3D MOUTH TRACKING FROM A COMPACT MICROPHONE ARRAY CO-LOCATED WITH A CAMERA

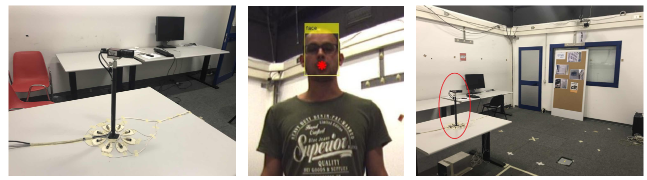
Xinyuan Qian[1], *Alessio Xompero*[1], *Alessio Brutti*[2], *Oswald Lanz*[2], *Maurizio Omologo*[2], *Andrea Cavallaro*[1]

[1]Centre for Intelligent Sensing, Queen Mary University of London, UK
[2]ICT-irst, Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

We address the problem of 3D audio-visual person tracking using a compact platform with co-located audio-visual sensors, without a depth camera. We present a face detection driven approach supported by 3D hypothesis mapping to image plane for visual feature matching. We then propose a video-assisted audio likelihood computation, which relies on a GCC-PHAT based acoustic map. Audio and video likelihoods are fused together in a particle filtering framework. The proposed approach copes with a reverberant and noisy environment, and can deal with person being occluded, outside the camera's Field of View (FoV), as well as not facing or far from the sensing platform. Experimental results show that we can provide accurate person tracking in both 3D and on image.

Index Terms— audio-visual fusion, particle filter, 3D person tracking, co-located sensor platform



(a) (b) (c)

Fig. 1. (a) The co-located audio-visual sensor platform consisting of an 8-element circular microphone array and a camera; (b) an example of mouth position estimate; and (c) the experimental environment.

1. INTRODUCTION

A fundamental task for scene understanding, human-machine and human-robot interaction is tracking the position of a person. Tracking can be carried out on the image plane [1–4], on a ground plane [5] or in 3D [6–9]. Methods for tracking a person in 3D mainly use distributed cameras and microphone arrays. However, the widespread use of smart-home devices, such as Google Home and Amazon Echo, as well as other robotic assistants, has triggered an increasing interest on platforms with co-located microphone arrays and cameras (see Fig. 1(a)). Only a handful of works focus on audio-visual 3D person tracking with small-size sensor configurations. For example [10] uses a single microphone pair in combination with stereo vision.

Unlike spatially distributed sensors, a compact and affordable configuration with a small number of co-located sensors facilitates audio-visual synchronization and calibration and can be used on a moving platform (e.g. a robot). However, using compact co-located sensors leads to important issues for person tracking. Besides traditional challenges like reverberation, background noise, random and abrupt person motion, other issues include person occlusions or outside the FoV of the camera, as well as a dependency on the distance from and orientation away from the microphones. Moreover the lack of depth information, due to the fact that sensors do not surround the person preventing triangulation to estimate its 3D position, is the most critical issue. In fact, neither a single RGB camera nor a circular microphone array can provide accurate 3D location estimates, especially under complex scenarios. We aim to exploit multi-modal information to improve tracking performance and to overcome the limitations of co-located sensor setups.

In this paper, we propose a novel approach for 3D person tracking using audio-visual signals captured by a co-located sensor platform consisting of an 8-element circular microphone array coupled with a camera. Unlike most of the state-of-the-art methods [11–13], our 3D tracker does not need a depth sensor. We extract three sources of information from the audio-visual streams. First, we estimate the 3D position of the mouth with a face detector. When a face detection is unavailable, we resort to a color-based measurement using a reference image, which, however, cannot provide information about the person distance from the platform. We then use audio as complementary information to strengthen the 3D position estimation, in particular when the face detector fails or the person is outside the FoV of the camera, and to eliminate distractors such as other people or false-positive detections. We use the previously estimated mouth height from the video to constrain the audio search space on a 2D plane and to reduce the audio uncertainties caused by the circular array to estimate the person distance from the platform. After the modality-dependent processing stages, information is fused and processed by a particle filter that estimates the 3D position of the person. Figure 2 shows the block diagram of the proposed method.

