AUDIO-VISUAL TRACKING BY DENSITY APPROXIMATION IN A SEQUENTIAL BAYESIAN FILTERING FRAMEWORK

Israel D. Gebru[1]*, Christine Evers*[2]*, Patrick A. Naylor*[2]*, Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes, France
[2]Imperial College London, Department of Electrical and Electronic Engineering, UK

ABSTRACT

The ability to explore and learn the surrounding environment is a major precondition for autonomous systems and applications including Human-Robot Interaction. Robot audition is particularly useful in situations where visual sensors suffer from limited Field of View or object occlusions. This is typically the case in scenarios where multiple talkers move freely within the environment surrounding the robot. Nevertheless, in enclosed environments, sound source localization is affected by reverberation of the sound waves off surrounding objects. Audio-visual fusion is therefore beneficial in order to disambiguate the positions of multiple moving talkers. This paper proposes a novel audio-visual tracking approach that exploits constructively both modalities in order to estimate the source trajectories in a joint state space. Recordings using a camcorder and microphone array are used to evaluate the proposed approach, demonstrating signficant improvements in tracking performance of the proposed audio-visual approach compared to two benchmark visual trackers.

***Index Terms*—** Motion estimation; Speech processing; Machine vision; Bayes methods; Audio-visual systems.

1. INTRODUCTION

Awareness of the surrounding environment is a prerequisite for interaction between humans and autonomous systems. In particular for Human-Robot Interaction (HRI), knowledge of the directions of sound sources in the surrounding acoustic environment is crucial in order to approach, look at, focus on, and engage with users.

However, in realistic conditions, speech radiated in enclosed environments is subject to reverberation due to reflections off surrounding walls and objects. Dominant early reflections therefore often lead to spurious false detections, whilst late reverberation causes localization errors. Furthermore, human talkers are highly dynamic sources, such that Directions-of-Arrival (DOAs) are both spatially- and time-varying. Therefore, source tracking approaches constructively exploit temporal models of the source dynamics to estimate and smooth the trajectory of corresponding DOAs. Whilst acoustic source tracking approaches, such as [1], propagate source trajectories through natural periods of speech inactivity within sentences, prolonged inactivity during dialogues often leads to track deletions.

Nevertheless, especially in the case of robotics, acoustic sensors are often coupled with synchronized camera systems. Therefore, visual information can be exploited constructively to disambiguate the practical challenges of acoustic signal processing. Recent contributions in the audio-visual community therefore utilize features extracted from visual signals to complement audio processing tasks and *vice versa*. The attention control system for mobile robots in [2, 3] uses separate but parallel audio and visual processing subsystems to identify salient events. In [4], a visual tracker is used to estimate the positions and velocities of people for blind source separation. Furthermore, a multi-person visual tracker is used for speaker diarization in multi-party dialogues in [5].

Nevertheless, to fully exploit the information of both modalities, audio-visual fusion – rather than disambiguation – is necessary. An audio-visual tracking system is proposed in [6, 7] that estimates the trajectories of talkers from the joint signals of several distributed cameras and a large microphone array. Nevertheless, joint audio-visual tracking approaches for compact configurations, such as for robot audition, are a novel contribution to the literature. The primary challenge for joint estimation from both modalities is that audio-visual observations are typically highly non-linear, non-Gaussian and multi-modal. The posterior Probability Density Function (pdf) is therefore analytically intractable. Nevertheless, classical approximations for non-linear systems, such as the Extended Kalman Filter (EKF) [8] and Unscented Kalman Filter (UKF) [9] are only valid for the estimation and propagation of unimodal Gaussian densities.

In this paper, we therefore propose to approximate the posterior density by a Gaussian Mixture Model (GMM). To ensure that a GMM representation can be retained over time, the predictive density is approximated by a Gaussian mixture model using the Unscented Transform (UT) [10]. The UT calculates the statistics of a random variable which undergoes a non-linear transformation and builds on the principle that it is easier to approximate a probability distribution than an arbitrary nonlinear function [10, 11]. Furthermore, we propose to use a density interpolation technique to approximate the observation likelihood function by a Gaussian Mixture (GM). To prevent the number of mixture from growing exponentially over time, a density approximation based on the Expectation Maximization (EM) algorithm is applied, resulting in a compact GM representation of the posterior pdf.

