Target Tracking in Heterogeneous Sensor Networks Using Audio and Video Sensor Fusion

Manish Kushwaha, Songhwai Oh[†], Isaac Amundson, Xenofon Koutsoukos, and Akos Ledeczi

Institute for Software Integrated Systems
Vanderbilt University
Nashville, TN, USA
{*manish.kushwaha,isaac.amundson, xenofon.koutsoukos,akos.ledeczi*}*@vanderbilt.edu*

[†]School of Engineering
University of California - Merced
Merced, CA, USA
songhwai.oh@ucmerced.edu

***Abstract*— Heterogeneous sensor networks (HSNs) with multiple sensing modalities are gaining popularity in diverse fields. Tracking is an application that can benefit from multiple sensing modalities. If a moving target emits sound then both audio and video sensors can be utilized. These modalities can complement each other in the presence of high background noise that impairs the audio or visual clutter affecting the video. Audio-video tracking can also provide cues for the other modality for actuation. In this paper, we describe an approach for target tracking in urban environments utilizing an HSN of mote class devices equipped with acoustic sensor boards and embedded PCs equipped with web cameras. Our system employs a Markov Chain Monte Carlo Data Association algorithm for tracking vehicles emitting engine noise. Experimental results from a deployment in an urban environment are used to demonstrate our approach.**

I. Introduction

Heterogeneous sensor networks (HSN) with multiple sensing modalities are gaining popularity in diverse fields because they can support multiple applications that may require diverse resources [25]. Multiple sensing modalities provide flexibility and robustness, however, different sensors may have different resource requirements in terms of processing, memory, or bandwidth (e.g., microphones vs. cameras). An HSN can have nodes with various capabilities for supporting several sensing tasks.

Multiple-target tracking is one such application that can benefit from multiple sensing modalities. Multiple-target tracking plays an important role in many areas of engineering such as surveillance [1], computer vision [6], network and computer security [7], and sensor networks [18]. If the targets are moving and emit some kind of sound then both audio and video sensors can be utilized. These modalities can complement each other in the presence of high background noise that impairs the audio or visual clutter affecting the video.

In this paper, we describe an approach for target tracking in urban environments utilizing an HSN of mote class devices equipped with acoustic sensor boards and embedded PCs equipped with web cameras. Our system employs a Markov Chain Monte Carlo Data Association (MCMCDA) algorithm [17] for tracking vehicles emitting engine noise. The paper also describes briefly the components of the system for audio processing, video processing, and multi-modal sensor fusion. Experimental results from a deployment in an urban environment are used to demonstrate our approach.

An overview on beamforming and its application for localization in sensor networks can be found in [5]. Beamforming methods have successfully been applied to detect single or even multiple sources in noisy and reverberant environments [4], [15]. Adaptive background-modeling methods for motion detection based on video include the work in [9] which modeled each pixel in a camera scene by an adaptive parametric mixture model of three Gaussian distribution and the adaptive nonparametric Gaussian mixture model to address background modeling challenges presented in [22]. Other techniques using high-level processing to assist the background modeling also have been proposed [11], [24]. Work in multimodal target tracking and multimodal sensor fusion using audio-video data includes object localization and tracking based on Kalman filtering [23] as well as particle filtering approaches [3], [2].

The rest of the paper is organized as follows. The next section describes the overall system architecture including the a description of the audio and the video processing approach. Next the multimodal sensor fusion in presented in section III. The multiple-target tracking algorithm is presented in section IV. The experiment and its evaluation is described in Section V followed by a summary of related work. Finally, we discuss lessons learned and future directions in section VI.

II. Architecture

The architecture of our system is shown in Figure 1. The HSN consists of audio sensors that perform beamforming and video sensors that detect moving objects. All nodes are time synchronized to allow sensor fusion. The sensor fusion node contains circular buffers that store timestamped measurements. A sensor fusion scheduler triggers periodically and generates a fusion timestamp which is

