Target Tracking in Urban Environments using Audio-Video Signal Processing in Heterogeneous Wireless Sensor Networks

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

Institute for Software Integrated Systems
Vanderbilt University
Nashville, TN, USA
manish.kushwaha@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. In this paper, we describe an approach for target tracking in urban environments utilizing a wireless HSN of mote class devices equipped with acoustic sensor boards and embedded PCs equipped with web cameras. Our system uses acoustic beamforming and motion detection for audio and video sensors, respectively. We also employ MCMCDA algorithm for data association and tracking. 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 [26]. 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 [2], computer vision [7], network and computer security [8], and sensor networks [19]. 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 [18] 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 acoustic beamforming and its application for localization in sensor networks can be found in [6]. Beamforming methods have successfully been applied to detect single or even multiple acoustic sources in noisy and reverberant environments [5], [16]. Adaptive background-modeling methods for motion detection based on video include the work in [10], which modeled each pixel in a camera scene by an adaptive parametric mixture model of three Gaussian distributions and the adaptive nonparametric Gaussian mixture model to address background modeling challenges presented in [23]. Other techniques using high-level processing to assist the background modeling also have been proposed [12], [25]. Work in multimodal target tracking and multimodal sensor fusion using audio-video data includes object localization and tracking based on Kalman filtering [24] as well as particle filtering approaches [4], [3].

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. Multimodal sensor fusion is presented in section III. The multiple-target tracking algorithm is presented in section IV. Experimental setup 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 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.

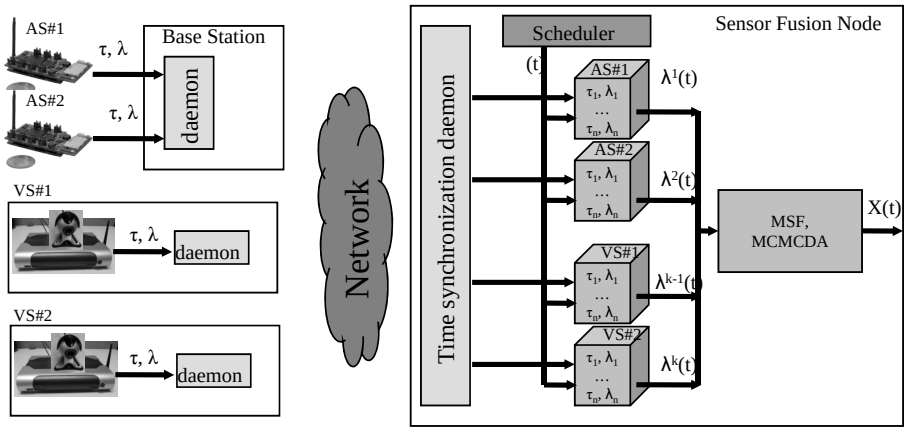

Fig. 1. Multimodal tracking system architecture

Audio Beamforming: Beamforming can be used to determine the direction(s) of arrival and the location(s) of acoustic source(s) [6], [16]. 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

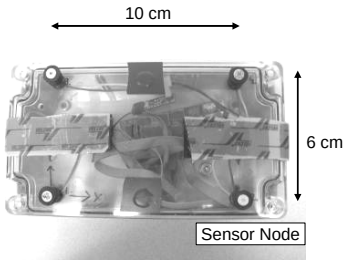

Fig. 2. Sensor Node Showing the Microphones

assumed to be on the same two-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. We use the motion detection algorithm using the background-foreground segmentation approach described in [12], which is based on an adaptive background mixture model and provides robust performance and low complexity in a wide range of situations. The dataflow in Figure 3 shows the motion detection algorithm and its components. 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

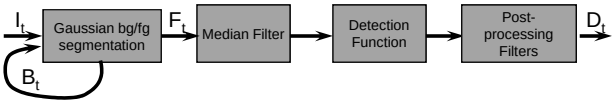

Fig. 3. Data-flow diagram of real-time motion detection algorithm

function is similar to the horizontal intensity accumulation function (IAF) defined in [13]. The top of Fig. 4 shows an example video detection function. In addition, 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 Logitech QuickCam Pro 4000 cameras attached to OpenBrick-E Linux embedded PCs. The motion detection algorithm is implemented using the OpenCV 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 comprise an ad-hoc 802.15.4 network while the video sensors, the mote-PC gateway, and the sensor fusion node form an ad-hoc 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. [9], [11]) but they cannot be applied directly to HSNs. To synchronize the entire network, we integrated existing protocols that provide high accuracy and low overhead for a specific network [1]. We used Elapsed Time on Arrival (ETA) [15] to synchronize the mote network and RBS [9] 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 sensor models and fusion algorithms for audio and video sensors. We use nonparametric sensor models for both the audio and video sensors.

