from M microphone pairs to obtain an acoustic map (GCF [25]) that represents the plausibility of an active sound source being at a given spatial position. The GCC-PHAT [20, 21] at microphone pair m at time t is:

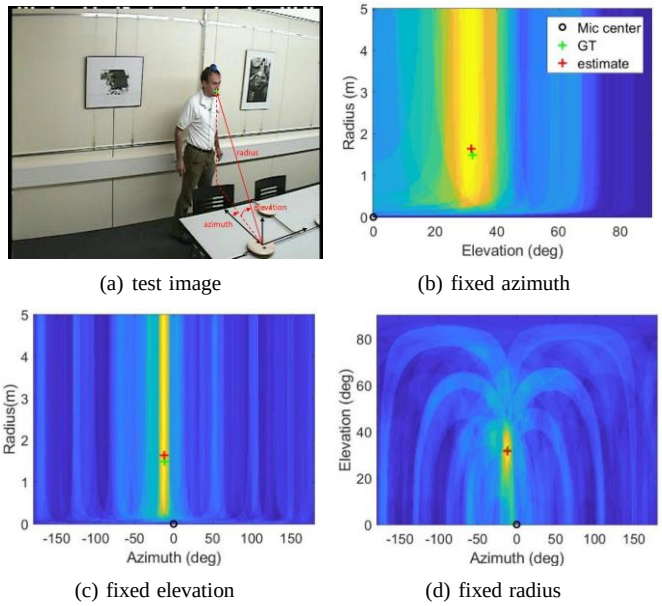
$$C_m(\tau, t) = \int_{-\infty}^{+\infty} \frac{S_{m_1}(t,f) S^{*}_{m_2}(t,f)}{|S_{m_1}(t,f)| \cdot S^{*}_{m_2}(t,f)} e^{j2\pi f\tau} df, \tag{9}$$

where f indicates frequency, S_{m_1} and S_{m_2} are the Short-Time-Fourier-Transform (STFT) computed at the m^{th} pair with microphones m_1 and m_2, τ indicates the inter-microphone time delay and $*$ is the complex conjugate. Ideally, $C_m(\tau, t)$ exhibits a peak when τ equals the actual Time Difference of Arrival (TDoA). The GCF value in $\mathbf{p}$ is thus derived from the GCC-PHAT computed at all the M microphone pairs:

$$g(\mathbf{p}, t) = \frac{1}{M} \sum_{m=1}^{M} C_m(\tau_m(\mathbf{p}), t), \tag{10}$$

where $\tau_m(\mathbf{p})$ is the TDoA expected at the m^{th} microphone pair if the emitting source is in $\mathbf{p}$. The position of the sound emission can be estimated by picking the peak of $g(\mathbf{p}, t)$.

However, the performance of GCF is sensitive to the microphone array configuration. A small planar microphone array (as in our case) cannot estimate the speaker height without



(a) test image (b) fixed azimuth

(c) fixed elevation (d) fixed radius

Fig. 7. Sample GCF results in spherical coordinates. (a) The reference spherical coordinates (in red), whose origin is at the center of the microphone array; (b)-(d): GCF computed in 2D when fixing, respectively, azimuth, elevation and radius at the ground truth The smaller the yellow region, the more certain the GCF (○ microphone array center; + SSL estimate; + GT).

microphone pairs spanning the vertical dimension. Moreover, the small inter-microphone distance compounds the errors in distance estimation. Fig. 7 exemplifies this problem with GCF in spherical coordinates, whose origin is at the center of the array. Each subplot shows a 2D acoustic map when the third coordinate is given by the ground truth. It is evident that the localization would be very accurate if the speaker-array distance were known (Fig. 7(d)).

To address this problem we propose a novel localization approach that uses visual 3D position estimates of mouths to suggest the most likely speaker-height plane to reduce the audio working space from 3D to 2D. This video-assisted GCF, $g_v(\mathbf{p}, t)$, is formulated as:

$$g_v(\mathbf{p}, t) = \frac{1}{M} \sum_{m=1}^{M} C_m \quad \tau_m(\mathbf{p}|o_{t',i}^{d,z}), t \quad , \tag{11}$$

where $\mathbf{p}|o_{t',i}^{d,z}$ is the projection of a generic 3D point $\mathbf{p}$ on the 2D plane defined by the height of the mouth, $q_{t',i}^{d,z}$, estimated from the last associated face detection $\mathbf{f}_{t',i}^{d}$ (that occurred at frame t'), and $\tau_m(\mathbf{p}|o_{t',i}^{d,z})$ is the corresponding TDoA. Thus, we define the audio likelihood as:

$$L(\mathbf{s}_t \mid \mathbf{p}) = \begin{cases} g(\mathbf{p}, t) & \text{if } \max g(\mathbf{p}, t) \quad \theta^a \\ \mathrm{U}(\mathbf{p}) & \text{otherwise,} \end{cases} \tag{12}$$

where θ^a is a threshold relying on the likelihood peak value to detect speech activity at the candidate states [4, 31, 33, 57].

IV. 3D Audio-visual tracking

The individual likelihoods proposed in the previous section support multi-modal tracking through a Sequential Importance Resampling Particle Filter (SIR-PF) [58]. We instantiate an SIR-PF for each target i with a repulsion mechanism that ensures that multiple filters do not collapse on a single target.

Let $\mathbf{p}_{t,i}^{(n)}$ be particle n ($n = 1, \ldots, N$) of target i at time t, whose state is $\mathbf{p} = (x, y, z)^{\intercal}$, where x, y, z are the world coordinates. Assuming conditional independence across the modalities, the multimodal likelihood equals the product of the individual ones [4, 5, 10, 18, 31, 33]. We, therefore, compute the weight of each particle as:

$$\omega_{t,i}^{(n)} \propto \left[L^a(\mathbf{s}_t|\mathbf{p}_{t,i}^{(n)}) L^v(J_t|\mathbf{p}_{t,i}^{(n)}) \right] \psi_{t,i}^{(n)}, \tag{13}$$

where L^v and L^a are defined in Eq. 8 and Eq. 12. Distance between targets can be used to overcome observations being corrupted during occlusions [34]. We want to force particles of a target near another target to be suppressed and to favour the resampling of particles farther away in the state space. The result of this process is that particles will appear as if they were drifting away from the other target as if a repulsion force was applied. The term $\psi_{t,i}^{(n)}$ implements this repulsion process on the particles. When their distance is smaller than b, particle weights are reduced as:

$$\psi_{t,i}^{(n)} = 2^{\frac{1}{b}\min(\min_{\tilde{i}} ||\mathbf{p}_{t,i}^{(n)} - \hat{\mathbf{p}}_{t-1,\tilde{i}}||_2, b)} - 1, \tag{14}$$

where b is the minimum allowed distance between mouths, $\tilde{i}$ indicates the identity of any other targets. $\psi_{t,i}^{(n)}$ can be seen as a notch filter applied to the likelihood. Weights are normalized as $\sum_{n=1}^{N} \omega_{t,i}^{(n)} = 1$.