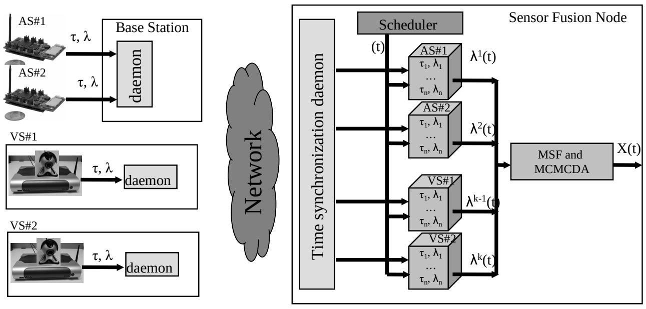

Fig. 1. Multimodal tracking system architecture

used to retrieve the sensor measurement values from the sensor buffers with timestamps closest to the generated fusion timestamp. The retrieved sensor measurement values are then used for multimodal fusion and estimation and tracking. Next, we briefly describe the main components of the system.

Audio Beamforming: Beamforming can be used to determine the direction(s) of arrival and the location(s) of acoustic source(s) [4]. A typical delay-and-sum beamformer divides the sensing region into directions, or *beams*. For each beam, assuming the source is located in that direction, the microphone signals are delayed according to the phase-shift and summed together into a composite signal. The square-sum of the composite signal, or the beam energy is computed for each of the beams, and are collectively called the *beamform*. The beam with maximum energy indicates the direction of the acoustic source.

In our system, the audio sensor node is a MICAz mote with an onboard Xilinx XC3S1000 FPGA chip that is used to implement the beamformer. The board supports four independent analog channels. A small beamforming array of four microphones arranged in a $10cm \times 6cm$ rectangle was placed on the sensor node, as shown in Fig. 2. The sources are assumed to be on the same two-

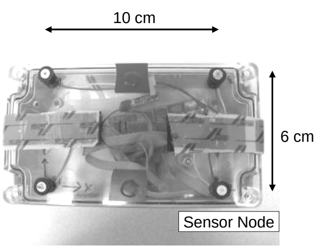

Fig. 2. Sensor Node Showing the Microphones

dimensional plane as the microphone array, thus it is sufficient to perform planar beamforming by dissecting the angular space into M equal angles, providing a resolution of $360/M$ degrees. In the experiments, the sensor boards were configured to perform simple delay-and-sum-type beamforming in real time with $M = 36$ beams, and an angular resolution of 10 degrees per beam.

Motion Detection Using Video: Video tracking systems aim at detecting moving objects and track their movements in a complex environment. A simple approach to motion detection from video data is via frame differencing. It compares each incoming frame with a background model and classifies the pixels of significant variation into the foreground. The foreground pixels are then processed for identification and tracking. We have implemented a motion detection algorithm using the background-foreground segmentation approach described in [11] which is a based on an adaptive background mixture model and provides robust performance and low complexity in a wide range of situations. Our sensor fusion method (Section III) utilizes only the angle of moving objects, thus we compute a simple detection function similar to the beam angle concept in audio beamforming. The detection function value for each beam direction is simply the number of foreground pixels in that direction. This detection function is similar to the horizontal intensity accumulation function (IAF) defined in [12]. In our experiments, we gathered video data of vehicles from multiple video sensors from an urban street setting. The data contained a number of real-life artifacts such as vacillating backgrounds, shadows, sunlight reflections and glint. The algorithm described above was not able to filter out such artifacts from the detections. We implemented two post-processing filters to improve the detection performance to remove undesirable persistent background and sharp spikes caused by sunlight reflections and glint.

The video sensors are based on Logitech QuickCam Pro 4000 cameras attached to OpenBrick-E Linux embedded PCs. The motion detection algorithm is implemented using OpenCV (open source computer vision) library. Our motion detection algorithm implementation runs at 4 frames-per-second and 320×240 pixel resolution. The number of beam angles is $M = 160$.

Time Synchronization: The audio sensors form an 802.15.4 network while the video sensors, the mote-PC gateways, and the sensor fusion node form a peer-to-peer 802.11b wireless network. In order to fuse audio and video sensor data for tracking moving objects, all the sensor nodes must have a common notion of time. Several synchronization protocols have emerged for wireless sensor networks (e.g. [8], [10]) but they cannot applied directly to HSNs. To synchronize the entire network, we integrated existing protocols that provide high accuracy and low overhead for a specific network. We used Elapsed Time on Arrival (ETA) [14] to synchronize the mote network and RBS [8] to synchronize the PC network. To synchronize a mote with a PC in software, we adopted the underlying methodology of ETA and applied it to serial communication. We evaluated synchronization accuracy using the pairwise difference method. Two motes timestamped the arrival of an event beacon, and forwarded the timestamp to the network sink, via one mote and two PCs. The average error over the 3-hop HSN was $101.52\mu s$, with a maximum of $1709\mu s$ which is sufficient for our application.

