Target Tracking using a Joint Acoustic Video System

Volkan Cevher, *Member, IEEE,* Aswin C. Sankaranarayanan, *Student Member, IEEE,* James H. McClellan, *Fellow, IEEE,* and Rama Chellappa, *Fellow, IEEE.*

***Abstract*— In this paper, a multi-target tracking system for collocated video and acoustic sensors is presented. We formulate the tracking problem using a particle filter based on a state space approach. We first discuss the acoustic state space formulation whose observations use a sliding window of direction-of-arrival estimates. We then present the video state space that tracks a target's position on the image plane based on online adaptive appearance models. For the joint operation of the filter, we combine the state vectors of the individual modalities and also introduce a time delay variable to handle the acoustic-video data synchronization issue, caused by acoustic propagation delays. A novel particle filter proposal strategy for joint state space tracking is introduced, which places the random support of the joint filter where the final posterior is likely to lie. By using the Kullback-Leibler divergence measure, it is shown that the joint operation of the filter decreases the worst case divergence of the individual modalities. The resulting joint tracking filter is quite robust against video and acoustic occlusions due to our proposal strategy. Computer simulations are presented with synthetic and field data to demonstrate the filter's performance.**

I. Introduction

Recently, hybrid nodes that contain an acoustic array collocated with a camera were proposed for vehicle tracking problems [1]. To intelligently fuse information coming from both modalities, novel strategies for detection and data association have to be developed to exploit the multi modal information. Moreover, the fused tracking system should be able to sequentially update the joint state vector that consists of multiple target motion parameters and relevant features (e.g., shape, color and so on), which is usually only partially observable by each modality.

It is well known that acoustic and video measurements are complementary modalities for object tracking. Individually, the acoustic sensors can detect targets [2]–[4], regardless of the bearing with low power consumption, and the video sensors can provide reliable high-resolution localization estimates [5], regardless of the target range, with high power consumption. Hence, by fusing the acoustic and video modalities, we (i) achieve tracking robustness at low acoustic signal-to-noise ratios (*SNR*) or during video occlusion, (ii) improve target counting/confirmation, and (iii) design algorithms that permit a power vs. performance trade-off for hybrid node management.

In the literature, one finds that fusion of acoustic and video modalities has been applied to problems such as tracking of humans under surveillance and smart videoconferencing. Typically, the sensors are a video camera and an acoustic array (not necessarily collocated). In [6], the acoustic time delay-of-arrivals (TDOA's), derived from the peaks of the generalized cross-correlation function, are used along with active contours to achieve robust speaker tracking with fast lock recovery. In [7], jump Markov models are used for tracking humans using audio-visual cues, based on foreground detection, image-differencing, spatiospectral covariance matrices, and training data. The work by Gatica-Perez *et al.* [8] demonstrates that particle filters, whose proposal function uses audio cues, have better speaker tracking performance under visual occlusions.

The videoconferencing papers encourage the fusion of acoustics and video; however, the approaches in these papers do not extend to the outdoor vehicle tracking problem. They omit the audio-video synchronization issue that must be modeled to account for acoustic propagation delays. In vehicle tracking problems, average target ranges of 100-600m result in acoustic propagation delays in the range of 0.3-2s. Acoustics and video asynchronization causes biased localization estimates that can lead to filter divergence. This is because the bias in the fused cost function increases the video's susceptibility to drift in the background. In addition, motion models should adaptively account for any rapid target motion. Moreover, the visual appearance models should be calculated online as opposed to using trained models for tracking. Although fixed image templates (e.g., wire-frames in [6], [8], [9]) are very useful for face tracking, they are not effective for tracking vehicles in outdoor environments. Adaptive appearance models are necessary for achieving robustness [10]–[12].

To track vehicles using acoustic and video measurements, we propose a particle filtering solution that can handle multiple sensor modalities. We use a fully joint tracker, which combines the video particle filter tracker [11] and a modified implementation of the acoustic particle filter tracker [13] at the state space level. We emphasize that combining the output of two particle filters is different from formulating one fully joint filter [14] or one interacting filter [1] (e.g., one modality driving the other). The generic proposal strategy described in [15] is used to carefully combine the optimal proposal strategies for the individual acoustic and video state spaces such that the random support of the particle filter is concentrated where the final posterior of the joint state

V. Cevher, A. C. Sankaranarayanan, and R. Chellappa are with the Center for Automation Research, Department of ECE, University of Maryland, College Park, MD 20742

J. H. McClellan is with the Center for Signal and Image Processing, School of ECE, Georgia Institute of Technology, Atlanta, GA 30332-0250.

Prepared through collaborative participation in the Advanced Sensors Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-02-0008.

space lies. The resulting filter posterior has a lower Kullback-Leibler distance to the true target posterior than any output combination of the individual filters.

The joint filter state vector includes the target heading direction $\phi_k(t)$, the logarithm of velocity over range $Q_k(t) = \log(v_k/r_k(t))$, observable only by the acoustics; target shape deformation parameters $\{a_1, a_2, a_3, a_4\}_k$, the vertical 2D image plane translation parameter $\eta_k(t)$, observable only by the video; and the target DOA $\theta_k(t)$, observable by both modalities. The subscript k refers to the k^{th} target. We also incorporate a time delay variable $\tau_k(t)$ into the filter state vector to account for acoustic propagation delays needed to synchronize the acoustic and video measurements. This variable is necessary to robustly combine the high resolution video modality with the lower resolution acoustic modality and to prevent biases in the state vector estimates.

The filter is initialized using a matching pursuit strategy to generate the particle distribution for each new target, one at a time [13], [16]. A partitioning approach is used to create the multiple target state vector, where each partition is assumed to be independent. Moreover, the particle filter importance function independently proposes particles for each target partition to increase the efficiency of the algorithm at moderate increase in computational complexity.

The organization of the paper is as follows. Sections II and III present the state space formulation of the individual modalities. Section IV describes a Bayesian framework for the joint state space, and Sect. V introduces the proposal strategy for the fully joint particle filter tracker. Section VI discusses the audio-video synchronization issue and presents our solution. Section VII details the practical aspects of the proposed tracking approach. Finally, Sect. VIII gives experimental results using synthetic and field data.

II. Acoustic State Space

The acoustic state space, presented in this section, is a modified form of the one used in [17]. we choose this particular acoustic state space because of its flexible observation model that can handle (i) multiple target harmonics, (ii) acoustic propagation losses, and (iii) time-varying frequency characteristics of the observed target acoustic signals, without changing the filter equations. Figure 1 shows the behavior of the acoustic state variables for a two-target example using simulated data.

A. State Equation

The acoustic state vector for target k has three elements $x_k(t) \triangleq [\ \theta_k(t)\ ,\ Q_k(t)\ ,\ \phi_k(t)\ ]^T$, where $\theta_k(t)$ is the k^{th} target DOA, $\phi_k(t)$ is its heading direction, and $Q_k(t)$ is its logarithm of the velocity-range ratio. The angular parameters $\theta_k(t)$ and $\phi_k(t)$ are measured counterclockwise with respect to the x-axis.

The state update equation is derived from the geometry imposed by the locally constant velocity model. The resulting state update equation is nonlinear [18], [19]:

$$x_k(t+\tau) = h_\tau\left(x_k(t)\right) + u_k(t), \tag{1}$$

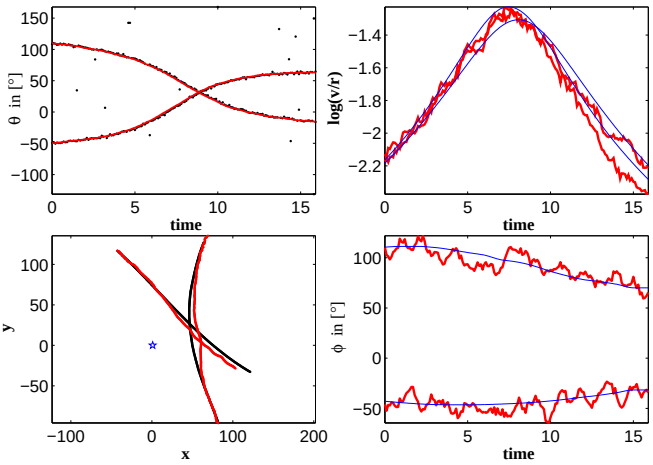


Fig. 1. (*Top Left*) Particle filter DOA tracking example with two targets. (*Bottom Left*) True track vs. calculated track. Note that the particle filter track is estimated using the filter outputs and the correct initial position. The particle filter jointly estimates the target heading (*Bottom Right*) and the target velocity over range ratio (*Top Right*), while estimating the target bearing. Note that the heading estimates typically tend to be much noisier than the DOA estimates.

where $u_k(t) \sim \mathcal{N}(0, \Sigma_u)$ with $\Sigma_u = \text{diag}\{\sigma^2_{\theta,k}, \sigma^2_{Q,k}, \sigma^2_{\phi,k}\}$ and $h_\tau(x_k(t)) =$

$$\begin{bmatrix} \tan^{-1}\left\{\frac{\sin\theta_k(t)+\tau\exp Q_k(t)\sin\phi_k(t)}{\cos\theta_k(t)+\tau\exp Q_k(t)\cos\phi_k(t)}\right\} \\ Q_k(t) - \frac{1}{2}\log\{1 + 2\tau\exp Q_k(t)\cos(\theta_k(t)-\phi_k(t)) + \\ \tau^2\exp(2Q_k(t))\} \\ \phi_k(t) \end{bmatrix}. \tag{2}$$

Reference [19] also discusses state update equations based on a constant acceleration assumption.

