Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 591340, 19 pages
doi:10.1155/2010/591340

Research Article

Acoustic Sensor-Based Multiple Object Tracking with Visual Information Association

Jinseok Lee,[1] Sangjin Hong,[1] Nammee Moon,[2] and Seong-Jun Oh[3]

[1] *Department of Electrical and Computer Engineering, Stony Brook University—SUNY, Stony Brook, NY 11794-2350, USA*
[2] *Hoseo Graduate School of Venture, Hoseo University, Seoul 137-867, Republic of Korea*
[3] *College of Information and Communications, Korea University, Seoul 136-701, Republic of Korea*

Correspondence should be addressed to Seong-Jun Oh, seongjun@korea.ac.kr

Received 6 February 2010; Revised 31 May 2010; Accepted 27 July 2010

Academic Editor: Walter Kellermann

Object tracking by an acoustic sensor based on particle filtering is extended for the tracking of multiple objects. In order to overcome the inherent limitation of the acoustic sensor for the simultaneous multiple object tracking, support from the visual sensor is considered. Cooperation from the visual sensor, however, is better to be minimized, as the visual sensor's operation requires much higher computational resources than the acoustic sensor-based estimation, especially when the visual sensor is not dedicated to object tracking and deployed for other applications. The acoustic sensor mainly tracks multiple objects, and the visual sensor supports the tracking task only when the acoustic sensor has a difficulty. Several techniques based on particle filtering are used for multiple object tracking by the acoustic sensor, and the limitations of the acoustic sensor are discussed to identify the need for the visual sensor cooperation. Performance of the triggering-based cooperation by the two visual sensors is evaluated and compared with a periodic cooperation in a real environment.

1. Introduction

Tracking multiple objects has been of great interest in numerous surveillance-required areas applied in diverse fields such as military, industry, medical, and mining fields [1, 2]. Among a variety of sensors deployed in a surveillance system, an acoustic sensor is widely used since it allows easy and quick deployment with a less computational complexity as well as a broad sampling range [3, 4]. Acoustic sensor-based object tracking is widely studied with several approaches. A time-delay estimation method aims at measuring the time delays of arrival signals at receivers [5]. A beamforming method uses a frequency-averaged output power of a steered beamformer [6]. A bearings-only tracking method aims at estimating position, velocity, and possibly some extra features by measuring the angles of the objects [7]. Using particle filtering's state-space approach, the object localization from an acoustic sensor is studied in [8], where the problem of multipath reflection of the acoustic signal is considered. Throughout this paper, we use a micropower gradient flow acoustic localizer, where four microphones measure interaural time differences of an object [3].

Despite the easy deployment of the acoustic sensor, there are several difficult issues when one acoustic sensor tracks multiple objects. Multiple objects and multiple measurements are randomly and inconsistently mapped especially when an acoustic sensor receives the bearing estimates with negligibly small difference [9]. In addition, the number of measurements is varying when the objects do not transmit sound wave, new objects come into an acoustic sensing range, or objects move out an acoustic sensing range. The varying number of measurements gives inconsistent measurement sequences to the acoustic sensor-based estimator [10]. Furthermore, when measurements are corrupted by the noise or the dynamic models are incorrect, the estimation performance is more severely degraded for the case of the multiple object tracking.

In order to overcome the limitations of the acoustic sensor-based estimation for multiple objects' tracking, the visual sensor-based estimation can be combined [11–15]. In [11], the visual sensor mainly tracks the objects, and the acoustic sensor partially supports the estimation when the tracked objects are occluded. This method is experimentally shown in a video conferencing environment. In the problem

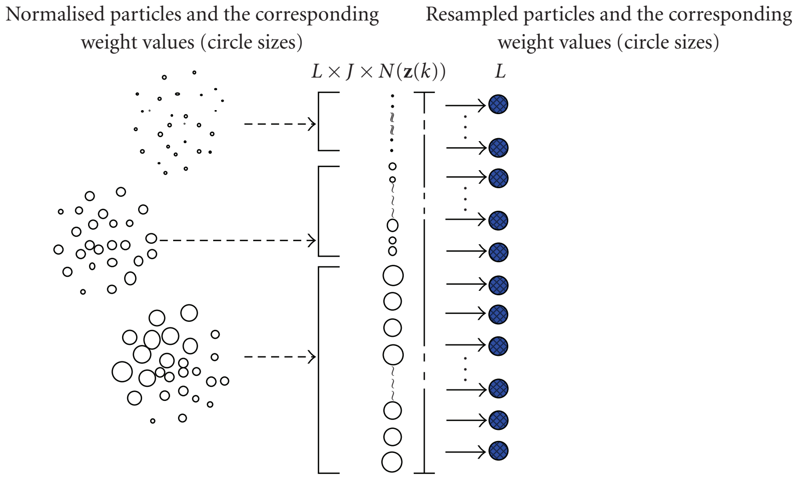

Normalised particles and the corresponding weight values (circle sizes)
Resampled particles and the corresponding weight values (circle sizes)
$L \times J \times N(\mathbf{z}(k))$
L

Figure 1: Resampling of $L \times J \times N(\mathbf{z}(k))$ particles to L particles.

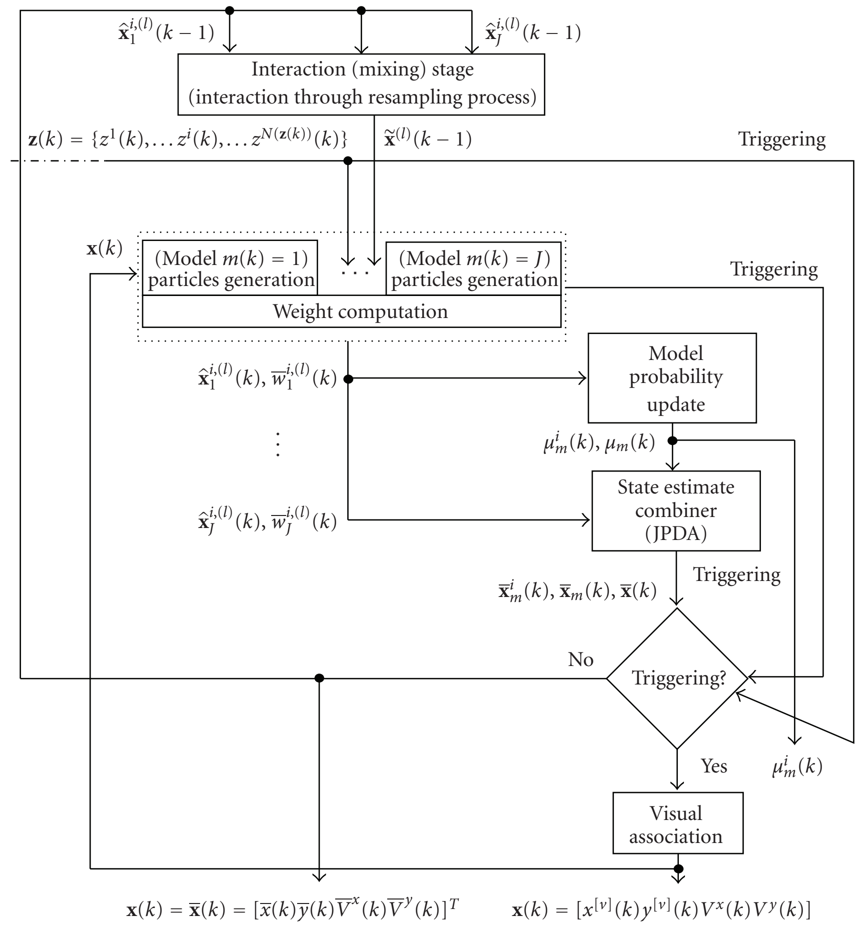
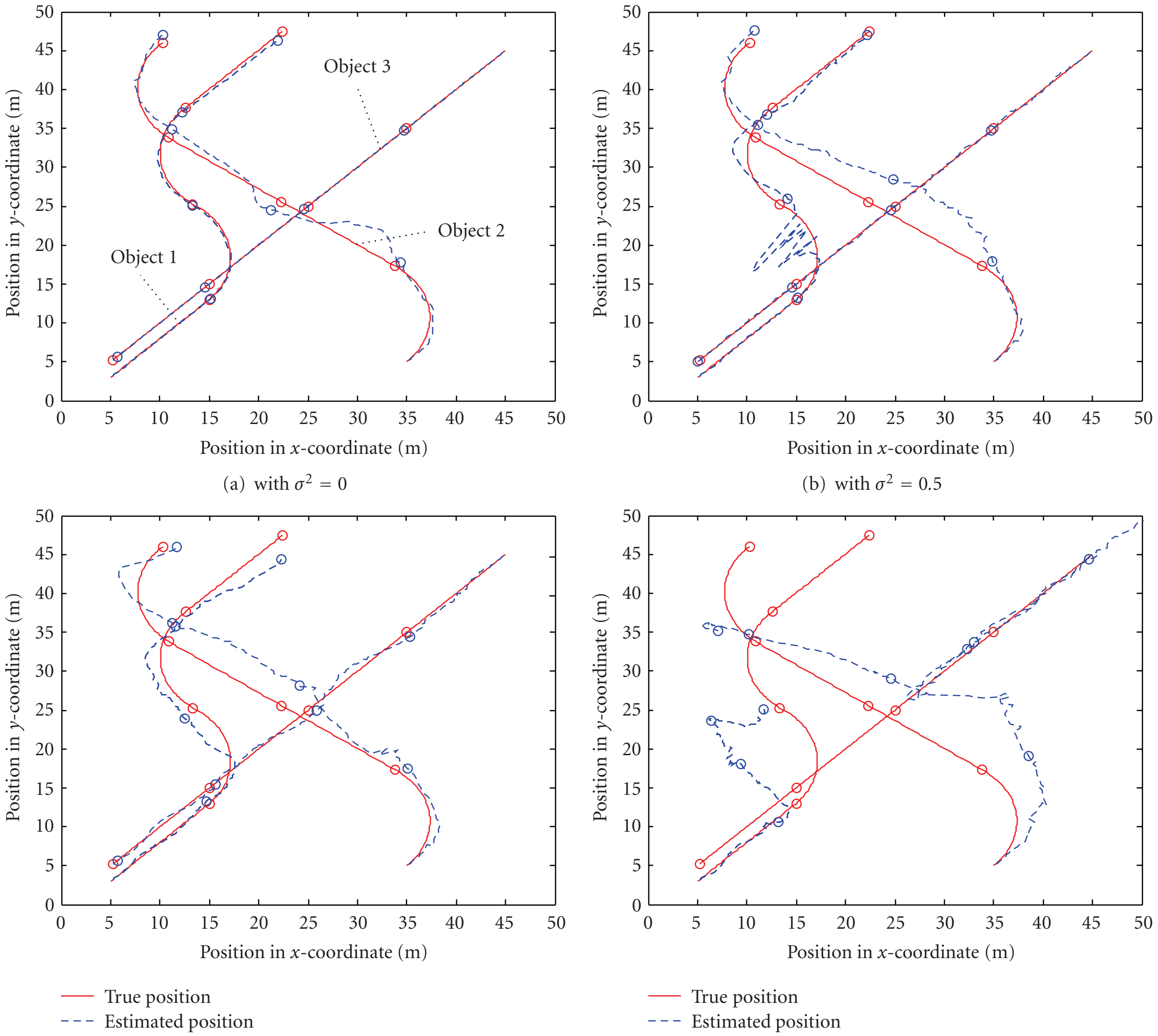