To encourage particles to explore the state space and to facilitate target re-identification after a target loss, the propagation in the prediction step is defined as:

$$\mathbf{p}_{t,i}^{(n)} = \mathbf{p}_{t-1,i}^{(n)} + 3^{\kappa}\mathbf{q}, \quad \kappa \in \{0, 1\} \tag{15}$$

where $\mathbf{q}$ is sampled from a zero-mean Gaussian with diagonal covariance matrix Σ_q. We use a higher prediction speed for low-scoring hypotheses: if a particle weight is in the lower 10%, then $\kappa = 1$; otherwise $\kappa = 0$.

Finally, the position of target i at time t, $\hat{\mathbf{p}}_{t,i}$, is estimated as:

$$\hat{\mathbf{p}}_{t,i} = \sum_{n=1}^{N} \omega_{t,i}^{(n)} \mathbf{p}_{t,i}^{(n)}, \tag{16}$$

which is an approximation to the expectation in Eq. 1. At each iteration, new particles are drawn from the discrete set $\{\mathbf{p}_{t,i}^{(n)}, \omega_{t,i}^{(n)}\}_{n=1}^{N}$ using weighted re-sampling [58]. Algorithm 1 summarizes the proposed tracker.

We associate a detected face bounding box $\mathbf{f}_t^d$ to target i considering the last position estimate $\hat{\mathbf{p}}_{t-1,i}$, through a greedy strategy (as used in [59]) with a discriminative visual model that approximates an optimal single-frame solution (Algorithm 2). We first derive a matching score matrix A for each target-detection pair (i, d) using the face and the torso of the target:

$$A_t(i, d) = L_{\text{det}}^{v}(J_t|\hat{\mathbf{p}}_{t-1,i}) [L_{\text{HSV}}^{v}(J_t|\mathbf{o}_{t,i}^{d}) + L_{\text{HSV}}^{v}(J_t|\mathbf{o}_{t,i}^{'d})], \tag{17}$$

where $\mathbf{o}_{t,i}^{'d} = \mathbf{o}_{t,i}^{d} - 0.4\mathbf{z}$ is the 3D torso point derived from the mouth location estimate $\mathbf{o}_{t,i}^{d}$ shifted along the vertical axis $\mathbf{z}$. To evaluate $L_{\text{HSV}}^{v}(J_t|\mathbf{o}_{t,i}^{'d})$ we use a torso spatiogram model of

Algorithm 1: The AV3T tracker

Initialize:
$t, i, \lambda, \Delta t, \theta^a, b, \{\boldsymbol{p}_{t_0,i}^{(n)}, \omega_{t_0,i}^{(n)}\}_{n=1}^{N}, T, \Lambda, W, H, N, B, M, \mathbf{q}$
while $t \leq T$ **do**
  $\boldsymbol{\rho}_t^d = [\mathbf{I}_{2\times 2}, \Lambda]\mathbf{f}_t^d$ %mouth estimate
  $\mathbf{o}_t^d = \Psi(\boldsymbol{\rho}_t^d; w, h, W, H)$ %image-to-3D projection
  compute $D_{t,i}$ with **Algorithm 2** %data association
  $\boldsymbol{p}_{t,i}^{(n)} = \boldsymbol{p}_{t-1,i}^{(n)} + 3^{\kappa}\mathbf{q}$ %propagate particles
  if $D_{t,i} \neq \emptyset$ **then**
    $L^v = L_{det}^v(\boldsymbol{J}_t|\boldsymbol{p}_{t,i}^{(n)})$ %discriminative model
  else
    if *target is visible* **then**
      $L^v = L_{HSV}^v(\boldsymbol{J}_t|\boldsymbol{p}_{t,i}^{(n)})$ %generative model
    else
      $L^v = \mathrm{U}(\boldsymbol{p}_{t,i}^{(n)})$ %uniform distribution
    end
  end
  $L^a = g_v(\boldsymbol{p}_{t,i}^{(n)}, t)$ %video-assisted GCF
  if $L^a \leq \theta^a$ **then**
    $L^a = \mathrm{U}(\boldsymbol{p}_{t,i}^{(n)})$ %uniform distribution
  end
  $\psi_{t,i}^{(n)} = 2^{\frac{1}{b}\min(\min_i ||\boldsymbol{p}_{t,i}^{(n)} - \hat{\mathbf{p}}_{t-1,i}||_2, b)} - 1$ %distance function
  $\omega_{t,i}^{(n)} \propto L^a L^v \psi_{t,i}^{(n)}$
  $\omega_{t,i}^{(n)} = \omega_{t,i}^{(n)} / \sum_{n=1}^{N} \omega_{t,i}^{(n)}$ %weights normalization
  $\hat{\mathbf{p}}_{t,i} = \sum_{n=1}^{N} \omega_{t,i}^{(n)} \boldsymbol{p}_{t,i}^{(n)}$ %3D position estimate
  Re-sample $N\omega_{t,i}^{(n)}$ particles from $\{\boldsymbol{p}_{t,i}^{(n)}, \omega_{t,i}^{(n)}\}_{n=1}^{N}$
  $t = t + 1$
end

Algorithm 2: Greedy data association

I: set of target indices
D_t: set of detections at time t

$A_t(i, d)$: score for each target-detection pair (i, d)
$D_{t,i}$: set of detections associated to target i

compute $A_t(i, d)$ with Eq. 17, $\forall i \in I$, $\forall d \in D_t$
while $I \neq \emptyset \wedge D_t \neq \emptyset$ **do**
  $(i^*, d^*) = \underset{i\in I, d\in D_t}{\mathrm{argmax}} A_t(i, d)$
  if (i^*, d^*) *satisfies Eq. 18* **then**
    $D_{t,i^*} \leftarrow d^*$
    $I = I \setminus i^*$ %$\setminus$ indicates exclude
  end
  $D_t = D_t \setminus d^*$
end

