VEHICLE DETECTION AND TRACKING USING ACOUSTIC AND VIDEO SENSORS

Rama Chellappa[†]*, Gang Qian*[‡] *and Qinfen Zheng*[†]

[†]Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD, 20742-3275
{rama, qinfen}@cfar.umd.edu

[‡]Department of Electrical Engineering and
Arts, Media and Engineering Program
Arizona State University
Tempe, AZ 85287-8706
gang.qian@asu.edu

ABSTRACT

Multimodal sensing has attracted much attention in solving a wide range of problems, including target detection, tracking, classification, activity understanding, speech recognition, etc. In surveillance applications, different types of sensors, such as video and acoustic sensors, provide distinct observations of on-going activities. In this paper, we present a fusion framework using both video and acoustic sensors for vehicle detection and tracking. In the detection phase, a rough estimate of target direction-of-arrival (DOA) was first obtained using acoustic data through beam-forming techniques. This initial DOA estimate designates approximate target location in video. Given the initial target position, the DOA is refined by moving target detection using the video data. Markov Chain Monte Carlo techniques are then used for joint audio-visual tracking. A novel fusion approach has been proposed for tracking, based on different characteristics of audio and visual trackers. Experimental results using both synthetic and real data are presented. Improved tracking performance has been observed by fusing the empirical posterior probability density functions obtained using both types of sensors.

1. INTRODUCTION

Joint audio-visual tracking has attracted much attention recently e.g. [1, 2]. However, it's not clear how to efficiently fuse audio and visual data for outdoor scenarios using low quality video and acoustic sensors. In this paper, we present a computational framework for joint audio-visual vehicle detection and tracking using Markov Chain Monte Carlo (MCMC) techniques. Low quality cameras with narrow field of view were used as video sensing devices. A novel fusion approach has been proposed for tracking, according to different characteristics of audio and visual trackers. Given a ground vehicle, its direction-of-arrival (DOA), heading direction and ratio of speed to its range are parameters which describe its position and dynamics. Both video cameras and acoustic array have been used as sensing devices to estimate these parameters. Each type of sensors has its own advantages as well as disadvantages. Acoustic array has complete field of sensing (360 degrees) and beam-forming techniques are ready for a rough estimate the DOA. However, tracking accuracy using acoustic array is limited and it has difficulties to identify the number of the vehicles in the field. Video cameras provide accurate DOA estimates and can easily find out the number of moving targets, but it has limited field of view. The estimation of vehicle heading direction is usually difficult when the target image size is small and no enough features can be extracted and reliably tracked. These disadvantages become dominate when low quality video and acoustic sensors are used in outdoor environments. In this paper, we propose a data fusion framework based on the Markov Chain Monte Carlo techniques. It combines detection and tracking results from co-centered acoustic array and video camera. By using both synthetic and real data, we have shown that the proposed fusion framework improves the overall vehicle detection and tracking performance.

2. SENSOR AND SYSTEM MODELS

The acoustic array we used contains a number of sensors uniformly distributed along a circular track. The camera center coincides with the circular acoustic array center. We assume the full knowledge of the sensor calibrations, such as camera focal length, sensor number and locations as well as the size of the acoustic array.

2.1. Acoustic Sensors

Let P be the number of acoustic sensors and K the number of vehicles in the field. Following the notation in [3, 5], the state parameters for the k^{th} target at time instant t is given by

$$x_k(t) = [\theta_k(t), q_k(t), \phi_k(t)]^T \quad (1)$$

where $\theta_k(t)$, $q_k(t)$ and $\phi_k(t)$ are the DOA, logarithmic ratio of the k^{th} target speed to its range, and the target heading direction, respectively. DOA is measured clockwise from the Y axis, while the heading directions counter clockwise from the X axis. Figure 1 shows the geometry of the problem with both video and acoustic sensors, where $p(t)$ is the vehicle position at time t. Using an acoustic array, a steering vector describes the complex array response for a target at DOA θ. The steering vector of the k^{th} target is given by

$$d(\theta_k) = \left[e^{-j2\pi f_k \beta_k^T z_1} \quad \cdots \quad e^{-j2\pi f_k \beta_k^T z_P} \right]^T \quad (2)$$

and $\beta_k=(1/c)[\cos(\theta_k),\sin(\theta_k)]^T$. z_l is the lth acoustic sensor location. c is sound speed. The array output for chirp signals is

$$y(t) = D(t)s(t) + n(t) \quad (3)$$

where $n(t)$ is an additive observation noise and $s(t)$ is the signal vector. $D(t)$ is formed by the steering vectors of different targets.