Sensor Model: Let $\lambda(\theta)$ denote the detection function (i.e. acoustic beamform for audio and video detection function for video) the nonparametric DOA sensor model for a single 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 $w = (\theta_i - \theta)/(\theta_i - \theta_{i-1})$.

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 [17], 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 the sound source from the sensor. We extended the above likelihood function by incorporating energy in the empirical variance. The modified likelihood function for both audio and video sensor models can be expressed as

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

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 the sensor and the detection function value at the cell. Since the sensor models are nonlinear and nonparametric, it is reasonable to use a nonparametric representation for the likelihood functions which are represented as discrete grids in 2D space. Figure 4 shows an example video detection function and the corresponding likelihood function.

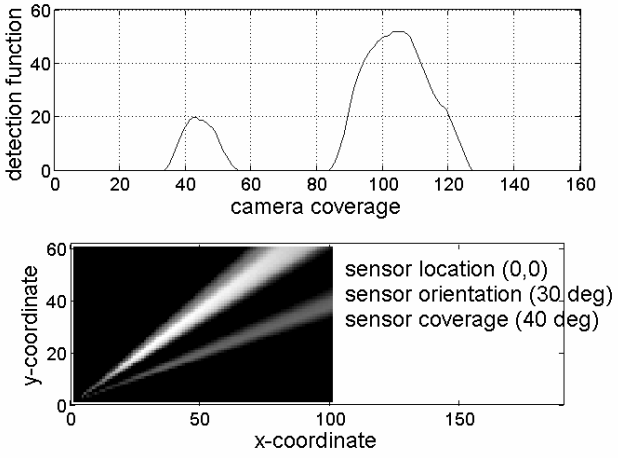

Fig. 4. An example video detection function and the corresponding likelihood function

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

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

or as the weighted-summation fusion

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

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 sensors, we compute target observations, which are used by the multitarget tracking and data association algorithm described in the 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 for 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 [21], 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 [20], [14]. Although particle filters are highly effective in single-target tracking, it is reported that they provide poor performance in multi-target tracking [14]. This is because a fixed number of particles is insufficient to represent the posterior distribution with exponentially increasing complexity (due to the data association problem). As shown in [14], 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 [14] or a detection algorithm [20]. But a single-scan approach

Number of beams in audio beamforming, M_{audio}	36
Number of angles in video detection M_{video}	160
Sensing region (meters)	25×20
Cell size (meters)	0.5×0.5

TABLE I

PARAMETERS USED IN EXPERIMENTAL SETUP

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) [22] 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 [18]. The simulation study in [18] 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 readers to [18].

V. EVALUATION

The deployment of the multi-modal target tracking system is shown in Figure 5. We employ 6 audio sensors and 6 video sensors deployed on either side of a road. The sensor network covers a small section of the road in front of a building shown at the bottom of the figure. It also covers an entrance/exit to a multi-story garage shown at the top of the figure by the two-sided arrow. Typically the targets (i.e, vehicles) move in the direction given by the arrows. 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 automatically 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 place the sensor nodes at marked locations. The camera orientations are manually calibrated by using known landmarks in the camera field-of-view. The audio sensors are placed on 1 meter high tripods to minimize audio clutter near the ground.

Table I presents the parameter values that we use in our tracking system. Sensor likelihood functions are calculated by discretizing the sensing region into the specified cell-sized grid. The tracked vehicles were part of an uncontrolled experiment. The vehicles were traveling on road at 15-30 mph speed.

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.

