This paper proposes a novel audio-visual tracking approach that exploits constructively audio and visual modalities in order to estimate trajectories of multiple people in a joint state space. The tracking problem is modeled using a sequential Bayesian filtering framework. Within this framework, we propose to represent the posterior density with a Gaussian Mixture Model (GMM). To ensure that a GMM representation can be retained sequentially over time, the predictive density is approximated by a GMM using the Unscented Transform. While a density interpolation technique is introduced to obtain a continuous representation of the observation likelihood, which is also a GMM. Furthermore, to prevent the number of mixtures from growing exponentially over time, a density approximation based on the Expectation Maximization (EM) algorithm is applied, resulting in a compact GMM representation of the posterior density. Recordings using a camcorder and microphone array are used to evaluate the proposed approach, demonstrating significant 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 images 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 [5,7] 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 GMM 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 GMM. To prevent the number of mixtures from growing exponentially over time, a density approximation based on the Expectation Maximization (EM) algorithm is applied, resulting in a compact GMM 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 5.

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, \( X_t \in \mathbb{R}^D \), with initial distribution, \( P(X_0) \), represents the two-dimensional (2D) location of \( N \) 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, \( P(X_t|X_{t-1}) \). The audio-visual observations at \( t \), denoted by \( Z_t = \{Z_t^a, Z_t^v\} \), are conditionally independent from all other variables given the state \( X_t \) and are generated according to the pdf:

\[
P(Z_t|X_t) = P(Z_t^a|X_t)P(Z_t^v|X_t),
\]

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

\[
X_t = f(X_{t-1}, V_t) \quad \text{(process model)} \tag{2}
\]
\[
Z_t^a = h_a(X_t, U_t^a) \quad \text{(auditory observation model)} \tag{3}
\]
\[
Z_t^v = h_v(X_t, U_t^v) \quad \text{(visual observation model)} \tag{4}
\]

where \( V_t \) denotes the process noise that drives the dynamic system through a nonlinear state transition function, \( f \), and where \( U_t^a, U_t^v \) denote the auditory and visual observation noise corrupting the observation of the system state through the nonlinear functions \( h_a \) and \( h_v \), respectively.

The face detector proposed in [12], and implemented in OpenCV [13], is used to obtain the visual observations, \( Z_t^v \). However, the face detector is only reliable when a frontal face is presented and uninformative if a person turned his/her face away from the camera. The multi-source localization approach in [14] is used to obtain auditory observations, \( Z_t^a \), 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, \( Z_{1:t-1} = \{Z_1, \ldots, Z_{t-1}\} \) up to time-step \( t-1 \), the current posterior at \( t \) is predicted using the state transition model and prior, \( P(X_t|X_{1:t-1}) \), as:

\[
P(X_t|Z_{1:t-1}) = \int P(X_t|X_{t-1})P(X_{1:t-1}|Z_{1:t-1}) \, dX_{t-1}. \tag{5}
\]

As the audio-visual observations, \( Z_t \), become available at \( t \), the state can be updated using Bayes’s theorem, i.e.,

\[
P(X_t|Z_{1:t}) = \frac{1}{C}P(Z_t|X_t)P(X_t|Z_{1:t-1}), \tag{6}
\]

where

\[
C = P(Z_t|Z_{1:t-1}) = \int P(Z_t|X_t)P(X_t|Z_{1:t-1}) \, dX_t
\]

is a normalization constant. The posterior \( P(X_t|Z_{1:t}) \), also referred to as the filtering distribution, is used as the prior distribution at the next time-step \( t+1 \).