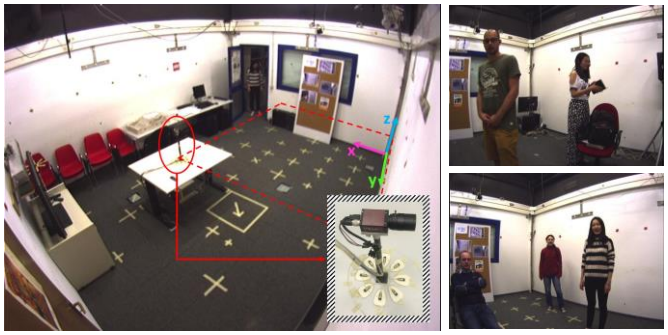


Fig. 8. Left: Recording environment of the CAV3D dataset. Yellow markers on the ground are used to calibrate the corner cameras. A close-view of the co-located sensor (surrounded by the red ellipse) is inserted on the bottom right of the picture: the camera is around 48 *cm* above the microphone array. The region covered by the camera's FoV is within the red dashed line. The world coordinates *x, y, z* are originated at the top-right room corner, which are marked as magenta, green and blue respectively; Right: keyframes.

target i instead of its head model. Then, we iteratively select the pair with the maximum score until no further valid pair is available. This data association process has a lower computational cost than the Hungarian algorithm [60], which makes associations in polynomial time; Multiple Hypothesis Tracking (MHT) [61], which considers multiple possible associations over the several past frames and has the highest complexity; and Joint Probabilistic Data Association Filter (JPDAF) [62], whose complexity grows exponentially with the number of targets.

A gating stage ensures that the associated detection is within a neighborhood of the target [63]:

$$||\boldsymbol{\rho}_{t,i}^d - \mathbf{p}_{t|\Delta t,i}||_2 \leq \lambda \sqrt{w_{t,i}^2 + h_{t,i}^2}, \quad (18)$$

where λ controls the size of the neighborhood.

V. The CAV3D dataset

Most audio-visual datasets focus on image-plane tracking [12, 38, 39, 42]. Datasets with 3D ground truth are either collected from spatially distributed sensors [43–45], or contain slowly moving or stationary targets [40, 41]. To overcome the limitations of these datasets [64], we collected CAV3D, a dataset recorded from a Co-located Audio-Visual platform for 3D tracking. CAV3D contains up to three simultaneous speakers captured in a $4.77 \times 5.95 \times 4.5$ m room. The sensing platform, placed on a table, consists of a camera co-located with an 8-element circular microphone array ($K = 8$) of 20-*cm* diameter. The 8-channel audio signals were recorded at 96 kHz (24 bits). Videos (768×1024 pixels) were recorded at 15 frames per second (fps) with a CCD color camera whose FoV is about 90°. The recording environment and keyframes are shown in Fig. 8.

The dataset includes 20 sequences whose duration varies from 15 to 80 *s* and are organized in three sessions, namely: CAV3D-SOT (9 sequences with a single speaker), CAV3D-SOT2 (6 sequences with a single speaker but two people present in the scene) and CAV3D-MOT (5 sequences with simultaneous speakers). The speakers perform different actions, undergo occlusions, have non-frontal views, enter/exit the camera's FoV, and have long non-speech periods. The room has strong reverberation (about 0.7 s [65]), background noise (*e.g.* from an air conditioner), human-made noise (*e.g.* clapping and stomping).

Parameters of the co-located platform sensor models were calibrated in the following way. We used markers with known, manually measured 3D position to align the camera model

to the world coordinate system (see Fig. 8). The camera follows a conventional pinhole model with radial distortion correction and we used standard procedures from OpenCV to estimate their parameters in three steps: (i) intrinsic parameters from a sequence of images showing a planar chessboard pattern in front of the camera at different poses; (ii) extrinsic parameters from 3D-2D point correspondences of 3D scene markers and their manually annotated 2D image coordinates; and (iii) a non-linear multi-view optimization that is detailed in the next paragraph. As for the audio acquisition chain, 3D microphone positions in the world reference system were manually measured with cm precision. Pre-amp gains were tuned to ensure the same sensitivity and dynamics across different channels. A complete description of the calibration process with implementation details is provided at the dataset webpage that is referenced at the end of next paragraph.

The speech/non-speech frames were annotated with Transcriber[1], including the speaker identities of individual segments. We also equipped the room with four additional cameras at the top corners to facilitate the annotation process. These cameras were hardware-triggered to ensure frame-level synchronization with the audio-visual sensing platform. We annotated mouth positions on each frame of each additional camera: frames were displayed sequentially in a graphical user interface, with a superimposed 50×50 cropped candidate region centered at the position annotated at the previous frame, for the annotator to update the replicated mouth location with a mouse click. Next, using scene markers with known 3D position, we initialized calibration parameters for each camera using Zhang's method [66]. This calibration was used to back-project to 3D rays each timestamped annotation tuple. We then computed the spatial least-squares intersection of the rays using Singular Value Decomposition. This intersection provides an estimate of the 3D mouth location associated with each annotation tuple. Finally, we run an optimization based on Sparse Bundle Adjustment [67] to obtain 3D trajectories, accurate calibration, and an algorithmic correction of the manual annotations by minimizing the re-projection error on all views and sequences simultaneously. This joint optimization provides more accurate, high-quality annotations on the image plane and is available with the dataset for evaluation. The dataset is available to the research community[2].

VI. Results

We compare the proposed AV3T with [10], [33] and with individual audio and video pipelines. In addition to 3D tracking, we also consider performance on the image plane and compare our results with the audio-assisted visual tracker in [1]. As datasets we use CAV3D and AV16.3. AV16.3 has no co-located sensors but has small circular microphone arrays, cameras with calibration information and mouth ground truth location in 3D. We use the first circular microphone array (we did not observe performance changes when considering the other array) and each (of the three) corner cameras individually. We use sequences with the 3D ground truth: for Single Object Tracking (SOT) we use seq08, 11, 12; whereas for Multiple Object Tracking (MOT) we use seq18, 19, 24, 25 and 30.