2. PROBLEM FORMULATION

We aim to track the 3D position, $\mathbf{p}_t$, of a person over time t, given audio signals, $\mathbf{s}_t$, captured by an 8-microphone circular array and frames, I_t, recorded by a RGB camera. In a sequential estimation, this task consists in first evaluating a probability $P(\mathbf{p} \mid \mathbf{s}_{1:t}, I_{1:t})$ of hypotheses $\mathbf{p}$ conditioned on past and current observations and then inferring the target state from P, e.g. via expectation:

$$\hat{\mathbf{p}}_t = \mathbb{E}_P(\mathbf{p} \mid \mathbf{s}_{1:t}, I_{1:t})\,. \tag{1}$$

When the signal formation $\mathbf{p} \mapsto \mathbf{s}, I$ is non-linear, incomplete and non-invertible as in our case, a common choice is a Bayesian model. Using Bayes rule and the total probability theorem, the Chapman-Kolmogorov recursion modelling, P is fully specified by

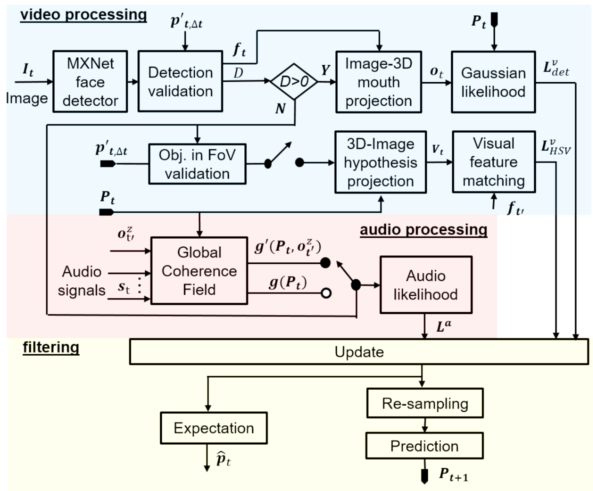


Fig. 2. Block diagram of the proposed audio-visual 3D tracker.

a data likelihood L, a first–order dynamics Q and an initial density dP_0 [14]:

$$P(\mathbf{p} \mid \mathbf{s}_{1:t}, I_{1:t}) \propto L(\mathbf{s}_t, I_t \mid \mathbf{p}) \int Q(\mathbf{p} \mid \mathbf{q})\; dP(\mathbf{q} \mid \mathbf{s}_{1:t-1}, I_{1:t-1})\,. \tag{2}$$

The only requirement is on L and Q to be evaluable point-wise, yielding a model that is flexible and computationally attractive if combined with sampling methods.

We realize Eq. 2 with a Particle Filter (PF) [14], which maintains a non-parametric representation of P by propagating a set of N independently and identically distributed (iid) samples (*particles*) from P, i.e.,

$$\{\mathbf{p}_t^{(1)}, ..., \mathbf{p}_t^{(N)}\} \overset{\text{iid}}{\sim} P(\mathbf{p} \mid \mathbf{s}_{1:t}, I_{1:t})\,. \tag{3}$$

This is achieved in two steps by (i) sampling from the prior mixture $\sum_n^N Q(\mathbf{p} \mid \mathbf{p}_{t-1}^{(n)})$ and (ii) re-sampling with probability $\propto L(\mathbf{s}_t, I_t \mid \mathbf{p})$. As common in multi-modal tracking, we assume conditional independence between modalities given the target state. The re-sampling probability is thus the product of the audio likelihood $L^a(\mathbf{s}_t \mid \mathbf{p})$ and the video likelihood $L^v(I_t \mid \mathbf{p})$.

Our solution comprises the modelling of the individual likelihoods, L^v, L^a (Sec. 3.1 and 3.2), and the propagation scheme and model Q (Sec. 3.3).

3. PROPOSED METHOD

3.1. Visual observation

Our person tracker is driven by a face detector, which allows us to derive the 3D mouth position with simple geometric considerations using prior knowledge of the typical size of a human face[1].