$\hat{\mathbf{x}}_1^{i,(l)}(k-1)$
$\hat{\mathbf{x}}_J^{i,(l)}(k-1)$
Interaction (mixing) stage
(interaction through resampling process)
$\mathbf{z}(k) = \{z^1(k), \dots z^i(k), \dots z^{N(\mathbf{z}(k))}(k)\}$
$\tilde{\mathbf{x}}^{(l)}(k-1)$
Triggering
$\mathbf{x}(k)$
(Model $m(k) = 1$) particles generation
. . .
(Model $m(k) = J$) particles generation
Triggering
Weight computation
$\hat{\mathbf{x}}_1^{i,(l)}(k), \overline{w}_1^{i,(l)}(k)$
Model probability update
$\mu_m^i(k), \mu_m(k)$
$\hat{\mathbf{x}}_J^{i,(l)}(k), \overline{w}_J^{i,(l)}(k)$
State estimate combiner (JPDA)
Triggering
$\overline{\mathbf{x}}_m^i(k), \overline{\mathbf{x}}_m(k), \overline{\mathbf{x}}(k)$
No
Triggering?
Yes
$\mu_m^i(k)$
Visual association
$\mathbf{x}(k) = \overline{\mathbf{x}}(k) = [\overline{x}(k)\overline{y}(k)\overline{V}^x(k)\overline{V}^y(k)]^T$
$\mathbf{x}(k) = [x^{[v]}(k)y^{[v]}(k)V^x(k)V^y(k)]$

Figure 2: IMM-PF data flow and the visual sensor cooperation.

of identifying the speaker inside a cluttered meeting room, audiovisual information from multiple acoustic and video sensors are combined in [12], where it is shown that the audiovisual multimodal framework outperforms the audio-only system in most scenarios. In [13], the acoustic-visual combining method is presented with the iterative decoding algorithm from the theory of turbo codes and factor graphs. This method computes the likelihood values both from the acoustic sensor and the visual sensor, and the one with a higher likelihood is selected for a more accurate estimation. In [14, 15], data from acoustic and visual sensors are simultaneously combined. In [14], a way



(a) with $\sigma^2 = 0$

(b) with $\sigma^2 = 0.5$

(c) with $\sigma^2 = 1.5$

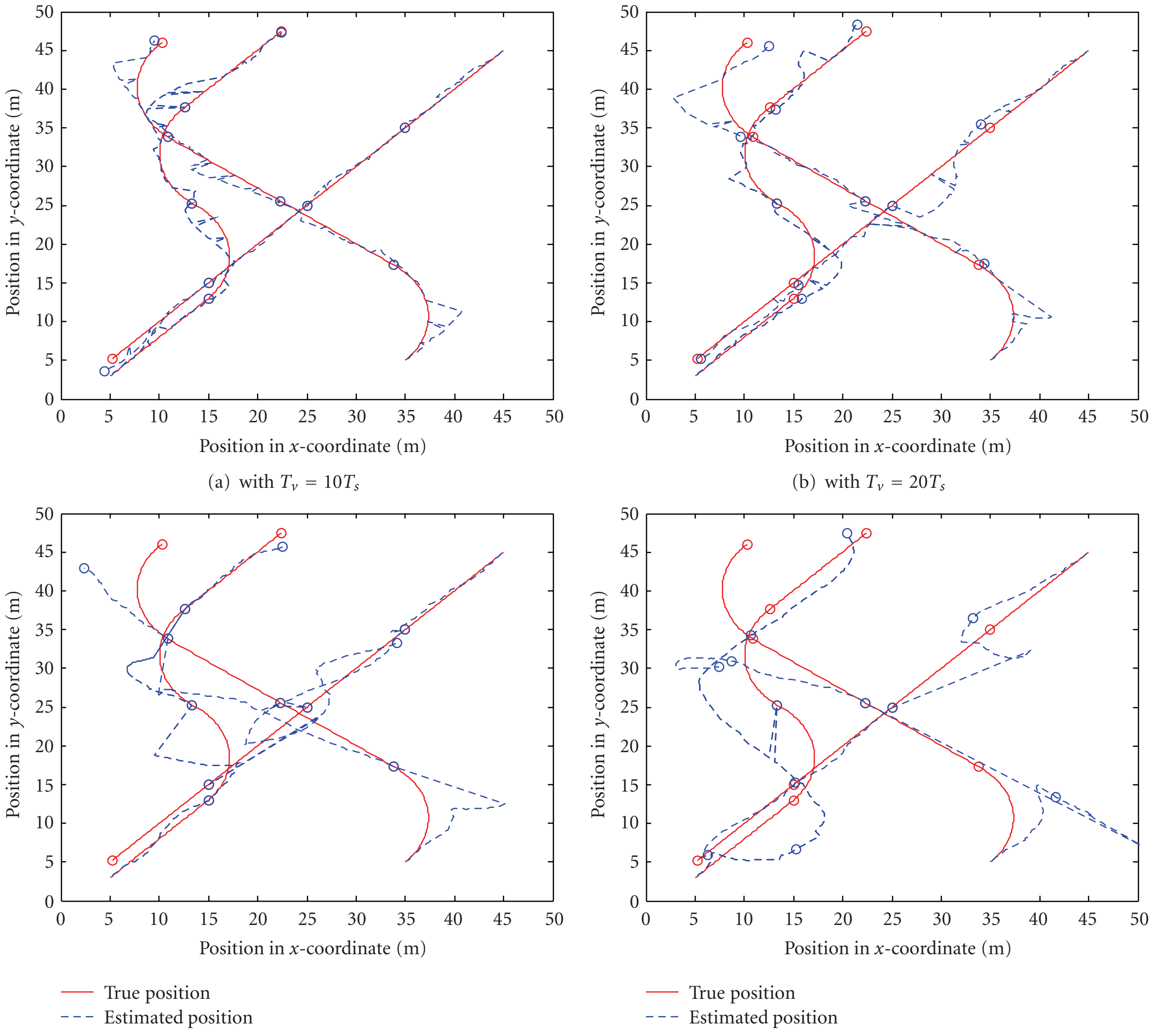
(d) with $\sigma^2 = 5.0$

Figure 3: The estimation only with an acoustic sensor for various measurement noise variances σ^2: 0, 0.5, 1.5, and 5.0.

of jointly processing different sources of information is presented using cooperative Hidden Markov Models (HMMs) with appearance models whereas in [15], a particle filter tracker is applied for both acoustic and video observation regarding the overlapped state-space models. Our interest is to minimize the resources from visual sensor since the visual sensor-based object localization requires much higher computational complexity [16, 17], and the visual sensor is assumed to be deployed for other purposes, so the visual sensor cannot dedicate its operation to support one acoustic sensor. A similar joint tracking can be found in [18, 19] with a specific application in mind. A large number of sensors are used in a heterogeneous sensor network to cover a large area in [18] whereas the concert hall application is considered in [19]. We take the approach where an acoustic sensor mainly tracks the multiple objects and the visual sensor cooperation is performed only when the acoustic sensor has a difficulty. Therefore, the cooperation is triggered by the acoustic sensor.

The acoustic sensor-based estimation is performed with bearings-only tracking developed by the sequential Monte Carlo methods known as the particle filter. In the fields of wireless communications, navigation systems, sonar, and robotics applications, the particle filtering is adopted as an emerging powerful tool for solving nonlinear and non-Gaussian problems [20–23]. The particle filters are generally used for an estimation and/or a detection of dynamic system parameters or states in real-time application. While the particle filter with an acoustic sensor tracks multiple objects, the visual sensor detects the objects and localizes their positions when the acoustic sensor triggers for the visual sensor cooperation.

The remainder of the paper is organized as follows. In Section 2, the object tracking by the acoustic sensor with



(a) with $T_v = 10T_s$

(b) with $T_v = 20T_s$

(c) with $T_v = 50T_s$

(d) with $T_v = 100T_s$

FIGURE 4: Visual sensor cooperation performance for periodic cooperations with T_v: $10T_s$, $30T_s$, $50T_s$, and $100T_s$ based on the result with measurement variance 5.0 in Figure 3(d) (500 particles are used in the simulation).

the multimodel and multimeasurement particle filtering is introduced as a background. In Section 3, several issues in tracking multiple objects based on an acoustic sensor are presented, and the corresponding triggering conditions are proposed for the visual sensor cooperation. In addition, the performance of the proposed visual sensor cooperation is compared with a periodic visual sensor cooperation. In Section 4, we verify the visual sensor cooperation with real data through the experiment. Our contribution is summarized and the final remarks are given in Section 5.

2. Background

2.1. Object Tracking with Multimodel and Multimeasurement. The acoustic sensor's object tracking is performed with bearings-only measurements. A bearings-only tracking is to estimate object positions and velocities with a sequence of noisy bearing measurements [7, 24]. For an object of interest, its state at discrete time k, $k \in \{1, 2, \ldots\}$, is described by

$$\mathbf{x}(k) = \mathbf{F}^{(m(k))}\mathbf{x}(k-1) + \mathbf{w}(k-1), \tag{1}$$

$$z(k) = H(\mathbf{x}(k)) + v(k), \tag{2}$$

where $\mathbf{x}(k)$ denotes the state vector of the object as $[x(k)y(k)V^x(k)V^y(k)]^T$ and $z(k)$ is the corresponding bearing measurement for the object. $[x(k), y(k)]$ is the 2-dimensional location of the object at time k, and $[V^x(k)V^y(k)]$ is the x- and y-directional velocity of the object at time k. $H(\mathbf{x})$ is the bearing measurement function for state vector $\mathbf{x}$ as $H(\mathbf{x}(k)) = \arctan((y(k))/(x(k)))$. The noise random process $\mathbf{w}(k-1)$ and measurement noise $v(k)$ are modeled as zero-mean independent Gaussian.

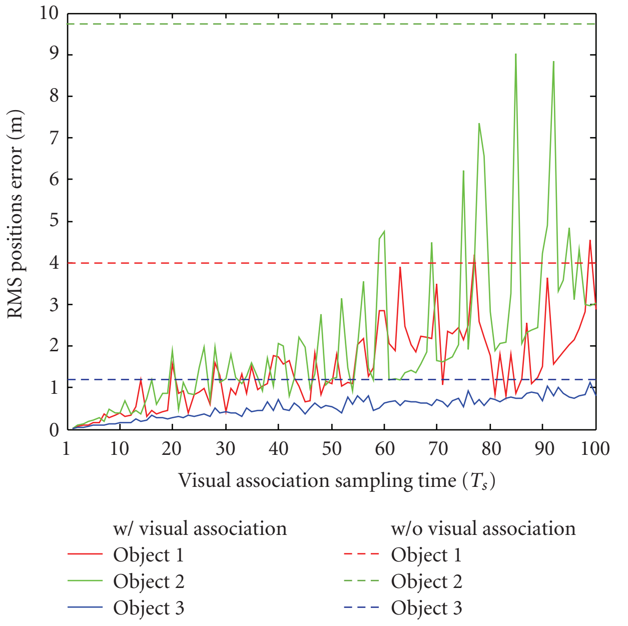


Figure 5: RMS position error for periodic cooperation with T_v: $1T_s$ to $100T_s$. (500 particles are used in the simulation).