B. Observation Equation

The observations $\mathbf{y}_{t,f} = \{y_{t-m\tau,f}(p)\}_{m=0}^{M-1}$ consist of a batch of DOA estimates from a beamformer, indexed by m. Hence, the acoustic data of window-length T is segmented into M segments of length τ, equal to a single video frame duration (typically $\tau = 1/30$s). The target motion should satisfy the constant velocity assumption during a window-length T. For ground targets, $T = 1$s is a reasonable choice. Each of these segments is processed by a beamformer, based on the temporal frequency structure of the observed target signals, to calculate possible DOA estimates. This procedure can be repeated F times for each narrow-band frequency indexed by f (Fig. 2). Note that only the peak locations are kept in the beamformer power pattern. Moreover, the peak values, indexed by p, need not be ordered or associated with peaks from the previous time in the batch and the number of peaks retained can be time-dependent.

The sliding batch of DOA's, $\mathbf{y}_{t,f}$, is assumed to form a normally distributed cloud around the true target DOA tracks. In addition, only one DOA is present for each target at each frequency f or the target is missed: multiple DOA measurements imply the presence of clutter or other targets. We also assume that there is a constant detection probability for each target denoted by κ^f, which might depend on the particular

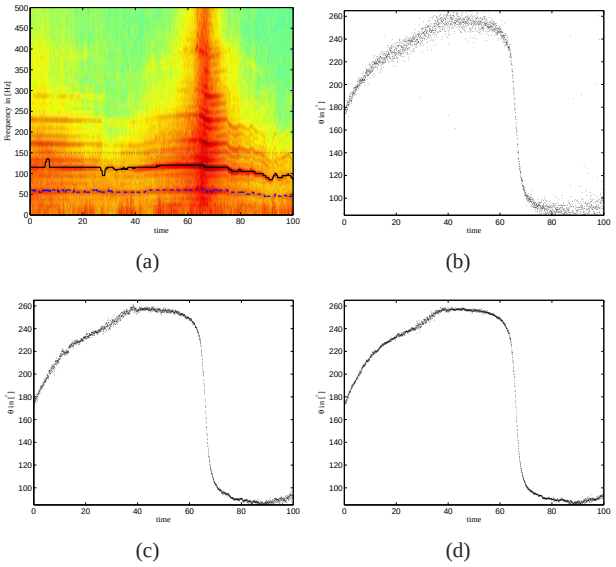


Fig. 2. A 10-element uniform circular microphone array is used to record a target's acoustic signal, while it is moving on an oval track (refer to Fig. 11). The acoustic array's inter-microphone distance is 1.1m. Hence, the maximum beamforming frequency without aliasing is approximately 150Hz. The acoustic sampling frequency is 44100Hz. (a) The time-frequency plot of the received signal. We estimated the bearing track of the vehicle using the MVDR beamformer [2], where the beamforming frequencies are chosen to be the dashed line for (b), the solid line for (c), and the dotted line for (d). For each acoustic bearing estimate, 1470 acoustic data samples are used, corresponding to 30 bearing estimates per second. The bearing tracks in (b-d) are indexed by $f = 1, 2, 3$ in the acoustic state space derivation and $F = 3$.

frequency f. If the targets are also simultaneously identified, an additional partition dependency, i.e., κ_k^f, is added.

For a given target, if we assume that the data is only due to its partition and clutter (hence, the DOA data corresponding to other targets are treated as clutter), we can derive the observation likelihood for the state $\mathbf{x}_t = \left[\ x_1^T(t),\ x_2^T(t),\ \ldots\ , x_K^T(t)\ \right]^T$ [17] as:

$$p(\mathbf{y}_t|\mathbf{x}_t) = \prod_{k=1}^{K} p(\mathbf{y}_t|x_k(t)) = \prod_{k=1}^{K}\prod_{f=1}^{F}\prod_{m=0}^{M-1} \left\{ \kappa_{0,1}^f \left(\frac{\gamma}{2\pi}\right)^{P_{m,f}} + \kappa_{1,1}^f \left(\frac{\gamma}{2\pi}\right)^{P_{m,f}-1} \sum_{p=1}^{P_{m,f}} \frac{\psi_{t,m,f}\left(p\middle|x_k\right)}{P_{m,f}} \right\}, \tag{3}$$

where the parameters $\kappa_{n,K}^f$ ($\sum_n \kappa_{n,K}^f = 1$) are the elements of a detection (or confusion) matrix, $p = 0, 1, \ldots, P_{m,f}$ for each f and m, and $\gamma \gg 1$ is a constant that depends on the maximum number of beamformer peaks P, the smoothness of the beamformer's steered response, and the number of targets K. The function ψ in (3) is derived from the assumption that the associated target DOA's form a Gaussian distribution around the true target DOA tracks:

$$\psi_{t,m,f}\left(p_i\middle|x_i\right) = \frac{1}{\sqrt{2\pi\sigma_\theta^2(m,f)}} \exp\left\{ -\frac{\left(h_{m\tau}^\theta(x_i(t)) - y_{t+m\tau,f}(p_i)\right)^2}{2\sigma_\theta^2(m,f)} \right\}, \tag{4}$$

where the superscript θ on the state update function h refers only to the DOA component of the state update and $\sigma_\theta^2(m,f)$ is supplied by the beamformer, using the curvature of the DOA power pattern at the peak location.

III. Video State Space

In this section, we give the details of the video state space. This video state space is also described in greater detail in [11]. We assume that the camera is stationary and is mounted at the center of the acoustic microphone array, at a known height above the ground. We also assume that the camera calibration parameters are known, which allows us to convert a location on the image plane to a DOA estimate while having the same reference axis as the acoustic state space. Figure 3 demonstrates a video tracker based on state space described in this section.

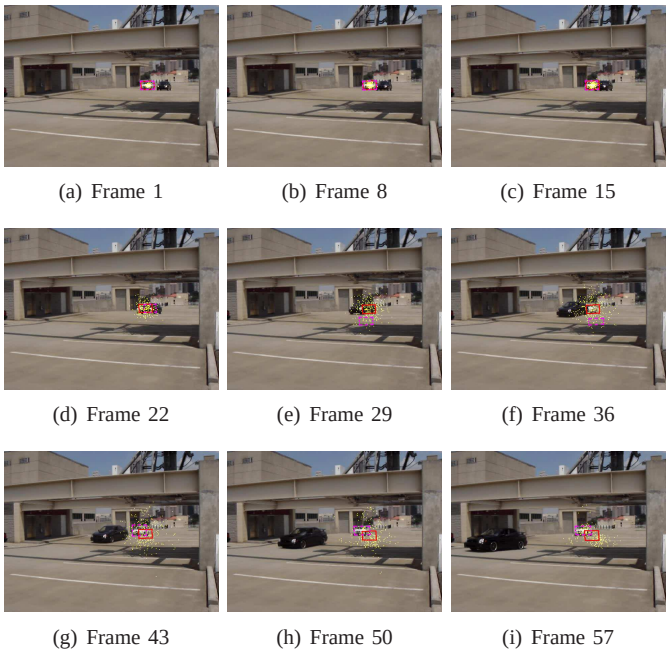


Fig. 3. Intensity based visual tracking of the white car using the particle filter based on the video state space described in this section. The solid box shows the mean of the posterior, whereas the dashed box shows the location of the mode of the posterior. The dot cloud depicts spatial particle distribution. In this scenario, the white car is occluded for 1 second corresponding to 30 video frames. The particle spread during occlusion increases because the robust statistics measure [11] renders the likelihood function non-informative. The filter quickly locks back to the target after occlusion.