Figure 6 shows the target tracking result for four different representative vehicle tracks. Tracks 1 and 2 show vehicles going into and out of the garage. Tracks 3 and 4 show vehicles going straight on the road.

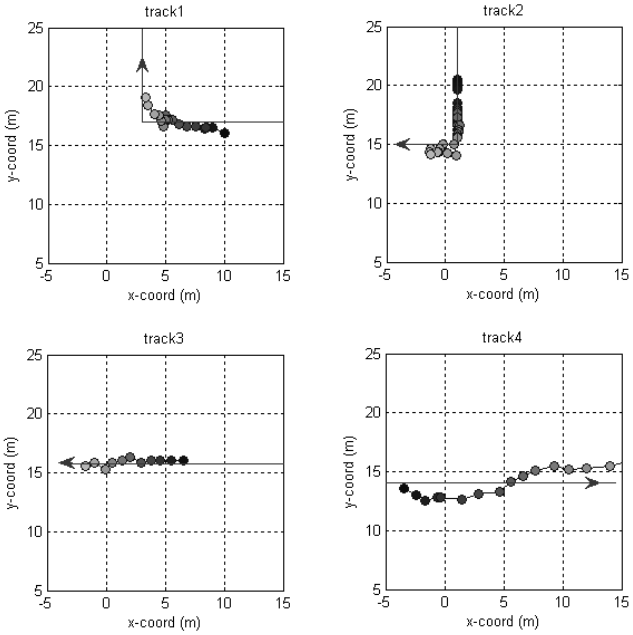

Fig. 6. Target Tracking

An average target tracking error based on ground truth video and estimated target tracks is estimated to be 2 meters. 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. We should note that the entire target tracking system works online in real-time at 4Hz. The average latency for detection is 4 seconds.

VI. CONCLUSIONS

We have developed a multimodal tracking system using an HSN consisting of six mote audio nodes and six PC camera nodes. Our system employs an MCMCDA framework for tracking multiple targets based on fused measurements from audio beamforming and video motion detection. Time synchronization across the HSN enables 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

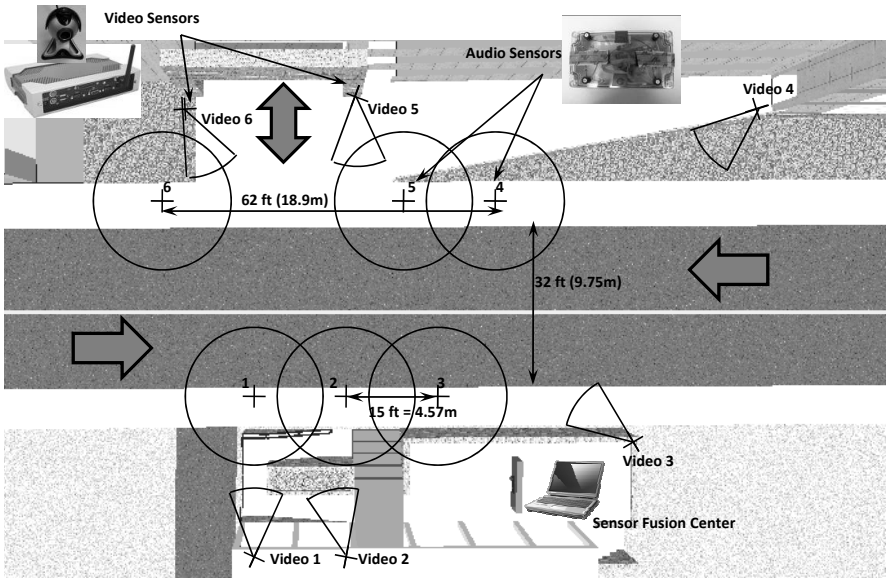

Fig. 5. Experimental setup

expand our HSN using additional mote class devices equipped with cameras.

VII. Acknowledgments

This work is partially supported by ARO MURI W911NF-06-1-0076.

References

[1] I. Amundson, B. Kusy, P. Volgyesi, X. Koutsoukos, and A. Ledeczi. Time synchronization in heterogeneous sensor networks. In *International Conference on Distributed Computing in Sensor Systems (DCOSS'08)*, 2008.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.

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

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

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

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

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