This paper is structured as follows: Section 2 introduces the audio-visual state and measurement models. The proposed tracking framework is detailed in Section 3. Experimental results are evaluated in Section 4, and conclusions drawn in Section 6.

2. SYSTEM MODEL

Consider a stream of synchronized sensory inputs, i.e., an image sequence and multi-channel microphone signals. Let t denote the time-step index of the audio-visual stream of data. The audio-visual tracking problem is modeled as a discrete-time Dynamic State Space Model (DSSM). The hidden system state, $\mathbf{X}_t \in \mathbb{R}^{\mathrm{D}}$, with initial distribution, $\mathrm{P}(\mathbf{X}_0)$, represents the two-dimensional (2D) location of

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

N human talkers on the image plane at time step t. The system state evolves over time as an unobserved first-order Markov process according to the state transition model given by the conditional pdf, $\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$. The audio-visual observations at t, denoted by $\mathbf{Z}_t = \{\mathbf{Z}_t^a, \mathbf{Z}_t^v\}$, are conditionally independent from all other variables given the state $\mathbf{X}_t$ and are generated according to the pdf:

$$\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t) = \mathrm{P}(\mathbf{Z}_t^a|\mathbf{X}_t)\mathrm{P}(\mathbf{Z}_t^v|\mathbf{X}_t), \tag{1}$$

where $\mathbf{Z}_t^a$ denotes the auditory observations extracted from the audio frame and $\mathbf{Z}_t^v$ denotes the visual observations extracted from the image frame at time-step t. The DSSM can also be written as a set of system equations representing the process and observation models as:

$$\mathbf{X}_t = \mathbf{f}(\mathbf{X}_{t-1}, \mathbf{V}_t) \quad \text{(process model)} \tag{2}$$
$$\mathbf{Z}_t^a = \mathbf{h}_a(\mathbf{X}_t, \mathbf{U}_t^a) \quad \text{(auditory observation model)} \tag{3}$$
$$\mathbf{Z}_t^v = \mathbf{h}_v(\mathbf{X}_t, \mathbf{U}_t^v) \quad \text{(visual observation model)} \tag{4}$$

where $\mathbf{V}_t$ denotes the process noise that drives the dynamic system through a nonlinear state transition function, $\mathbf{f}$, and where $\mathbf{U}_t^a, \mathbf{U}_t^v$ denotes the auditory and visual observation noise corrupting the observation of the system state through the nonlinear observation functions $\mathbf{h}_a$ and $\mathbf{h}_v$, respectively.

The face detector proposed in [12], and implemented in OpenCV,[1] is used to obtain the visual observations, $\mathbf{Z}_t^v$. However, the face detector is only reliable when a frontal face is presented and uninformative if a person turned away from the camera. The multi-source localization approach in [13] is used to obtain auditory observations, $\mathbf{Z}_t^a$, in the form of DOA estimates in azimuth and inclination.

A major challenge in audio-visual processing is the representation of auditory and visual observations in a common space. In this paper, a geometric transformation [3] is therefore applied to the DOAs in order to map the source directions from spherical space to a pixel position on the image plane. Therefore, auditory DOAs, visual facial detections, and the desired system states can be treated in the same mathematical space.

3. PROPOSED METHOD

Given all available observations $\mathbf{Z}_{1:t-1} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_{t-1}\}$ up to time-step $t-1$, the current posterior at t is predicted using the state transition model and prior, $\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})$, as:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})\,\mathrm{d}\mathbf{X}_{t-1}, \tag{5}$$

As the audio-visual observations, $\mathbf{Z}_t$, become available at t, the state can be updated using Bayes's theorem, i.e.,

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t}) = \frac{1}{C}\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) \tag{6}$$

where

$$C = \mathrm{P}(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}) = \int \mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1})\,\mathrm{d}\mathbf{X}_t \tag{7}$$

is a normalization constant. The posterior $\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t})$, also referred to as the filtering distribution, is used as the prior distribution at the next time-step $t+1$.

