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

Xinyuan Qian¹, Alessio Bruttí², Maurizio Omologo², Andrea Cavallaro¹

¹Centre for Intelligent Sensing, Queen Mary University of London, UK
²ICT-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; NC: number of cameras; NM: number of microphone arrays x number of microphones in each array; R: reference; Prop.: proposed method.

<table>
<thead>
<tr>
<th>Ref</th>
<th>C</th>
<th>NC</th>
<th>NM</th>
<th>R</th>
<th>Fusion</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>[13]</td>
<td>Dis.</td>
<td>1</td>
<td>1 × 8</td>
<td>Image</td>
<td>Hybrid</td>
<td>PF</td>
</tr>
<tr>
<td>[16]</td>
<td>Dis.</td>
<td>3</td>
<td>1 × 8</td>
<td>Image</td>
<td>Hybrid</td>
<td>PF</td>
</tr>
<tr>
<td>[9]</td>
<td>Dis.</td>
<td>1</td>
<td>4 × 2</td>
<td>Ground</td>
<td>Late</td>
<td>KF, PF</td>
</tr>
<tr>
<td>[17]</td>
<td>Dis.</td>
<td>5</td>
<td>5 × 2</td>
<td>Ground</td>
<td>Late</td>
<td>KF</td>
</tr>
<tr>
<td>[10]</td>
<td>Dis.</td>
<td>4</td>
<td>4 × 4</td>
<td>3D</td>
<td>Late</td>
<td>KF</td>
</tr>
<tr>
<td>[14]</td>
<td>Dis.</td>
<td>5</td>
<td>3 × 4</td>
<td>3D</td>
<td>Hybrid</td>
<td>KF, PF</td>
</tr>
<tr>
<td>[18]</td>
<td>Dis.</td>
<td>4</td>
<td>3 × 4</td>
<td>3D</td>
<td>Late</td>
<td>PF</td>
</tr>
<tr>
<td>[15]</td>
<td>Dis.</td>
<td>4</td>
<td>7 × 4</td>
<td>3D</td>
<td>Late</td>
<td>PF</td>
</tr>
<tr>
<td>[19]</td>
<td>Loc.</td>
<td>1</td>
<td>1 × 2</td>
<td>Image</td>
<td>Late</td>
<td>KF, PF</td>
</tr>
<tr>
<td>[20]</td>
<td>Loc.</td>
<td>1</td>
<td>1 × 2</td>
<td>Image</td>
<td>Late</td>
<td>PF</td>
</tr>
<tr>
<td>Prop.</td>
<td>Loc.</td>
<td>1</td>
<td>1 × 8</td>
<td>3D</td>
<td>Late</td>
<td>PF</td>
</tr>
</tbody>
</table>
spherical coordinates. The origin of the spherical coordinates is the centre of the circular microphone array.

Let the state vector \( s \) be defined as

\[
    s = (\theta, \varphi, r, v_y, v_\varphi, v_r)^T,
\]

where \( (\theta, \varphi, r) \) indicates azimuth, elevation and radius; \( v_r \) is the velocity in radius direction; and \( v_y \) and \( v_\varphi \) 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 \( \varepsilon_{th} \). 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

\[
    \phi_m(t, f) = \frac{S_{m_1}(t, f) S_{m_2}(t, f)^*}{|S_{m_1}(t, f)||S_{m_2}(t, f)|},
\]

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

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

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

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

\[
    G(t, p) = \frac{1}{M} \sum_{m=0}^{M-1} C_m(t, \delta_m(p)),
\]

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

\[
    p_i^t = \arg \max_{p \in \mathcal{P}} G(t, p),
\]

where \( \mathcal{P} \) denotes the set of points of the 3D grid under analysis and \( p \) 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 \( x_t \) be

\[
    x_t = (u, v, w, h)^T_i,
\]

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 \( \alpha_t \) (i.e. middle point of the bounding box \( x_t \), Fig. 2(a)) of the target from the image plane to the 3D world coordinates:

\[
    p_i^t = H \alpha_t,
\]

where \( p_i^t \) is the estimated mouth position from the video and \( H \) 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 \( x_t \) using the Bhattacharyya distance:

\[
    D_t = \sqrt{1 - \sum_{u=1}^{U} \sqrt{r(u)q(u)}},
\]