Let $\mathbf{f}_t^d = [u, v, w, h]^T$ be the bounding box of the d^{th} detected face ($d = 1, \ldots, D$) at time t, where (u, v) is the position of the top left corner and (w, h) are width and height. We geometrically extract the mouth position, $\boldsymbol{\rho}_t^d = [u + 0.5w, v + 0.75h]^T$, and then use the pinhole camera model and camera calibration information [15] to obtain its 3D location. We determine the scaling factor by modelling the shape of a face with a rectangle oriented towards the camera and the prior knowledge on the face width W to obtain via image-to-3D back-projection[2] the 3D mouth position: $\mathbf{o}_t^d = \Psi\left[\boldsymbol{\rho}_t^d; w, W\right]$. We validate the output of the face detector with:

$$||\boldsymbol{\rho}_t^d - \mathbf{p}'_{t,\Delta t}||_2 \leq \lambda\sqrt{w^2 + h^2} \tag{4}$$

where λ controls the acceptable error range and $\mathbf{p}'_{t,\Delta t}$ is the average estimated mouth position on image plane in the last Δt frames.

We use spherical coordinates to better model the higher inaccuracy in the distance estimation, which is based on the hypothesised face width W. Let $\tilde{\mathbf{o}}_t^d$ and $\tilde{\mathbf{p}}$ be the estimated mouth position and a generic 3D point in spherical coordinates. Assuming a Gaussian distribution of the estimates, we evaluate the likelihood of the hypothesis p as:

$$L^v_{\text{det}}(I_t \mid \mathbf{p}) = \sum_{d=1}^{D} \exp\left[-\left(\tilde{\mathbf{o}}_t^d - \tilde{\mathbf{p}}\right)\Sigma_v^{-1}\left(\tilde{\mathbf{o}}_t^d - \tilde{\mathbf{p}}\right)^T\right], \tag{5}$$

where Σ_v accounts for the different estimation accuracy in the three spherical coordinates.

When the face is not visible or the face detector fails when the person is inside the camera's FoV, we resort to a generative model and evaluate a color-based likelihood. First, we map each 3D hypothesis (particle) to the image plane by creating a bounding box using a 3D hyperrectangle oriented towards the camera $\mathbf{v} = \Phi\left[\mathbf{b}\left(\mathbf{p}; W, H\right)\right]$, where $\mathbf{b}(\mathbf{p}; W, H)$ is the 3D rectangle created from a generic 3D point $\mathbf{p}$ with face width W and height H and Φ indicates the 3D-to-image projection. Then, we compare the color features of the bounding box with a reference image (which is updated to the last detection $\mathbf{f}_{t'}$) of the person using a Hue-Saturation-Value (HSV) spatiogram [16]. We measure the similarity $L^v_{\text{HSV}}(I_t \mid \mathbf{p})$ between two spatiograms using [17], which is derived from the Bhattacharyya coefficients.

Finally, we define the *visual likelihood* as:

$$L^v(I_t \mid \mathbf{p}) = \begin{cases} L^v_{\text{det}}(I_t \mid \mathbf{p}) & \text{if } D > 0 \\ L^v_{\text{HSV}}(I_t \mid \mathbf{p}) & \text{if } \mathbf{p}'_{t,\Delta t} \in I^{0.9} \\ 1/N & \text{otherwise,} \end{cases} \tag{6}$$

where $I^{0.9}$ is a rectangular crop corresponds to the central 90% region of the image. It is used with $\mathbf{p}'_{t,\Delta t}$ to indicate whether the person is inside camera's FoV.

3.2. Video driven acoustic observations

Acoustic source localization can be accomplished by combining the information of M microphone pairs to obtain acoustic maps that represent the plausibility of an active sound source to be at a given spatial position [18]. Let the source be in $\mathbf{p}$ and $\tau_m(\mathbf{p})$ be the expected Time Difference of Arrival (TDoA) between the microphones of the m^{th} pair. If $C_m(\cdot)$ is the Generalized Cross Correlation PHAse Transform (GCC-PHAT) function computed at the m^{th} microphone pair [19, 20], then the Global Coherence Field (GCF) can be evaluated at each position $\mathbf{p}$ as [21]:

$$g(\mathbf{p}) = \frac{1}{M}\sum_{m=0}^{M-1} C_m\left(\tau_m(\mathbf{p})\right). \tag{7}$$

[1]Size variations of the human face are much smaller than those of other body parts (e.g. upper-body), thus allowing a more accurate 3D inference.

[2]The back-projection error is stable when $W \in [0.13, 0.15]$ m.