[1] Open Source Computer Vision Library. See `http://opencv.org/`

3.1. Predicted pdf

Assume that the prior pdf at t is given by a GMM with n_{t-1} components, i.e.,

$$\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) = \sum_{i=1}^{n_{t-1}} \pi_{t-1}^i \mathcal{N}(\mathbf{X}_{t-1}; \mathbf{X}_{t-1}^i, \mathbf{\Sigma}_{t-1}^i), \tag{8}$$

where $\{\pi_{t-1}^i\}_{i=1}^{n_{t-1}}$ are the mixing weights with $\sum_{i=1}^{n_{t-1}} \pi_{t-1}^i = 1$, $\{\mathbf{X}_{t-1}^i\}_{i=1}^{n_{t-1}}$ are the mean vectors with $\mathbf{X}_{t-1}^i \in \mathbb{R}^{\mathrm{D}}$ and $\{\mathbf{\Sigma}_{t-1}^i\}_{i=1}^{n_{t-1}}$ are the covariance matrices with $\mathbf{\Sigma}_{t-1}^i \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$. The UT enables the propagation of the means and covariance matrices through a nonlinear function. To calculate the statistics of $\mathbf{X}_t$ that undergoes a nonlinear transformation, a set of $2\mathrm{D}+1$ weighted samples or sigma points, $\{\boldsymbol{S}_{(i,j)}\}_{j=0,i=1}^{2\mathrm{D},n_{t-1}}$, are carefully chosen so that they capture the mean and covariance of the system state. A selection scheme that satisfies this requirement is [11, 14]:

$$\begin{aligned}
&\boldsymbol{S}_{(i,0)} = \mathbf{X}_{t-1}^i, \qquad w_{(i,0)} = \tfrac{\lambda}{(\mathrm{D}+\lambda)}, \qquad j = 0 \\
&\boldsymbol{S}_{(i,j)} = \mathbf{X}_{t-1}^i - \Big(\sqrt{(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i}\Big)_j, \qquad j = 1, \ldots, \mathrm{D} \\
&\boldsymbol{S}_{(i,j)} = \mathbf{X}_{t-1}^i + \Big(\sqrt{(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i}\Big)_{j-\mathrm{D}}, \quad j = \mathrm{D}+1, \ldots, 2\mathrm{D} \\
&w_{(i,j)} = \tfrac{1}{2(D+\lambda)} \quad j = 1, \ldots, 2\mathrm{D},
\end{aligned} \tag{9}$$

where $w_{(i,j)}$ is the weight associated with the jth sigma point such that $\sum_{j=0}^{2\mathrm{D}} w_{(i,j)} = 1$ and λ is a scaling parameter and $\Big(\sqrt{(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i}\Big)_j$ is the j^{th} row of the matrix square root of $(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i$. As the random variable undergoes a non-linear transformation, these points are propagated through this non-linear function and are used to reconstruct the new means and covariance matrices. The estimated mean, $\bar{\mathbf{X}}_t^i$, and covariance, $\bar{\mathbf{\Sigma}}_t^i$, of the i^{th} Gaussian component in the predicted distribution are approximated using a weighted sample mean and covariance of the sigma points. Finally, the GMM approximating the predicted pdf is given by:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \sum_{i=1}^{n_{t-1}} \pi_t^i \mathcal{N}(\mathbf{X}_t; \bar{\mathbf{X}}_t^i, \bar{\mathbf{\Sigma}}_t^i), \tag{10}$$

where

$$\pi_t^i = \pi_{t-1}^i, \qquad \bar{\mathbf{X}}_t^i = \sum_{j=0}^{2\mathrm{D}} w_{(i,j)}\boldsymbol{S}_{(i,j)}, \quad \text{and}$$

$$\bar{\mathbf{\Sigma}}_t^i = \sum_{j=0}^{2\mathrm{D}} w_{(i,j)}\Big(\boldsymbol{S}_{(i,j)} - \bar{\mathbf{X}}_t^i\Big)\Big(\boldsymbol{S}_{(i,j)} - \bar{\mathbf{X}}_t^i\Big)^\top + \mathbf{Q}_t.$$

Here, $\mathbf{Q}_t$ is the process noise convariance matrix.