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

---

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

**Fig. 2.** (a) Estimated mouth position (red filled circle) and 3 x 3 sub-images division of the upper-body detection result \( x_t \), (b) proposed sensor configuration for a mobile platform, (c) a generic 3D point \( p \) in spherical coordinates.
Let $N$ be the total number of particles. Particles are first initialized as $s_0 \sim p(x_0)$ with equal weights: $\omega_0(n) = \frac{1}{N}$ for $n = 1, ..., N$. We assume that for each particle the variations of $\theta$, $\varphi$ and $r$ are independent and particles are propagated (predicted) as:

$$s^{(n)}_t = Fs^{(n)}_{t-1} + q^{(n)}_t,$$

where $s^{(n)}_t$ is the state of the $n^{th}$ particle at time-frame $t = 1, ..., T$; $q^{(n)}_t$ is the Gaussian-distributed prediction noise with zero-mean and covariance $Q$, $q^{(n)}_t \sim N(0, Q)$. $F$ is a $6 \times 6$ 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 $\omega^{(n)}_t$ is updated according to a new observation set. By assuming audio-visual observations are independent from each other, each $\omega^{(n)}_t$ is proportional to the product of the individual likelihoods:

$$\omega^{(n)}_t \propto p(Z_t^{(a)}|s^{(n)}_t)p(Z_t^{(v)}|s^{(n)}_t),$$

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

$$p(Z_t^{(a)}|s^{(n)}_t) \propto \exp \left[-(Z_t^{(a)} - s_t^{(n)})^T \Lambda^{a}_t^{-1} (Z_t^{(a)} - s_t^{(n)})\right],$$

$$p(Z_t^{(v)}|s^{(n)}_t) \propto \exp \left[-(Z_t^{(v)} - s_t^{(n)})^T \Lambda^{v}_t^{-1} (Z_t^{(v)} - s_t^{(n)})\right],$$

where $\Lambda^{a}_t$ and $\Lambda^{v}_t$ are diagonal matrices with the single coordinate variances i.e. $\Lambda^{a}_t = diag(\sigma_{\theta}^2, \sigma_{\varphi}^2, \sigma_{r}^2)$. We make the following standard deviations adaptive to the reliability measures:

$$\sigma_{\theta}^2 = \alpha_{\theta}^2/g^{\max},$$

$$\sigma_{\varphi}^2 = \alpha_{\varphi}^2/(1 - D_t),$$

where $\alpha_{\theta}^2$ and $\alpha_{\varphi}^2$ are user-defined constants indicating uncertainty of reliabilities. The larger their value, the broader the particle distribution. For the audio, we make $\sigma_{\theta}^2$ inversely proportional to the GCF peak value $G^{\max}$ and set $\sigma_{\varphi}^2$, $\sigma_{r}^2$ to constants. For the video, we make $\sigma_{\varphi}^2$ inversely proportional to $(1 - D_t)$ where $D_t$ is the Bhat-tacharyya distance between the two colour histograms in Eq. 8. If one of the cues is unavailable, i.e. non-speech (VAD output $\lambda = 0$) or empty detection, the corresponding likelihood element is discarded.

Finally, the estimated target position $p_t$ is computed as:

$$p_t = \sum_{n=1}^{N} \omega^{(n)}_t s^{(n)}_t,$$

Next, in the re-sampling step, particles are selected according to their assigned weights $\omega^{(n)}_t$. Particles with relatively high weights are duplicated while those with low weights are discarded [11].
ers (A-A VT) 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 of $\theta$ and $\phi$ estimation (in %) with different microphone pair combinations. (MG: microphone pair geometry, the grid resolution is $1^\circ \times 1^\circ \times 0.1$m).

<table>
<thead>
<tr>
<th>MG</th>
<th>seq01</th>
<th>seq03</th>
<th>seq15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_\theta$ (%)</td>
<td>$P_\phi$ (%)</td>
<td>$P_\theta$ (%)</td>
</tr>
<tr>
<td>Fig. 4(b)</td>
<td>51.16</td>
<td>37.15</td>
<td>59.47</td>
</tr>
<tr>
<td>Fig. 4(d)</td>
<td>55.24</td>
<td>38.90</td>
<td>69.32</td>
</tr>
<tr>
<td>Fig. 4(e)</td>
<td>56.60</td>
<td>37.82</td>
<td>68.48</td>
</tr>
<tr>
<td>Fig. 4(f)</td>
<td>74.98</td>
<td>65.71</td>
<td>83.49</td>
</tr>
</tbody>
</table>

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 $(\theta, \phi, r)$ information for adaptive 3D target tracking in robotics.

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