While a position estimate of the sound emission can be obtained from the maximum of the GCF acoustic map, when a compact microphone array is employed, GCF fails to provide accurate 3D estimations, in particular along the range dimension. This problem can be circumvented if some knowledge about the mouth height is available. Therefore, we propose a video-driven GCF, $g'(\mathbf{p}, o^z_{t'})$, which is computed by projecting a generic 3D point $\mathbf{p}$ into the 2D plane through the mouth height $o^z_{t'}$, estimated from the most recent face detection.

Finally, we define the *audio likelihood* as:

$$L^a(\mathbf{s}_t \mid \mathbf{p}) = \begin{cases} g(\mathbf{p}) & \text{if } D > 0,\ \max_{\mathbf{p}} g(\cdot) \geq \vartheta_a \\ g'(\mathbf{p}, o^z_{t'}) & \text{if } D = 0,\ \max_{\mathbf{p}} g'(\cdot) \geq \vartheta_a \\ 1/N & \text{otherwise,} \end{cases} \tag{8}$$

where $g(\cdot)$ is the previous g related variable in the brace. ϑ_a is a threshold used to remove not reliable audio observations due to pauses, presence of noise or narrow-band spectral content. In case of multiple detections, we select $o^z_{t'}$ as the closest one to the 3D point $\mathbf{p}$ under analysis.

3.3. Prediction

Given the likelihoods defined in Section 3.1 and 3.2 and assuming conditional independence across the modalities, an approximation of the posterior in Eq. 2 is obtained from the particle set at time $t-1$ as described in Sec. 2, by sampling the random variable $\mathbf{p}$ from

$$\{\mathbf{p}^{(1)}_t, ..., \mathbf{p}^{(N)}_t\} \overset{\text{iid}}{\sim} L^a(\mathbf{s}_t \mid \mathbf{p}) L^v(I_t \mid \mathbf{p}) \sum_{n=1}^{N} \mathcal{N}\left(\mathbf{p}; \mathbf{p}^{(n)}_{t-1}, 3^{\kappa}\Sigma_r\right) \tag{9}$$

Here, we model first-order dynamics Q (Eq. 2) as a mixture of Gaussian distributions whose covariance matrix Σ_r is diagonal and κ is 1 if the likelihood product is in the lower 10% (higher prediction speed for low-scoring hypotheses) and 0 otherwise.

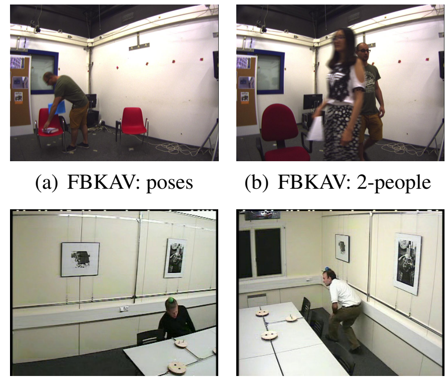
Finally, the 3D position estimate of the mouth is the empirical expectation that approximates Eq. 1:

$$\hat{\mathbf{p}}_t = \frac{1}{N} \sum_{n=1}^{N} \mathbf{p}^{(n)}_t \tag{10}$$

4. EXPERIMENTS

We evaluate the proposed tracker on two datasets and compare it against the audio-visual trackers in [22] and in [3], as well as with trackers using the individual modalities, namely Audio-Only (AO) and Video-Only (VO). To account for the probabilistic nature of the PF framework, we consider the average Mean Absolute Error (MAE) (in m) for 10 runs and the Tracking Rate (TR), which is the percentage of frames where the error is smaller than 0.4 m.

Datasets. We use the publicly available AV16.3 [23] to allow a comparison with the literature and we collect a new one with co-located sensors. In AV16.3, the video is captured by 3 cameras at 25 Hz with resolution of 360×288 pixels and audio is recorded at 16 kHz using two 8-element circular microphone arrays with 10 cm radius. In our experiments we use only one camera and one microphone array from the sequences $seq08$, $seq11$ and $seq12$. Moreover, we recorded the FBKAV dataset with co-located audio-visual sensors and 3D labelling to overcome the lack of a public dataset with these properties. The co-located sensors consist in a Allied Marlin F-080C camera and an 8-element circular array of omnidirectional microphones with 10 cm radius (Fig. 1(a)), positioned on



(a) FBKAV: poses (b) FBKAV: 2-people

(c) AV16.3: seq11-cam1 (d) AV16.3: seq12-cam3

Fig. 3. Key frames from FBKAV (a-b) and AV16.3 dataset (c-d).