A. Implementation details

AV3T detects faces with an MXNet implementation of the light Convolutional Neural Network (CNN)[3] [68]. The other AV3T parameters were defined on a small set of sequences not used for testing: the number of points in STFT is 2^{12} on AV16.3 and 2^{15} on CAV3D; the speech activity threshold, θ^a, is 0.1 for AV16.3 and 0.03 for CAV3D[4]; the number of bins per channel is $B = 8$ [1, 3, 17]; $(W, H) = (15, 20)$ cm is an approximate average size of a face's central region; the face validation parameter $\lambda = 2.5$ and the time lag $\Delta t = 3$ is set to avoid large instant tracking error on data association; the number of microphone pairs is $M = 28$ [33]; the number of particles per target is $N = 100$; the prediction matrix $\Sigma_q = diag(1,\ 1,\ 0.5)$ m/s when target is inside the camera's FoV and is divided by 10 when it is outside; the update matrix in the discriminative model is $\Sigma_v = diag(2^\circ,\ 2^\circ,\ 0.4m)$; and $b = 20$ cm indicates the minimum feasible distance between two mouth estimates under a side-by-side face situation. Note that except for the STFT points and the voice activity thresholds which depends on the sampling rate, we use the same parameters for both datasets. Because the number of targets is constant and known, filters are initialized at the ground-truth positions at time t_0 with added Gaussian noise. Given the probabilistic nature of the PF, all results are averaged over 10 runs.

B. Performance measures

We use as performance measures Track Loss Rate (TLR) and Mean Absolute Error (MAE) in 3D and on the image plane.

TLR is the percentage of frames with a track loss. We declare a target to be lost if, in 3D, the error is above 30 cm and, on the image plane, if the error is larger than 1/30 of the length of the image diagonal or if only the ground truth or the estimate is inside the FoV.

The MAE in 3D (in m) is defined as:

$$\varepsilon_{3d} = \frac{1}{|I|T}\sum_{i=1}^{|I|}\sum_{t=1}^{T}||\hat{\mathbf{p}}_{t,i} - \mathbf{p}_{t,i}||_2, \tag{19}$$

where $|I|$ is the total number of targets and T is the total number of frames. The MAE on the image plane (in pixels) is defined as:

$$\varepsilon_{img} = \frac{1}{|I|\tilde{T}}\sum_{i=1}^{|I|}\sum_{t=1}^{\tilde{T}}||\hat{\mathbf{p}}'_{t,i} - \mathbf{p}'_{t,i}||_2, \tag{20}$$

where $\tilde{T}$ is the total number of frames where both the estimates and the ground truth is inside the FoV, $\hat{\mathbf{p}}'_{t,i}$ is the estimated position of the target and $\mathbf{p}'_{t,i}$ is the ground-truth position.

[1]Transcriber: http://trans.sourceforge.net/en/presentation.php

[2]The CAV3D dataset and the code of AV3T will be available for download at: https://ict.fbk.eu/units/speechtek/CAV3D

[3]https://github.com/tornadomeet/mxnet-face

[4]Note that different audio parameters are due to different audio sampling frequencies in the two datasets.

TABLE III
MAE (m) for SSL estimates on AV16.3 and CAV3D, single speaker sequences. For simplicity, '·' represents the variables of a function. '^{s}' indicates results are computed at the speaker's ground truth height plane. Key – $g(\cdot)$: GCF computed in 3D; $L^a_{TDoA}(\cdot)$: maximum TDoA based likelihood; $g_v(\cdot)$: video-assisted GCF; $g(\cdot)^s$: GCF on ground truth height plane (*upper bound*).

	$g(\cdot)$	$L^a_{TDoA}(\cdot)^s$	$g_v(\cdot)$	$g(\cdot)^s$
CAV3D	.85	.56	.53	*.47*
AV16.3	.48	.31	.31	*.19*

To simplify the notation, we will use ε to represent either ε_{3d} or ε_{img}. Moreover, since ε would be considerably affected by the large errors due to target losses [69], we introduce ε' that denotes the MAE computed on frames when tracking is successful *i.e.*estimates are located within 30 cm in 3D from the target. In summary, TLR and ε represent the % of frames that a tracker follows the target and the precision of the position estimates, while ε' is a compound measure of the two errors.

C. Evaluation of AV3T and its components

We quantify the contribution of each component in terms of performance and compare with alternative solutions. In particular, we motivate the adoption of the acoustic map and the use of the generative visual likelihood in combination with face detections.

A TDoA-based likelihood [5, 6, 10, 18, 57] could be an alternative to our GCF likelihood (Eq. 11). For M microphone pairs, the TDoA-based likelihood is:

$$L^a_{\text{TDoA}}(\mathbf{s}_t \mid \mathbf{p}) = \exp\left(-\frac{1}{2M\sigma_\tau^2}\sum_{m=1}^{M}(\tau_m(\mathbf{p}) - \hat{\tau}_m)^2\right). \quad (21)$$

where $\hat{\tau}_m$ is the estimated TDoA corresponding to the peak of the GCC-PHAT $C_m(\tau, t)$, and the standard deviation σ_τ represents the estimation uncertainty. The TDoA likelihood estimates the most likely time delay for each microphone pair and the final results are sensitive to inaccuracies at individual pairs, especially when the speaker is far from the microphone array. Conversely, GCF postpones any decisions to when the results from all microphone pairs have been combined. Moreover, the TDoA likelihood relies on the noise standard deviation σ_τ , which is more sensitive to varying acoustic environment. Fig. 9 shows audio likelihood maps for GCF and TDoA, computed at the ground-truth speaker-height plane: GCF has a better localization accuracy than TDoA.

Tab. III compares SSL results on AV16.3 and CAV3D when speaker-height information is available. Adding a prior on speaker-height substantially increases performance (column 1 vs. column 2-4). Moreover, the video-assisted GCF likelihood outperforms the TDoA likelihood (column 2 vs column 3) without using the ground truth information, thus confirming what is shown in Fig. 9. Finally, the 3D video estimates can be used to suggest the most likely speaker-height (column 1 vs column 3), but a considerable margin is still available if the ground-truth speaker-height is used (column 4).