3.2. Likelihood

The observation likelihood is a measure for evaluating which configuration of $\mathbf{X}_t$ best matches the observations $\mathbf{Z}_t$ at time-step t. In order to efficiently explore the possible configuration space of $\mathbf{X}_t$ and obtain a good approximation for the posterior distribution, m samples $\{\mathbf{X}_i\}_{i=1}^m$, are generated from the GM distribution given in (10). For each sample $\mathbf{X}_i$, the audio-visual likelihood l_i is computed as:

$$l_i = \beta_1 \times l_{app}(\mathbf{X}_i) + \beta_2 \times l_a(\mathbf{X}_i) + \beta_3 \times l_v(\mathbf{X}_i) \tag{11}$$

where l_{app} is function that evaluate appearance similarity between tracked targets and image regions indicated by the hypothesis state configuration $\mathbf{X}_i$, l_a is a function that measure the distance between $\mathbf{X}_i$ and auditory observations, and l_v measures the distance between $\mathbf{X}_i$ and visual observations, and $\{\beta_i\}_{i=1}^3$ are parameters that control

the influence of the likelihood functions. We note that the likelihood in (11) increases as audio-visual observations become available. However, the likelihood does not increase for missing detections. This property makes the model robust to absence of audio-visual observations e.g., in time of missed face detections and speech inactivity.

To obtain a continuous approximation of the likelihood given the discrete samples $\{\mathbf{X}_i\}_{i=1}^m$ and associated likelihood $\{l_i\}_{i=1}^m$, the Radial Basis Function (RBF) [15] is used for interpolation. Therefore, a Gaussian kernel is assigned to each sample $i = 1, \ldots, m$, such that the likelihood of $\mathbf{X}_j$ induced by the i^{th} kernel is given by

$$\mathrm{P}_i(\mathbf{X}_j) = \mathcal{N}(\mathbf{X}_j; \mathbf{X}_i, \mathbf{P}_i), \tag{12}$$

where the sample location, $\mathbf{X}_i$, is used as the mean and the covariance, or the kernel bandwidth, $\mathbf{P}_i$, is set to k-nearest neighbors (KNN) distance, i.e.,[2]

$$\mathbf{P}_i = c\,\mathrm{diag}\big(\mathrm{KNN}_1^i(k), \ldots, \mathrm{KNN}_\mathrm{D}^i(k)\big)\mathbf{I}, \tag{13}$$

where c is a constant that depends on the number of samples and the dimensionality, $\mathbf{I}$ is the D-dimensional identity matrix, and $\mathrm{KNN}_j^i(k)$ is the KNN distance of sample i in the j^{th} dimension.

Therefore, the observation likelihood $\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)$ is approximated by $n_t \ll m$ Gaussians, such that

$$\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t) = \sum_{i=1}^{n_t} w_i \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^i, \mathbf{P}_t^i). \tag{14}$$

The weight, w_i, of kernel $i = 1, \ldots, n_t$, is computed by solving the constrained non-negative least square problem [15]:

$$\begin{aligned} &\arg\min_{\mathbf{w}} \ \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2 \\ &\text{subject to elements of } \mathbf{w} \geq 0 \end{aligned} \tag{15}$$

where $\mathbf{A} \in \mathbb{R}^{m \times m}$ is a design matrix with each element (i, j) given by $\mathrm{P}_i(\mathbf{X}_j)$, the matrix $\mathbf{b} \in \mathbb{R}^{m \times 1}$ contains l_i for each row $i = 1, \ldots, m$, and $\mathbf{w} = (w_1, w_2, \ldots, w_m)$ is the kernel weight vector.