$$D(t)=[d(\theta_1(t)) \cdots d(\theta_k(t))] \quad (4)$$

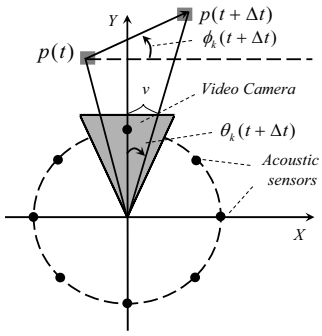

Figure 1. Sensor geometry and target motion parameters

2.2. Video Camera

Vehicles moving on the ground plane are tracked by static cameras. Based on this context knowledge, three parameters are used to describe the target motion in the image plane,

$$x_k^{(V)}(t) = \left[s_k(t), a_k(t), b_k(t)\right]^T \quad (5)$$

where s is the scaling factor, representing the size changes of the target on the image plane. It is determined by the distance of the target from the camera. *(a,b)* are the translation of the target in the horizontal and vertical directions in the image plane. Suppose that at the initial time instant the image coordinate of a point on the target being tracked is (u_0, v_0). Then, the image coordinate of the same point on the vehicle after motion at time t is given by

$$\begin{bmatrix} u_t \\ v_t \end{bmatrix} = s(t) \cdot \left(\begin{bmatrix} u_0 \\ v_0 \end{bmatrix} - \begin{bmatrix} a(t) \\ b(t) \end{bmatrix} \right) \quad (6)$$

Given the target position in the image plane, which is the centroid of the target bounding box, it is straightforward to find out the target DOA. Assume the image coordinate of the target centroid is *(u,v)*, the related DOA is given by

$$\theta = \arctan(v / f) \quad (7)$$

with f being the focal length of the camera. Assume that over a small time interval Δt, the vehicle translation along the Y axis (camera looking direction) is small, then the logarithmic ratio of the target speed to its range can be approximated by

$$q = \log(\Delta d) - \log(\Delta t \cdot f) \quad (8)$$

where Δd is the centroid displacement over Δt.

3. VEHICLE DETECTION

Vehicle detection is achieved by fusing detection results from both acoustic and video processing. The fusion framework is illustrated by Figure 2. At first, rough DOA estimates are obtained by using narrow-band beam-forming techniques such as multiple signal classification (MUSIC). At the same time, image background of the video camera is modeled by a mixed-Gaussian distribution. Since our video and acoustic sensors are co-centered, this rough DOA from acoustic processing corresponds to a vertical strip in the images captured by the video camera. By applying background subtraction to this vertical strip in the image, moving vehicles can be detected so that refined DOA estimates can be obtained.

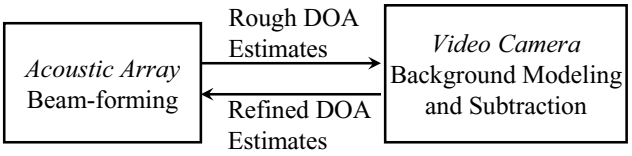

Figure 2. Fusion for target detection

4. TRACKING

Once targets are detected, both acoustic data and video are used for target tracking. According to the data sampling difference between video and acoustic sensor, the tracking results from different sensors are informed to each other as guidance for target tracking in the next time instant. Both acoustic and video tracking deploy MCMC techniques such as particle filter to achieve robustness against observation. For acoustic tracking, we utilized an implementation of the algorithms described by [5].