$\mathbf{F}^{(m)}$ is the 4×4 state-transition matrix for model m, $m \in \{1, 2, \ldots, J\}$, where J is the number of the hypothesized models [1, 25] and $m(k)$ is the model index at time k for the object in tracking. The model plays an important role to estimate an object state by representing complicated object motion with mathematical expression. Various mathematical models of object motion have been developed for both practitioners and researchers in the tracking community [26]. In this paper, we adopted constant velocity model, clockwise coordinated turn model, and anticlockwise coordinated turn model. For the object of interest, the model switching is governed by a finite-state Markov chain according to the switching probabilities $\text{Prob}[m(k) = v \mid m(k-1) = u]$ of switching from model u to v, $u, v \in \{1, 2, \ldots, J\}$. Note that this switching probabilities are not needed in the following estimation. As there are multiple bearing measurements, let $\mathbf{z}(k)$ denote a set of measurements as $\{z^1(k), z^2(k), \ldots, z^{N(\mathbf{z}(k))}(k)\}$, where $z^i(k)$ is the ith measurement and $N(\mathbf{z}(k))$ is the number of bearing measurements at time k. Also, define $\mathbf{z}^i(1 : k)$ as the set of measurements up to and including time k as $\{z^i(1), z^i(2), \ldots, z^i(k)\}$, where $i = 1, 2, \ldots, N(\mathbf{z}(k))$. Note that as the unlabeled measurements are received by an acoustic sensor, it is not known which measurement index corresponds to the object of interest. Furthermore, the correspondences between the objects and the measurements are not consistent—the relationship changes over time.

The goal of object tracking is to estimate the state of the object and the probability that the object's model index is m at time k for the given history of observations. More specifically, based on the particle filtering, the following items are estimated:

(i) conditional probability density function (pdf) of the object's state $\mathbf{x}(k)$ at time k given the history of observation up to time k; $p(\mathbf{x}(k) \mid \mathbf{z}(1 : k))$,

(ii) conditional expected state when the model index is m at time k; $\overline{\mathbf{x}}_m(k)$,

(iii) unconditional probability that the object's model index is m at time k; $\mu_m(k)$,

where $m \in \{1, 2, \ldots, J\}$ and $\sum_{m=1}^{J} \mu_m(k) = 1$. Conditional expected means and the probabilities are not directly used for the object tracking but they are used to trigger the visual sensor cooperation discussed in the next section. As we use the particle filtering technique for the state estimation, the conditional pdf is estimated with many particles in the state space, where each particle is of equal conditional probability density through the sequential importance resampling (SIR) algorithm [27]. L, $L \gg 1$, particles are updated for every new observation, and the estimation is done as follows. L resampled particles are given, and they represent the conditional pdf, $p(\mathbf{x}(k-1) \mid \mathbf{z}(1 : k-1))$. Then, there is a set of new $N(\mathbf{z}(k))$ measurements $\{z^1(k), z^2(k), \ldots, z^{N(\mathbf{z}(k))}(k)\}$. From these measurements and the given L particles, we want to obtain

(i) L resampled particles representing $p(\mathbf{x}(k) \mid \mathbf{z}(1 : k))$,

(ii) conditional mean vector, $\overline{\mathbf{x}}_m(k)$ and the unconditional probabilities of the object's model, $\mu_m(k)$, where $m \in \{1, 2, \ldots, J\}$, then eventually the mean vector estimate $\overline{\mathbf{x}}(k)$ as the weighted sum.

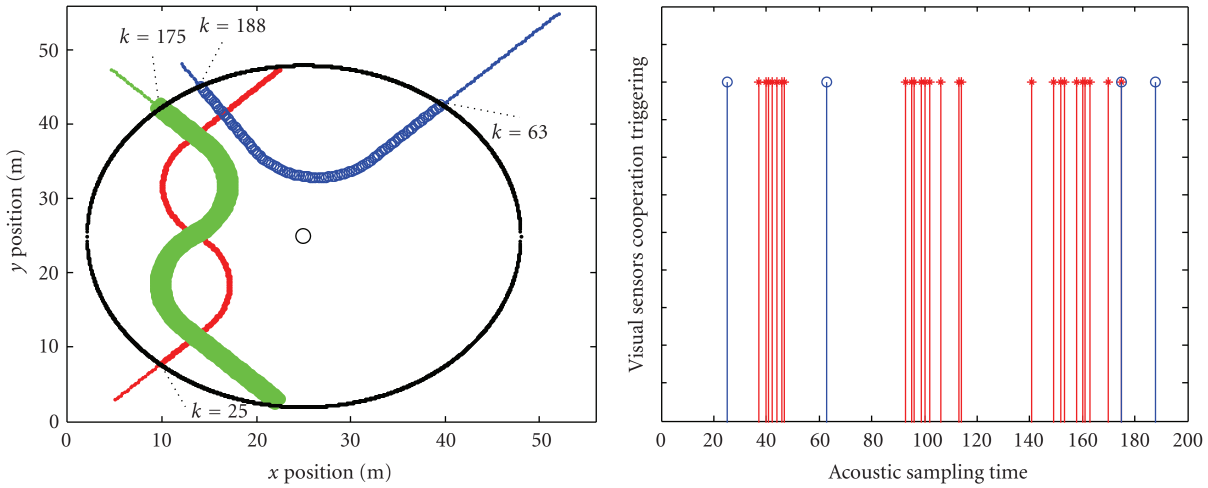
2.2. Multiple Model Particle Filter with Visual Sensor Cooperation. The state estimation is done by the interacting multiple model particle filter (IMM-PF) framework [28]. The IMM estimator is a state-estimation algorithm for a system represented by Markovian switching model with multiple model indices. In the particle filtering stage at time k, $L \times J$, particles $\hat{\mathbf{x}}_m^{(l)}(k)$, for $l \in \{1, 2, \ldots, L\}$ and $m \in \{1, 2, \ldots, J\}$, are drawn from the previous a posteriori density function $p(\mathbf{x}(k-1) \mid \mathbf{z}(1 : k-1))$ for each model m as follows:

$$\hat{\mathbf{x}}_m^{(l)}(k) = \mathbf{F}^{(m)}\tilde{\mathbf{x}}^{(l)}(k-1) + \mathbf{n}_m^{(l)}(k), \quad \text{for } l \in \{1, 2, \ldots, L\}, \quad m \in \{1, 2, \ldots, J\}, \tag{3}$$

where $\tilde{\mathbf{x}}^{(l)}(k-1)$ is the resampled particles at time $k-1$ and $\mathbf{n}_m^{(l)}(k)$'s are identically distributed independent Gaussian zero-mean noise. The predicted bearing measurements to particles $\hat{\mathbf{x}}_m^{(l)}(k)$'s are obtained as

$$\hat{z}_m^{(l)}(k \mid k-1) = H\left(\hat{\mathbf{x}}_m^{(l)}(k)\right) = \arctan\left(\frac{\hat{y}_m^{(l)}(k)}{\hat{x}_m^{(l)}(k)}\right), \tag{4}$$

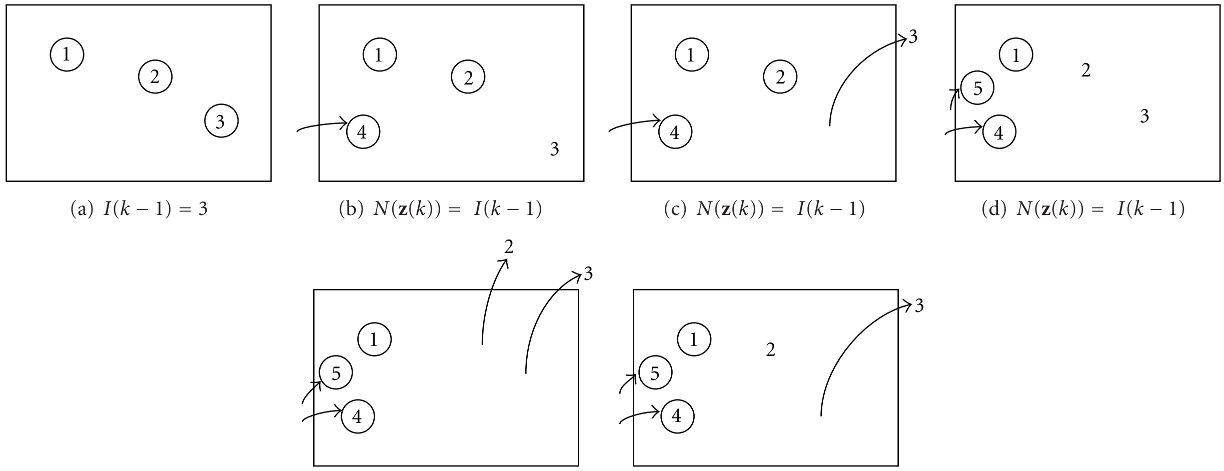
for $l \in \{1, 2, \ldots, L\}$ and $m \in \{1, 2, \ldots, J\}$, where $(\hat{y}_m^{(l)}(k), \hat{x}_m^{(l)}(k))$ is the lth particle's 2-dimensional position of $\hat{\mathbf{x}}_m^{(l)}(k)$ with model m. Note that there are $L \times J$ predicted measurements for the object of interest. These $L \times J$ predicted



(a) The 23-meter radius acoustic sensing range and object trajectories

(b) Triggering timings where "o" is by the varying number of objects and "$*$" is by the measurements resolution problem

Figure 6: The triggering timings based on system dynamics.



(a) $I(k-1) = 3$

(b) $N(\mathbf{z}(k)) = I(k-1)$

(c) $N(\mathbf{z}(k)) = I(k-1)$

(d) $N(\mathbf{z}(k)) = I(k-1)$

(e) $N(\mathbf{z}(k)) = I(k-1)$

(f) $N(\mathbf{z}(k)) = I(k-1)$

Figure 7: Examples when the triggering based on (14) does not work.