III. MULTIMODAL SENSOR FUSION

This section describes different sensor models and sensor fusion algorithms for audio and video sensors. We use (i) a nonparametric model for the audio sensors and (ii) a parametric mixture-of-Gaussian model for the video sensors.

Audio Sensor Model: Let $\lambda(\theta)$ denote the audio detection function (or beamform) that represent the energy of the acoustic source for each beam. The nonparametric DOA sensor model for a single audio sensor is the piecewise linear interpolation

$$\lambda(\theta) = w\lambda(\theta_{i-1}) + (1-w)\lambda(\theta_i), \text{ if } \theta \in [\theta_{i-1}, \theta_i]$$

where $w = (\theta_i - \theta)/(\theta_i - \theta_{i-1})$.

Video Sensor Model: The video detection algorithm captures the angle of one or more moving objects and is parametrized as a mixture-of-Gaussian

$$\lambda(\theta) = \sum_{i=1}^{n} a_i f_i(\theta)$$

where n is the number of components, $f_i(\theta)$ is the probability density function, and a_i is the mixing proportion for component i. Each component is a Gaussian density function parametrized by μ_i and σ_i^2

$$f_i(\theta) = \mathcal{N}(\theta|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(\theta-\mu_i)^2}{2\sigma_i^2}\right)$$

The component parameters μ_i, σ_i^2 and a_i are calculated from the detection function.

Likelihood Function: A likelihood function of the form

$$p(z|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\theta)^2}{2\sigma^2}\right)$$

for DOA sensors is presented in [16], where θ is calculated from the geometry of the sound source position x and the sensor position ζ. The variance σ^2 is an empirical function of distance of sound source from the sensor. We extended above likelihood function by incorporating energy in the variance. The modified likelihood function for above audio and video sensor models can be expressed as,

$$p(\lambda(\theta)|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta_{peak}-\theta)^2}{2\sigma^2}\right) \qquad (1)$$

where θ is calculated from the geometry of the target position x and the sensor position ζ, $\theta_{peak} = \arg\max \lambda(\theta)$ is the peak location closest to θ, $\lambda(\theta)$ is the sensor detection function described above, and $\sigma^2 = f(\lambda(\theta), x)$ is the variance which is a function of distance from sensor and the detection function value at the cell. Since the sensor models are nonlinear, it is reasonable to use a nonparametric representation for the likelihood functions which are represented as discrete grids in 2D space similar to [16]. Figure 3 shows an example video detection function and the corresponding likelihood function.

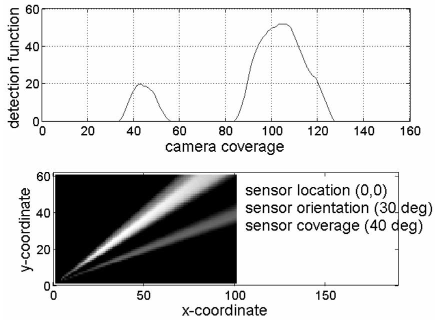

Fig. 3. An example video detection function and the corresponding likelihood function

The combined likelihood function from multiple sensors can be calculated either as product fusion

$$p(z|x) = \prod_{k=1,\ldots,K} p_k(z|x)$$

or as weighted-summation fusion

$$p(z|x) = \sum_{k=1,\ldots,K} w_k \cdot p_k(z|x)$$

of the individual sensor likelihood functions, where K is the number of sensors. Since we are using a common likelihood function for both audio and video modalities, the multimodal likelihood functions can be combined in a seamless manner.

Using the combined likelihood function of all relevant sensor, we compute target observations which is used by the multitarget tracking and data association algorithm described in next section (see section IV). The target observations are generated from the likelihood function using a peak detection algorithm that detects all the local maxima of the two-dimensional function.