4.1. Video Tracking

We developed a video tracking algorithm using a motion-encoded particle filter [6]. In the motion-encoded particle filter, motion detection is applied to guide the particle distributions for target tracking from one time instant to the next. Assume that at time t, target k is tracked by particles $\{x^{(i)}\}$ with weights $\{w^{(i)}\}$. To continue tracking at time $t+1$, n_i predicted particles will be drawn based on $x^{(i)}$, according to the state dynamics. n_i is proportional to $w^{(i)}$, the corresponding weight of $x^{(i)}$. Suppose that there are m potential objects detected by background subtraction technique. They indicate possible locations of this targets at time $t+1$. Hence, the n_i predicted particles generated from $x^{(i)}$ should be distributed among these m candidates according to certain distribution. Let $n_i^{(j)}$ be the number of particles assigned to potential target j, $j=1,\ldots,m$ and $n_i^{(j)}$ can be computed by

$$n_i^{(j)} = n_i p(d_j) / \sum_{j=1}^{m} p(d_j) \quad (9)$$

where d_j is the Euclidean distance between the center of the j^{th} potential target and that of the target being tracked by $x^{(i)}$ at time t. $n_i^{(j)}$ is proportional to a distribution $p(d_j)$, which could be a truncated Gaussian with zero mean. Intuitively, the closer the potential target is to the current target position represented by $x^{(i)}$, the more likely this is the right target and more particles should be assigned to this potential target.

4.2. Tracking Fusion

Observation noises have different effects on acoustic and visual tracking. For example, in our experiments, we have observed larger variances presented in the empirical pdf of DOA obtained using acoustic data, comparing with video tracking. On the other hand, video tracking can be disturbed easily by things such occlusion, low contrast between background and vehicle, which might occur fairly often in real scenarios. Therefore, the fusion goal is to improve the tracking accuracy, with the ability to combat target occlusion. The following fusion approach has been used to achieve this goal. Initially, targets are tracked separately using acoustic and video data. The tracking fusion occurs periodically with a period of ΔT. Let $\{x_{v,t}^{(i)}, w_{v,t}^{(i)}\}$ and $\{x_{a,t}^{(i)}, w_{a,t}^{(i)}\}$ be the target motion parameter samples and weights at time t from video and acoustic tracking, respectively. The fusion contains two steps. First, video motion estimates at

previous time instant *t-1* are used in the prediction of acoustic tracking (as illustrated by Figure 3). Then, the weights of these new samples are evaluated using acoustic data. Equation (7) and (8) are used to convert video motion parameters to acoustic motion parameters. In our implementation, half new acoustic motion samples are obtained based on the video motion samples and the other half are drawn based on previous acoustic motion samples. Let N be the number of particles used in acoustic tracking. The detailed algorithm for fusing video tracking results in acoustic tracking is as follows.

Fusion of Video Motion Samples and Weights in Acoustic Tracking

1. Transform parameters. Suppose the initial position of the current tracked target is (u_0, v_0). For video motion sample, $x_{v,t-1}^{(i)}=[s_{t-1}^{(i)},a_{t-1}^{(i)},b_{t-1}^{(i)}]$, the corresponding acoustic motion parameters are given by

$$\theta_v^{(i)} = \arctan((v_0 + b_{t-1}^{(i)})/f)$$

$$q_v^{(i)} = 0.5\log((a_{t-1}^{(i)} - a_{t-1-\Delta t}^{(i)})^2 + (b_{t-1}^{(i)} - b_{t-1-\Delta t}^{(i)})^2) - \log(\Delta t \cdot f)$$

   Draw $\phi_v^{(i)}$ from uniform distribution in $(d_v, d_v]$, where the boundaries can be determined by motion vector in the image plane and the change of the scale parameters.
2. Draw $N/2$ motion samples from $\{\theta_v^{(i)}, q_v^{(i)}, \phi_v^{(i)}\}$ according to weights $\{w_{v,t-1}^{(i)}\}$. Draw another $N/2$ motion samples from $(\{x_{a,t-1}^{(i)}\}$, according to weights $\{w_{a,t-1}^{(i)}\}$.
3. Add dynamic noises to these samples to obtain prediction of acoustic motion sample at time t.

When a vehicle is occluded visually, the tracking will be solely performed using acoustic data. Similarly, no acoustic data will be observed when the vehicle stopped with engine turned off and only video data will be used.