A. State Equation

The video state vector for target k has six elements: four affine deformation parameters $\mathbf{a}_k(t) = \left[\ a_{k,1,t}\ ,\ \ldots\ ,\ a_{k,4,t}\ \right]^T$, a vertical 2-D translation parameter $\eta_k(t)$, and the target DOA $\theta_k(t)$: $x_k(t) \triangleq \left[\ \mathbf{a}_k^T(t)\ ,\ \eta_k(t)\ ,\ \theta_k(t)\ \right]^T$. The affine deformation parameters linearly model the object rotation, shear and scaling (affine minus translation), whereas the translation parameter and the DOA account for the object translation, all on the image plane. The state update equation consists of a predictive shift and a diffusion component:

$$x_k(t) = h_\tau\left(x_k(t-\tau)\right) + u_k(t) = \hat{x}_k(t-\tau) + \nu_k(t) + u_k(t), \tag{5}$$

where $\nu_k(t)$ is an adaptive velocity component, affecting only $\eta_k(t)$ and $\theta_k(t)$ in the state vector. It is calculated

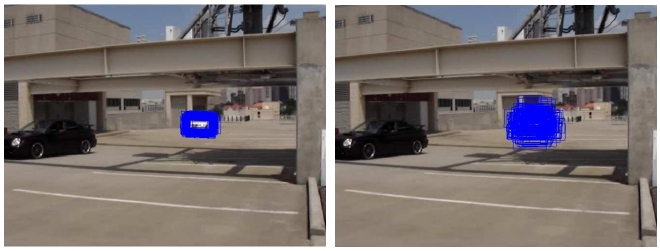
using a first-order linear prediction method on two successive frames; $\hat{x}_k(t-\tau)$ is the maximum *a posteriori* estimate of the state at time $t-\tau$; and $u_k(t)$ is an adaptive noise component, calculated by measuring the difference between the updated appearance and the calculated appearance at time t, as described in [11]. Note that the video state mode estimates $\hat{x}_k(t-\tau)$ are stored in the memory, because they are later used for adaptively determining a time delay variable for acoustic-video synchronization.

The state equation is constructed so that it can effectively capture rapid target motions. The adaptive velocity component accounts for the object's shift within the image frame, whereas the adaptive noise term captures its drift around its motion. Hence, the adaptive velocity model simply encodes the object's inertia into the tracker and generates particles that are tightly centered around the object of interest for improved efficiency (Fig. 4). If we do not account for the object's shift using the adaptive noise component, we need to increase the variance of the drift component to capture the actual movement of the object. Hence, we may start to lose our focus on the target as shown in Fig. 4(b) without the adaptive velocity component. In this case, if the background is somewhat similar to the target, it is automatically injected into the appearance models through the EM algorithm. Hence, the background also becomes part of the tracked object, thereby creating local minima to confuse the tracker in its later iterations.

The adaptive noise variance is based on residual motion errors generated by the adaptive velocity component. It decreases when the quality of the prediction from the adaptive velocity component is high, and increases when the prediction is poor. Finally, when the tracker is visually occluded (occlusion is defined in the next subsection), the target motion is characterized using a Brownian motion and $\nu_k(t) = 0$ is enforced. Hence, during an occlusion, the state dynamics changes to the following form:

$$x_k(t) = x_k(t-\tau) + u_k(t). \tag{6}$$

We avoid the use of the adaptive velocity model during occlusion because the object motion may change significantly during an occlusion.



(a) with the adaptive velocity model (b) without the adaptive velocity model

Fig. 4. Comparison of the proposed particles when the adaptive velocity model is used. Note that the particles are tightly clustered around the target when we use the adaptive velocity model. In contrast, without velocity prediction, we need to use more particles to represent the same posterior, because most particles have very low weights.

B. Observation Equation

The observation model is a mixture of following adaptive appearance models: a wandering $\mathcal{W}_t$, a stable $\mathcal{S}_t$, and an optional fixed template model $\mathcal{F}_t$. The wandering model $\mathcal{W}_t$ captures transient appearance changes based on two successive frames, whereas the stable model $\mathcal{S}_t$ encodes appearance properties that remain relatively constant over a large number of frames (Fig. 5). The fixed template $\mathcal{F}_t$ is useful for tracking recognized targets, however it is not considered any further in this paper. The adaptive observation model in this paper uses the pixel intensity values for these appearance models for computational efficiency as suggested in [11]. Although the image intensity values are typically not robust to changes in illumination, the appearance model described here can adapt to changes in illumination. However, it is still possible to lose track if there are sudden changes in illumination. We use a very simple model to circumvent this problem. We normalize the mean and the variance of the appearance as seen by each particle. This makes our tracker immune to uniform scaling of the intensities. If we know that the illumination changes are severe, we can adopt an alternative feature at the expense of computation without chancing our filter mechanics, such as the spatial phase data of the object [12] that is more robust to illumination changes.

The observation model is dynamically updated by an online expectation maximization (EM) algorithm that adaptively calculates the appearance parameters $\{\mu_{i,t}, \sigma^2_{i,t}\}$, $(i = w, s)$ of the appearance models $A_t = \{\mathcal{W}_t, \mathcal{S}_t\}$, and the model mixture probabilities $m_{i,t}$, $(i = w, s)$ for each pixel [20], [21]. The details of the EM algorithm for calculating the mixture probabilities and model parameters can be found in [11], [12]. Omitting the details of the derivations, the observation likelihood is given by the following expression:

$$p(\mathbf{y}_t|\mathbf{x}_t) = \prod_{k=1}^{K}\prod_{j=1}^{d}\left\{\sum_{i=w,s} m_{i,t}\mathrm{N}(\mathcal{T}_k(y_t(j)); \mu_{i,t}(j), \sigma^2_{i,t}(j))\right\}, \tag{7}$$

where $\mathcal{T}_k$ is the affine transformation that extracts the image patch of interest by using the state vector $x_k(t)$; d is the number of pixels in the image patch; and $\mathrm{N}(x; \mu, \sigma^2)$ is the density

$$\mathrm{N}(u; \mu, \sigma^2) \propto \exp\left\{-\rho\left(\frac{u-\mu}{\sigma}\right)\right\}, \tag{8}$$

where u is normalized to have unit variance, and

$$\rho(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \le c; \\ c|u| - \frac{1}{2}c^2, & \text{o/w.} \end{cases} \tag{9}$$

The function $\rho(\cdot)$ is Huber's criterion function, which is commonly used for outlier rejection [22]. It provides a compromise between mean estimators that are susceptible to outliers and median estimators that are usually robust to outliers. The constant c is used to determine the outlier pixels that cannot be explained by the underlying models. Furthermore, methods from robust statistics allow us to formally decide when the tracker is *visually occluded*, which implies that the particle with the highest likelihood has more than 50% of its pixels, which are classified as outliers by the appearance model. This

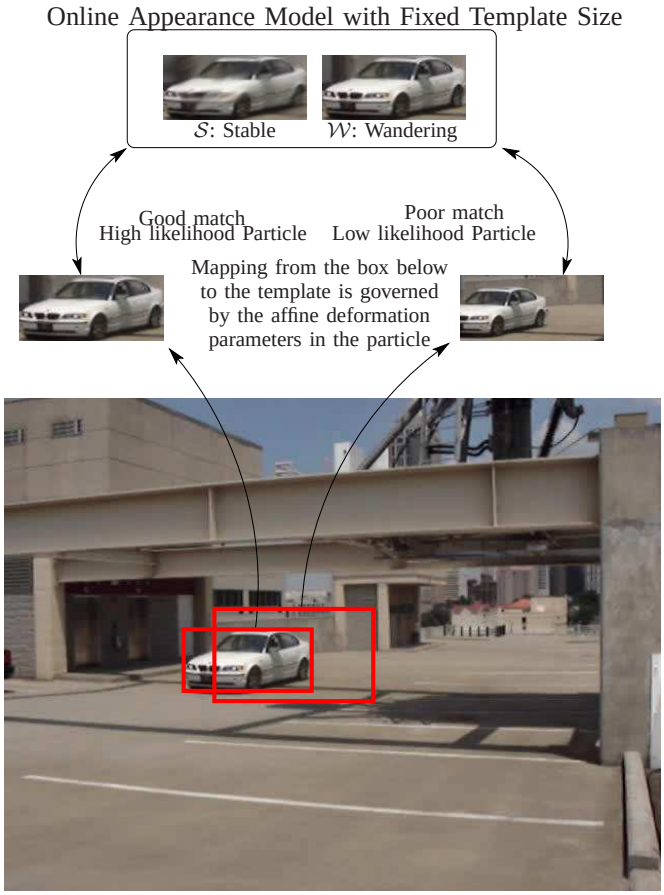


Fig. 5. The online appearance model is illustrated. The model has two components: $\mathcal{S}$ (stable) and $\mathcal{W}$ (wandering). The stable model temporally integrates the target image in its bounding box using a forgetting factor. On the other hand, the wandering model uses two-frame averages. Note that each model uses a fixed size image template that is updated by an online EM algorithm [11]. To determine a particle's likelihood, an image patch is first determined using the particle elements. Then, the patch is mapped back to the template domain using the affine transformation parameters, where it is compared with the updated appearance model. This operation requires interpolation and contributes to most of the filter's computational complexity.

criterion is discussed in greater detail in [11].