IV. MULTIPLE-TARGET TRACKING

The essence of the multi-target tracking problem is to find a track of each object from the noisy measurements. If the sequence of measurements associated with each object is known, multi-target tracking reduces to a set of state estimation problems, for which many efficient algorithms are available. Unfortunately, the association between measurements and objects is unknown. The *data association* problem is to work out which measurements were generated by which objects; more precisely, we require a partition of measurements such that each element of a partition is a collection of measurements generated by a single object or clutter. Due to this data association problem, the complexity of the posterior distribution of the states of objects grows exponentially as time progresses. It is well-known that the data association problem is NP-hard [20], so we do not expect to find efficient, exact algorithms for solving this problem.

In order to handle highly nonlinear and non-Gaussian dynamics and observations, a number of methods based on particle filters has been recently developed to track multiple objects in video [19], [13]. Although particle filters are highly effective in single-target tracking, it is reported that they provide poor performance in multi-target tracking [13]. It is because a fixed number of particles is insufficient to represent the posterior distribution with the exponentially increasing complexity (due to the data association problem). As shown in [13], [26], an efficient alternative is to use Markov chain Monte Carlo (MCMC) to handle the data association problem in multi-target tracking.

For our problem, there is an additional complexity. We do not assume the number of objects is known. A *single-scan* approach, which updates the posterior based only on the current scan of measurements, can be used to track an unknown number of targets with the help of trans-dimensional MCMC [26], [13] or a detection algorithm [19]. But a single-scan approach cannot maintain tracks over long periods because it cannot revisit previous, possibly incorrect, association decisions in the light of new evidence. This issue can be addressed by using a *multi-scan* approach, which updates the posterior based on both current and past scans of measurements. The well-known *multiple hypothesis tracking* (MHT) [21] is a multi-scan tracker, however, it is not widely used due to its high computational complexity.

A newly developed algorithm, called Markov chain Monte Carlo data association (MCMCDA), provides a computationally desirable alternative to MHT [17]. The simulation study in [17] showed that MCMCDA was computationally efficient compared to MHT with heuristics (*i.e.*, pruning, gating, clustering, N-scan-back logic and k-best hypotheses). In this paper, we use the online version of MCMCDA to track multiple objects in a 2-D plane. Due to the page limitation, we omit the description of the algorithm in this paper and refer interested readers to [17].

V. Evaluation

The deployment of the multi-modal target tracking system is shown in Figure 4. We employ 6 audio sensors and 3 video sensors deployed on either side of a road. The complex urban street environment presents many challenges including gradual change of illumination, sunlight reflections from windows, glints due to cars, high visual clutter due to swaying trees, high background acoustic noise due to construction and acoustic multipath effects. The objective of the system is to detect and track vehicles using both audio and video under these conditions.

Sensor localization and calibration for both audio and video sensors is required. In our experimental setup, we manually placed the sensor nodes at marked locations and orientations. The audio sensors were placed on 1 meter high tripods to minimize audio clutter near the ground.

We gathered audio and video detection data for a total duration of 43 minutes. Table I presents the parameter values that we use in our tracking system. Sensor likelihood functions were calculated by discretizing the sensing region in specified cell-sized grid. The tracked vehicles were part of an uncontrolled experiment. The vehicles were traveling on road at 15-30 mph speed.

Number of beams in audio beam-forming, M_{audio}	36
Number of angles in video detection M_{video}	160
Sensing region (meters)	35×20
Cell size (meters)	0.5×0.5

TABLE I
Parameters used in experimental setup

In our simulations we experimented with six different approaches. We used audio-only (A), video-only (V) and audio-video (AV) sensor observations for sensor fusion. For each of these data sets, the combined likelihood was computed either as the weighted-sum or product of individual sensor likelihood functions.

The ground truth is estimated post-facto based on the video recording by a separate camera. The standalone ground truth camera was not part of any network, and had the sole responsibility of recording video. For evaluation of tracking accuracy, the center of mass of the vehicle is considered to be the true location.

A. Single Target

We shortlisted 9 vehicle tracks where there was only a single target in the sensing region. The average duration of tracks was 3.75 sec with 2.75 sec minimum and 4.5 sec maximum. Figure 5 shows the target tracking result for two different representative vehicle tracks. The figure also shows the raw observations obtained from multimodal sensor fusion and peak detection algorithm.