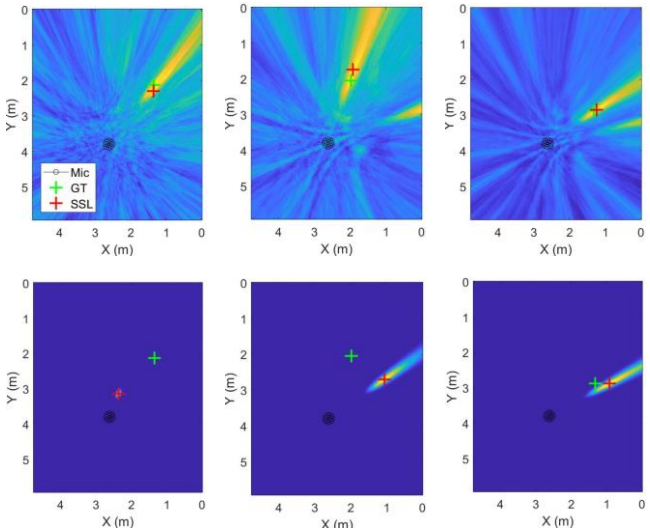


Fig. 9. Acoustic maps computed at the ground-truth speaker-height plane. First row: speaker position and pose; second row: GCF map; third row: TDoA map. Yellow (blue) corresponds to higher (lower) probability of a source being present (microphone; + SSL estimate; + ground truth).

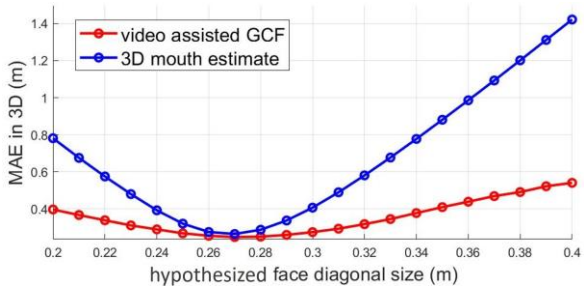


Fig. 10. Localization error for image-to-3D projection (blue) and video-assisted GCF SSL (red) when varying the length of the diagonal of the face bounding box on AV16.3.

Fig. 10 compares the localization accuracy of the 3D video estimates (blue) and the video-assisted audio estimates (red) under varying lengths of the diagonal of the face bounding box. The video estimates derived from the image-to-3D mouth projection (Eq. 3) are very sensitive to the hypothesized face size that affects the scaling factor estimation which, in turn, leads to inaccurate depth estimations. While video-assisted GCF also depends on the face size, which determines the height estimation, the corresponding sensitivity is lower.

Let the Face Detection Rate (FDR) measures the ratio between the number of frames with a detection (including true and false positives) and the overall number of frames. To validate the generative model combined with face detections in the video likelihood, Tab. IV compares video-only tracking on AV16.3 (targets always inside the camera's FoV), using the discriminative model only (VO$^-$), and both the discriminative and generative models (VO). With the generative model, TLR in 3D decreases from 62.3% to 54.28% on SOT and from 70.01% to 55.63% on MOT. Improvements are observed also

TABLE IV
Video-only tracking results on AV16.3 with the discriminative model only (VO^-) and both discriminative and generative model (VO). KEY – FDR: Face Detection Rate; TRL: Track Loss Rate; ε: MAE on all frames; ε': MAE on successfully tracked frames.

			Image		3D	
			VO^-	VO	VO^-	VO
SOT	FDR= 44.0%	TLR	41.78	8.78	62.30	54.28
		ε	37.4±9.5	8.1±.9	.74±.10	.43±.05
		ε'	6.7±.4	5.3±.1	.16±.01	.15±.01
MOT	FDR= 46.8%	TLR	59.94	14.57	70.01	55.63
		ε	50.4±14.0	15.8±6.1	.75±.14	.50±.10
		ε'	7.1±.9	5.1±.4	.16±.02	.14±.02

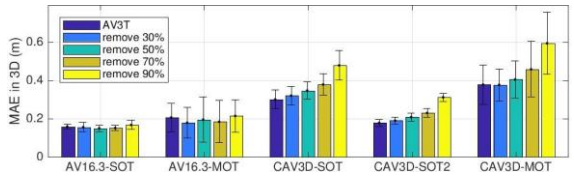
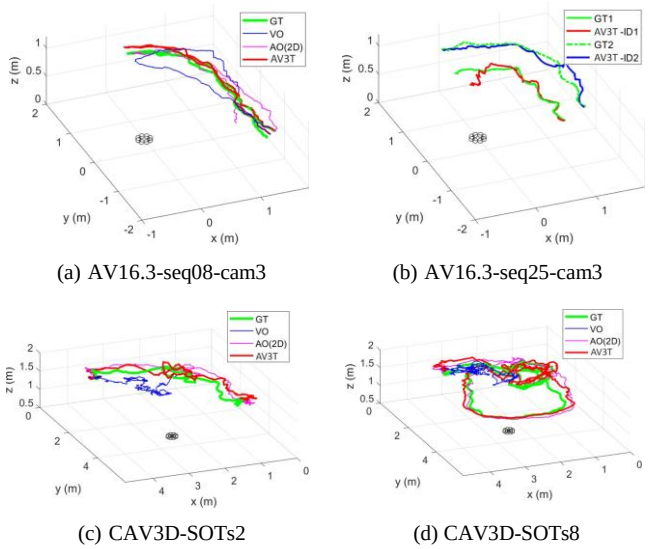


Fig. 11. The sensitivity of AV3T to the number of face detections.



(a) AV16.3-seq08-cam3 (b) AV16.3-seq25-cam3

(c) CAV3D-SOTs2 (d) CAV3D-SOTs8

Fig. 12. 3D trajectories from AV16.3 (top) and CAV3D (bottom). Green indicates the ground-truth trajectories.

in terms of MAE. (Note that for less than half of the frames a detection is available.) On the image plane, TLR decreases from 41.78% to 8.78% on SOT and from 59.94% to 14.57% on MOT. These results confirm that using only face detections is insufficient in realistic conditions.

We also investigate the sensitivity of AV3T to the number of available face detections. Fig. 11 shows the influence on the average MAE in 3D of randomly removing detections (30, 50, 70, and 90%) on CAV3D and AV16.3. The face detection rates on CAV3D-SOT, CAV3D-SOT2, CAV3D-MOT sequences are 71.0%, 99.4% and 90.1%, respectively; and equal to 44.0% on AV16.3-SOT and 46.8% on AV16.3-MOT (the higher detection rate in CAV3D is due to the higher image resolution). Removing detections in AV16.3 has little influence on the 3D tracking accuracy as the audio and video scenarios have comparable difficulty levels (and both are simpler than in the CAV3D dataset). Instead in CAV3D, where the audio scenarios are more challenging than the video ones (strong room reverberation, rapidly moving distant speakers who are not oriented towards the platform), the face detector plays an important role. Note that, however, the audio-visual results are always superior to the unimodal results (as we will see in Tab. V and Tab. VI).