a table in a room of size $4.77 \times 5.95 \times 4.5$ m. The room reverberation time is 0.7 s [18] and audio signals are recorded at 96 kHz. Video is captured at 15 Hz with resolution of 1024×768 pixels. Synchronization and calibration are obtained manually. To generate annotation data with an accuracy error of less than 10 cm, we use *SmarTrack* [24]. To do so, we complemented the dataset with recordings using a spatially distributed sensor set-up consisting of four Allied cameras at the corners of the room. We use four sequences and each of them lasts for around one minute: (1) 'easy': the person moves around, mostly in the FoV, speaking towards the sensor platform; (2) '2-people': the person always talks, moving around while another silent person enters in the FoV; (3) 'behind': the person enters the FoV, walks behind the camera while talking and finally re-enters the FoV; (4) 'poses': the person always talking in the FoV, but in a variety of challenging poses (i.e. not oriented towards the sensors, bending over). Fig.3 shows sample frames of the two datasets.

Implementation details. We detect faces using a freely available software based on an MXNet implementation of light CNN[3] [25]. The face width is $W = 0.14$ m and height is $H = 0.18$ m. The spatiogram in the HSV color space has 8 bins per channel; $\lambda = 2.5$, $\Delta t = 3$ and Σ_v in Eq. 5 is a diagonal matrix with elements (2°,2°,0.4 m). We compute GCC-PHAT using a 2^{10}-point and a 2^{15}-point Hanning window in AV16.3 and in FBKAV, respectively. The overlapping factor between two consecutive windows is set to provide one-to-one audio-visual frame correspondence. The validation threshold ϑ_a in Eq. 8 is set to 0.1 in AV16.3, and 0.03 in FBKAV. Different parameter settings are used because of their different sampling frequency. All the microphone pairs (M=28) within the 8-element circular array are used to obtain the GCF acoustic map. Finally, the diagonal elements of the prediction matrix Σ_r are set equivalent to $(1, 1, 0.5)$ m/s and we use 100 particles to perform 3D tracking.

Discussion. Table 1 shows the results of the proposed Audio-Visual (AV) 3D tracking with AO and VO only. For AO, we consider only 2D tracking fixing the mouth height to 1.5 m. In 'easy', both AO and VO perform similarly to AV one, with AO using knowledge of the person's mouth height. In '2-people', both VO and AV perform well thanks to the face validation stage which removes false positives from the other silent person. In 'behind', neither AO nor VO performs satisfactorily, because the person is outside the FOV for half of the sequence and has long speech pauses when inside the

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

Table 1. 3D tracking results on FBKAV, in comparison with [22].

	MAE (m)			
	AO (2D)	VO	[22]	AV
easy	0.13±.01	0.15±.01	0.31±.01	0.15±.01
2-people	0.32±.04	0.18±.01	0.50±.01	0.18±.01
behind	0.43±.04	1.07±.43	0.52±.01	0.26±.02
poses	0.95±.03	0.33±.02	0.80±.01	0.42±.02
Avg.	0.46±.03	0.43±.12	0.53±.01	0.25±.01

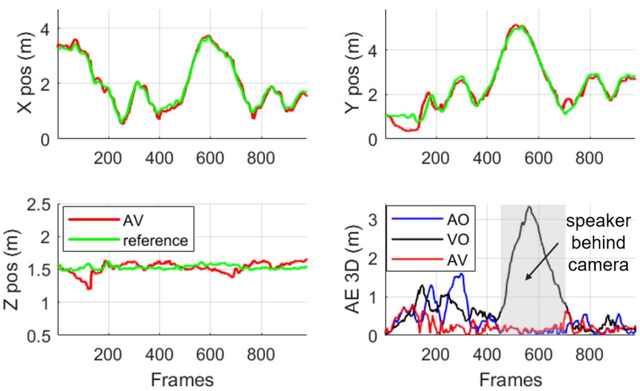


Fig. 4. FBKAV-'behind': Comparison between the reference and AV results in individual X,Y and Z coordinates. Comparison of 3D Absolute Error (AE) at frame t among AO, VO and AV.