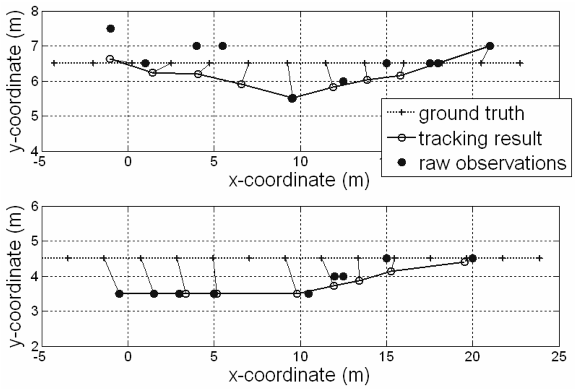

Fig. 5. Target Tracking (a) no missed detection (b) with missed detections

Figure 6 shows average tracking errors for all vehicle tracks for weighted-sum fusion approach. The missing bars indicate that the data association algorithm was not able to successfully estimate a track for the target. Figure 7 averages tracking errors for all the tracks to compare different tracking approaches. Table II compares average tracking errors and fraction of estimated tracks across likelihood fusion and modality dimension. Table III shows the reduction in tracking error for AV approach over audio-only and video-only approach. For summation fusion, the AV approach was able to reduce tracking error

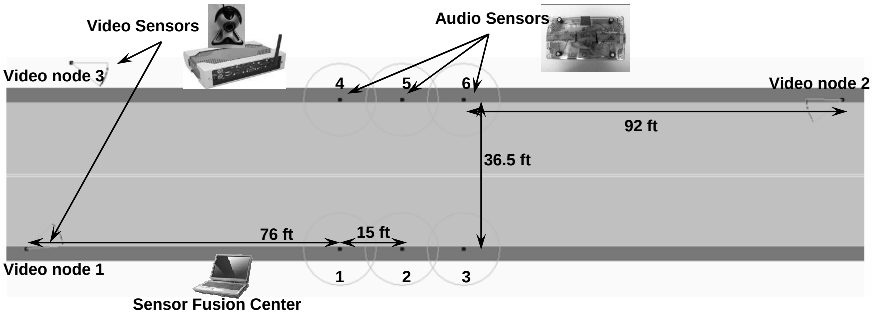

Fig. 4. Experimental setup

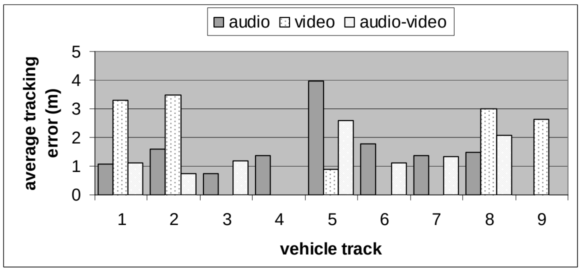

Fig. 6. Tracking errors (weighted sum fusion)

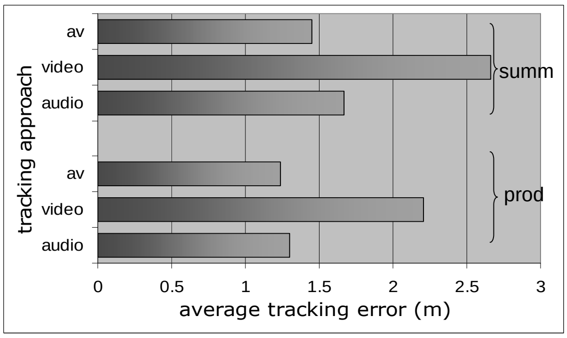

Fig. 7. Average tracking errors for all estimated tracks

by an average of 0.26 m and 1.04 m for audio and video approach respectively. The AV approach improved accuracy for 57% and 75% of the tracks for audio and video approach respectively. For rest of the tracks, error was either increased or remained same. Similar results are presented for product fusion in table III. In general, AV approach improved either on audio or video or both approaches.