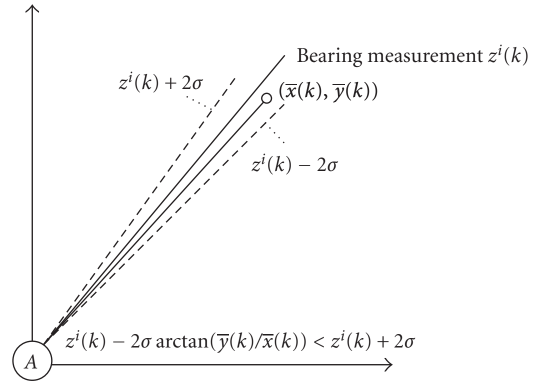

Bearing measurement $z^i(k)$
$z^i(k) + 2\sigma$
$(\overline{x}(k), \overline{y}(k))$
$z^i(k) - 2\sigma$
$z^i(k) - 2\sigma \arctan(\overline{y}(k)/\overline{x}(k)) < z^i(k) + 2\sigma$
A

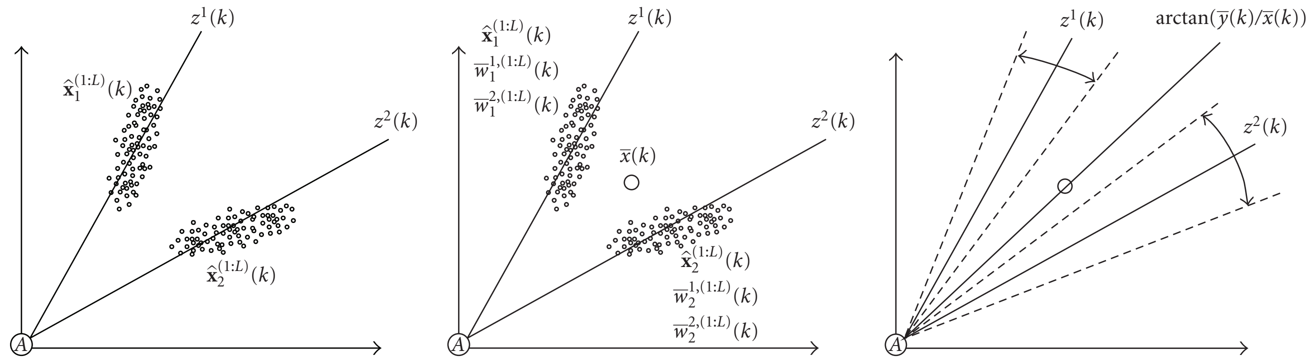
Figure 8: 95% confidence true bearing ranges-based triggering method.

measurements lead to the weight evaluation from the set of actual measurements $\mathbf{z}(k)$

$$\overline{w}_m^{i,(l)}(k) = d\left(z^i(k) - \hat{z}_m^{(l)}(k \mid k-1)\right), \qquad (5)$$

for $i \in \{1, 2, \ldots, N(\mathbf{z}(k))\}$, $l \in \{1, 2, \ldots, L\}$, and $m \in \{1, 2, \ldots, J\}$, where $d(\cdot)$ is the particle weight evaluation function from the Gaussian probability density function [23, 29]. Since each particle $\hat{\mathbf{x}}_m^{(l)}(k)$ is assigned with $N(\mathbf{z}(k))$ weights, there are $L \times J \times N(\mathbf{z}(k))$ weights. $\overline{w}_m^{i,(l)}(k)$ denotes the (unnormalized) weight of the lth particle in model m for given measurement $z^i(k)$. These $L \times J \times N(\mathbf{z}(k))$ weights are normalized as follows:

$$w_m^{i,(l)}(k) \frac{\overline{w}_m^{i,(l)}(k)}{\sum_{i'=1}^{N(\mathbf{z}(k))} \sum_{m'=1}^{J} \sum_{l'=1}^{L} \overline{w}_{m'}^{i',(l')}(k)}, \qquad (6)$$

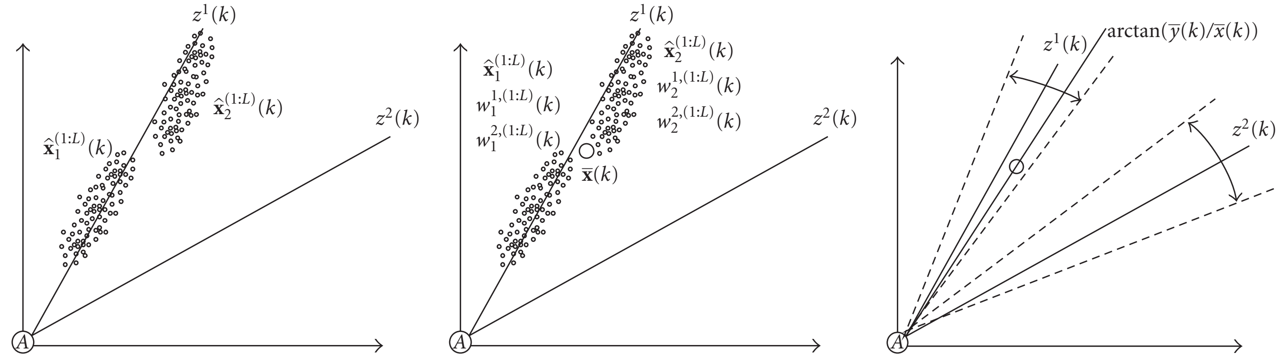

$z^1(k)$
$\hat{\mathbf{x}}_1^{(1:L)}(k)$
$z^2(k)$
$\hat{\mathbf{x}}_2^{(1:L)}(k)$
A
$\hat{\mathbf{x}}_1^{(1:L)}(k)$
$\overline{w}_1^{1,(1:L)}(k)$
$\overline{w}_1^{2,(1:L)}(k)$
$z^1(k)$
$\overline{x}(k)$
$z^2(k)$
$\hat{\mathbf{x}}_2^{(1:L)}(k)$
$\overline{w}_2^{1,(1:L)}(k)$
$\overline{w}_2^{2,(1:L)}(k)$
A
$z^1(k)$
$\arctan(\overline{y}(k)/\overline{x}(k))$
$z^2(k)$
A

(a) Particles generation and unlabeled measurements update

(b) Particles weights evaluation and object state estimation

(c) 95 % confidence true bearing ranges and the bearing of the estimated position

Figure 9: Deviated estimation example with multiple models and multiple measurements.


$z^1(k)$
$\hat{\mathbf{x}}_2^{(1:L)}(k)$
$\hat{\mathbf{x}}_1^{(1:L)}(k)$
$z^2(k)$
A
$z^1(k)$
$\hat{\mathbf{x}}_1^{(1:L)}(k)$
$w_1^{1,(1:L)}(k)$
$w_1^{2,(1:L)}(k)$
$\hat{\mathbf{x}}_2^{(1:L)}(k)$
$w_2^{1,(1:L)}(k)$
$w_2^{2,(1:L)}(k)$
$z^2(k)$
$\overline{\mathbf{x}}(k)$
A
$z^1(k)$
$\arctan(\overline{y}(k)/\overline{x}(k))$
$z^2(k)$
A

(a) Particles weights evaluation and object state estimation

(b) Particles weights evaluation and object state estimation

(c) 95% confidence true bearing ranges and the bearing of the estimated position

Figure 10: Deviated estimation example where the triggering condition in (15) is not enough.

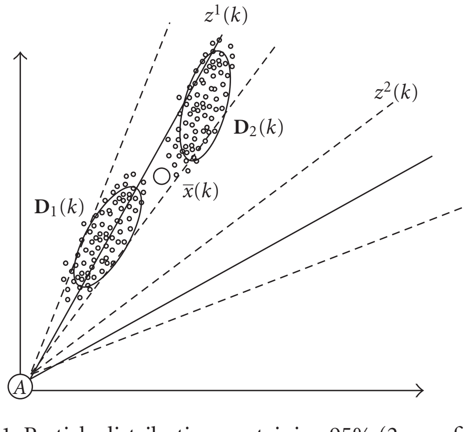
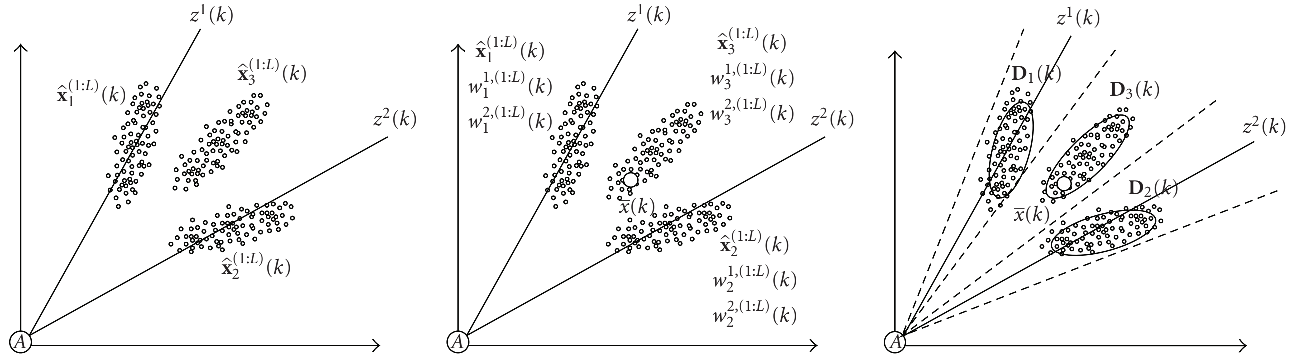


Figure 11: Particle distribution containing 95% (2σ confidence) of the particles assuming they are Gaussian distributed.

for $i \in \{1,2,\dots,N(\mathbf{z}(k))\}$, $l \in \{1,2,\dots,L\}$, and $m \in \{1,2,\dots,J\}$. The SIR algorithm is used to obtain $\widetilde{\mathbf{x}}^{(l)}(k)$'s, $l \in \{1,2,\dots,L\}$ with the equal conditional probability density from $\hat{\mathbf{x}}_m^{(l)}(k)$ particles with $w_m^{i,(l)}(k)$ weights. Note that there are $L \times J$ particles, and each particle has $N(\mathbf{z}(k))$ weight values. However, in order to apply the SIR algorithm, each particle has to have only one weight. Each particle is identically copied $N(\mathbf{z}(k))$ times to have the same number of weights, then the SIR algorithm is applied as in Figure 1, where $L \times J \times N(\mathbf{z}(k))$ particles are transformed to L resampled particles. Each circle in Figure 1 illustrates the weight of the particle. The resampled particles are assigned with an equal weight of $1/L$. The particles distribution with the resampled particles $\widetilde{\mathbf{x}}^{(l)}(k)$ with each corresponding weight value $1/L$ represents the conditional pdf of $p(\mathbf{x}(k) \mid \mathbf{z}(1:k))$. The resampled particles, $\widetilde{\mathbf{x}}^{(l)}(k)$, are used for generating particles $\hat{\mathbf{x}}_m^{(l)}(k+1)$ as in (3) for time $k+1$.

In order to estimate the final estimated state vector denoted as $\overline{\mathbf{x}}(k)$, the joint probability density association (JPDA) method is used, which makes use of all $L \times J \times N(\mathbf{z}(k))$ particles. $\overline{\mathbf{x}}(k)$ can also be obtained from the resampled L particles, but using the original $L \times J \times N(\mathbf{z}(k))$ particles can give a better mean estimate of the state. The JPDA technique uses a weighted average of all the measurements falling inside an object track's validation region to update the object state [30]. In addition, the weighted average of all possible J