Fig. 12 shows 3D trajectories on AV16.3 and CAV3D. Fig. 12(a) compares different modalities when the speaker walks forward and backward in the room. The video trajectory (blue line) is far from the ground truth (green line) because of varying detection sizes on profile and frontal faces for the image-to-3D projection (Eq. 3). However, since the video-assisted GCF is not sensitive to the detection size (see Fig. 10), the AO(2D) (tracking on the speaker ground truth height plane, in magenta) and AV3T (red) results are unaffected. Fig. 12(b) shows the tracking of two speakers (marked as blue and red), which are very close to the ground truth trajectories. Fig. 12(c-d) compares different modalities in CAV3D in situations when the VO trajectory (blue) is bounded by the FoV whereas AO(2D) (magenta) and AV3T (red) can follow the target outside the FoV as the speaker is active.

D. Comparisons with state-of-the-art methods

We finally compare AV3T against state-of-the-art (SoA) methods in 3D [10, 33] and on the image plane [1, 33]. We use the results reported in [1] for AV16.3 and conducted new experiments on CAV3D by using the image ground-truth as the audio observation, which eliminates the influence of the two different SSL methods, *i.e.* [47] for [1] and [25] for ours. Moreover, because [1] cannot track target outside the FoV, we re-initialize their tracker at the image ground-truth when it re-enters the scene. Frames without a target are not considered in the error computation. For [10] and [33] we include the proposed likelihoods in our PF implementation to compare with the same tracking parameters. For [10] we use the image ground-truth again instead of using noisy visual observations from the head tracker [51] and therefore its results should be considered an upper bound for the method.

Tab. V shows that AV3T considerably outperforms the other methods (and the unimodal approaches) in terms of TLR on CAV3D. The tracking accuracy in CAV3D-SOT2 is better than in CAV3D-SOT, which includes abrupt speaker-orientation changes and more challenging actions, such as clapping, stomping and arranging objects. Note how the performance of [1] decreases considerably in CAV3D. Note that it is not possible to compare with [33] that is not a multi-target tracker. Tab. VI shows that the availability of spatially distributed sensors in AV16.3 facilitates 3D tracking and all methods perform considerably better than in CAV3D. In AV16.3-SOT, AV3T outperforms [33] the unimodal trackers in terms of TLR in 3D, and achieves a slightly higher TLR (13.3%) than [10], which however uses the image ground-truth and benefits from the triangulation of distributed sensors. Moreover, AV3T outperforms individual modalities on the image plane and is more accurate than [1] in the successfully tracked frames. In

TABLE V
Performance scores (the smaller, the better) on CAV3D. [10]^s uses image ground truth as the visual observation. [1]^s uses image ground truth as the audio observation.

		Image plane					3D				
		[1]^s	[33]	AO (2D)	VO	AV3T	[10]^s	[33]	AO (2D)	VO	AV3T
SOT	TLR	29.5±12.4	25.0±1.2	52.2 ± 4.7	38.4 ± 17.5	**7.0 ± 3.6**	84.8±5.4	68.7±2.9	56.5±4.4	47.3±13.5	**31.8±3.5**
	ε	60.0±34.1	38.2±2.3	60.3 ± 6.9	80.2±103.0	**16.5 ± 8.6**	.84 ± .15	.50 ± .02	.52 ± .08	.76 ± .34	**.30 ± .05**
	ε	24.5±30.5	15.5 ± .4	27.7 ± 1.2	12.7 ± 1.1	**12.2 ± .3**	.17 ± .02	.20 ± .01	.17 ± .01	.16 ± .01	**.16 ± .01**
SOT2	TLR	33.0±18.5	23.0 ± .9	38.3 ± 3.9	13.4 ± 7.6	**4.0 ± 1.6**	85.2±4.5	62.9±2.8	43.6±4.9	20.1 ± 7.1	**11.1±3.1**
	ε	81.7±73.5	53.4±2.6	48.0 ± 6.0	36.5 ± 27.2	**20.8 ± 5.4**	.75 ± .07	.47 ± .02	.37 ± .07	.31 ± .12	**.18 ± .02**
	ε	23.7±64.5	13.3 ± .3	25.0 ± .6	12.0 ± .2	**11.7 ± .2**	.17 ± .02	.20 ± .01	.15 ± .01	.14 ± .01	**.14 ± .00**
MOT	TLR	16.0±10.0	-	59.4 ± 11.5	37.1 ± 7.1	**11.2 ± 5.9**	77.7±8.1	-	70.2±9.0	56.6 ± 6.2	**35.7±6.6**
	ε	59.3±33.9	-	155.7±60.6	127.9±60.1	**24.8±23.7**	.92 ± .23	-	1.03±.27	1.05 ± .22	**.43 ± .12**
	ε	17.6±27.4	-	19.9 ± 2.1	12.2 ± 1.3	**10.1 ± .6**	.16 ± .02	-	.16 ± .02	**.14 ± .02**	.15 ± .01

TABLE VI
Performance scores (the smaller, the better) on AV16.3. [10]^s uses image ground truth as the visual observation.

		Image plane					3D				
		[1]	[33]	AO (2D)	VO	AV3T	[10]^s	[33]	AO (2D)	VO	AV3T
SOT	TLR	-	48.2 ± 3.8	48.1 ± 6.0	9.0 ± 1.9	**8.5 ± 2.6**	**10.4 ± 3.4**	29.2 ± 3.7	34.9 ± 8.9	52.7 ± 5.5	13.3 ± 4.3
	ε	11.8±.2	19.9 ± 1.6	24.1 ± 5.7	8.2 ± 1.1	**7.7 ± 1.3**	**.15 ± .01**	.25 ± .02	.28 ± .07	.41 ± .05	.16 ± .02
	ε	-	8.5 ± .3	7.6 ± .5	**5.3 ± .1**	**5.3 ± .1**	.12 ± .01	.14 ± .01	.15 ± .01	.16 ± .01	**.11 ± .01**
MOT	TLR	-	-	56.6 ± 9.4	15.5 ± 9.0	**9.2 ± 6.0**	37.7 ± 5.6	-	44.9 ± 1.2	56.3 ± 9.8	**15.8 ± 8.9**
	ε	11.2±.1	-	38.4 ± 9.2	17.9 ± 8.8	**10.1 ± 3.7**	.31 ± .03	-	.48 ± .12	.52 ± .11	**.21 ± .07**
	ε	-	-	7.7 ± .9	5.1 ± .4	**4.9 ± .3**	.14 ± .01	-	.15 ± .02	.15 ± .02	**.11 ± .01**

AV16.3-MOT, AV3T outperforms [10] in 3D in terms of TLR. The average 3D error of AV3T during tracking is 11 cm.