Audio and video modalities are able to track vehicles successfully, though they suffer from poor performance in presence of high background noise and clutter. In

		Average error (m)	Tracks estimated
Fusion	Summ	1.93	74%
	Prod	1.58	59%
Modality	Audio	1.49	89%
	Video	2.44	50%
	AV	1.34	61%

TABLE II

AVERAGE TRACKING ERRORS AND TRACKS ESTIMATED

	Summ		Prod	
	Average error reduction (m)	Tracks improved	Average error reduction (m)	Tracks improved
Audio	0.26	57%	0.14	100%
Video	1.04	75%	0.90	67%

TABLE III

AVERAGE REDUCTION IN TRACKING ERROR FOR AV OVER AUDIO AND VIDEO-ONLY FOR ALL ESTIMATED TRACKS

general, audio sensors are able to track vehicles with good accuracy, but they suffer from high uncertainty and poor sensing range. As expected, fusing the two modalities gives better performance. There are some cases where audio tracking performance is better than fusion. This is because of the poor performance of video tracking. Video cameras were placed at an angle along the road to maximize coverage of the road. This makes video tracking very sensitive to camera calibration errors and camera placement. Also, an occasional obstruction in front of a camera confused the tracking algorithm which took a while to recover. We plan to use automatic camera calibration and more robust camera placements in our next set of experiments to improve video tracking performance.

Fusion based on product of likelihood functions gives better performance but it is more vulnerable to sensor conflict and errors in sensor calibration, etc. The weighted-sum approach is more robust to conflicts and sensor errors, but it suffers from high uncertainty. The average tracking error of 2 meters is reasonable considering the fact that a vehicle is not a point source, and the cell size used in fusion is 0.5 meters.

B. Multiple Targets

Many tracks with multiple moving vehicles in the sensing region were recorded during the experiment. Most of them had vehicles moving in the same direction. Only a few tracks included multiple vehicles crossing each other. Figure 8 shows the multiple target tracking result for three vehicles where two of them were crossing each other. Fig. 8(a) shows the three tracks with the ground truth, while Fig. 8(b) shows the x-coordinate of the tracks with time. The average tracking errors for the three tracks were 1.29m, 1.60m and 2.20m. Fig. 8 shows the result when only video data from three video sensors was used. Multiple target tracking with audio data could not distinguish between targets when they were crossing each

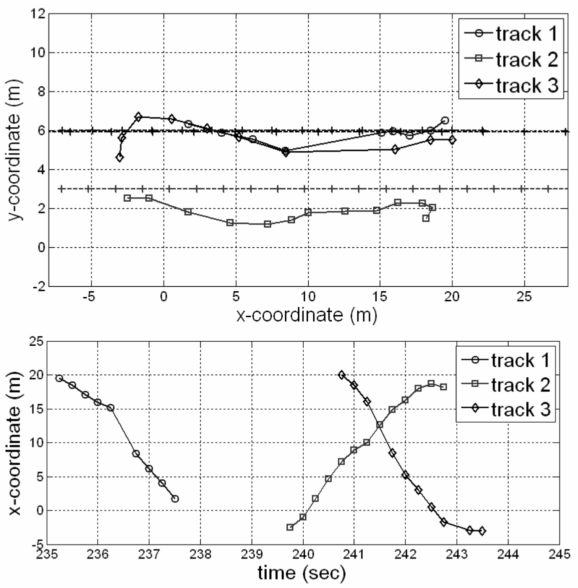

Fig. 8. Multiple Target Tracking (a) XY plot (b) X-Coordinate with time

other. This is due to the fact that beamforming is done assuming acoustic signals are generated from a single source. Acoustic beamforming methods exist for detecting and estimating multiple targets [5]. We plan to use those methods in our next set of experiments.

VI. Conclusions

We have developed a multimodal tracking system using an HSN consisting of six mote audio nodes and 3 PC camera nodes. Our system employs Markov Chain Monte Carlo Data Association framework for tracking multiple targets based on fused measurements from audio beamforming and video motion detection. Time synchronization across the HSN allows the fusion of the sensor measurements. We have deployed the HSN and evaluated the performance by tracking moving vehicles in an uncontrolled urban environment. Fusion of audio and video measurements can improve the tracking performance. The main direction of our future work is to improve robustness of the tracking system. An important challenge toward this direction is addressing sensor conflict that can degrade the performance of any fusion method and needs to be carefully considered. Scalability is also an important aspect that has to be addressed, and we plan to expand our HSN using additional mote class devices equipped with cameras.