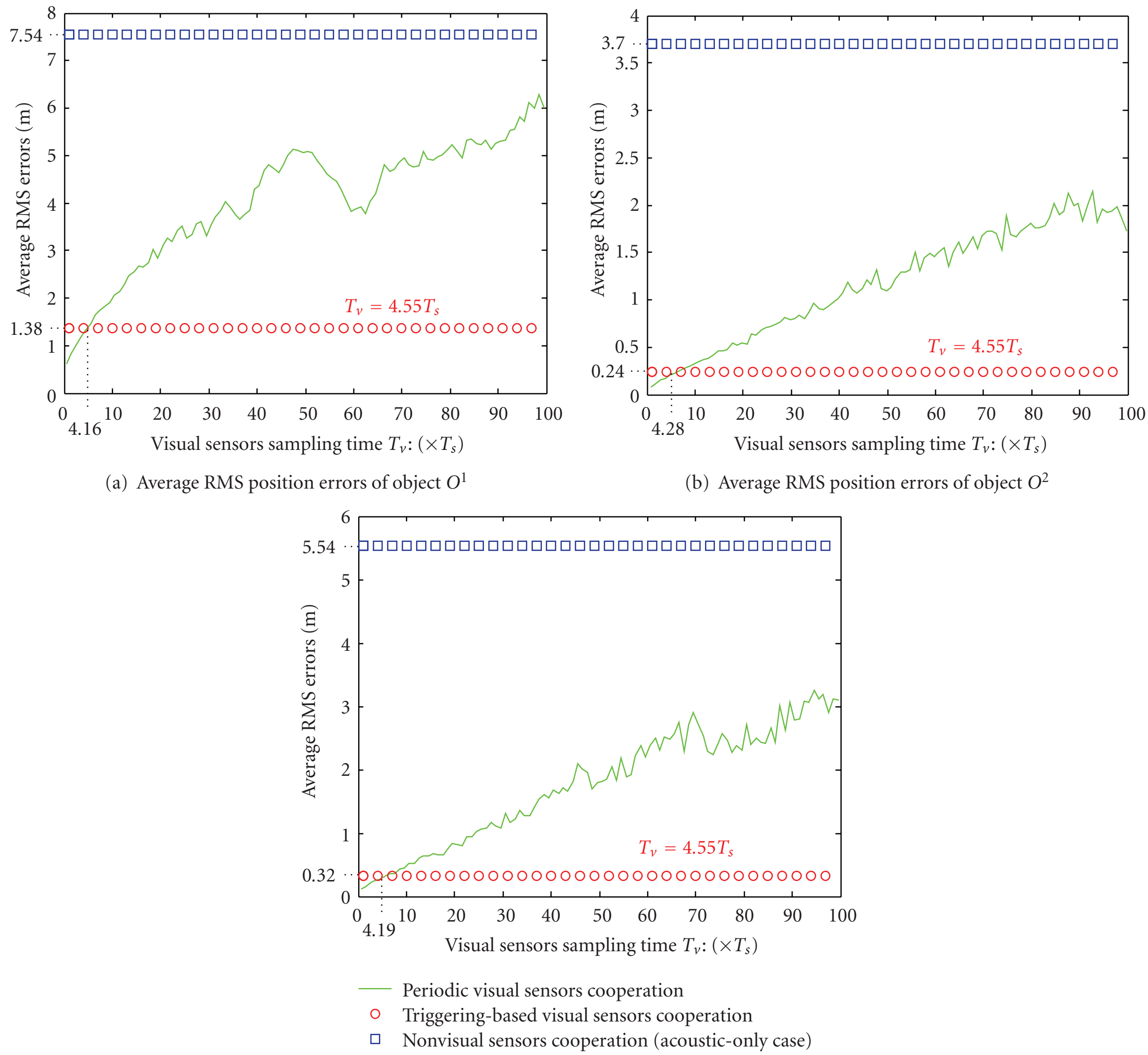


(a) Particles weights evaluation and object state estimation

(b) Particles weights evaluation and object state estimation

(c) 95% confidence true bearing ranges and the bearing of the estimated position

Figure 12: Deviated estimation example where both conditions (15) and (16) should be considered for the visual sensor cooperation.



(a) Average RMS position errors of object O^1

(b) Average RMS position errors of object O^2

(c) Average RMS position errors of object O^3

Figure 13: Average RMS position errors with three cooperation approaches. For the periodic visual sensor cooperation, the period varies from $1T_s$ to $100T_s$.

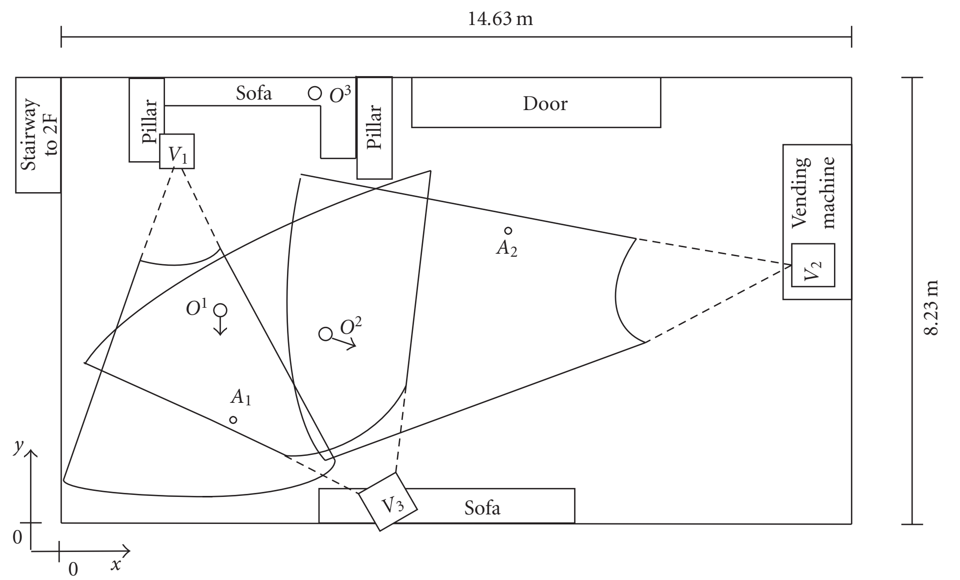


Figure 14: The visual sensor cooperation with an acoustic sensor-based estimation is experimented in an indoor environment with size 14.63 m × 8.23 m.

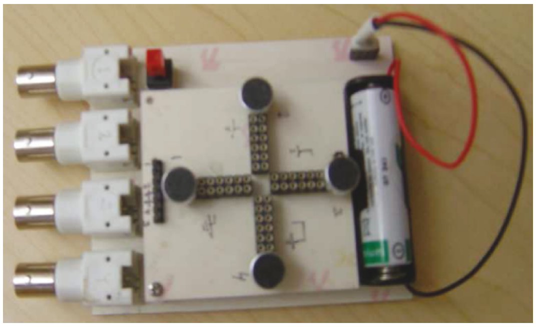


Figure 15: Four microphones measure interaural time differences of an object, and the azimuth and elevation angles are derived by scaling the speed of wave propagation and the unit dimensions of the microphones array.

models is also applied for estimating $\overline{\mathbf{x}}(k)$. First, $\overline{\mathbf{x}}_m^i(k)$'s, the conditional means of the state given each measurement $z^i(k)$ over the particles set, $\hat{\mathbf{x}}_m^{(l)}(k)$'s of model m is obtained as

$$\overline{\mathbf{x}}_m^i(k) = \sum_{l=1}^{L} \hat{\mathbf{x}}_m^{(l)}(k) \cdot w_m^{i,(l)}(k), \quad \text{for } i \in \{1,2,\dots,N(\mathbf{z}(k))\}, \quad m \in \{1,2,\dots,J\}. \tag{7}$$

Then, $\overline{\mathbf{x}}_m(k)$'s, $m \in \{1,2,\dots,J\}$, the conditional means of the state for model m is obtained as

$$\overline{\mathbf{x}}_m(k) = \sum_{i=1}^{N(\mathbf{z}(k))} \overline{\mathbf{x}}_m^i(k) \cdot \mu_m^i(k), \tag{8}$$

where $\mu_m^i(k)$ represents the probability that the model index is m given the measurement $z^i(k)$, and it is obtained as

$$\mu_m^i(k) = \frac{\sum_{l=1}^{L} w_m^{i,(l)}(k)}{\sum_{m=1}^{J}\left(\sum_{l=1}^{L} w_m^{i,(l)}(k)\right)}, \quad \text{for } i \in \{1,2,\dots,N(\mathbf{z}(k))\}, \quad m \in \{1,2,\dots,J\}. \tag{9}$$

Finally, the mean state vector estimate $\overline{\mathbf{x}}(k)$ is obtained as

$$\overline{\mathbf{x}}(k) = \sum_{m=1}^{J} \overline{\mathbf{x}}_m(k) \cdot \mu_m(k), \tag{10}$$

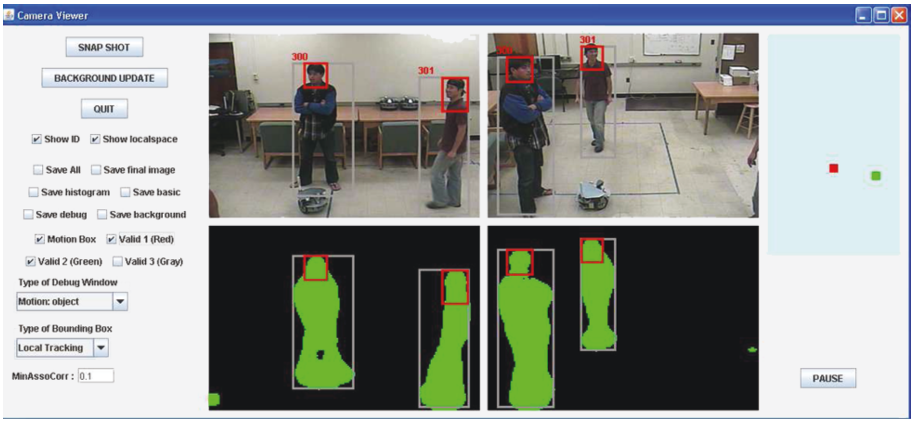
where $\mu_m(k)$'s, $m \in \{1,2,\dots,J\}$, is the probability that the object's model index is m, and it is obtained as

$$\mu_m(k) = \frac{\sum_{i=1}^{N(\mathbf{z}(k))} \mu_m^i(k)}{\sum_{m=1}^{J}\left(\sum_{i=1}^{N(\mathbf{z}(k))} \mu_m^i(k)\right)}. \tag{11}$$

Let $(x^{[v]}(k), y^{[v]}(k))$ denote the visually localized position of the triggered object at time k. Then, if the cooperation is performed at time k, the final estimated state vector $\overline{\mathbf{x}}(k) = [\overline{x}(k)\overline{y}(k)\overline{V}^x(k)\overline{V}^y(k)]^T$ of the object of interest is replaced by $[x^{[v]}(k)y^{[v]}(k)\overline{V}^x(k)\overline{V}^y(k)]$. Figure 2 illustrates the acoustic sensor-based IMM-PF data flow incorporated with the visual sensor cooperation, where the triggering conditions can be from measurements and/or estimated results with the particle filtering.

3. Effect of Visual Sensor Cooperation

In this section, the triggering conditions of the visual sensor cooperation are discussed. As a reference, unconditional periodic triggering is discussed in Section 3.1, and we show that additional triggering conditions are needed unless



☑ Show ID ☑ Show localspace

□ Save All □ Save final image

□ Save histogram □ Save basic

□ Save debug □ Save background

☑ Motion Box ☑ Valid 1 (Red)

☑ Valid 2 (Green) □ Valid 3 (Gray)

Type of Debug Window: Motion: object

Type of Bounding Box: Local Tracking

MinAssoCorr : 0.1

FIGURE 16: Visual sensors-based tracking demo: two visual sensors simultaneously detect, identify, and localize multiple objects.

the cooperation period is sufficiently small. The acoustic sensor-based estimation can have difficulties from two different perspectives—the system dynamics and the estimation performance. These two issues can be considered as two different triggering conditions. First, due to the system dynamics, the number of objects in tracking and the number of measurements in the acoustic sensor can be different. If so, the acoustic sensor cannot track multiple objects correctly, and the support from the visual sensor is needed. There can be several cases for the system dynamics, and they are discussed in Section 3.2. The performance degradation of the object tracking by the acoustic sensor, in our application the particle filter's performance, can be overcome by the support from the visual sensor even when the number of tracked objects and the number of measurements are the same. In this case, the performance of the estimation can be a condition for the triggering, and they are discussed in Section 3.3. Performance improvement by having the two triggering conditions is presented by the simulation in Section 3.4. Throughout this section, the acoustic-based bearing measurements are simulated instead of real bearing estimates. In addition, the visually localized position of the triggered object is assumed to be given. On the condition, we will first evaluate and analyze the performances of the acoustic sensor-based particle filter and our visual sensor cooperation method.