5. EXPERIMENT RESUTLS

Both synthetic and real data have been used to test the proposed fusion framework for target detection and tracking using acoustic and video sensors. Synthetic videos were produced through the projection of 3D vehicle models to the image plane based on the model position in 3D space. Assuming Lambertian surfaces, texture maps of the vehicles with any desired viewing and lighting angles can be produced. Given camera and acoustic array calibration, the visual and audio data of one or more vehicles at any position along pre-designated tracks can be easily created. In one of our experiments, two vehicles traveled across the field of view of the camera. The truck moves away from the camera and the tank toward the camera. A total of 180 frames of video and audio data were generated. In this experiment, the video algorithm detected these vehicles and sent the counts and initial state estimates to the acoustic processing algorithm. Then the acoustic processing algorithm used the initial state input from the camera and kept tracking of these two vehicles. Figure 4 shows the detection results and the related templates, marked by bounding boxes and labeled by target IDs. Table 1 gives the initial estimates and ground-truth values of the DOAs, heading angles and relative speeds of these two targets.

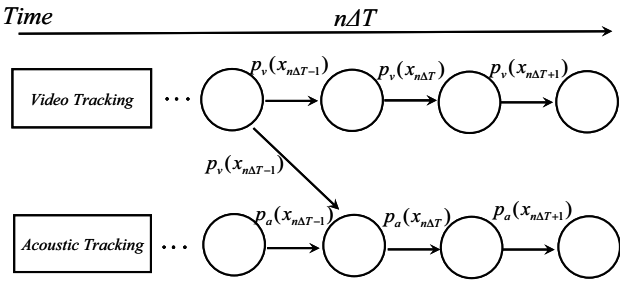

Figure 3. One fusion snapshot at ΔT during tracking

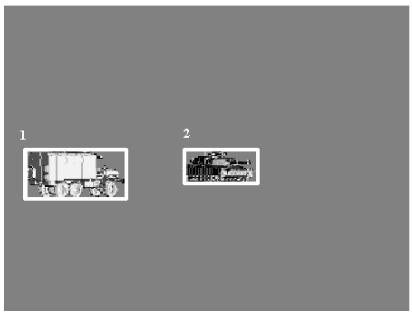

Figure 4. Detected targets from videos

Table 1: Initial Estimates from Video and Ground Truth of Acoustic State Parameters for Target 1 (T1) and Target 2 (T2)

	T1: Initial Estimates /Ground Truth	T2:
θ	-9.3°/-10.79°	1°/0.097°
q	-3.69/-3.37	-3.62/-3.40
ϕ	Uniformly distributed in [-45°,45°]/ 37.79°	Uniformly distributed in [-45,45]/ -39.85°

One example using real data is also included in this paper. Figure 5 shows a bird's eye-view of the monitored field, with the trajectory of a vehicle moving from upper-left to lower-right. The arrow indicates the camera looking direction. The camera has a narrow vertical field of view of 5.15°. The pixel size of the images is 480 × 720. The acoustic array contains 4 sensors, with a diameter of one meter. Video and acoustic data collected over ten seconds were used to test the algorithm. Figure 6 shows the vehicle detection and tracking using video. The empirical posterior probability density functions (pdf) of the DOA at the last time instant using acoustic data, video and after sensor fusion are shown in figure 7 from top to bottom, respectively. The minimum mean square error (MMSE) estimates of DOA at different time can be found by computing the conditional mean of the corresponding empirical pdf (Figure 8). It can be clearly seen that the MMSE estimates with sensor fusion are the closest to the ground truth. The performance improvement on target heading direction and the ratio of speed to target range is not obvious by fusing both types of sensors. It is not surprising since these two parameters are related to the velocity of the vehicle, which is a higher order motion parameter, comparing with DOA. One of the reasons for no obvious improvement from fusion is the lack of accurate mapping between acoustic and video

tracking along the dimension of these two parameters in the state space.

6. CONCLUSIONS

A sensor fusion framework for vehicle detection and tracking using video and acoustic data is presented in this paper. Due to the sensor fusion, the detection takes advantages of both the omni-direction sensing field of the acoustic sensors as well as the detection accuracy of the video camera. Video motion samples are deployed to guide the prediction step in acoustic tracking. Improved DOA estimates have been obtained by using the proposed fusion framework.