Deciding on whether or not an object is occluded is an arduous task. However, this task is alleviated when we also track the appearance. Our decision is based on the outlier statistics and is reliable. We provide a Monte Carlo run of the occlusion decision in the simulations section to show the reliability of our occlusion strategy. We show that the variability of the occlusion detection is rather small once a threshold is chosen. Further examples of this occlusion strategy can be found in [11]. The influence of an error on this decision is discussed in our observation model. If we are late in declaring an occlusion, the appearance of the occluding object injects itself into the target appearance, thereby causing local minima in the tracking algorithm. However, given the complexity of the problem, one should not expect superlative performance for all the possible cases.

Another issue in handling occlusion is the change in the appearance of the target during occlusion. This could happen due to changes in global illumination, changes in the pose of the target, or dramatic changes in the projected target size on the image plane. Recovery of visual tracking cannot be guaranteed, except when these changes are not severe. In cases, where the track is recovered, we update the appearance model using the appearance associated with the particle with maximum likelihood. We say that track has been regained after occlusion, when the tracker is not visually occluded (as defined before) for a fixed set of frames (ten frames for the experiments in the paper).

IV. Bayesian Framework for Tracking the Joint State Space

In this section, a Bayesian framework is described for combining the acoustic ($\mathcal{S}_1$) and video ($\mathcal{S}_2$) state spaces that share a common state parameter. The results below can be generalized to time-varying systems including nuisance parameters. It is assumed that the state dimensions are constant even if the system is time-varying. Define

$$\begin{aligned} \mathcal{S}_{\mathrm{i}}: \qquad x_{i,t} &= \begin{bmatrix} \chi_t \\ \psi_{i,t} \end{bmatrix} \sim q_i(x_{i,t}|x_{i,t-1}) \\ y_{i,t} &\sim f_i(y_{i,t}|x_{i,t}), \end{aligned} \tag{10}$$

where the observed data in each space is represented by $\{y_{i,t}, i = 1,2\}$, $\chi_t = \theta_t$ (overlapping state parameter), $\psi_{1,t} = [\ Q(t)\ ,\ \phi(t)\ ]^T$, and $\psi_{2,t} = \left[\ \mathbf{a}^T(t)\ ,\ \eta(t)\ \right]^T$. The state transition density functions $q_i(\cdot|-)$ are given by (1) and (5). The observations are explained through the density functions $f_i(\cdot|-)$, given by (3) and (7). The observation sets y_i are modeled as statistically independent given the state through conditionally independent observation densities. This assumption is justified in our problem: for example, a vehicle's time-frequency signature is independent of its colors or textures. In most cases, it may be necessary to verify this assumption mathematically for the problem at hand [14], [23] by using the specific observation models.

To track the joint state vector $x_t = [\chi_t, \psi_{1,t}, \psi_{2,t}]$ with a particle filter, the following target posterior should be determined:

$$\begin{aligned} p(x_t|x_{t-1}, y_{1,t}, y_{2,t}) &\propto p(y_{1,t}, y_{2,t}|x_t)p(x_t|x_{t-1}) \\ &= \pi_t(y_{1,t}, y_{2,t})\pi_{t-1}(x_t), \end{aligned} \tag{11}$$

where $\pi_s(\cdot) = p(\cdot|x_s)$. Note that the Markovian property is enforced in (11). That is, given the previous state and the current data observations, the current state distribution does not depend on the previous state track and the previous observations.

Equation (11) allows the target posterior to be calculated up to a proportionality constant, where the proportionality is independent of the current state x_t. The first pdf on the right hand side of (11) is called the joint-data likelihood and can be simplified, using the conditional independence assumption on the observations:

$$\pi_t(y_{1,t}, y_{2,t}) = f_1(y_{1,t}|x_{1,t})f_2(y_{2,t}|x_{2,t}). \tag{12}$$

The second pdf in (11), corresponding to a joint state update, requires more attention. State spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ may have different updates for the common parameter set since

they had different models.[1] This poses a challenge in terms of formulating the common state update for x_t. Instead of assuming a given analytical form for the joint state update as in [14], we combine the individual state update marginal pdfs for the common state parameter as follows:

$$\pi_{t-1}(\chi_t) = c p_1(\chi_t)^{o_1} p_2(\chi_t)^{o_2} r(\chi_t)^{o_3}, \tag{13}$$

where $c \geq 1$ is a constant, $p_i(\chi_t) \triangleq p(\chi_t|x_{i,t-1})$ is the marginal density, the probabilities o_i for $i = 1, 2$ ($\sum_i o_i = 1$) define an ownership of the underlying phenomenon by the state models, and $r(\chi_t)$ is a (uniform/reference) prior in the natural space of the parameter χ_t [24] to account for unexplained observations by the state models.

If we denote the Kullback-Leibler distance as D, then

$$D(\alpha(\chi_t)||\pi_{t-1}(\chi_t)) = -\log c + \sum_i o_i D(\alpha(\chi_t)||p_i(\chi_t)) \tag{14}$$

where α is the unknown true χ_t distribution. Hence, $D(\alpha||\pi_{t-1}) \leq \max_i\{D(\alpha||p_i)\}$. $\pi_{t-1}(\chi_t)$ always has a smaller KL distance to the true distribution than the maximum KL distance of $p_i(\chi_t)$. This implies that (13) alleviates the worst case divergence from the true distribution [25]. Hence, this proves that one of the trackers does assist the other in this framework.

The ownership probabilities, o_i, can be determined using an error criteria. For example, one way is to monitor how well each partition $x_{i,t}$ in x_t explains the information streams $y_{i,t}$ through their state-observation equation pair defined by $\mathcal{S}_i$, (10). Then, the respective likelihood functions can be aggregated with an exponential envelope to recursively solve for the o_i's (e.g., using an EM algorithm). In this case, the target posterior will be dynamically shifting towards the better self-consistent model while still taking into account the information coming from the other, possibly incomplete, model, which might be temporarily unable to explain the data stream.

If one believes that both models explain the underlying process equally well regardless of their self-consistency, one can set $o_1 = o_2 = 1/2$ to have the marginal distribution of χ_t resemble the product of the marginal distributions imposed by both state spaces. The proposal strategy in the next section is derived with this assumption on the ownership probabilities, because, interestingly, it is possible to show that assuming equal ownership probabilities along with (13) leads to the following conditional independence relation on the state spaces:

$$\pi_{t-1}(x_{1,t})\pi_{t-1}(x_{2,t}) = q_1(x_{1,t}|x_{1,t-1})q_2(x_{2,t}|x_{2,t-1}). \tag{15}$$

Equation (15) finally results in the following update equation:

$$\begin{aligned}
\pi_{t-1}(x_t) &= \pi_{t-1}(\psi_{1,t}, \psi_{2,t}|\chi_t)\pi_{t-1}(\chi_t) \\
&= \pi_{t-1}(\psi_{1,t}|\chi_t)\pi_{t-1}(\psi_{2,t}|\chi_t)\pi_{t-1}(\chi_t) \\
&= \frac{\pi_{t-1}(x_{1,t})\pi_{t-1}(x_{2,t})}{\pi_{t-1}(\chi_t)} \\
\Rightarrow \pi_{t-1}(x_t) &= \frac{q_1(x_{1,t}|x_{1,t-1})q_2(x_{2,t}|x_{2,t})}{\pi_{t-1}(\chi_t)},
\end{aligned} \tag{16}$$

where

$$\pi_{t-1}(\chi_t) \propto \left[\iint q_1(x_{1,t}|x_{1,t-1})d\psi_{1,t} q_2(x_{2,t}|x_{2,t})d\psi_{2,t}\right]^{1/2}. \tag{17}$$

V. Proposal Strategy

A proposal function, denoted as $g(x_t|x_{t-1}, y_t)$, determines the random support for the particle candidates to be weighted by the particle filter. Two very popular choices are (i) the state update $g \propto q_i(x_t|x_{t-1})$ and (ii) the full posterior $g \propto f_i(y_t|x_t)q_i(x_t|x_{t-1})$. The first one is attractive because it is analytically tractable. The second one is better because it incorporates the latest data while proposing particles, and it results in less variance in the importance weights of the particle filter since, in effect, it directly samples the posterior [26], [27]. Moreover, it can be analytically approximated for faster particle generation by using local linearization techniques (see [27]), where the full posterior is approximated by a Gaussian. The analytical form of the proposal functions for acoustic and video state spaces, obtained by local linearization of the posterior, is given by

$$g(x_t|x_{t-1}, y_t) \sim \mathcal{N}(\mu_g, \Sigma_g), \tag{18}$$

where the Gaussian density parameters are

$$\begin{aligned}
\Sigma_g &= \left(\Sigma_y^{-1} + \Sigma_u^{-1}\right)^{-1}, \\
\mu_g &= \Sigma_g\left(\Sigma_y^{-1}x_{\text{mode}} + \Sigma_u^{-1}h_\tau(x(t-\tau))\right),
\end{aligned} \tag{19}$$

and where x_{mode} is the mode of the data likelihood, and $\Sigma_y^{-1}(k)$ is the Hessian of data likelihood at x_{mode}. The details of these proposal functions can be found in [11], [13]. Hence, in either way of proposing particles, one can assume that an analytical relation for g_i, defining the support of the actual posterior for each state space, can be obtained.