3.1. Periodic Visual Sensor Cooperation. Suppose that the visual sensor periodically localizes the object positions and supports the acoustic sensor-based estimation every visual sampling time T_v. Note that we define acoustic sampling time as T_s. In order to verify the effect of the periodic visual sensor cooperation, the tracking environment with three objects and an acoustic sensor are used as follows.

(i) Objects O^1, O^2, and O^3 are initially positioned at (50 m,30 m), (35 m,50 m), and (45 m,45 m), respectively. Trajectories of the three objects are shown in Figure 3(a). Note that the simulation dimension is set to 50 m × 50 m for simulation analysis only, and the dimension is practically adjusted in the next section for the real data experiment.

(ii) Each object trajectory is sampled by 200 acoustic bearing data, and $T_s = 1$ second.

(iii) Three models are considered – constant velocity $\mathbf{F}^{(1)}$, clockwise coordinated turn $\mathbf{F}^{(2)}$, and anticlockwise coordinated turn $\mathbf{F}^{(3)}$ with manoeuvre rotation acceleration 0.01 m/s^2 [26]. They are

$$\mathbf{F}^{(1)} = \begin{pmatrix} 1 & 0 & T_s & 0 \\ 0 & 1 & 0 & T_s \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{F}^{(p)} = \begin{pmatrix} 1 & 0 & \dfrac{\sin\left(\mathfrak{R}_k^{(p)} T_s\right)}{\mathfrak{R}_k^{(p)}} & -\dfrac{\left(1-\cos\left(\mathfrak{R}_k^{(p)} T_s\right)\right)}{\mathfrak{R}_k^{(p)}} \\ 0 & 1 & \dfrac{\left(1-\cos\left(\mathfrak{R}_k^{(p)} T_s\right)\right)}{\mathfrak{R}_k^{(p)}} & \dfrac{\sin\left(\mathfrak{R}_k^{(p)} T_s\right)}{\mathfrak{R}_k^{(p)}} \\ 0 & 0 & \cos\left(\mathfrak{R}_k^{(p)} T_s\right) & -\sin\left(\mathfrak{R}_k^{(p)} T_s\right) \\ 0 & 0 & \sin\left(\mathfrak{R}_k^{(p)} T_s\right) & \cos\left(\mathfrak{R}_k^{(p)} T_s\right) \end{pmatrix}, \tag{12}$$

where $p = 2,3$ and $\mathfrak{R}_k^{(p)}$ is the model-dependent turning rates expressed as

$$\mathfrak{R}_k^{(2)} = \frac{\alpha}{\sqrt{\left(V^x(k-1)\right)^2 + \left(V^y(k-1)\right)^2}},$$
$$\mathfrak{R}_k^{(3)} = \frac{-\alpha}{\sqrt{\left(V^x(k-1)\right)^2 + \left(V^y(k-1)\right)^2}}, \tag{13}$$

with α being the factor determining the rotation degree as 1 m/s^2.

(iv) Measurement noise variance σ^2 varies from 0.0 to 5.0, which corresponds to $v(k)$ in (2).

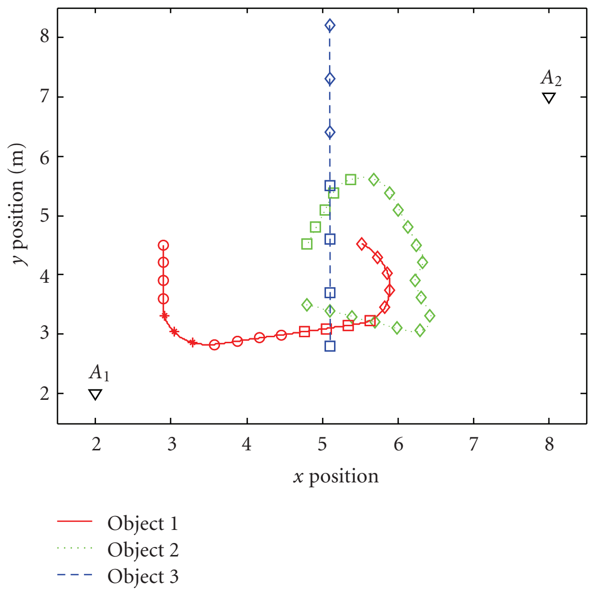


Figure 17: Three objects trajectories and the positions of the two acoustic sensors.

Figure 3 shows the performance of the acoustic sensor based estimation for various noise variances, where 500 particles are used for each object in each model. As the noise variance increases, the estimation has a higher Root-Mean-Square (RMS) position error. Especially with the noise variance of 5.0, the RMS position error of each object is 3.98, 9.80, and 1.08, respectively. Under the same condition with the noise variance of 5.0 in Figure 3(d), the visual sensor periodically supports the acoustic sensor-based estimation by updating the localized object position for each object. The effect of the periodic visual sensor supporting different sampling time T_v is shown in Figures 4 and 5. In Figure 4, the estimated trajectories are shown for different visual sensor's sampling times T_v: $10T_s$, $20T_s$, $50T_s$, and $100T_s$. Figure 5 shows the average RMS position errors with visual sensor's sampling time T_v from $1T_s$ to $100T_s$ through 1,000 time trials, respectively.

From the results shown in Figures 4 and 5, it is difficult to find an optimal visual sensor's sampling time T_v. It can only be seen that the estimated object position becomes more accurate as the visual sensor's sampling time T_v is close to the acoustic sampling time T_s. Even when the acoustic sensor estimates an object's position close to the true position, the visual sensor may unnecessarily support the acoustic sensor through the periodic cooperation. In order to efficiently use the precious visual sensor cooperation, it has to be triggered only when the cooperation is necessary. Furthermore, the periodic cooperation does not efficiently support the acoustic sensor-based estimation against deviated estimation, measurement resolution problem, and a varying number of objects. These issues are discussed in the following subsections.

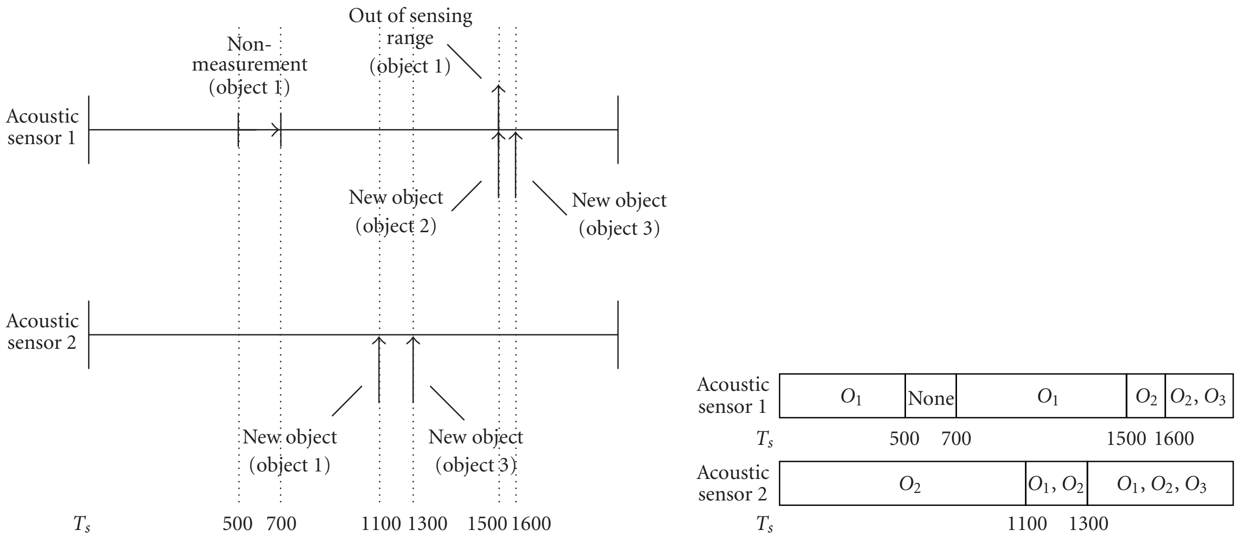
3.2. Triggering Based on System Dynamics. An acoustic sensor can have a difficulty in measuring multiple bearing measurements when the difference is negligibly small—the acoustic sensor has a limited resolution of $\Delta z_{\text{critical}}$ [3, 31]. The bearing measurement difference of two objects less than $\Delta z_{\text{critical}}$ can cause an acoustic sensor to recognize only one sound wave by merging the multiple incoming sound waves. Let $I(k)$ denote the number of objects estimated by the acoustic sensor at time k. Then, if the acoustic sensor cannot differentiate the objects, the number of measurements at time k, $N(\mathbf{z}(k))$ and the number of estimated objects at time $k-1$ become unequal as

$$N(\mathbf{z}(k)) \neq I(k-1). \tag{14}$$

The visual sensor cooperation should be triggered in case of (14). Once the visual sensor supports the acoustic sensor-based estimator with the visually localized positions at time k, the number of estimated objects $I(k)$ is updated and verified with $N(\mathbf{z}(k+1))$ for time $k+1$.

Together with the measurement resolution problem, an acoustic sensor also has a difficulty in estimating the state with a varying number of objects/measurements positioned within the measurable range of the acoustic sensor. The number of measurements is varying when the objects do not transmit sound wave, new objects come into an acoustic sensing range, or objects move out an acoustic sensing range. Then, similarly to the measurement resolution problem, the number of measurements at time k and the number of estimated objects at time $k-1$ become unequal as in (14). More specifically, in the varying number of objects/measurements, if $N(\mathbf{z}(k)) < I(k-1)$, objects move out of acoustic sensing range, or/and an acoustic sensor does not receive bearing measurements from objects at time k. On the other hand, if $N(\mathbf{z}(k)) > I(k-1)$, new objects are moving into the acoustic sensing range at time k. That is, the varying number of objects/measurements can also be triggered for the visual sensor cooperation with the same condition as (14). After the visual sensor cooperation, $I(k)$ is updated and verified with $N(\mathbf{z}(k+1))$ for time $k+1$.

Consider the environment shown in Figure 6(a), where the acoustic sensor is positioned at $(25\,\text{m}, 25\,\text{m})$, while the bearing sources are sampled every 1 second during 200 seconds period with the noise variance of 3. Object 1 and object 2 start from $(5\,\text{m}, 3\,\text{m})$ and $(22\,\text{m}, 3\,\text{m})$ with initial velocities of $(0.2\,\text{m/s}, 0.2\,\text{m/s})$ and $(-0.2\,\text{m/s}, 0.2\,\text{m/s})$, respectively. Two objects are with model $\mathbf{F}^{(1)}$ except at time $k = 51T_s$, $101T_s$, and $151T_s$. Their models at those times are $\mathbf{F}^{(2)}$ or $\mathbf{F}^{(3)}$ defined in (12), and the resulting trajectories are shown in Figure 6(a). Object O^1 is moving into the acoustic sensing range at time $25T_s$ and object O^2 is moving out at time $175T_s$. The new object O^3, is moving in the acoustic sensing range at time $63T_s$ and moving out at time $188T_s$. Object O^3 is initially with model $\mathbf{F}^{(1)}$, and it changes to $\mathbf{F}^{(2)}$ and returns to $\mathbf{F}^{(1)}$ at time $101T_s$ and $151T_s$, respectively. Figure 6(b) shows the triggering timings based on the system dynamics including the measurement resolution problem and the varying number of objects. For better understanding, "o" is marked when the triggering timing is caused by the varying number of objects, while "$\ast$" is marked when it is caused by the measurement resolution problem.



