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

Xinyuan Qian\textsuperscript{1}, Alessio Xompero\textsuperscript{1}, Alessio Brutti\textsuperscript{2}, Oswald Lanz\textsuperscript{2}, Maurizio Omologo\textsuperscript{2}, Andrea Cavallaro\textsuperscript{1}

\textsuperscript{1}Centre for Intelligent Sensing, Queen Mary University of London, UK
\textsuperscript{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

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, \(p_t\), of a person over time \(t\), given audio signals, \(s_t\), 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 \(P(p | s_{1:t}, I_{1:t})\) of hypotheses \(p\) conditioned on past and current observations and then inferring the target state from \(P\), e.g. via expectation:

\[
\hat{p}_t = \mathbb{E}_P(p | s_{1:t}, I_{1:t}).
\]  

When the signal formation \(p \mapsto s, I\) 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
a data likelihood $L$, a first-order dynamics $Q$ and an initial density $dP_0$ [14]:

$$P(p | s_{1:t}, I_{1:t}) \propto L(s_t, I_t | p) \int Q(q | s_{1:t-1}, I_{1:t-1}) dP(q | s_{1:t-1}, I_{1:t-1}).$$

(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.,

$$\{p_1^{(1)}, \ldots, p_N^{(N)}\} \overset{\text{iid}}{\sim} P(\mathbf{s}_{1:t}, I_{1:t}).$$

(3)

This is achieved in two steps by (i) sampling from the prior mixture $\sum_{i=1}^{N} Q(p | p_i^{(n)})$ and (ii) re-sampling with probability $\propto L(s_t, I_t | p)$. 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 $L^a(s_t | p)$ and the video likelihood $L^v(I_t | p)$.

Our solution comprises the modelling of the individual likelihoods, $L^a, L^v$ (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 face1.

Let $\mathbf{f}_d = [u, v, w, h]^T$ 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. Geometrically extract the mouth position, $\mathbf{p}_d = [u + 0.5w, v + 0.75h]^T$, 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-projection2 the 3D mouth position: $\mathbf{o}_d^3 = \Phi[\mathbf{p}_d, w, W]$. We validate the output of the face detector with:

$$||\mathbf{o}_d^3 - \mathbf{p}_d^{t},\Delta t||_2 \leq \lambda \sqrt{w^2 + h^2}$$

(4)

where $\lambda$ controls the acceptable error range and $\mathbf{p}_d^{t},\Delta t$ is the average estimated mouth position on image plane in the last $\Delta t$ frames.

We use spherical coordinates to better model the higher inaccuracies in the distance estimation, which is based on the hypothesised face width $W$. Let $\tilde{o}_d^3$ and $\mathbf{p}$ 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:

$$L^a_{\text{det}}(I_t | \mathbf{p}) = \sum_{d=1}^{D} \exp \left[ - (\tilde{o}_d^3 - \mathbf{p}) \Sigma_v^{-1} (\mathbf{o}_d^3 - \mathbf{p})^T \right].$$

(5)

where $\Sigma_v$ 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 particle of the person using a Hue-Saturation-Value (HSV) spatiogram [16]. We measure the similarity $f$ between two spatiograms using [17], which is derived from the Bhattacharyya coefficients.

Finally, we define the visual likelihood as:

$$L^v(I_t | \mathbf{p}) = \begin{cases} L^v_{\text{det}}(I_t | \mathbf{p}) & \text{if } D > 0 \\ L^v_{\text{HSV}}(I_t | \mathbf{p}) & \text{if } \mathbf{p}_d^{t},\Delta t \in \mathcal{A} \text{ otherwise} \\ 1/N & \text{otherwise} \end{cases}$$

(6)

where $\mathcal{A}$ is a rectangular crop corresponds to the central 90% region of the image. It is used with $\mathbf{p}_d^{t},\Delta t$ 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 $p$ and $\tau_m(p)$ 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 $p$ as [21]:

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

(7)

1Size 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.

2The 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, circumvented if some knowledge about the mouth height is available, in particular along the range dimension. This problem can be from the maximum of the GCF acoustic map, when a compact 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 table in a room of size 4.77×5.95×4.5 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, in a variety of challenging poses (i.e. not oriented towards the sensors, bending over). Fig.3 shows sample frames of the two datasets.

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 \( p \) from

\[
\{p^{(1)}, \ldots, p^{(N)}\} \sim L^a(s_{t} | p)L^v(f_{t} | p) \sum_{n=1}^{N} \left( p; p^{(n)}, \kappa \right) \]

Here, we model first-order dynamics \( Q \) (Eq. 2) as a mixture of Gaussian distributions whose covariance matrix \( \Sigma_r \) is diagonal and \( \kappa \) 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:

\[
p_{t} = \frac{1}{N} \sum_{n=1}^{N} p^{(n)}
\]

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

\[3][https://github.com/tornadomeet/mxnet-face]
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>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>08</td>
<td>1</td>
<td>0.15</td>
<td>0.12</td>
<td>10.75 ± 0.13</td>
<td>4.31 ± 0.20</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.24</td>
<td>0.11</td>
<td>7.33 ± 0.09</td>
<td>4.66 ± 0.09</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.20</td>
<td>0.09</td>
<td>9.85 ± 0.12</td>
<td>5.34 ± 0.13</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.31</td>
<td>0.33</td>
<td>14.66 ± 0.34</td>
<td>8.15 ± 0.71</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.29</td>
<td>0.14</td>
<td>14.01 ± 0.12</td>
<td>7.48 ± 0.53</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.26</td>
<td>0.12</td>
<td>13.96 ± 0.23</td>
<td>6.64 ± 0.15</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.41</td>
<td>0.26</td>
<td>12.49 ± 0.16</td>
<td>6.86 ± 0.42</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.51</td>
<td>0.17</td>
<td>10.81 ± 0.24</td>
<td>10.67 ± 2.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.47</td>
<td>0.20</td>
<td>11.86 ± 0.24</td>
<td>9.71 ± 3.20</td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td>0.32</td>
<td>0.17</td>
<td>11.75 ± 0.19</td>
<td>7.09 ± 0.83</td>
</tr>
</tbody>
</table>

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