Figure 6 describes the proposal strategy used for the joint state space. Each state space has a proposal strategy described by the analytical functions $\{g_i, i = 1, 2\}$ defined over the whole state spaces. Then, the proposal functions of each state g_i are used to propose particles for the joint space by carefully combining the supports of the individual posteriors. First, marginalize out the parameters $\psi_{i,t}$:

$$\hat{g}_i(\chi_t|x_{i,t-1}, y_{i,t}) = \int g_i(x_{i,t}|x_{i,t-1}, y_{i,t})d\psi_{i,t}. \tag{20}$$

The functions, $\hat{g}_i$, describe the random support for the common state parameter χ_t and can be combined in the same way as the joint state update (13). Hence, the following function

$$\hat{g}(\chi_t|x_{t-1}, y_{1,t}, y_{2,t}) \propto [\hat{g}_1(\chi_t|x_{1,t-1}, y_{1,t})\hat{g}_2(\chi_t|x_{2,t-1}, y_{2,t})]^{1/2} \tag{21}$$

[1]There is no exact state update function for all targets. Individual state spaces may employ different functions for robustness, which is the case in our problem.

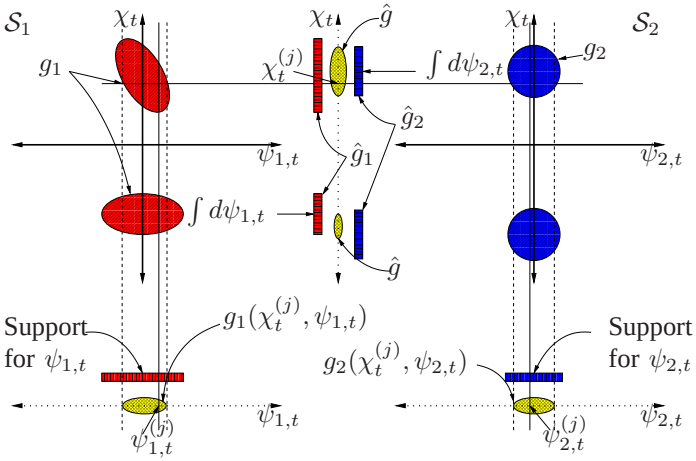


Fig. 6. The supports, g_i's, for the posterior distribution in each state space, $\mathcal{S}_i$, are shown on the axes χ_t vs. $\psi_{i,t}$. Particles for the joint state are generated by first generating χ_t's from the combined supports of the marginal distributions of χ_t. Then, the $\psi_{i,t}$'s are sampled from the g_i's as constrained by the given χ_t realization.

can be used to generate the candidates $\chi_t^{(j)}$ for the overlapping state parameters. Then using $\chi_t^{(j)}$, one can generate $\psi_{i,t}^{(j)}$ from $g_i(\chi_t^{(j)}, \psi_{i,t}|x_{i,t-1}, y_{i,t})$ and form $x_t^{(j)} = [\chi_t^{(j)}, \psi_t^{(j)}, \varphi_t^{(j)}]$.

In general, Monté-Carlo simulation methods can be used to simulate the marginal integrals in this section [28]. Here, we show how to calculate the marginal integrals of the state models. Simulation of the other integrals are quite similar. Given $\chi_t^{(j)}$, draw M samples using $\psi_{i,t}^{(m)} \sim g_i(\chi_t^{(j)}, \psi_{i,t}|x_{i,t-1}, y_{i,t})$.[2] Then,

$$\int q_1(\chi_t^{(j)}, \psi_{i,t}|x_{1,t-1})d\psi_{i,t} \approx \frac{1}{M}\sum_{m=1}^{M} \frac{q_1(\chi_t^{(j)}, \psi_{i,t}^{(m)}|x_{1,t-1})}{g_1(\chi_t^{(j)}, \psi_{i,t}^{(m)}|x_{1,t-1}, y_{1,t})}. \quad (22)$$

The pseudo-code for the joint strategy is given in Table I. Finally, the importance weights for the particles generated by the joint strategy described in this section can be calculated as follows:

$$w^{(j)} \propto \frac{p(x_t^{(j)}|x_{t-1}, y_{1,t}, y_{2,t})\hat{g}(\chi_t^{(j)}|x_{t-1}, y_{1,t}, y_{2,t})}{g_1(\chi_t^{(j)}, \psi_{1,t}^{(j)}|x_{1,t-1}, y_{1,t})g_2(\chi_t^{(j)}, \psi_{2,t}^{(j)}|x_{2,t-1}, y_{2,t})}. \quad (23)$$

VI. Time Delay Parameter

The joint acoustic video particle filter sequentially estimates its state vector at video frame rate, as the acoustic data arrives. Hence, the joint filter state estimates are delayed with respect to the actual event that produces the state, because the acoustic information propagates much slower than the video information. Although it is possible to formulate a filter so that estimates are computed as the video data arrives, the resulting filter cannot use the delayed acoustic data. Hence, it is not considered here. The adaptive time delay estimation also allows position tracking on the ground plane. However, small errors in the time delay estimates translate into rather large errors in target range estimates, resulting in large errors in target position estimates. Hence, the main reason for estimating time delay is to ensure the stability of the joint filter.

[2]It is actually not necessary to draw the samples directly from $g_i(\chi_t^{(j)}, \psi_{i,t}|-)$. An easier distribution function approximating only q_i can be used for simulating the marginalization integral (22).

TABLE I
Pseudo Code for Joint Proposal Strategy

i. Given the state update q_i and observation relations f_i for the individual state spaces $\{\mathcal{S}_i, i = 1, 2\}$, determine analytical relations for the proposal functions g_i's. For the individual proposal functions g_i, it is important to approximate the true posterior as close as possible because these approximations are used to define the random support for the final joint posterior. For this purpose, Gaussian approximation of the posterior (18) or linearization of the state equations can be used [27].
ii. Determine the support for the common state parameter χ_t using (21). The expression for $\hat{g}$ may have to be approximated or simulated to generate candidates $\chi_t^{(j)}$, $j = 1, 2, \ldots, N$ where N is the number of particles.
iii. Given $\chi_t^{(j)}$,
- calculate the marginal integrals by using (22) to determine g_i,
- generate $\psi_{i,t}^{(j)} \sim g_i(\chi_t^{(j)}, \psi_{i,t}|x_{i,t-1}, y_{i,t})$,
- form $x_t^{(j)} = [\chi_t^{(j)}, \psi_{1,t}^{(j)}, \psi_{2,t}^{(j)}]$, and
- calculate the importance weights, $w^{(j)}$'s, using (23).

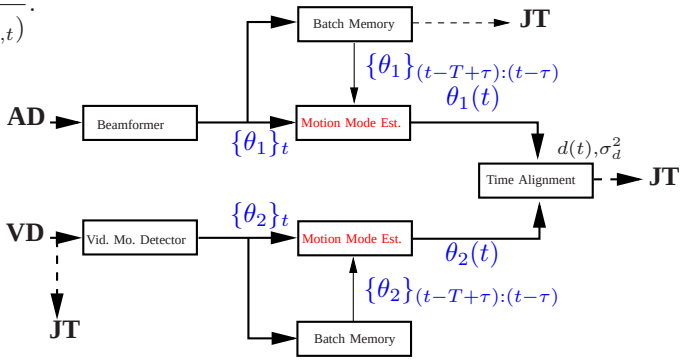


Fig. 7. At time t, τ seconds of acoustic data (**AD**) and a frame of video data (**VD**) are processed to obtain possible target DOA's $\{\theta_i\}_t$. This preprocessing is done by a beamformer block and a video motion detector block, respectively. With the guidance of the joint tracker (**JT**), these DOA's are used to determine the DOA mode tracks, $\theta_i(t)$ (Fig. 8), to estimate the time delay $d(t)$. The estimated time delay parameters are then used in the proposal function of the joint tracker.

To synchronize the audio-video information, we add an additional time delay variable $d_k(t)$ for each target k to form an augmented joint filter state:

$$x_k(t) \triangleq \left[\ \mathbf{a}_k^T(t)\ ,\ \eta_k(t)\ ,\ \theta_k(t)\ ,\ Q_k(t)\ ,\ \phi_k(t)\ ,\ d_k(t)\ \right]^T. \quad (24)$$

The time delay $d_k(t)$ is defined geometrically as:

$$d_k(t) = ||\boldsymbol{\xi} - \boldsymbol{\chi}_k\left(t - d_k(t)\right)||/c, \quad (25)$$

where $\boldsymbol{\xi} = [s_x, s_y]^T$ is the hybrid node position in $2D$, $\boldsymbol{\chi}_t = [x_{k,target}(t), y_{k,target}(t)]^T$ is the k^{th} target position, and c is the speed of sound. Using the geometry of the problem, it is possible to derive an update equation for $d_k(t)$:

$$d_k(t+\tau) = d_k(t)\exp\{u_{d,k}(t)\} \sqrt{1 + 2\tau\exp\{Q_k(t)\}\cos\left(\theta_k(t) - \phi_k(t)\right) + \tau^2\exp\{2Q_k(t)\}}, \tag{26}$$

where the Gaussian state noise $u_{d,k}(t)$ is injected as multiplicative.