FoV. In this case the proposed audio-visual tracker outperforms the two individual modalities. Fig.4 illustrates the AV tracking results of sequence 'behind' in individual coordinates and its superiority over AO and VO. The sequence 'poses' includes very challenging audio situations with the person arranging objects and facing away from the microphone array. As a result, the performance of AO considerably deteriorates with respect to the other sequences, in particular along the range dimension, and affects the AV tracking, which performs slightly worse than VO. Overall, an average 3D error of 0.25 m was obtained on the four sequences, which outperforms [22].

Table 2 reports the TR and the face detection rate. The results are in line with what reported in Table 1. Note that although the proposed method heavily relies on the face detector for the visual likelihood, the VO and AV results are always superior.

Fig.5 quantifies the sensitivity to the face detection results and helps analyse the impact of the other likelihoods. In 'easy', both modalities perform well and the accuracy is unaffected by the removal of face detection results. For the other sequences, the MAE in 3D increases when the number of detections removed, thus leading to a performance close to the AO (2D) case. This deterioration becomes evident only if at least 50% of the detections are removed.

For AV16.3, we report results in 3D as well as on the image plane to compare them with the audio-visual tracker in [3]. Results on seq08, seq11 and seq12 over three different camera views are given in Table 3. For [22], we fit the audio-visual likelihoods into our PF framework with the same parameters used for tracking. Additionally, we replace the Viola-Jones upper-body detector [26] with the MXNet face detector. The overall 3D tracking accuracy is improved from 0.32 m to 0.17 m thanks to our likelihood computation method and fusion. When tracking on the image plane, the proposed method also outperforms the accuracy of [3] in every sequence and the MAE is improved from 11.75 to 7.09 pixels.

Table 2. % of face detection rate (DR= # of true positives / total # of frames) and TR on FBKAV.

	DR (%)	TR (%)			
		AO (2D)	VO	[22]	AV
easy	70.94	98.03	98.88	74.08	97.17
2-people	80.25	75.18	94.80	44.56	93.81
behind	48.41	62.69	48.24	40.88	81.05
poses	48.02	14.41	71.81	15.08	64.64
Avg.	61.91	62.58	78.43	43.65	84.17

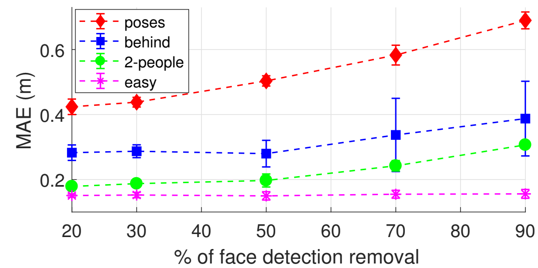


Fig. 5. Influence of randomly removing a percentage of detections on 3D tracking performance.

Table 3. Audio-visual tracking results in 3D and on the image plane on AV16.3(cam$\in$ 1, 2, 3 is the camera index). Standard deviation is reported for the image plane only, in 3D it is always small (<0.04).

		MAE (m)		MAE (pixels)	
seq	cam	[22]	AV	[3]	AV
08	1	0.15	0.12	10.75 ± 0.13	4.31 ± 0.20
	2	0.24	0.11	7.33 ± 0.09	4.66 ± 0.09
	3	0.20	0.09	9.85 ± 0.12	5.34 ± 0.13
11	1	0.31	0.33	14.66 ± 0.34	8.15 ± 0.71
	2	0.29	0.14	14.01 ± 0.12	7.48 ± 0.53
	3	0.26	0.12	13.96 ± 0.23	6.64 ± 0.15
12	1	0.41	0.26	12.49 ± 0.16	6.86 ± 0.42
	2	0.51	0.17	10.81 ± 0.24	10.67 ± 2.00
	3	0.47	0.20	11.86 ± 0.24	9.71 ± 3.20
Avg.		0.32	0.17	11.75 ± 0.19	7.09 ± 0.83

5. CONCLUSION

We propose a novel audio-visual tracker capable of performing 3D person tracking using a small-size co-located audio-visual set up, without any depth sensor. The system is supported by a face detector, by 3D hypothesis mapping, and by video assisted audio likelihood computation. Thanks to the complementary use of audio and visual signals, we were able to outperform significantly our previous method [22], under all the addressed experimental conditions. In particular, it is worth noting that the audio modality contributes to system robustness when the person is outside the FoV for a long time, while the video modality plays a key role, for instance to suggest the most likely mouth height where to compute a 2D acoustic map.

