3D AUDIO-VISUAL SPEAKER TRACKING WITH AN ADAPTIVE PARTICLE FILTER

Xinyuan Qian[1], *Alessio Brutti*[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 propose an audio-visual fusion algorithm for 3D speaker tracking from a localised multi-modal sensor platform composed of a camera and a small microphone array. After extracting audio-visual cues from individual modalities we fuse them adaptively using their reliability in a particle filter framework. The reliability of the audio signal is measured based on the maximum Global Coherence Field (GCF) peak value at each frame. The visual reliability is based on colour-histogram matching with detection results compared with a reference image in the RGB space. Experiments on the AV16.3 dataset show that the proposed adaptive audio-visual tracker outperforms both the individual modalities and a classical approach with fixed parameters in terms of tracking accuracy.

Index Terms— audio-visual fusion, adaptive weighting, particle filter, 3D speaker tracking

1. INTRODUCTION

Tracking a target with a localised (co-located) multi-modal sensor platform is desirable for autonomously navigating robots and human-robot interaction. A moving speaker is an important type of target, which can be tracked using audio [1] or video [2], or by fusing the two modalities exploiting the complementarity of audio and video signals [3]. However, in a changing environment, appropriately weighting each modality dynamically is still an open research problem.

The target state can be estimated by the on-board sensors in different state spaces, such as the ground plane for path planning [4], the image plane for face recognition [5] and the 3D world coordinates for navigation or grasping [6]. Locating a target in 3D offers important information to analyse the target as well as the interactions between the robot and the environment. However, only a few works have addressed the problem of tracking a speaker in 3D using a localised sensor platform [7][8].

A Kalman Filter (KF) can be used for late audio-visual fusion to track a speaker on the ground plane under the assumptions of Gaussian noise and linear state functions [9]. Face locations detected from different camera views and Direction of Arrival (DoA) estimations can be incorporated in an Extended Kalman Filter (EKF) to update the estimates of the target 3D location [10]. Particle Filters (PF) [11] are applicable for non-linear models to fuse multi-sensor data for tracking. DoA information from audio processing can assist video for joint speaker diarisation and tracking on the image plane [12]. Similarly, the DoA estimations can be mapped to the image plane to constrain and update the speaker track from video [13]. The main limitation is that the performance decreases considerably when the acoustic environment worsens as this approach projects the particles from visual tracking towards the estimated DoA line. Additionally, this tracking operates on the image plane only. Face candidates from three Viola-Jones detectors are validated to reduce false positives using the probability scores from a Gaussian Mixture Model (GMM). Camera calibration information helps to estimate the target 3D position from video, which is fused with the estimated Time Difference of Arrival (TDOA) generated from a Generalized Cross Correlation with Phase Transform (GCC-PHAT) approach for audio-visual 3D location update in PF [14]. Multiple microphone arrays and cameras can be distributed around a room to jointly track with a PF the speaker 3D position and head orientation [15]. An RGB histogram is computed for the visual likelihood to fuse with the audio likelihood from GCC-PHAT based TDOA estimation. A comparison of state-of-the-art methods for audio-visual speaker tracking is shown in Table 1.

In this paper, we propose a PF that estimates the azimuth, elevation and radius of a moving speaker using a co-located circular microphone array and a standard camera. We use the Global Coherence Field (GCF, aka SRP-PHAT [21]) peak value and Bhattacharyya distance between colour histograms to adapt the weights of audio-visual cues. The proposed audio-visual fusion algorithm can be implemented on an independent robotic platform without the use of ambient sensors. The block diagram of the proposed 3D audio-visual speaker tracker is shown in Fig.1.

2. PROPOSED METHOD

Let the audio and video signals be captured by a circular microphone array and a camera (Fig.2(b)). Depending on the sensor geometry, the estimation of azimuth, elevation and radius leads to different reliabilities. In particular, the radius estimate is in general less accurate. Therefore, we separate those elements and track the speaker in

Table 1. Summary of state-of-the-art methods for audio-visual speaker tracking. KEY – C: sensor configuration type; Dis: distributed; Loc: localised; N_C: number of cameras; N_M: number of microphone arrays $\times$ number of microphones in each array; R: reference; Prop.: proposed method.