We suppress the partition dependence on the variables from now on for brevity. Figure 7 illustrates the mechanics of time delay estimation. To determine $d(t)$, we first determine the mode of the acoustic state vector within a batch period of T seconds. Given the calculated acoustic data mode, which is also used in the proposal stage of the particle filter, $x_{1,\text{mode}}(t)$, an analytical relation for acoustic DOA track $\theta_1(t)$ (Fig. 8) is determined, using the state update function (2). This functional estimate $\theta_1(t)$ of the acoustic DOA's and acoustic data is used to determine an average variance of the DOA's $\widetilde{\sigma}^2_{1,\theta}$ around the functional, between times t and $t-T$. Note that $\widetilde{\sigma}^2_{\theta}$ is estimated using the missing and spurious data assumptions similar to the ones presented in Sect. II.

Next, we search the stored mode estimates of the video state, which is used in the video state update function (5), to determine $M = T/\tau$ (i.e., the number of video frames per second) closest video DOA estimates. These DOA's are used, along with the constant velocity motion assumption, to determine a functional estimate $\theta_2(t)$ of the DOA track and an average DOA variance $\widetilde{\sigma}^2_{2,\theta}$, based on the video observations, as shown in Fig. 8. The observation likelihood for the time delay variable $d(t)$ is approximated by the following Gaussian:

$$p(d(t)|\mathbf{y}_{1,t}, \mathbf{y}_{2,t}) \approx \mathcal{N}\left(\mu_d \left(1 + Te^{Q_{\text{mode}}} \cos\left[(\theta_1(t-T) + \theta_1(t))/2 - \phi_{\text{mode}}\right] + T^2 e^{2Q_{\text{mode}}}/4\right)^{\frac{1}{2}}, \sigma_d^2\right), \tag{27}$$

where the mean is the average distance between the functional inverses of $\theta_1(t)$ and $\theta_2(t)$:

$$\mu_d = \left| \frac{\int_{\theta_1(t)}^{\theta_1(t-T)} \left[\theta_1^{-1}(\theta') - \theta_2^{-1}(\theta')\right] d\theta'}{\theta_1(t) - \theta_1(t-T)} \right|. \tag{28}$$

The variance σ_d^2 is determined by dividing the average DOA variances by the functional slope average:

$$\sigma_d^2 = \left| \frac{\theta_1(t) - \theta_1(t-T)}{\int_{t-T}^{t} \frac{\partial\theta_1(t')}{\partial t'} dt'} \right| \widetilde{\sigma}^2_{1,\theta} + \left| \frac{\theta_1(t) - \theta_1(t-T)}{\int_{t-T}^{t} \frac{\partial\theta_2(t')}{\partial t'} dt'} \right| \widetilde{\sigma}^2_{2,\theta}. \tag{29}$$

In the joint filter, the particles for the time delay parameter are independently proposed with a Gaussian approximation to the full time delay posterior, using (26) and (27) [27].

VII. Algorithm Details

The joint acoustic-video particle filter tracker code is given in Table II. In the following subsections, we discuss other practical aspects of the filter.

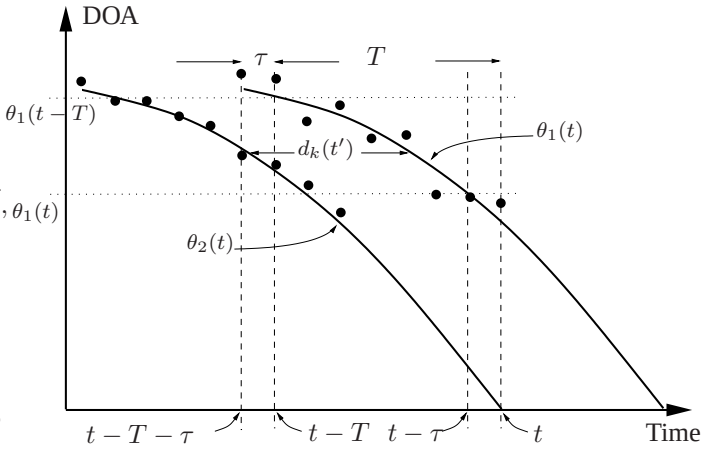


Fig. 8. The time delay $d_k(t)$ between the acoustic and video DOA tracks, $\theta_1(t)$ and $\theta_2(t)$, respectively.

A. Initialization

The organic initialization algorithms for the video and acoustic trackers are employed to initialize the joint filter. The joint filter initialization requires an interplay between the modalities, because the state vector is only partially observable by either modality. In most cases, the video initializer is cued by the acoustics, because the video modality consumes significantly more power. Below, we describe the general case where each modality is turned on.

Briefly, the organic initialization algorithms work as follows. In video, motion cues and background modeling are used to initialize target appearance models, $\mathbf{a}_k^T(t)$, $\eta_k(t)$, and $\theta_k(t)$ by placing a bounding box on targets and by coherent temporal processing of the video frames [11]. In acoustics, the temporal consistency of the observed DOA's is used to initialize target partitions by using a modified Metropolis-Hastings algorithm [13], [29].

To initialize targets, a matching-pursuit idea is used [13], [16]. The most likely target is initialized first and then its corresponding data is gated out [30]. Note that the target motion parameters alleviate the data association issues between the video and acoustic sensors, because both modalities are collocated. Hence, the overlapping state parameter θ is used to fuse the video shape parameters and acoustic motion parameters.

When a target is detected by the organic initialization algorithms, the time delay variable is estimated using the scheme described in Sect. VI. The initialization scheme in [13] is used to determine the target motion parameters, where the video DOA mode estimates are used as an independent observation dimension to improve the accuracy. Finally, a target partition is deleted by the tracking algorithm at the proposal stage if both acoustic and video modalities do not see any data in the vicinity of the proposed target state.

B. Multi Target Posterior

The joint filter treats the multiple targets independently, using a partition approach. The proposal and particle weighting of each target partitions are independent. This allows

a parallel implementation of the filter where a new single target tracking joint filter is employed for each new target. Hence, the complexity of the filter increases linearly with the number of targets. Note that for each target partition, it is crucial that data corresponding to the other target partitions are treated as clutter. This approach is different from the joint probability density association (JPDA) approach that would be optimal for assigning probabilities to each partition by adding mixtures that consist of data permutations and partition combinations [30]. In JPDA, no data would be assigned to more than one target. However, in our approach, the same DOA might be assigned to multiple targets.

Notably, it is shown in [13] that the independence assumption in this paper for the joint state space is reasonable for the acoustic tracker. There is a slight performance degradation in bearing estimation, when the targets cross; however, it is not noticeable in most cases. Moreover, the JPDA approach is not required by the video tracker. When the targets cross, if the targets are not occluding each other as their DOA's cross, the vertical 2-D translation parameter $\eta_k(t)$ resolves the data association issue between the partitions. The motion parameters also resolve the data association, similar to the acoustic tracker, to alleviate the filter performance. If there is occlusion, it is handled separately using robust statistics as described below.

C. Occlusion Handling

In video, if the number of outlier pixels, defined in (9), is above some threshold, occlusion is declared. In that case, the updates on the appearance model and the adaptive velocity component in the state update (5) are stopped. The current appearance model is kept and the state is diffused with increasing diffusion variance. The data likelihood for the occluded target is set to 1 for an uninformative response under the influence of robust statistics. Similarly, the acoustic data likelihood is set to 1 when the number of DOA's within the batch gate of a partition is less than some threshold (e.g., M/2).

VIII. Simulations

Our objective with the simulations is to demonstrate the robustness and capabilities of the proposed tracker. We provide two examples. In the first example, a vehicle is visually occluded and the acoustic mode enables track recovery. In the second example, we provide joint tracking of two targets and provide time delay estimation results.

A. Tracking through Occlusion

Figure 9 shows the tracking results for a car that is occluded by a tree. The role of the DOA variable in the state space is crucial for this case. In the absence of information from any one of the modalities, the DOA still remains observable and is estimated from the modality that is not occluded. However, the rest of the states corresponding to the failed modality remains unobservable, and the variance of the particles along these dimensions continues to increase as the occlusion persists. Hence, it is therefore sometimes necessary to use an increasing number of particles to regain track until the failed modality is rectified.