3.3. Posterior pdf

Both the predicted density in (10) and the observation likelihood in (14) are represented by GMs. Therefore, the posterior pdf in (6) is equivalent to a product of two GMs, and is therefore equivalent to a GM with an exponentially increasing number of components, i.e.,

$$\begin{aligned} &\left(\sum_{i=1}^{n_{t-1}} \pi_t^i \mathcal{N}(\mathbf{X}_t; \bar{\mathbf{X}}_t^i, \bar{\mathbf{\Sigma}}_t^i)\right)\left(\sum_{j=1}^{n_t} \tau_t^j \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^j, \mathbf{P}_t^j)\right) \\ &\quad = \sum_{i=1}^{n_{t-1}} \sum_{j=1}^{n_t} w_t^{ij} \mathcal{N}(\boldsymbol{\mu}_t^{ij}, \mathbf{\Sigma}_t^{ij}) \end{aligned} \tag{16}$$

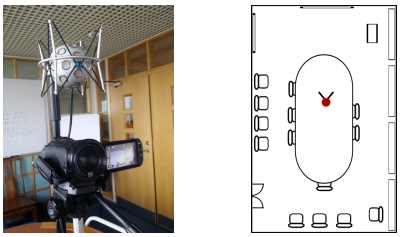
where

$$w_t^{ij} = \frac{\pi_t^i \tau_t^j \exp\left(-\frac{1}{2}(\mathbf{X}_t^j - \bar{\mathbf{X}}_t^i)^\top \left(\bar{\mathbf{\Sigma}}_t^i + \mathbf{P}_t^j\right)^{-1} (\mathbf{X}_t^j - \bar{\mathbf{X}}_t^i)\right)}{(2\pi)^{(\mathrm{D}/2)} \left|\bar{\mathbf{\Sigma}}_t^i + \mathbf{P}_t^j\right|^{1/2}} \tag{17}$$

$$\mathbf{\Sigma}_t^{ij} = \left(\left(\bar{\mathbf{\Sigma}}_t^i\right)^{-1} + \left(\mathbf{P}_t^j\right)^{-1}\right)^{-1} \tag{18}$$

$$\boldsymbol{\mu}_t^{ij} = \mathbf{\Sigma}_t^{ij}\left(\left(\bar{\mathbf{\Sigma}}_t^i\right)^{-1} \bar{\mathbf{X}}_t^i + \left(\mathbf{P}_t^j\right)^{-1} \mathbf{X}_t^j\right) \tag{19}$$

[2]The approach can be naturally extended to more complex approaches to kernel bandwidth selection, discussed in, e.g., [16].



(a) **(b)**

Fig. 1: Camera-microphone and recording room setup. (a) The camera-microphone setup used to recording test scenarios. (b) Recording room schematic (the position of the camera-microphone pair is shown with a red dot.

The resulting density function in (16) is a weighted mixture of Gaussians with $n_{t-1} \times n_t$ components.

To mitigate an exponential explosion in the number of components, we first note that typically many of the kernels correspond to weights close to zero and are therefore negligible. To remove stochastically irrelevant kernels, the weighted data EM in [17] is used to fit a GMM to (16). The optimal number of GM components, m_t, for fitting is selected in a principled manner based on the information-theoretic Minimum Message Length (MML) principle. We note that although the weighted data EM is not a general density approximation method, it does preserve significant mode locations and approximates well-separated GMs accurately.

Thus, the final posterior distribution at time-step t is given by:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t}) = \sum_{i=1}^{m_t} \pi_t^i \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^i, \mathbf{\Sigma}_t^i), \tag{20}$$

where m_t is the number of components [17].