Ref	C	N_C	N_M	R	Fusion	Algorithm
[13]	Dis.	1	1×8	Image	Hybrid	PF
[16]	Dis.	3	1×8	Image	Hybrid	PF
[9]	Dis.	1	4×2	Ground	Late	KF, PF
[17]	Dis.	5	5×2	Ground	Late	KF
[10]	Dis.	4	4×4	3D	Late	KF
[14]	Dis.	5	3×4	3D	Hybrid	KF, PF
[18]	Dis.	4	3×4	3D	Late	PF
[15]	Dis.	4	7×4	3D	Late	PF
[19]	Loc.	1	1×2	Image	Late	KF, PF
[20]	Loc.	1	1×2	Image	Late	PF
Prop.	Loc.	1	1×8	3D	Late	PF

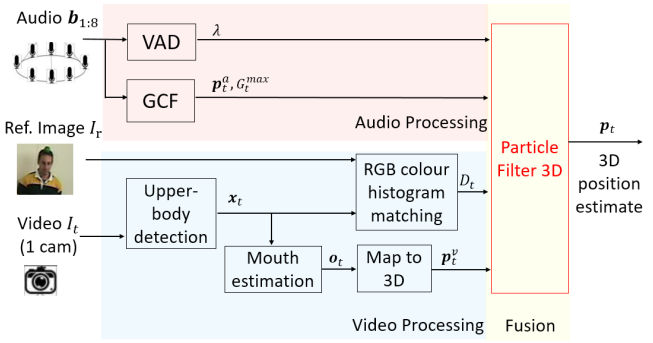


Fig. 1. Block diagram of the proposed 3D speaker tracker using audio-visual signals. A Voice Activity Detector (VAD) is first applied to find speech segments and the sound source location is estimated from the GCF algorithm. An upper-body detector is used to find the target position on the image plane. The mouth position of the target (the sound source) is projected into 3D with known camera calibration information. Finally, the estimated mouth positions from independent audio and video signals are fused in a particle filter framework where dynamic weights are based on the reliability of each modality. ($\mathbf{b}_{1:8}$: audio signals captured by the 8-element circular microphone array; λ : binary output of VAD; G_t^{max} : GCF peak value at time t; $\boldsymbol{x}_t$: upper body detection bounding box; $\boldsymbol{o}_t$: mouth position; $\boldsymbol{p}_t^a$,$\boldsymbol{p}_t^v$ and $\boldsymbol{p}_t$: estimated target 3D position from audio, video and audio-visual cues).

spherical coordinates. The origin of the spherical coordinates is the centre of the circular microphone array.

Let the state vector $\mathbf{s}$ be defined as

$$\boldsymbol{s} = (\theta, \varphi, r, v_\theta, v_\varphi, v_r)^T, \tag{1}$$

where (θ,φ,r) indicates azimuth, elevation and radius; v_r is the velocity in radius direction; and v_θ and v_φ are angular speeds.

2.1. Audio processing

We use Voice Activity Detection (VAD) to consider an audio segment as speech when its average power is beyond a threshold e_{th}. This task could alternatively be done by making use of SNR and Zero Crossing Rate [22] (ZCR) and the highest GCC-PHAT response [15].

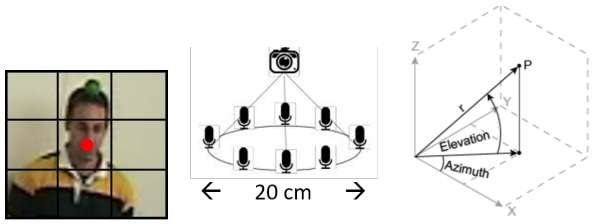
The speaker position can be estimated in four steps [23], namely normalised crosspower-spectrum estimation, coherence measure (aka GCC-PHAT) estimation, 3D acoustic map generation and speaker position estimation where a 3D acoustic map represents probabilistically the likely position of a sound source.