The video modality regains the track immediately, as the target comes out of occlusion. The spread of particles (the dot cloud in Fig. 9) gives an idea of the observability of the vertical location parameter on the image plane. Further, the dramatic reduction in this spread as the target comes out of occlusion, demonstrates the previously unobservable visual components recovering the track. It is also interesting to compare the spread of particles in Fig. 9 with the pure visual tracking example in Fig. 3, where the spread of particle increases isotropically on the image plane, due to complete occlusion. Hence, the joint tracking reduces the uncertainty through the second modality. For this example, the simulation parameters are given in Table III. The acoustic bearing data is generated by adding Gaussian noise to the bearing track that corresponds to the ground truth. The acoustic bearing variance is 4 degrees between $t = 1$s to $t = 5$s, when the vehicle engine is getting occluded by the tree. It is 2 degrees when the vehicle engine is not occluded.

Figure 10 shows the results of a Monte-Carlo run, where the filter is rerun with different acoustic noise realizations. The threshold for declaring an occlusion is set as 40%. Figure 10(a) shows the joint bearing estimate results whereas Fig. 10(b) and (c) show the acoustics-only and video-only tracking results, respectively. In Fig. 10(a), there is a small positive bias in the bearing estimates at the end due to the target's pose change. As can be seen in Fig. 9(h) and (i), the rear end of the vehicle is visible after the vehicle comes out of the occlusion. The online appearance model locks on the front of the vehicle, whose appearance was stored before the occlusion. Hence, the rear end of the vehicle is ignored, causing the bias. We see in Fig. 10(c) that the video-only tracker cannot handle this persistent occlusion without the help of the acoustics.

Note the time evolution of the estimate variances shown in Figs. 10(d) and (e) for the joint tracker and the acoustics-only tracker. When the video modality is unable to contribute, the variance of the estimate approaches acoustics-only results. When the video recovers, the estimate variance drops sharply. Figures 10(f) and (g) show the distribution of the vertical displacement parameter. When the occlusion is over at $t = 6$s, the video quickly resolves its ambiguity in the vertical displacement (Fig. 10(g)), whereas the variance of the vertical displacement in Fig. 10(f) increases linearly with time due to divergence. Figures 10(h) and (i) demonstrate the occlusion probability of the target.

TABLE III
Simulation Parameters

Number of particles, N	1000
$\varphi(t)$ noise Σ_φ	diag $[0.02, 0.002, 0.002, 0.2, 2]$
θ noise $\sigma_{\theta,k}$	1°
Q noise $\sigma_{Q,k}$	0.05s^{-1}
ϕ noise $\sigma_{\phi,k}$	4°
Video Measurement noise σ_θ	$0.1,^\circ$
App. Model Template Size	15×15 (in pixels)
Beamformer batch period, τ	$\frac{1}{30}$s
Frame Size	720×480

TABLE II
JOINT ACOUSTIC VIDEO PARTICLE FILTER TRACKER PSEUDO-CODE

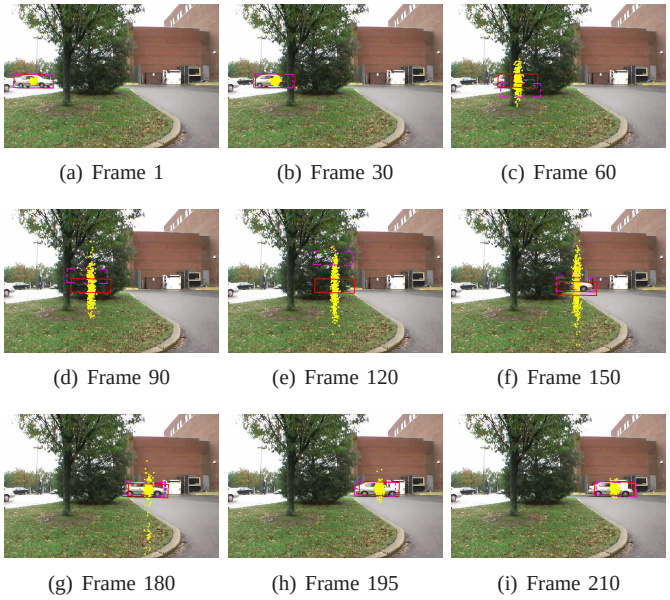
1. For each particle i ($i = 1, 2, \ldots, N$) and each partition k ($k = 1, 2, \ldots, K$)
   - Sample the time delay $d_k^{(i)}(t) \sim g_d(d_k(t)|y_{1,t}, y_{2,t}, x_k^{(i)}(t-T))$, where $g_d(\cdot)$ is the Gaussian approximation to (26) and (27).
   - Using the procedure illustrated in Table I, sample $\chi_k^{(i)}(t)$, $\psi_k^{(i)}(t)$, and $\varphi_k^{(i)}(t)$ from $x_k^{(i)}(t-T)$ with the time synchronized acoustic and video data $y_{1,t}$ and $y_{2,t-d^{(i)}(t)}$.
2. Calculate the weights $w_t^{*(i)}$ using (23). Determine visual and acoustic occlusions by looking at the likelihood estimates of each particle: $p(y_{1,t}|\chi^{(i)}(t), \psi^{(i)}(t))$ (acoustics) and $p(y_{2,t}|\chi^{(i)}(t), \varphi^{(i)}(t))$ (video).
   - A particle is *visually occluded* if a sufficient number of pixels in the template are outliers for the appearance model. The number of outlier pixels is calculated by (7) and (9): the number of terms in the summation for which $\rho(u)$ function is evaluated on the region $|u| > c$. If the number of such pixels is higher than 50%, it is claimed that the appearance, as hypothesized by the particle, is visually occluded.
   - If the particle that has the maximum video likelihood is visually occluded, then declare that the target has been occluded for the frame. In this case, the states represented by $\varphi(t)$ are unobservable and their sampling is done separately as in [11].
   - Similarly, a particle is *acoustically occluded*, if the observation DOA's $\mathbf{y}_{1,t}$ differ significantly from the value of DOA hypothesized by the mode particle. By counting the DOA's $y_{1,t+m\tau}$ in the gate of the hypothesized DOA's $h_{m\tau}(x^{(i)}(t))$, we declare an acoustic occlusion. If more than half the DOA observations in the batch are termed occluded, the particle is labeled as acoustically occluded.
   - If the particle that has the maximum acoustic likelihood is acoustically occluded, then we term the estimation at time t to be acoustically occluded. In this case, the states $\psi(t)$ are unobservable and are sampled separately as in [13].
   - When a particle is occluded, the corresponding time delay is sampled from (26).
3. Calculate the weights using (23) and normalize.
4. Perform the estimation [27]: $E\{f(\mathbf{x}_t)\} = \sum_{i=1}^{N} w_t^{(i)} f(\mathbf{x}_t^{(i)})$.
5. Resample the particles: Only states that are observable participate in resampling. For example, if the observations are visually occluded then the states $\varphi(t)$ are not resampled. Similarly, if the observations are acoustically occluded, then the states $\psi(t)$ are not resampled.
   - Heapsort the particles in a ascending order according to their weights: $\mathbf{x}_t^{(i)} \rightarrow \tilde{\mathbf{x}}_t^{(i)}$.
   - Generate $\omega \sim \mathcal{U}[0, 1)$.
   - For $j = 1, 2, \ldots, N$
     a. $u^{(j)} = \frac{j-\omega}{N}$,
     b. Find i, satisfying $\sum_{l=1}^{i-1} \tilde{w}_t^{(i)} < u^{(j)} \leq \sum_{l=1}^{i} \tilde{w}_t^{(i)}$,
     c. Set $\mathbf{x}_t^{(j)} = \tilde{\mathbf{x}}_t^{(i)}$.
6. Update the appearance model with the appearance corresponding to the particle with maximum likelihood, if this likelihood value exceeds the threshold. The appearance model is not updated during visual occlusion. Finally, we reinitialize the appearance model when the tracker is visually unoccluded for 10 consecutive frames, after visual occlusions of at least one second.

B. Time Delay Estimation

We performed a simulation with the time delay variable on a synthetically constructed multi-target data set. The simulation parameters are given in Table IV. The temporal tracks of two targets are shown in Fig. 11. The simulation parameters are given in Table II. The results of the DOA and time delay estimation are shown in Fig. 12. The filter handles multiple targets independently by treating the data of the other target as clutter. Note the variance of the time delay estimates decreases as the targets get closer to the hybrid sensors. It is important to account for this time delay, because filter instability occurs due to the estimation biases when filtered with the unsynchronized data.

IX. CONCLUSIONS

In this paper, we presented a particle filter tracker that can exploit acoustic and video observations for target tracking by merging different state space models that overlap on a common parameter. By the construction of its proposal function, the filter mechanics render the particle filter robust against target occlusions in either modality, when used with Huber's robust statistics criterion function. The presented filter also demonstrates a scheme for adaptive time-synchronization of the multi



(a) Frame 1 (b) Frame 30 (c) Frame 60

(d) Frame 90 (e) Frame 120 (f) Frame 150

(g) Frame 180 (h) Frame 195 (i) Frame 210

Fig. 9. Joint tracking of a vehicle that is occluded by a tree. The particle cloud at each frame represents the discrete support of the posterior distribution of the vehicle position in the image plane. Note that the particle spread during the occlusion increases along the vertical axis. This spread suddenly decreases, once occlusion is gone. The target is occluded in frames 40 to 180.