(a) Nonmeasurement, new object appearance, and movement out of sensing range with respect to each sensor

(b) Measured objects from each sensor A_1 and A_2

Figure 18: Measured objects over time.

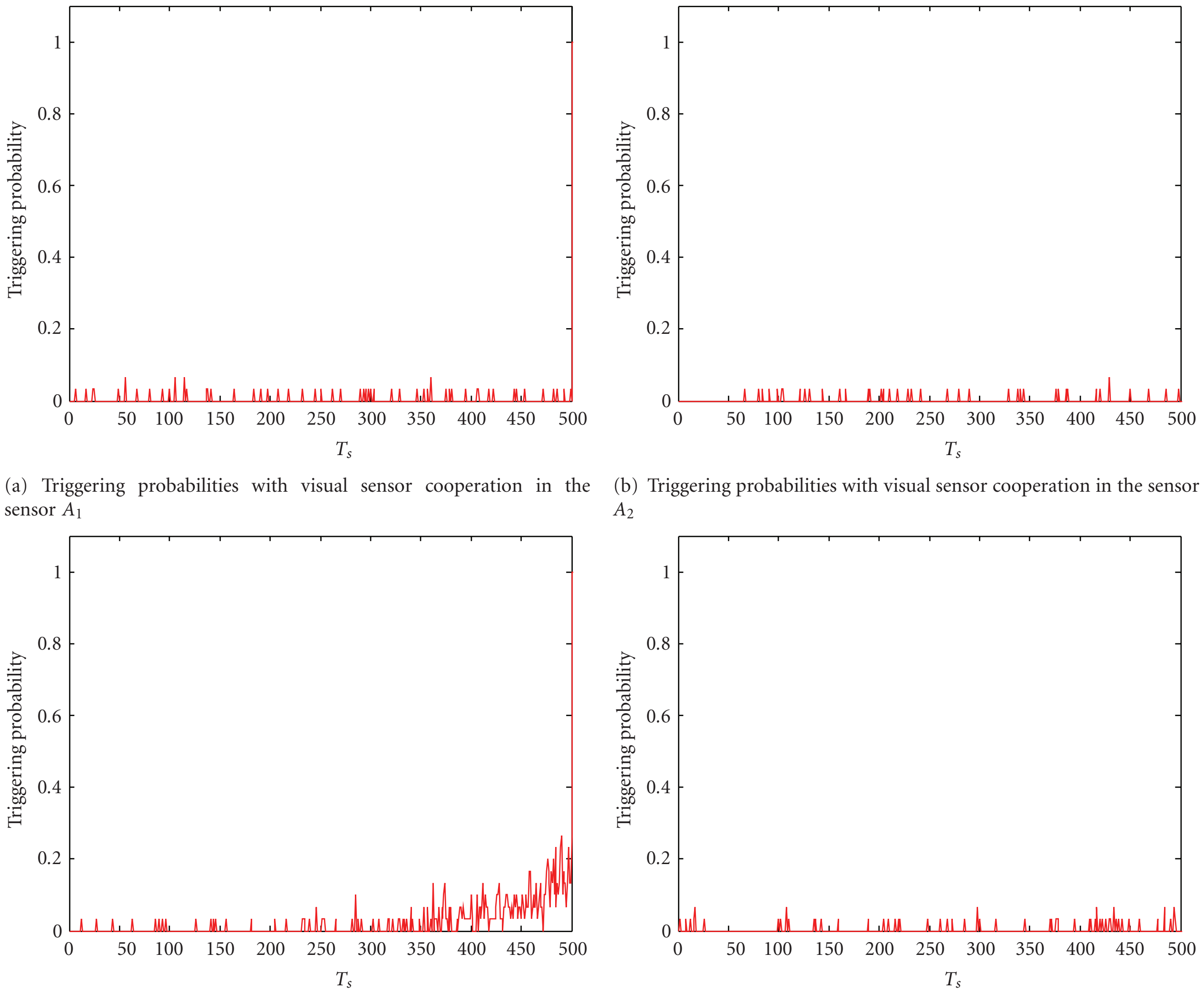
3.3. Triggering Based on Estimation Performance. The triggering based on (14) cannot trigger the visual sensor cooperation for a simultaneous varying number of objects or measurements. Figure 7 shows several examples. Given the three objects in Figure 7(a), Figure 7(b) shows that a new object O^4 moves into the acoustic sensing range, while O^3 bearing measurement is not received by an acoustic sensor. In this case, the number of objects $I(k-1)$ and the number of measurements $N(\mathbf{z}(k))$ are the same even though the number of objects is varying and the cooperation of a visual sensor is needed. Similarly, the condition in (14) does not trigger a cooperation either for the case in Figure 7(c), where the new object O^4 moves into the acoustic sensing range, while O^3 moves out the acoustic sensing range. Figures 7(d), 7(e), and 7(f) also illustrate similar cases, where the cooperation is not triggered despite the need. The cases in Figures 7(b) through 7(f) should trigger the visual sensor cooperation by considering the estimation performance at the particle filtering state.

The triggering based on the estimation performance is to find the triggering timing with the deviated estimation at the particle-filtering stage, while the triggering based on the system dynamics is to find the triggering timing with the inconsistency between $I(k)$ and $N(\mathbf{z}(k))$. The deviated estimation is caused by the cases in Figure 7 or an incorrect interaction between the measurement and the predicted particles. It is nontrivial to evaluate how the estimated position is deviated from a true object position because an acoustic sensor receives only the bearing measurements, and the triggering should be based on the difference between the angle from the estimated position and the bearing measurement. Let $(\overline{x}(k), \overline{y}(k))$ be the estimated position of an object and a bearing measurement $z^i(k)$, $i \in \{1, 2, \ldots, N(\mathbf{z}(k))\}$ with noise variance σ^2 are given as illustrated in Figure 8. Assuming that the bearing measurement $z^i(k)$ follows the Gaussian distribution, its range between $z^i(k) - 2\sigma$ and $z^i(k) + 2\sigma$ contains 95% (2σ confidence) of the true bearing. Then, the estimated position $(\overline{x}(k), \overline{y}(k))$ is considered as a deviation if the following condition is satisfied

$$\begin{gathered}\arctan\left(\frac{\overline{y}(k)}{\overline{x}(k)}\right) < z^i(k) - 2\sigma \quad \text{or} \\ \arctan\left(\frac{\overline{y}(k)}{\overline{x}(k)}\right) > z^i(k) + 2\sigma, \\ \forall\, i \in \{1, 2, \ldots, N(\mathbf{z}(k))\}.\end{gathered} \tag{15}$$

This means that if no bearing measurement falls within ±2σ of the estimated angle, then the visual sensor is triggered for the cooperation. Note that the measurement variance σ^2 is known from the acoustic sensor's performance characteristics.

The 95% confidence true bearing range plays an important role to evaluate the deviated estimation, especially for estimating multiple object states with multiple models, $I(k) > 1$ and $J > 1$. Consider the estimation with multiple objects and two models. Figure 9 illustrates a deviated estimation example with simplified sequential steps from particles generation to object state estimation. In Figure 9(a), two-model-based particles $\widehat{\mathbf{x}}_1^{(1:L)}(k)$ and $\widehat{\mathbf{x}}_2^{(1:L)}(k)$ are generated for an object, and the unlabeled measurements $z^1(k)$ and $z^2(k)$ are updated. Suppose that measurement $z^1(k)$ is obtained from the object of interest while measurement $z^2(k)$ is obtained from another object. Suppose also that $\widehat{\mathbf{x}}_1^{(1:L)}(k)$ is generated close to $z^1(k)$ and $\widehat{\mathbf{x}}_2^{(1:L)}(k)$ is generated close to $z^2(k)$. Then, in Figure 9(b), particles' weights for model 1 given measurement $z^1(k)$, $\overline{w}_1^{1,(1:L)}$ and particles' weights for model 2 given the measurement $z^2(k)$, $\overline{w}_2^{2,(1:L)}$ are evenly dominating for the particles $\widehat{\mathbf{x}}_1^{(1:L)}(k)$ and $\widehat{\mathbf{x}}_2^{(1:L)}(k)$,



(a) Triggering probabilities with visual sensor cooperation in the sensor A_1

(b) Triggering probabilities with visual sensor cooperation in the sensor A_2

(c) Triggering probabilities without visual sensor cooperation in the sensor A_1

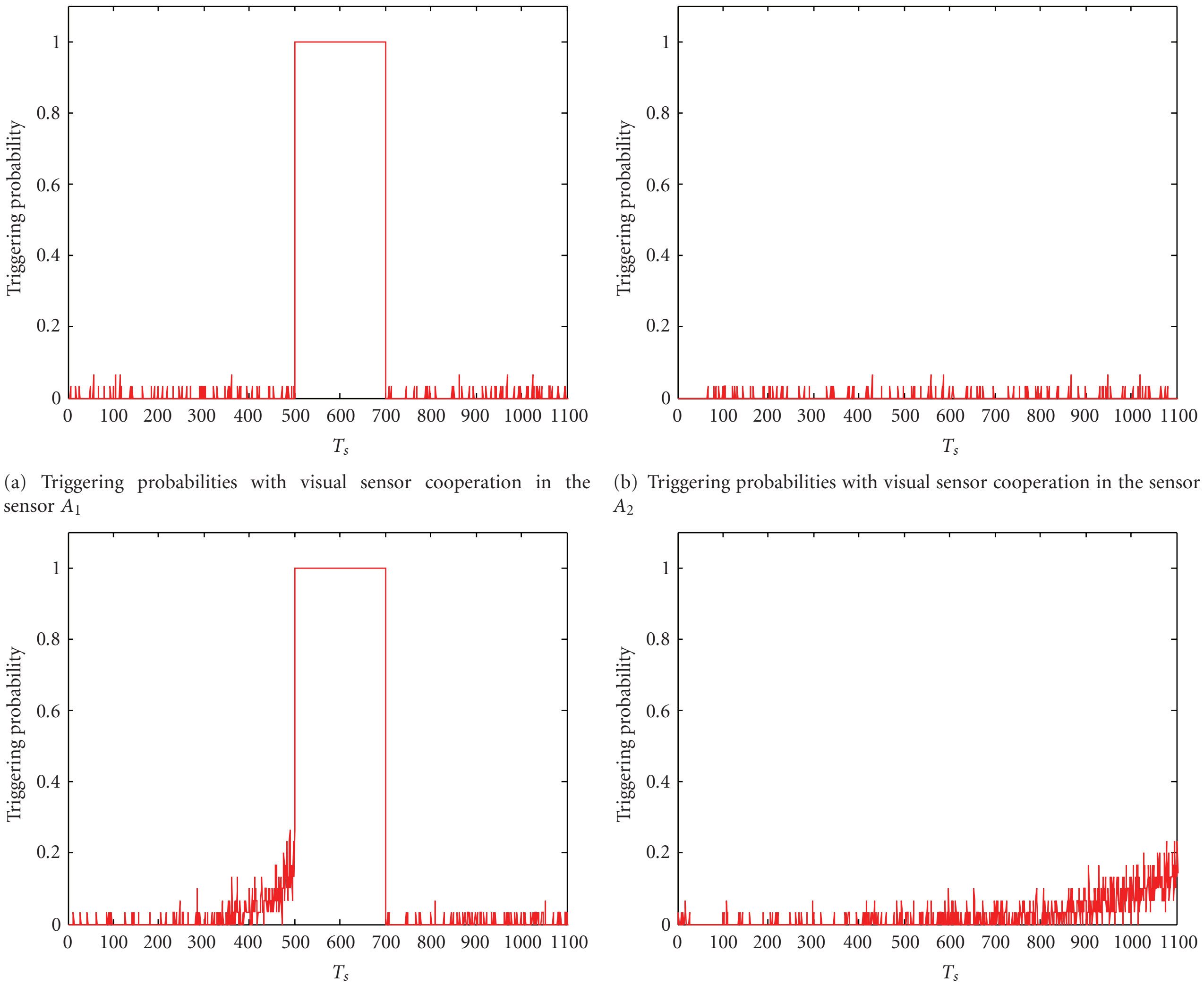
(d) Triggering probabilities without visual sensor cooperation in the sensor A_2