If $S_{m_1}(t,f)$ and $S_{m_2}(t,f)$ are the spectra of audio signals from the m^{th} microphone pair (m_1, m_2), computed by applying Fourier Transform to the corresponding windowed segments, centred at time t, then the *normalised crosspower-spectrum* $\phi_m(t,f)$ can be estimated as

$$\phi_m(t,f) = \frac{S_{m_1}(t,f) S^*_{m_2}(t,f)}{|S_{m_1}(t,f)|\,|S^*_{m_2}(t,f)|}, \tag{2}$$

where f means frequency and $*$ indicates complex conjugate. The *coherence measure* C_m can be computed as

$$C_m(t,\tau) = \int_{-\infty}^{+\infty} \phi_m(t,f) e^{j2\pi f\tau} df, \tag{3}$$



(a) (b) (c)

Fig. 2. (a) Estimated mouth position (red filled circle) and 3×3 sub-images division of the upper-body detection result $\boldsymbol{x}_t$, (b) proposed sensor configuration for a mobile platform, (c) a generic 3D point $\boldsymbol{p}$ in spherical coordinates.

the inverse transform of Eq. 2, which represents the similarity between two segments for time lag τ at the m^{th} microphone pair.

If $\boldsymbol{p}$ is a generic 3D point, the *3D acoustic map* $G(t,\boldsymbol{p})$ can be generated as

$$G(t,\boldsymbol{p}) = \frac{1}{M} \sum_{m=0}^{M-1} C_m(t, \delta_m(\boldsymbol{p})), \tag{4}$$

where M is the number of used microphone pairs and $\delta_m(\boldsymbol{p})$ is the ideal TDOA at a 3D position $\boldsymbol{p}$ with respect to the m^{th} microphone pair. Finally, the *speaker position* (i.e. the sound source location) $\boldsymbol{p}_t^a$ is considered to be the point with maximum value in the map:

$$\boldsymbol{p}_t^a = \arg\max_{\boldsymbol{p} \in \boldsymbol{P}} G(t,\boldsymbol{p}). \tag{5}$$

where $\mathbf{P}$ denotes the set of points of the 3D grid under analysis and $\boldsymbol{p}$ is a single point within it.

2.2. Video processing

We use the Viola-Jones algorithm [24] for upper-body detection on the image plane. Let the corresponding bounding box $\boldsymbol{x}_t$ be

$$\boldsymbol{x}_t = (u, v, w, h)_t^T, \tag{6}$$

where (u,v) is the upper-left corner position; and w and h are the width and height of the bounding box.

By assuming the width of the shoulders and camera calibration information to be known, we map the approximate mouth position $\boldsymbol{o}_t$ (i.e. middle point of the bounding box $\boldsymbol{x}_t$, Fig. 2(a)) of the target from the image plane to the 3D world coordinates:

$$\boldsymbol{p}_t^v = \boldsymbol{H}\boldsymbol{o}_t, \tag{7}$$

where $\boldsymbol{p}_t^v$ is the estimated *mouth position* from the video and $\boldsymbol{H}$ is the re-projection matrix derived from the camera pinhole model [25].

We use colour information to measure the reliability of the information extracted from the video. We calculate the dissimilarity measure between a user-defined reference image I_r and the region surrounded by the detected bounding box $\boldsymbol{x}_t$ using the Bhattacharyya distance:

$$D_t = \sqrt{1 - \sum_{u=1}^{U} \sqrt{\boldsymbol{r}(u)\boldsymbol{q}(u)_t}}, \tag{8}$$

where $\boldsymbol{r}(u)$ and $\boldsymbol{q}(u)_t$ are $1 \times U$ vectors of the normalised colour histogram of the reference image and the current detection result.

U is the number of bins used by the histogram and $0 \leq D_t \leq 1$. We choose the RGB colour space [2] and divide the detected upper-body region into 3×3 sub-images (see Fig. 2(a)) for generating a concatenated histogram for similarity comparison.