E. Complexity and speed

The complexity of our method is linear to the overall number of particles N (each evaluation of likelihood terms and their fusion is independent per particle) except for the first three steps in the while loop in Algorithm 1 that are not particle operations: the first is a constant cost for face detection; the second has linear complexity in the number of detections returned; and the third is the computational complexity of Algorithm 2, which is upper-bounded (no association possible) by:

$$|D_t||I|c_A + |D_t|c_g + |D_t|c_0 + |I|\frac{|D_t|(|D_t|-1)}{2}c_0 \quad (22)$$

and lower-bounded (all i^*, d^* are valid associations) by

$$|D_t||I|c_A + Mc_g + 3Mc_0 + \sum_{k=0}^{M}(|D_t|-k)(|I|-k)c_0, \quad (23)$$

where $M = \min(|I|, |D_t|)$, c_A, c_g are the costs of evaluating Eq. 17, Eq. 18, and c_0 is a (negligible) cost of (i) one comparison (in argmax), (ii) set insertion to update D_{t,i^*} when the gating is passed and (iii) set reduction (of I, D_t).

Fig. 13 shows an empirical evaluation of the computation speed of our implementation. Experiments were run with a non-optimized MATLAB code on a 3.3 GHz Intel Xeon CPU (E31245). The left y-axis indicates that in order to get stable performance with AV3T, 50 particles suffice. When $N = 100$, the execution time is 0.14 spf (7.19 fps) on AV16.3 and 0.21 spf (4.77 fps) on CAV3D.

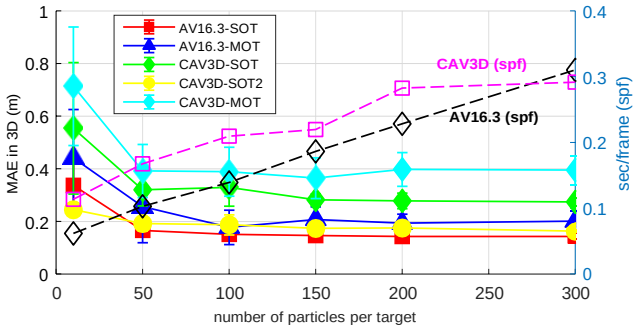


Fig. 13. Influence of the number of particles per target on tracking accuracy (solid lines) and on the execution time (dashed lines).

VII. Conclusion

We proposed AV3T, a novel 3D speaker tracker that uses audio-visual signals captured by a small and co-located sensing platform, without any depth sensor or any tracker applied before multi-modal fusion. AV3T estimates the 3D mouth position from face detections and models the likelihood in the camera's spherical coordinates based on the uncertainties derived from the image-to-3D projection from the camera. Moreover, AV3T uses video to suggest the most likely speaker-height plane for acoustic map computation and, during misdetections, uses a color-spatiogram-based generative model. The video-assisted SSL is more accurate than the 3D mouth estimates and less sensitive to errors in the hypothesized face size. We also contributed a new annotated audio-visual dataset, which we distribute to the research community.

We have identified three main directions for future work. The first direction is an extension to tracking a varying number of targets. The second direction is modeling varying head

orientations, which influences the expected face detection size that is usually smaller as profile than when frontal. The third direction is making the audio processing more robust as speech signals primarily contain reflections that cause larger TDoA estimates and lead to overestimating the distance of the speakers from the sensing platform.

ACKNOWLEDGEMENTS

We thank L. Cristoforetti, D. Giordani and A. Xompero for their help in the data collection.

REFERENCES

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

[2] 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, Nov 2010.

[3] H. Zhou, M. Taj, and A. Cavallaro, "Audiovisual tracking using STAC sensors," in *Proc. of Int. Conf. on Distributed Smart Cameras*, Vienna, Austria, Sept 2007.

[4] A. Brutti and O. Lanz, "A joint particle filter to track the position and head orientation of people using audio visual cues," in *Proc. of European Signal Processing Conf.*, Aalborg, Denmark, Aug 2010.

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

[6] D. Zotkin, R. Duraiswami, and L. S. Davis, "Multimodal 3D tracking and event detection via the particle filter," in *IEEE Workshop on Detection and Recognition of Events in Video*, Vancouver, BC, Canada, Jul 2001.

[7] M. Heuer, A. Al-Hamadi, B. Michaelis, and A. Wendemuth, "Multi-modal fusion with particle filter for speaker localization and tracking," in *Proc. of Int. Conf. on Multimedia Technology*, Zurich, Switzerland, Jun 2011.

[8] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 882–894, Jul 2010.

[9] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, vol. 38, no. 3, pp. 799–807, Jun 2008.

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

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

[12] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1086 – 1099, May 2017.

[13] V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Trans. on Multimedia*, vol. 9, no. 4, pp. 715–727, Jun 2007.

[14] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *Int. Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, Nov 2007.

[15] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun 2014.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. of Int. Conf. on Neural Information Proc. Systems*, Montreal, Canada, Dec 2015.

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

[18] H. Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 4, pp. 503–513, Sept 2008.

[19] U. Kirchmaier, S. Hawe, and K. Diepold, "Dynamical information fusion of heterogeneous sensors for 3D tracking using particle swarm optimization," *Information Fusion*, vol. 12, no. 4, Oct 2011.

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

[21] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, May 1997.

[22] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 4, pp. 1–19, Dec 2006.

[23] Q. Liu, W. Wang, T. deCampos, P. J. B. Jackson, and A. Hilton, "Multiple speaker tracking in spatial audio via PHD filtering and depth-audio fusion," *IEEE Trans. on Multimedia*, vol. 20, no. 7, pp. 1767–1780, Jul 2018.

[24] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. d. Campos, R. J. Hughes, D. Menzies, M. F. S. Gálvez, Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, "An audio-visual system for object-based audio: From recording to listening," *IEEE Trans. on Multimedia*, vol. 20, no. 8, pp. 1919–1931, Aug 2018.