4. EXPERIMENTAL SETUP

The camera-microphone setup shown in Fig. 1 is used to gather audio-visual recordings in order to evaluate the performance of the proposed model. The setup consists of a HD video camera and a 32 channel spherical microphone array (Eigenmike). The camera provides video frames with a resolution of 1920×1200 pixels at a rate of 25 Frames-Per-Second (FPS). The original audio signals are captured at 48 kHz, and are downsampled to 16 kHz. The acoustic impulse generated by a "clapperboard" is used for audio-visual synchronization.

Two scenarios are recorded in order to test the robustness of the proposed tracking model to dynamic scenes. The scenarios are referred to in the following as **S1** and **S2**. Both recordings are taken in the meeting room illustrated in Fig. 1b. In both scenarios, two talkers are speaking whilst moving within the room. The talkers occasionally occlude each other and move in and out of the camera field of view. In **S1**, the participants take speech turns, whilst the talkers speak simultaneously most of the time in **S2**. Scenario **S1** has 3000 video frames (2 minutes); **S2** has 2000 video frames (1 minute and 20 seconds). For both scenarios, the ground truth is identified by hand-annotating in each video frame the talker positions with bounding boxes around faces, 2D locations inferred from the bounding boxes, and target ID.

Face detection yields fully automatic initialization. The algorithm initializes a new tracker for a person that has subsequent detections with overlapping bounding boxes, which are neither oc-

cluded nor associated to an already existing tracker. The initial mean positions are drawn from a Gaussian distribution centered at the detection center. The initial size corresponds to the detection size. Appearance model based on RGB color histogram is initialized from image region around the initial bouding box. Conversely, if no enough detections are found for the same target within T consecutive frames, the track is automatically terminated. Since addition or termination of target either increases or decreases the system state dimension, this procedure is performed at the start of each-time step, prior to density approximation. The Hungarian algorithm [18] is used to obtain the optimal face detection to existing target association. The association cost matrix is based on the pairwise similarity of RGB histogram between existing targets and detections.

The common CLEAR MOT metrics [19] consisting of multiple metrics are used for evaluation. The Multiple Object Tracking Precision (MOTP) evaluates the intersection area over the union area of bounding boxes. The Multiple Object Tracking Accuracy (MOTA) calculates the accuracy composed of false negatives, false positives and identity switching. Furthermore, the Optimal MAss Transfer (OMAT) metric [20] and the Optimal Aub-Pattern Assignment (OSPA) metric [21] are evaluated in order to evaluate tracking performance independent of track labelling. These metrics compute set distances between the ground truth set of objects present in the scene and the set of objects estimated by the tracker.

5. EXPERIMENTAL VALIDATION

The performance of the proposed tracking model using auditory and visual observations (P-AV) is compared against the proposed tracking model without the auditory data (P-V) and the multi-person visual tracker (OV) in [22].

The results are summarized in Table 1, whilst the audio-visual tracking results for the proposed P-AV are illustrated for video frames 351, 352, 451, and 452 of S1 in Fig. 2. The full videos, Matlab code and additional examples are available online[3].

The comparison between P-AV and P-V highlights that constructive exploitation of the auditory observations leads to improved disambiguation of the states of the moving talkers. An improvement of 11.3 percentage points in MOTA is achieved using P-AV compared to P-V for S1. For S2, the improvement for the audio-visual tracker corresponds to 3.5 percentage points. The difference in results between S1 and S2 can be explained by the difficulty of the scenarios. In S2, the position of the talkers is mostly static, mainly affected by body and head rotations. Therefore, as facial occlusions are rare, face detection is highly reliable. In S1, frequent visual occlusions and obscurations as well as crossing speaker paths lead to less reliable visual detections. Furthermore, the talkers frequently speak simultaneously. Hence, audio DOAs for both talkers are exploited constructively for audio-visual disambiguation in frames where faces are visually occluded.

The results also highlight that the proposed approach outperforms the benchmark OV in both scenarios. This is due to propagation of multiple Gaussians for each target at each time-step by the proposed model, leading to increased robustness to measurement ambiguity. For S1, P-AV therefore results in an improvement of 37.9 and 32.7 percentage points in MOTA and MOTP respectively compared to OV. For S2, an improvement of 45.5 and 32.7 percentage points is achieved.