2.3. Audio-visual fusion

The audio and video localization results are fused in PF for the estimation of the final target 3D position. The fusion process is divided in four steps, namely initialization, prediction, update and re-sampling.

Let N be the total number of particles. Particles are first initialized as $\boldsymbol{s}_0^{(n)} \sim p(\boldsymbol{s}_0)$ with equal weights: $\omega_0^{(n)} = \frac{1}{N}$ for $n = 1, ..., N$. We assume that for each particle the variations of θ, φ and r are independent and particles are propagated (predicted) as:

$$\boldsymbol{s}_t^{(n)} = \boldsymbol{F}\mathbf{s}_{t-1}^{(n)} + \boldsymbol{q}_t^{(n)}, \tag{9}$$

where $\mathbf{s}_t^{(n)}$ is the state of the n^{th} particle at time-frame $t = 1, ..., T$; $\boldsymbol{q}_t$ is the Gaussian-distributed prediction noise with zero-mean and covariance $\boldsymbol{Q}$, $\boldsymbol{q}_t \sim \mathcal{N}(0, \boldsymbol{Q})$. $\boldsymbol{F}$ is a 6×6 prediction matrix which is represented as the first-order linear motion model.

In the update step both audio and video cues are used to evaluate the reliability of the particle, whose weight $\omega_t^{(n)}$ is updated according to a new observation set. By assuming audio-visual observations are independent from each other, each $\omega_t^{(n)}$ is proportional to the product of the individual likelihoods:

$$\omega_t^{(n)} \propto p(\boldsymbol{Z}_t^a|\boldsymbol{s}_t^{(n)})p(\boldsymbol{Z}_t^v|\boldsymbol{s}_t^{(n)}), \tag{10}$$

where $\mathbf{Z}_t^a$ and $\mathbf{Z}_t^v$ are single observations consisting of 3D location estimates in separate coordinates from individual modality and $\omega_t^{(n)}$ subjects to $\sum_{n=1}^{N} \omega_t^{(n)} = 1$. We assume the likelihoods in Eq. 10 follow Gaussian distribution with respect to the observation:

$$p(\mathbf{Z}_t^a|\boldsymbol{s}_t^{(n)}) \propto \exp\left[-(\mathbf{Z}_t^a - \mathbf{s}_t^{(n)})^T {\Lambda_t^a}^{-1} (\mathbf{Z}_t^a - \mathbf{s}_t^{(n)})\right], \tag{11}$$

$$p(\mathbf{Z}_t^v|\boldsymbol{s}_t^{(n)}) \propto \exp\left[-(\mathbf{Z}_t^v - \mathbf{s}_t^{(n)})^T {\Lambda_t^v}^{-1} (\mathbf{Z}_t^v - \mathbf{s}_t^{(n)})\right], \tag{12}$$

where Λ_t^a and Λ_t^v are diagonal matrices with the single coordinate variances i.e. $\Lambda_t^a = diag({\sigma_{\theta_t}^a}^2, {\sigma_{\phi_t}^a}^2, {\sigma_{r_t}^a}^2)$. We make the following standard deviations adaptive to the reliability measures:

$$\sigma_{\theta_t}^a = \alpha_\theta^a / G_t^{max}, \tag{13}$$

$$\sigma_t^v = \alpha^v / (1 - D_t), \tag{14}$$

where α_θ^a and α^v are user-defined constants indicating uncertainty of reliabilities. The larger their value, the broader the particle distribution. For the audio, we make $\sigma_{\theta_t}^a$ inversely proportional to the GCF peak value G_t^{max} and set $\sigma_{\varphi_t}^a, \sigma_{r_t}^a$ to constants. For the video, we make σ_t^v inversely proportional to $(1 - D_t)$ where D_t is the Bhattacharyya distance between the two colour histograms in Eq. 8. If one of the cues is unavailable, i.e. non-speech (VAD output λ= 0) or empty detection, the corresponding likelihood element is discarded.