[25] M. Omologo, P. Svaizer, and R. De Mori, "Acoustic transduction," in *Spoken Dialogue with Computer*. Academic Press, 1998, ch. 2, pp. 1–46.

[26] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2001.

[27] M. J. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 828–836, Jun 2003.

[28] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Audio-visual speech-turn detection and tracking," in *Int. Conf. on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, Aug 2015.

[29] O. Lanz, "Approximate Bayesian multibody tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1436–1449, Jul 2006.

[30] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. of Int. Conf. on Multimodal interfaces*, Trento, Italy, Oct 2005.

[31] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia, "A generative approach to audio-visual person tracking," in *Int. Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006.

[32] F. Keyrouz, U. Kirchmaier, and K. Diepold, "Three dimensional object tracking based on audiovisual fusion using particle swarm optimization," in *Proc. of Int. Conf. on Information Fusion*, San Diego, CA, USA, Oct 2008.

[33] X. Qian, A. Brutti, M. Omologo, and A. Cavallaro, "3D audio-visual speaker tracking with an adaptive particle filter," in *Proc.*

of IEEE Int. Conf. on Audio, Speech and Signal Processing, New Orleans, LA, USA, Mar 2017.

[34] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Trans. on Multimedia*, vol. 16, no. 3, pp. 864–880, Apr 2014.

[35] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer, "Audiovisual person tracking with a mobile robot," in *In Proc. of Int. Conf. on Intelligent Autonomous Systems*, 2004.

[36] M. Brandstein, *A Framework for Speech Source Localization Using Sensor Arrays*. Brown University, 1995.

[37] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *ICCV Workshop on Computer Vision for Audio-Visual Media*, Venice, Italy, Oct 2017.

[38] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Tracking the active speaker based on a joint audio-visual observation model," in *Proc. of Int. Conf. on Computer Vision Workshops*, Santiago, Chile, Dec 2015.

[39] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 718–731, Apr 2015.

[40] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Cech, K. Kulkarni, A. Deleforge, and R. P. Horaud, "RAVEL: An Annotated Corpus for Training Robots with Audiovisual Abilities," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 79–91, Mar 2013.

[41] E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. Horaud, "The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements," in *Proc. of Int. Conf. on Multimodal interfaces*, Chania, Greece, Oct 2008.

[42] M. Taj, "Surveillance Performance EValuation Initiative (SPEVI): Audiovisual people dataset," 2007. [Online]. Available: http://www.eecs.qmul.ac.uk/~andrea/spevi.html

[43] J. Carletta, "Announcing the AMI meeting corpus," *The ELRA Newsletter*, vol. 11, no. 1, pp. 3–5, 2006.

[44] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Journal on Language Resources and Evaluation*, vol. 41, no. 3, pp. 389–407, Dec 2007.

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

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

[47] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Philadelphia, PA, USA, May 2005.

[48] Y. Liu, A. Hilton, J. Chambers, Y. Zhao, and W. Wang, "Non-zero diffusion particle flow SMC-PHD filter for audio-visual multi-speaker tracking," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Calgary, Alberta, Canada, Apr 2018.

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

[50] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. on Information Theory*, vol. 21, no. 1, pp. 32–40, Jan 1975.

[51] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, Jun 1998.

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

[53] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Jan 2018.

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

[55] S. T. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, Jun 2005.

[56] C. O. Conaire, N. E. O'Connor, and A. F. Smeaton, "An improved spatiogram similarity measure for robust object localisation," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Honolulu, HI, USA, Apr 2007.

[57] Y. Tian, Z. Chen, and F. Yin, "Distributed Kalman filter-based speaker tracking in microphone array networks," *Applied Acoustics*, vol. 89, pp. 71–77, Aug 2015.

[58] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag: New York, 2001.

[59] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, Sept 2011.

[60] H. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quaterly*, pp. 83–97, 1955.

[61] D. B. Reid, "An algorithm for tracking multiple targets," *Trans. on Automatic Control*, vol. 24, no. 6, pp. 843–854, Dec 1979.

[62] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *Journal of Oceanic Engineering*, vol. 8, no. 3, pp. 173–184, Jul 1983.

[63] Y. Cai, N. de Freitas, and J. J. Little, "Robust visual tracking for multiple targets," in *Proc. of European Conf. on Computer Vision*, Berlin, Germany, May 2006.

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

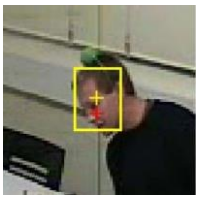
[65] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 11, pp. 1–17, Jan 2010.

[66] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.

[67] M. A. Lourakis and A. Argyros, "SBA: A software package for generic sparse bundle adjustment," *ACM Trans. Math. Software*, vol. 36, no. 1, pp. 29–58, Mar 2009.

[68] X. Wu, R. He, and Z. Sun, "A lightened CNN for deep face representation," *arXiv preprint arXiv:1511.02683*, 2015.

[69] E. Maggio and A. Cavallaro, *Video Tracking: Theory and Practice*. Wiley Publishing, 2011.



Xinyuan Qian received the B.Eng. (with First Class honors) and M.Sc. (with Distinction) degree in 2014 and 2015 respectively, both from the University of Edinburgh. She is currently pursuing the Ph.D. degree at Queen Mary, University of London, supervised by Prof. Andrea Cavallaro. From 2017 to 2018, she worked in Fondazione Bruno Kessler as a visiting student, supervised by Maurizio Omologo and Alessio Brutti. Her research interest mainly includes sound source localization, multi-modal sensor fusion and multi-speaker tracking.

Alessio Brutti is a tenured researcher at Fondazione Bruno Kessler, Trento, Italy. After graduating in Telecommunication engineering at the University of Padova, Padova, Italy, in 2001, in 2003 he joined the Center for Information and Communication Technologies of FBK. In 2006 he completed his Ph.D. in Computer Science at the University of Trento, Trento, Italy. His main research interests focus on multi-modal signal processing for biometrics and scene analysis, in particular for speech related applications as localization and tracking, speaker identification/diarization and speech enhancement.

PLACE PHOTO HERE

Oswald Lanz Biography text here.

PLACE PHOTO HERE

Maurizio Omologo Biography text here.

PLACE PHOTO HERE

Andrea Cavallaro Biography text here.