[3] http://team.inria.fr/perception/avtracking_by_dabf/

Table 1: Tracking results performance comparison. ↑ denotes higher scores indicate better results, and ↓ denotes lower scores indicate better results.

Sequence	Methods	MOTA (in %)↑	MOTP (in %)↑	OMAT↓	OSPA↓
S1	P-AV	83.6	89.1	186.6	112.7
	P-V	72.3	82.8	245.6	234.0
	OV[22]	45.7	56.4	378.6	367.8
S2	P-AV	89.8	84.8	201.4	198.0
	P-V	86.3	80.9	200.4	193.6
	OV[22]	44.3	47.6	356.7	367.8

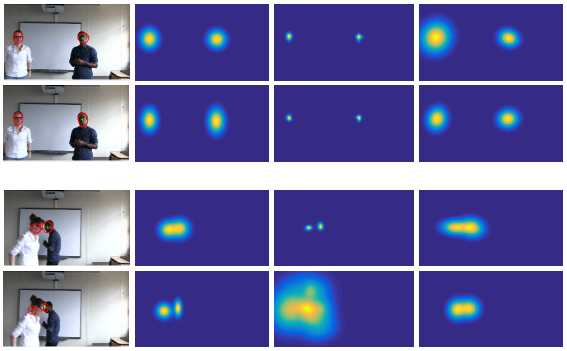


Fig. 2: Examples of results obtained on scenario S1. Each row shows results at different video frame: first and second row are from video frame(#351,#352); third and fouth row are from video frame(#451,#452). The first column shows the final GMM. The second, third and fourth columns show the posterior, likelihood and predictive probability density plots on the image space, respectively.

6. DISCUSSION AND CONCLUSION

This paper presented a novel approach to audio-visual tracking that jointly utilizes multiple DOAs obtained by sound source localization and facial detections to estimate the trajectories of multiple talkers in pixel space. Within a Bayesian framework, the posterior multi-source pdf is propagated using a GMM. To retain the Gaussianity over time, the non-linear observation likelihood function is approximated by a density interpolation technique. Furthermore, to avoid the exponential explosion of Gaussians, an EM algorithm is used to cluster Gaussians, thereby retaining only statistically relevant components. Results based on measurements with a camcorder and eigenmike demonstrate signficant improvements in performance for audio-visual tracking compared to using only visual data.

7. REFERENCES

[1] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 1206–1210.

[2] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, pp. 28–35.

[3] J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Čech, S. Wrede, and R. Horaud, "Online multimodal speaker detection for humanoid robots," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*. IEEE, 2012, pp. 126–133.

[4] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895–910, 2010.

[5] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *arXiv preprint arXiv:1603.09725*, 2016.

[6] V. Kılıç, 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.

[7] D. Gatica-Perez, G. Lathoud, J.-M. 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.

[8] B. D. Anderson and J. B. Moore, "Optimal filtering. 1979," 1979.

[9] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *AeroSense'97*. International Society for Optics and Photonics, 1997, pp. 182–193.

[10] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. Wan, "The unscented particle filter," in *NIPS*, vol. 2000, 2000, pp. 584–590.

[11] S. J. Julier and J. K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford, Tech. Rep., 1996.

[12] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[13] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*. IEEE, 2014, pp. 258–262.

[14] S. S. Haykin *et al.*, *Kalman filtering and neural networks*. Wiley Online Library, 2001.

[15] T. Poggio and F. Girosi, "A theory of networks for approximation and learning," DTIC Document, Tech. Rep., 1989.

[16] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690, 1991.

[17] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "Em algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[18] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[19] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[20] J. R. Hoffman and R. P. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 34, no. 3, pp. 327–336, 2004.

[21] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

[22] S. Ba, X. Alameda-Pineda, A. Xompero, and R. Horaud, "An on-line variational Bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.