TABLE IV
SIMULATION PARAMETERS

Number of particles, N	1000
θ noise $\sigma_{\theta,k}$	1°
Q noise $\sigma_{Q,k}$	0.05s^{-1}
ϕ noise $\sigma_{\phi,k}$	4°
Time delay d noise $\sigma_{d,k}$	0.2s
Acoustic Measurement noise σ_θ	1°
Video Measurement noise σ_θ	$0.1,^\circ$
Beamformer batch period, τ	$\frac{1}{30}$s

modal data for parameter estimation. The time delay variable is incorporated into the filter and is modeled as multiplicative. It is the authors' observation that without the time delay variable, the joint filter is susceptible to divergence.

X. ACKNOWLEDGEMENTS

The authors would like to thank Milind Borkar, Soner Ozgur, and Mahesh Ramachandran for their help in the collection of the data that generated Figs. 3, 4, 5, and 9. We also would like to thank the anonymous reviewers, whose comments improved the final presentation of the paper.

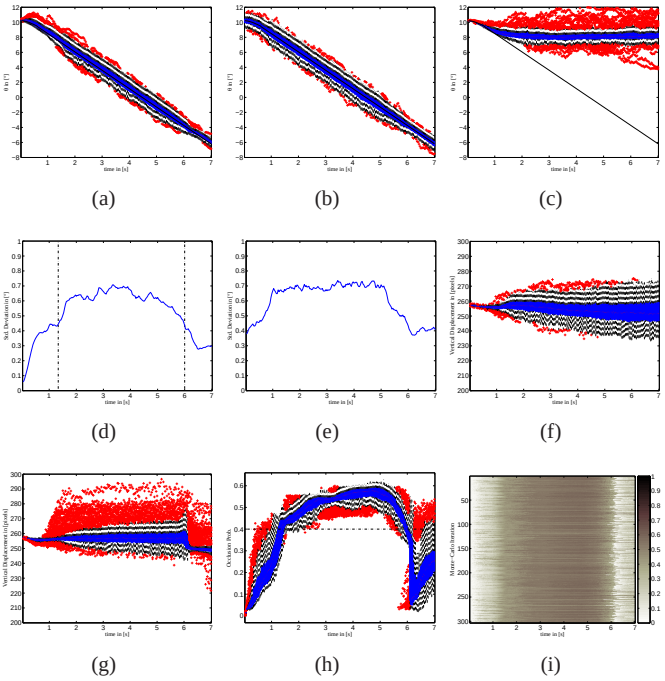
REFERENCES

[1] R. Chellappa, G. Qian, and Q. Zheng, "Vehicle detection and tracking using acoustic and video sensors," in *ICASSP 2004*, Montreal, CA, 17-21 May 2004.

[2] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, 1993.

[3] M. Wax and T. Kaliath, "Detection of signals by information theoretic criteria," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 387–392, April 1985.

[4] S. Valaee and P. Kabal, "Detection of signals by information theoretic criteria," *IEEE Trans. on Signal Processing*, vol. 52, pp. 1171–1178, May 2004.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 10. Results of 300 independent Monte-Carlo simulations of the experiment illustrated in Fig. 9. (a) MATLAB's boxplot of the estimated target DOA track with the joint tracker. The visual occlusion is between $t = 1$s and $t = 6$s. There is a small positive bias in the bearing estimates because of effect of the Brownian nature of the video state update equation in (13). (b) The estimated DOA track using acoustics-only. (c) The estimated DOA track using video-only. The video cannot handle the persistent occlusion by itself. (d-e) The time evolution of the estimate variances is shown for the joint filter and acoustics only, in their respective order. When the video is unable to provide information, the joint tracker' estimation performance becomes similar to the acoustics-only tracking results. The joint tracker's variance of the bearing estimate during the occlusion is slightly smaller than the acoustics-only variance because it is biased. (f) Vertical displacement is unobservable during the visual occlusion. Hence, the video-only estimate variance increases linearly with time. (g) Note the variance of the estimates dramatically reduces once the target becomes unoccluded, demonstrating the recovery speed of the tracker. (h) The occluded percentage of pixels, corresponding to the MAP particle. The gradual rise is attributed to the increasing partial occlusion as the car drives behind the tree, hence there is significant drop once the target comes out of occlusion. (i) Probability of occlusion for the Monte-Carlo runs. The track recovery after occlusion is robust as illustrated by the Monte-Carlo runs.

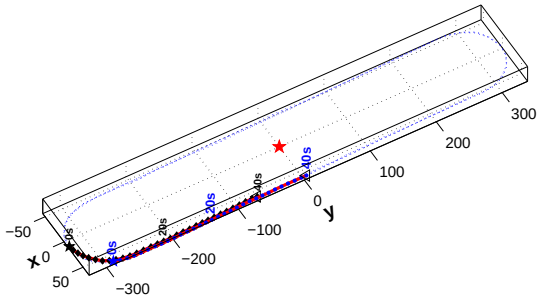
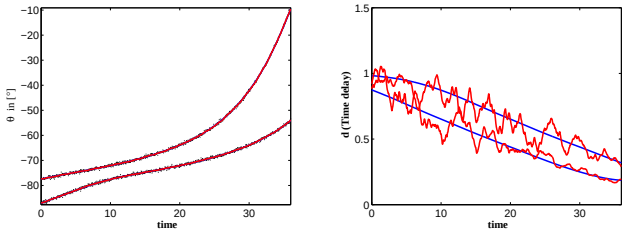


Fig. 11. Two targets follow an oval track (dotted line). The hybrid node is situated at the origin.



(a) (b)

Fig. 12. (a) Tracking of multiple targets with simulatenous estimation of time delay. (b) Estimated time delays. Note the reduction in the variance of the time delay estimates as the time delays get smaller.

[5] R. Hartley and A. Zisserman, *Multiple View Geometry in computer vision*, Cambridge University Press, 2003.

[6] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking," in *ICCV 2001*, 7–14 July 2001.

[7] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *ICASSP 2004*, Orlando, FL, 17-21 May 2004, vol. 5, pp. 881–884.

[8] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *ICIP 2003*, 14–17 Sept. 2003.

[9] M. Isard and A. Blake, *Active Contours*, Springer, 2000.

[10] B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans. Image Processing*, vol. 11, pp. 530–544, May 2002.

[11] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. Image Processing*, vol. 13, pp. 1491–1506, November 2004.

[12] A. D. Jepson, D. J. Fleet, and T. El-Maraghi, "Robust online appearance model for visual tracking," *IEEE Trans. on Pattern Anal. and Mach. Int.*, vol. 25, pp. 1296–1311, Oct. 1998.

[13] V. Cevher and J. H. McClellan, "Acoustic direction-of-arrival multi target tracking," under revision at *IEEE Trans. on SP*.

[14] I. Leichter, M. Lindenbaum, and E. Rivlin, "A probabilistic framework for combining tracking algorithms," in *CVPR 2004*, WDC, June 27–July 2 2004.

[15] V. Cevher and J. H. McClellan, "Proposal strategies for joint state space tracking with particle filters," in *ICASSP 2005*, Philadelphia, PA, 18–23 March 2005.

[16] S. Mallat and S. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[17] V. Cevher and J. H. McClellan, "An acoustic multiple target tracker," in *IEEE SSP 2005*, Bordeaux, FR, 17–20 July 2005.

[18] Y. Zhou, P. C. Yip, and H. Leung, "Tracking the direction-of-arrival of multiple moving targets by passive arrays: Algorithm," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2655–2666, October 1999.

[19] V. Cevher and J. H. McClellan, "General direction-of-arrival tracking with acoustic nodes," *IEEE Trans. on Signal Processing*, vol. 53, no. 1, pp. 1–12, January 2005.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[21] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, pp. 195–239, April 1984.

[22] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, pp. 73–101, March 1964.

[23] M. R. Liggings II, C. Y. Chong, I. Kadar, M. G. Alford, V. Vannicola, and S. Thomopoulos, "Distributed fusion architectures and algorithms for target tracking," *Proceedings of the IEEE*, vol. 85, pp. 95–107, Jan. 1997.

[24] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. R. Statist. Soc. B*, vol. 41, pp. 113–147, 1979.

[25] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society*, vol. 28, pp. 131–142, 1966.

[26] J. S. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems," *Journal of the American Statistical Association*, vol. 93, pp. 1032–1044, September 1998.

[27] A. Doucet, N. Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[28] B. D. Ripley, *Stochastic Simulation*, John Wiley & Sons Inc., 1987.

[29] V. Cevher and J. H. McClellan, "Fast initialization of particle filters using a modified Metropolis-Hastings algorithm: Mode-Hungry approach," in *ICASSP 2004*, Montreal, CA, 17–22 May 2004.

[30] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*, Academic-Press, 1988.