Finally, the estimated target position $\boldsymbol{p}_t$ is computed as:

$$\boldsymbol{p}_t = \sum_{n=1}^{N} \omega_t^{(n)} \boldsymbol{s}_t^{(n)}, \tag{15}$$

Next, in the re-sampling step, particles are selected according to their assigned weights $\omega_t^{(n)}$. Particles with relatively high weights are duplicated while those with low weights are discarded [11].

(a) (b) (c)

Fig. 3. Key frames of test sequences (a) seq01 (b) seq03 (c) seq15 of the AV16.3 corpus [26].

3. RESULTS

3.1. Experimental setup

We compare the proposed approach with [3] on the AV16.3 corpus [26], which provides synchronized audio-visual data together with camera calibration information. The audio signals are captured by an 8-element circular microphone array with a diameter of $20cm$. This dataset is the first step towards a localised array in a mobile platform and the camera calibration information allows us to project data from the image plane to 3D world. Audio signals were recorded at 16 kHz and video sequences were recorded at 25 Hz. Each frame is made of 280×360 pixels. To make one-to-one correspondence of audio and video frame, we use a 1024-point Hanning window with 0.375 overlapping factor for audio spectra computation and the GCF searching grid resolution is $1^\circ \times 1^\circ \times 0.1m$. The VAD threshold e_{th} is set empirically to 0.06. We assume the real width of the bounding box to be $0.58m$ for 3D re-projection, which corresponds to our measured average length of a human shoulder plus its margin distance to the bounding box. The colour histograms are calculated in the RGB space with $16 \times 16 \times 16$ bins. The number of particles is set to 300 based on the result of the evaluation in [18].

In the experiments we use the signals captured by camera 1 and microphone array 1 in sequence 01, 03 and 15 (key frames are shown in Fig. 3), which includes a single speaker. The influence of different microphone pair processing techniques on 3D SSL results are compared and the proposed audio-visual speaker tracking approach is evaluated against (1) audio-only tracking; (2) video-only tracking; and (3) audio-visual tracking with fixed parameters (implementation of [3]).

3.2. Influence of microphone pair selection

The precision of azimuth (θ) and elevation (φ) estimations using the GCF algorithm [23] with different microphone pair combinations are compared in Table 2, where P_θ and P_φ are percentage of correct estimations (whose mean absolute error (MAE) $\leq 5^\circ$) over all speech frames. We do not consider the radius here as it is outside the focus of this paper. From the results, we conclude that, for a circular array, the estimation of θ is more accurate than that of φ as the microphone locations are dense on the azimuth plane. By using all the microphone pairs (Fig. 4(f)) as we do, we can always get the best performance while using only eight adjacent microphone pairs (Fig. 4(b)) gives the worst performance.

3.3. Tracking results comparison

The MAE of 3D speaker tracking of audio-only (AT), video-only (VT), classical audio-visual tracking approach [3] with fixed parameters (F-AVT) and the proposed approach with adaptive paramet-

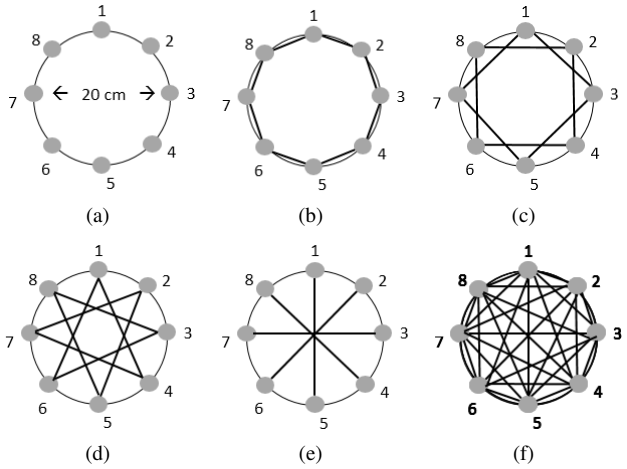


Fig. 4. (a) Top view of a uniform circular microphone array, (b)-(f) different combinations of microphone pairs. (grey filled circle: microphone, black line connection: microphone pair).