References

[1] Matthew J. Beal, Nebojsa Jojic, and Hagai Attias, "A graphical model for audiovisual object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 828–836, 2003.

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

[3] Volkan Kılıç, Mark Barnard, Wenwu Wang, and Josef Kittler, "Audio assisted robust visual tracking with adaptive particle filtering," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 186–200, 2015.

[4] Israel D. Gebru, Silèye Ba, Georgios Evangelidis, and Radu Horaud, "Audio-visual speech-turn detection and tracking," in *International Conference on Latent Variable Analysis and Signal Separation*, August 2015, pp. 143–151.

[5] Eleonora D'Arca, Neil M. Robertson, and James Hopgood, "Person tracking via audio and video fusion," in *Data Fusion & Target Tracking Conference: Algorithms & Applications*, May 2012, pp. 1–6.

[6] Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. on Applied Signal Processing*, vol. 2002, no. 1, pp. 1154–1164, 2002.

[7] Kai Nickel, Tobias Gehrig, Rainer Stiefelhagen, and John McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. of Int. Conf. on Multimodal Interfaces*, October 2005, pp. 61–68.

[8] Roberto Brunelli, Alessio Brutti, Paul Chippendale, Oswald Lanz, Maurizio Omologo, Piergiorgio Svaizer, and Francesco Tobia, "A generative approach to audio-visual person tracking," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 55–68.

[9] Alessio Brutti and Oswald Lanz, "A joint particle filter to track the position and head orientation of people using audio visual cues," in *Proc. of European Signal Processing Conference*, August 2010, pp. 974–978.

[10] Ulrich Kirchmaier, Simon Hawe, and Klaus Diepold, "Dynamical information fusion of heterogeneous sensors for 3D tracking using particle swarm optimization," *Information Fusion*, vol. 12, no. 4, pp. 275–283, 2011.

[11] Nagasrikanth Kallakuri, Jani Even, Yaileth Morales, Carlos Ishi, and Norihiro Hagita, "Probabilistic approach for building auditory maps with a mobile microphone array," in *Proc. of IEEE Int. Conf. on Robotics and Automation*, May 2013, pp. 2270–2275.

[12] Nagasrikanth Kallakuri, Jani Even, Yaileth Morales, Carlos Ishi, and Norihiro Hagita, "Using sound reflections to detect moving entities out of the field of view," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, November 2013, pp. 5201–5206.

[13] Jani Even, Yaileth Morales, Nagasrikanth Kallakuri, Jonas Furrer, Carlos Toshinori Ishi, and Norihiro Hagita, "Mapping sound emitting structures in 3d," in *Proc. of IEEE Int. Conf. on Robotics and Automation*, June 2014, pp. 677–682.

[14] Sanjeev M. Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb 2002.

[15] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.

[16] Stanley T. Birchfield and Sriram Rangarajan, "Spatiograms versus histograms for region-based tracking," in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, June 2005, pp. 1158–1163.

[17] Ciaran O. Conaire, Noel E. O'Connor, and Alan F. Smeaton, "An improved spatiogram similarity measure for robust object localisation," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, April 2007, pp. 1069–1072.

[18] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer, "Multiple source localization based on acoustic map deemphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 11, January 2010.

[19] Maurizio Omologo and Piergiorgio Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, April 1994, pp. 273–276.

[20] Charles Knapp and Glifford Carter, "The generalized correlation method for estimation of time delay," vol. 24, no. 4, pp. 320–327, 1976.

[21] Maurizio Omologo, Piergiogio Svaizer, and Renato De Mori, "Acoustic transduction," in *Spoken Dialogue with Computer*, Renato De Mori, Ed., chapter 2, pp. 1–46. Academic Press, 1998.

[22] Xinyuan Qian, Alessio Brutti, Maurizio Omologo, and Andrea Cavallaro, "3D audio-visual speaker tracking with an adaptive particle filter," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, March 2017, pp. 2896–2900.

[23] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez, "AV16.3: an audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*, pp. 182–195. Springer, June 2004.

[24] Oswald Lanz, "Approximate bayesian multibody tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1436–1449, July 2006, `http://tev.fbk.eu/smartrack`.

[25] Xiang Wu, Ran He, and Zhenan Sun, "A lightened CNN for deep face representation," *arXiv preprint arXiv:1511.02683*, 2015.

[26] Paul Viola and Michael Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.