#### 3.1. Predicted pdf

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

\[
P(X_{t-1}|Z_{1:t-1}) = \sum_{i=1}^{n_{t-1}} \pi_{t-1}^i \mathcal{N}(X_{t-1}; \bar{X}^i_{t-1}, \Sigma_{t-1}^i), \tag{8}
\]

where \( \{\pi_{t-1}^i\}_{i=1}^{n_{t-1}} \) are the mixing weights with \( \sum_{i=1}^{n_{t-1}} \pi_{t-1}^i = 1 \), \( \{\bar{X}^i_{t-1}\}_{i=1}^{n_{t-1}} \) are the mean vectors with \( \bar{X}^i_{t-1} \in \mathbb{R}^D \) and \( \{\Sigma_{t-1}^i\}_{i=1}^{n_{t-1}} \) are the covariance matrices with \( \Sigma_{t-1}^i \in \mathbb{R}^{D \times D} \). The UT enables the propagation of the means and covariance matrices through a nonlinear function. To calculate the statistics of \( X_t \) that undergoes a non-linear transformation, a set of \( 2D + 1 \) weighted samples or sigma points, \( \{S(u_{t,j}))_{j=0,1,\ldots,D} \), are carefully chosen so that they capture the mean and covariance of the system state. A selection scheme that satisfies this requirement is [15]:

\[
S(u_{t,0}) = X_{t-1}, \quad w(u_{t,0}) = \frac{1}{D+1}, \quad j = 0 \tag{9}
\]
\[
S(u_{t,j}) = X_{t-1} + (\sqrt{(D+\lambda)\Sigma_{t-1}^i})_j, \quad j = 1, \ldots, D \tag{10}
\]

where \( w(u_{t,j}) \) is the weight associated with the jth sigma point such that \( \sum_{j=0}^{2D} w(u_{t,j}) = 1 \) and \( \lambda \) is a scaling parameter and \( (\sqrt{(D+\lambda)\Sigma_{t-1}^i})_j \) is the \( j \)th row of the square root of \( (D+\lambda)\Sigma_{t-1}^i \). 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, \( \bar{X}_t \), and covariance, \( \Sigma_t \), of the \( j \)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:

\[
P(X_t|Z_{1:t}) = \sum_{i=1}^{n_{t-1}} \pi_{t}^i \mathcal{N}(X_t; \bar{X}^i_t, \Sigma_t^i), \tag{11}
\]

where

\[
\bar{X}^i_t = \sum_{j=0}^{2D} w(u_{t,j})S(u_{t,j}), \quad \text{and}
\]

\[
\Sigma_t^i = \sum_{j=0}^{2D} w(u_{t,j}) \left( S(u_{t,j}) - \bar{X}_t \right)^\top \left( S(u_{t,j}) - \bar{X}_t \right) + Q_t.
\]

Here, \( Q_t \) is the process noise covariance matrix.

#### 3.2. Likelihood

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

\[
l_t = \beta_1 l_{app}(X_t) + \beta_2 l_a(X_t) + \beta_3 l_v(X_t), \tag{12}
\]

where \( l_{app} \) is function that evaluate appearance similarity between tracked targets and image regions indicated by the hypothesis state configuration \( X_t \), \( l_a \) is a function that measure the distance between \( X_t \) and auditory observations, \( l_v \) measures the distance between \( X_t \) and visual observations. \( \{\beta_i\}_{i=1}^{3} \) are parameters that control the influence of the likelihood functions. We set a large value for \( \beta_i \) since the appearance similarity is more reliable compared to other two distance measures. 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 \( \{ X_i \}_{i=1}^{n_t} \) and associated likelihood \( \{ l_i \}_{i=1}^{n_t} \), the Radial Basis Function (RBF) [16] is used for interpolation. Therefore, a Gaussian kernel is assigned to each sample \( i = 1, \ldots, m \), such that the likelihood of \( X_i \) induced by the \( i \)th kernel is given by
\[
P_i(X_j) = \mathcal{N}(X_j; X_i, P_i),
\]
where the sample location, \( X_i \), is used as the mean and the covariance, or the kernel bandwidth, \( P_i \), is set to \( K \)-nearest neighbors (KNN) distance, i.e.
\[
P_i = c \text{ diag}(\text{KNN}_1(k), \ldots, \text{KNN}_m(k)) I,
\]
where \( c \) is a constant that depends on the number of samples and the dimensionality, \( I \) is the \( D \)-dimensional identity matrix, and \( \text{KNN}_i(k) \) is the KNN distance of sample \( i \) in the \( j \)th dimension.