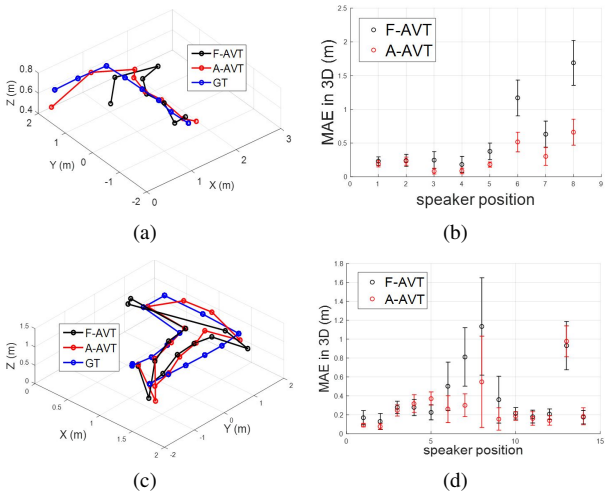


Fig. 5. Results on seq01 (top row) and seq03 (bottom row). Left column: 3D trajectories of ground truth (GT), A-AVT and F-AVT. (The average estimated position is plotted when the speaker remains stationary.) Right column: 3D MAE error and its standard deviation at different speaker positions.

Table 2. Precision of θ and φ estimation (in %) with different microphone pair combinations. (MG: microphone pair geometry, the grid resolution is $1^\circ \times 1^\circ \times 0.1m$).

MG	seq01		seq03		seq15	
	P_θ (%)	P_φ (%)	P_θ (%)	P_φ(%)	P_θ(%)	P_φ (%)
Fig. 4(b)	51.16	37.15	59.47	22.71	65.26	16.43
Fig. 4(c)	55.24	38.90	69.32	48.38	64.79	53.52
Fig. 4(d)	53.63	37.78	72.52	47.31	69.01	53.52
Fig. 4(e)	56.00	37.82	68.48	56.97	67.61	48.83
Fig. 4(f)	**74.98**	**65.71**	**83.49**	**77.21**	**84.51**	**64.79**

ers (A-AVT) is listed in Table 3 and illustrated in Fig. 5. Each experiment was run five times and we took the average as the final results. We set standard deviations σ_t^a and σ_t^v in Λ_t^a and Λ_t^v to (0.3, 0.3, 0.5) in separate spherical coordinates for the F-AVT. The value of α_θ^a was 0.03 and 0.05 which indicates 99.9% of observation error locates within $\pm 12^\circ$ and $\pm 20^\circ$. As the range of D is around 3 times of G^{max}, we make $\alpha^v \approx 3\alpha_\theta^a$ to normalise error distribution of individual audio and video modalities. The proposed A-AVT outperforms the accuracy of the individual modalities because of the restricted particle distributions by using the reliability measures in the likelihood computation. Additionally, in most cases, the proposed A-AVT shows a more stable performance with a reduced standard deviation of errors.

Because of the special sensor configuration, for both a single camera and a circular microphone array, errors are mainly caused by the radius estimation. Large errors in the middle parts of sequences are due to poor detection of the upper-body detector. False negatives occur when the speaker stands in a dark area and does not face the camera. Moreover, during a silent period the particle filter has no information to use, which makes the error larger.

4. CONCLUSION

We proposed a PF based framework for 3D speaker tracking using audio-visual signals from a circular microphone array and a standard camera. We separately weight in a late fusion phase the reliability of audio and video information, using colour histogram distance for images and GCF peak value for audio. Results show improved 3D tracking accuracy which indicates the potential feasibility of separately measuring the reliability level of (θ, φ, r) information for adaptive 3D target tracking in robotics.