7. REFERENCES

[1] J. Vermaak, M. Gangnet, A. Blake and P. Perez, "Sequential Monte Carlo Fusion of Sound and Vision for Speaker Tracking," Proc. ICCV 2001, vol I, pp. 741-748

[2] D. Zotkin, R. Duraiswami, L. Davis "Joint Audio-visual Tracking Using Particle Filters", EURASIP journal on Applied Signal Processing, vol. 2002(11), pp. 1154-1164

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

[4] M. Orton and W. Fitzgerald, "A Bayesian approach to tracking multiple targets using sensor arrays and particle filters," IEEE Trans.on Acoustics, Speech, and Signal Processing, vol. 50, no. 2, Feb.2002, pp. 216-223

[5] V. Cevher, J. H. McClellan, "Tracking of Multiple Wideband Targets Using Passive Sensor Arrays and Particle Filters," IEEE 10th Digital Signal Processing Workshop, October 13-16, 2002

[6] B. Li, R. Chellappa. "Simultaneous Tracking and Verification via Sequential Posterior Estimation", Proc. CVPR 2000, Hilton Head, SC, June 2000.

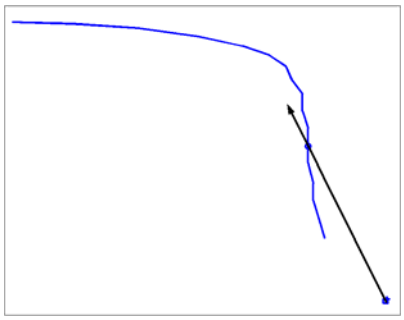

Figure 5: Vehicle trajectory and sensor location (arrow starting point) and camera looking direction (arrow direction)

Figure 6: Video tracking results

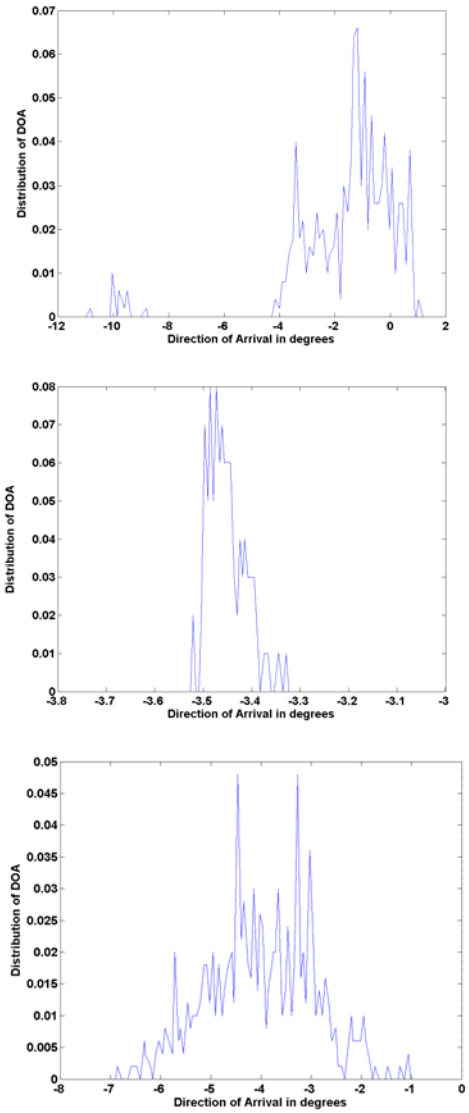

Figure 7: Empirical pdf of DOA using acoustic (top), video data (middle) and after fusion (bottom) at the last time instant

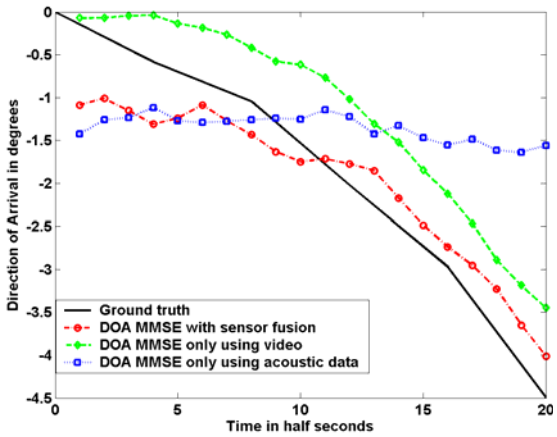

Figure 8. MMSE estimates before and after sensor fusion