Therefore, the observation likelihood \( P(Z_t|X_i) \) is approximated by \( n_t \ll m \) Gaussians, such that
\[
P(Z_t|X_i) = \sum_{i=1}^{n_t} w_i \mathcal{N}(X_j; X_i, P_i).
\]
The weight, \( w_i \), of kernel \( i = 1, \ldots, n_t \), is computed by solving the constrained non-negative least square problem [16]:
\[
\arg \min_{\mathbf{w}} \| \mathbf{A} \mathbf{w} - \mathbf{b} \|_2^2
\]
subject to elements of \( \mathbf{w} \geq 0 \)
where \( \mathbf{A} \in \mathbb{R}^{m \times m} \) is a design matrix with each element \((i, j)\) given by \( P_i(X_j) \), the matrix \( \mathbf{b} \in \mathbb{R}^{m \times 1} \) contains \( l_i \) for each row \( i = 1, \ldots, m \), and \( \mathbf{w} = [w_1, w_2, \ldots, w_m] ^\top \) is the kernel weight column-vector.

\[3.3.\text{Posterior pdf}\]

Both the predicted density in [14] and the observation likelihood in [1] 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.,
\[
\left( \sum_{i=1}^{n_t-1} \pi_i \mathcal{N}(X_j; \Sigma_i, \mu_i) \right) \left( \sum_{i=1}^{n_t} \pi_i \mathcal{N}(X_j; \Sigma_i, \mu_i) \right) = \sum_{j=1}^{n_t-1} \sum_{i=1}^{n_t} \pi_{ij} \mathcal{N}(\mu_{ij}, \Sigma_{ij}),
\]
where
\[
\pi_{ij} = \frac{\exp \left( -\frac{1}{2} (X_j - \mu_{ij})^\top (\Sigma_{ij} + P_i)^{-1} (X_j - \mu_{ij}) \right)}{(2\pi)^{D/2} |\Sigma_i + P_i|^{1/2}}
\]
\[
\Sigma_{ij} = \left( \Sigma_i + P_i \right)^{-1} + \left( \mu_{ij} - \mu_i \right) \left( \mu_{ij} - \mu_i \right)^\top
\]
\[
\mu_{ij} = \mu_i + \Sigma_i (\Sigma_i + P_i)^{-1} (X_j - \mu_i).
\]
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 [18] 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:
\[
P(X_t|Z_{1:t}) = \sum_{i=1}^{m_{t}} \pi_i \mathcal{N}(X_t; \Sigma_i, \mu_i),
\]
where \( m_{t} \) is the number of components [18].

\[4.\text{Experimental setup}\]

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 (Eigemike). 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. 15. 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 occluded 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. An appearance model based on RGB color histogram is initialized from the image region around the initial bounding box. Conversely, if not enough detections are found for the same target within \( T \) consecutive frames, the track is automatically terminated. Since addition
and termination of targets, respectively, increases and decreases the system state dimension, this procedure is performed at the start of each-time step, prior to density approximation. The Hungarian algorithm [19] is used to obtain the optimal face detection to existing target association. The association cost matrix is based on the pairwise similarity of the RGB histogram between existing targets and detections.

The common CLEAR MOT metrics [20] 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 [21] and the Optimal Sub-Pattern Assignment (OSPA) metric [22] are used 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-A V) is compared against the proposed tracking model without the auditory data (P-V) and the multi-person visual tracker (OV) in [23].

The results are summarized in Table 1 whilst the audio-visual tracking results for the proposed P-A V 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 [4].

The comparison between P-A V 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-A V 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-A V 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.

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 eigemike demonstrate significant improvements in performance for audio-visual tracking compared to using only visual data.

7. REFERENCES