References

[1] Y. Bar-Shalom and T. Fortmann. Tracking and data association. In *Mathematics in Science and Engineering Series 179 Academic Press*, 1988.

[2] V. Cevher, A. Sankaranarayanan, J. H. McClellan, and R. Chellappa. Target tracking using a joint acoustic video system. In *IEEE Transactions on Multimedia*, June 2007.

[3] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell. A probabilistic framework for multi-modal multi-person tracking. In *IEEE Workshop on Multi-Object Tracking*, 2003.

[4] J. Chen, L. Yip, J. Elson, H. Wang, D. Maniezzo, R. Hudson, K. Yao, and D. Estrin. Coherent acoustic array processing and localization on wireless sensor networks. In *Proceedings of the IEEE*, volume 91, pages 1154–1162, August 2003.

[5] J. C. Chen, K. Yao, and R. E. Hudson. Acoustic source localization and beamforming: theory and practice. In *EURASIP Journal on Applied Signal Processing*, pages 359–370, April 2003.

[6] I. Cox. A review of statistical data association techniques for motion correspondence. *International Journal of Computer Vision*, 10(1):53–66, 1993.

[7] G. Cybenko, V. Berk, V. Crespi, R. Gray, and G. Jiang. An overview of process query systems. In *Proc. of SPIE Vol. 5403, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense III*, Orlando, FL, April 2004.

[8] J. Elson, L. Girod, and D. Estrin. Fine-grained network time synchronization using reference broadcasts. In *Operating Systems Design and Implementation (OSDI)*, 2002.

[9] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *Conference on Uncertainty in Artificial Intelligence*, 1997.

[10] S. Ganeriwal, R. Kumar, and M. B. Srivastava. Timing-sync protocol for sensor networks. In *ACM SenSys*, 2003.

[11] P. KaewTraKulPong and R. B. Jeremy. An improved adaptive background mixture model for realtime tracking with shadow detection. In *Workshop on Advanced Video Based Surveillance Systems (AVBS)*, 2001.

[12] S. Karimi-Ashtiani and C. C. J. Kuo. Automatic real-time moving target detection from infrared video. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP'06)*, 2006.

[13] Z. Khan, T. Balch, and F. Dellaert. MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1805–1918, Nov. 2005.

[14] B. Kusy, P. Dutta, P. Levis, M. Maroti, A. Ledeczi, and D. Culler. Elapsed time on arrival: A simple and versatile primitive for time synchronization services. *International Journal of Ad hoc and Ubiquitous Computing*, 2(1), January 2006.

[15] A. Ledeczi, A. Nadas, P. Volgyesi, G. Balogh, B. Kusy, J. Sallai, G. Pap, S. Dora, K. Molnar, M. Maroti, and G. Simon. Countersniper system for urban warfare. *ACM Trans. Sensor Networks*, 1(2), 2005.

[16] J. Liu, J. Reich, and F. Zhao. Collaborative in-network processing for target tracking. In *EURASIP, Journal on Applied Signal Processing*, 2002.

[17] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for general multiple-target tracking problems. In *CDC*, 2004.

[18] S. Oh, L. Schenato, P. Chen, and S. Sastry. Tracking and coordination of multiple agents using sensor networks: System design, algorithms and experiments. *Proceedings of the IEEE*, 95(1):234–254, January 2007.

[19] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, 2004.

[20] A. Poore. Multidimensional assignment and multitarget tracking. In I. J. Cox, P. Hansen, and B. Julesz, editors, *Partitioning Data Sets*, pages 169–196. American Mathematical Society, 1995.

[21] D. Reid. An algorithm for tracking multiple targets. *IEEE Trans. Automatic Control*, 24(6):843–854, December 1979.

[22] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[23] N. Strobel, S. Spors, and R. Rabenstein. Joint audio video object localization and tracking. In *IEEE Signal Processing Magazine*, 2001.

[24] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *IEEE International Conference on Computer Vision*, 1999.

[25] M. Yarvis, N. Kushalnagar, H. Singh, A. Rangarajan, Y. Liu, and S. Singh. Exploiting heterogeneity in sensor networks. In *IEEE INFOCOM*, 2005.

[26] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.