Table 3. MAE of 3D speaker tracking in m with standard deviations in brackets. (AT: audio-only, VT: video-only, F-AVT: audio-visual with fixed parameters [3], A-AVT: proposal with adaptive parameters. $(\alpha_\theta^a, \alpha^v)$=$(0.03, 0.1)$ in Setup1 and $(0.05, 0.15)$ in Setup2).

	seq01		seq03	
Method	**Setup1**	**Setup2**	**Setup1**	**Setup2**
AT	.655 (.65)	.670 (.31)	.349 (.35)	.379 (.19)
VT	.528 (.52)	.522 (.52)	.828 (.82)	1.067 (1.37)
F-AVT	.458 (.46)	.510 (.45)	.324 (.32)	.502 (.40)
A-AVT	**.243** (.24)	**.247** (.20)	**.270** (.27)	**.300** (.29)

The major limitation of the proposed method is the inaccurate radius estimation (the main reason for errors). We will address this crucial aspect using the mobility of the robotic platform. Another limitation is that we still need to set the value of some parameters in adaptive fusion, which makes our approach less robust to environmental variations. Therefore, our future work will focus on improving the accuracy and robustness of 3D speaker tracking using adaptive audio-visual fusion under a noisy environment with a mobile platform.

5. REFERENCES

[1] A. Brutti, M. Omologo, and P. Svaizer, "A sequential monte carlo approach for tracking of overlapping acoustic sources," in *Proceedings of the European Signal Processing Conference*, 2009, pp. 2559–2563.

[2] E. Maggio and A. Cavallaro, "Multi-part target representation

for color tracking," in *Proceedings of IEEE International Conference on Image Processing*, 2005, vol. 1, pp. 729–732.

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

[4] Y. Wang and A. Cavallaro, "Prioritized target tracking with active collaborative cameras," in *Proceedings of IEEE International Conference on Advanced Signal and Video based Surveillance*, 2016, pp. 131–137.

[5] P. Tresadern, C. McCool, N. Poh, P. Matejka, A. Hadid, C. Levy, T. Cootes, and S. Marcel, "Mobile biometrics: joint face and voice verification for a mobile platform," *IEEE Pervasive Computing*, vol. 99, no. 1, pp. 79–87, 2012.

[6] R. B. Rusu, A. Holzbach, R. Diankov, G. Bradski, and M. Beetz, "Perception for mobile manipulation and grasping using active stereo," in *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, 2009, pp. 632–638.

[7] 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, pp. 275–283, 2011.

[8] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A Fink, and G. Sagerer, "Audiovisual person tracking with a mobile robot," in *Proceedings of International Conference on Intelligent Autonomous Systems*, march 2004, pp. 898–906.

[9] E. D'Arca, N. M Robertson, and J. Hopgood, "Person tracking via audio and video fusion," in *Proceedings of Data Fusion & Target Tracking Conference: Algorithms & Applications*. IET, 2012, pp. 1–6.

[10] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 118–121.

[11] M S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[12] I. D Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Audio-visual speech-turn detection and tracking," in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 143–151.

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

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

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

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

[17] M. Taj and A. Cavallaro, "Audio-assisted trajectory estimation in non-overlapping multi-camera networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3517–3520.

[18] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proceedings of ACM International Conference on Multimodal Interfaces*, 2005, pp. 61–68.

[19] H. Zhou, M. Taj, and A. Cavallaro, "Audiovisual tracking using STAC sensors," in *Proceedings of ACM/IEEE International Conference on Distributed Smart Cameras*, 2007, pp. 170–177.

[20] M. Heuer, A. Al-Hamadi, B. Michaelis, and A. Wendemuth, "Multi-modal fusion with particle filter for speaker localization and tracking," in *Proceedings of IEEE International Conference on Multimedia Technology*, 2011, pp. 6450–6453.

[21] M. Omologo, P. Svaizer, and R. De Mori, "Acoustic transduction," *Spoken Dialogue with Computers*, pp. 23–69, 1998.

[22] R. G Bachu, S. Kopparthi, B. Adapa, and B. D Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Society for Engineering Education Zone Conference Proceedings*, 2008, pp. 1–7.

[23] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4349–4352.

[24] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, 2001.

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

[26] G. Lathoud, J. Odobez, and D. Gatica-Perez, "AV16. 3: an audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*, pp. 182–195. Springer, 2004.