Figure 19: Triggering probabilities of the two sensors between $1T_s$ and $500T_s$.

respectively. According to the weights, the estimated object state $\overline{\mathbf{x}}(k)$ is obtained with the average of each model-based particles information. Finally, the bearing of the estimated position $\arctan(\overline{y}(k)/\overline{x}(k))$ strays off from the 95% confidence true bearing range of $z^1(k)$ as illustrated in Figure 9(c).

However, the 95% confidence true bearing range as in (15) does not necessarily trigger the visual sensor cooperation. Figure 10 illustrates another deviated estimation example, where the visual sensor cooperation cannot be triggered with the 95% confidence true bearing range from the condition in (15). Similarly to the example in Figure 9, suppose that measurement $z^1(k)$ is obtained from the object of interest while measurement $z^2(k)$ is obtained from another object. In Figure 10(a), two-model-based particles $\hat{\mathbf{x}}_1^{(1:L)}(k)$ and $\hat{\mathbf{x}}_2^{(1:L)}(k)$ for the object of interest are generated and both of them are at the angle close to $z^1(k)$. Then, in Figure 10(b), particles' weights for model 1 given measurement $z^1(k)$ and particles' weights for the model 2 given the measurement $z^1(k)$, $\overline{w}_1^{1,(1:L)}$ and $\overline{w}_2^{1,(1:L)}$ are evenly dominating for the particles $\hat{\mathbf{x}}_1^{(1:L)}(k)$ and $\hat{\mathbf{x}}_2^{(1:L)}(k)$, respectively. According to the weights, the estimated object state $\overline{\mathbf{x}}(k)$ is obtained with the average of each model-based particles. As illustrated in Figure 10(c), even though the estimated object state $\overline{\mathbf{x}}$(k) is deviated by the two models, the bearing of the estimated position $\arctan(\overline{y}(k)/\overline{x}(k))$ does not trigger the visual sensor cooperation from the condition in (15).

In order to overcome the limitation of the triggering with the 95% confidence true bearing range in (15), we consider an additional triggering condition based on predicted particles distribution. The particle distribution can be expressed with an ellipse representing the region, which contains 95% (2σ confidence) of the particles assuming that they are Gaussian distributed [32] in two dimensions. Denote the 95% confidence ellipse of $\hat{\mathbf{x}}_j^{(1:L)}(k)$ as $\mathbf{D}_j(k)$, where $j \in \{1,2,\ldots,M\}$ represents the model index.



(a) Triggering probabilities with visual sensor cooperation in the sensor A_1

(b) Triggering probabilities with visual sensor cooperation in the sensor A_2

(c) Triggering probabilities without visual sensor cooperation in the sensor A_1

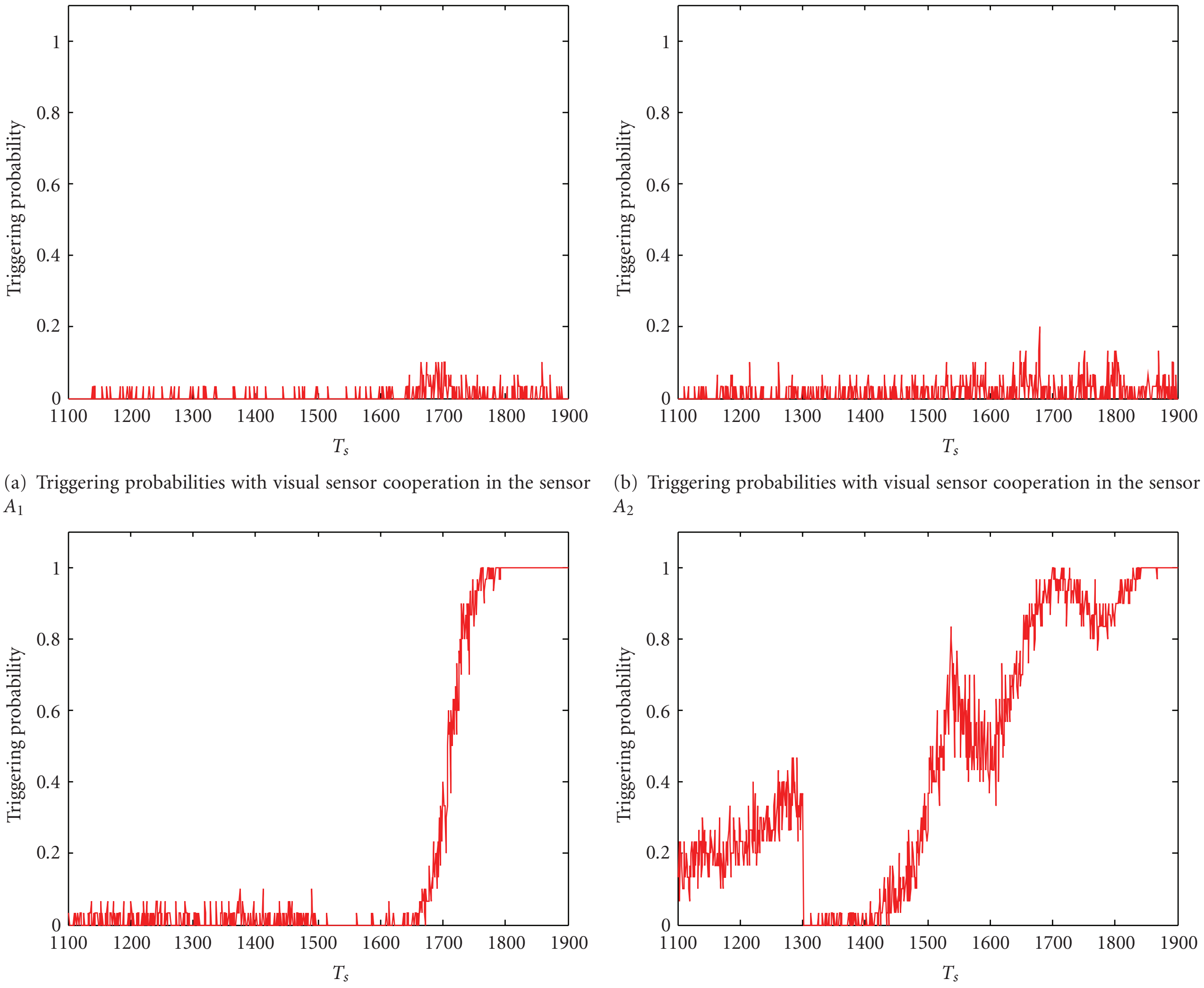
(d) Triggering probabilities without visual sensor cooperation in the sensor A_2

Figure 20: Triggering probabilities of two sensors between $501T_s$ and $1100T_s$.

Figure 11 illustrates the 95% confidence particles ellipses $\mathbf{D}_1(k)$ and $\mathbf{D}_2(k)$ corresponding to $\hat{\mathbf{x}}_1^{(1:L)}(k)$ and $\hat{\mathbf{x}}_2^{(1:L)}(k)$ in the deviated estimation example in Figure 10. If the estimated position $(\overline{x}(k), \overline{y}(k))$ is obtained outside the 95% confidence predicted particles ellipse as in Figure 10, then it is considered as a deviation. In a general form, the estimated position $(\overline{x}(k), \overline{y}(k))$ is considered as a deviation with the condition of

$$(\overline{x}(k), \overline{y}(k)) \notin \mathbf{D}_j(k), \quad \forall j \in \{1, 2, \ldots, M\}. \tag{16}$$

Even though the 95% confidence particles ellipses in the condition (16) is to overcome the limitation of the triggering with the 95% confidence true bearing range in the condition (15), these two conditions should be used together—at least one condition indicates a deviation, then the cooperation should be triggered. Figure 12 illustrates another deviated example, where the visual sensor cooperation are triggered not by (16) but by (15). Also, suppose that measurement $z^1(k)$ is obtained from the object of interest while measurement $z^2(k)$ is obtained from another object. In Figure 12(a), three-model-based particles $\hat{\mathbf{x}}_1^{(1:L)}(k)$, $\hat{\mathbf{x}}_2^{(1:L)}(k)$, and $\hat{\mathbf{x}}_3^{(1:L)}(k)$ for the object of interest are generated as, $\hat{\mathbf{x}}_1^{(1:L)}(k)$ are generated close to $z^1(k)$, and $\hat{\mathbf{x}}_2^{(1:L)}(k)$ are generated close to $z^2(k)$. Then, as illustrated in Figure 12(b), each model-based particles' weights $\overline{w}_1^{1,(1:L)}$, $\overline{w}_1^{2,(1:L)}$, $\overline{w}_2^{1,(1:L)}$, $\overline{w}_2^{2,(1:L)}$, $\overline{w}_3^{1,(1:L)}$, and $\overline{w}_3^{2,(1:L)}$ are evaluated corresponding to the unlabeled measurements $z^1(k)$, $z^2(k)$, and $z^3(k)$, where $\overline{w}_1^{1,(1:L)}$ and $\overline{w}_2^{2,(1:L)}$ are evenly dominating for the particles $\hat{\mathbf{x}}_1^{(1:L)}(k)$ and $\hat{\mathbf{x}}_2^{(1:L)}(k)$, respectively. Finally, the estimated object state $\overline{\mathbf{x}}(k)$ is obtained with the particles information averaged over model 1 and 2, which is close not to $\hat{\mathbf{x}}_1^{(1:L)}(k)$ but to $\hat{\mathbf{x}}_2^{(1:L)}(k)$. In this case, the estimated position $(\overline{x}(k), \overline{y}(k))$ is satisfied with (16), but the bearing of the estimated position $\arctan(\overline{y}(k)/\overline{x}(k))$ is not satisfied with (15) as illustrated in Figure 12(c). Thus, the 95% confidence particles ellipses in



(a) Triggering probabilities with visual sensor cooperation in the sensor A_1

(b) Triggering probabilities with visual sensor cooperation in the sensor A_2

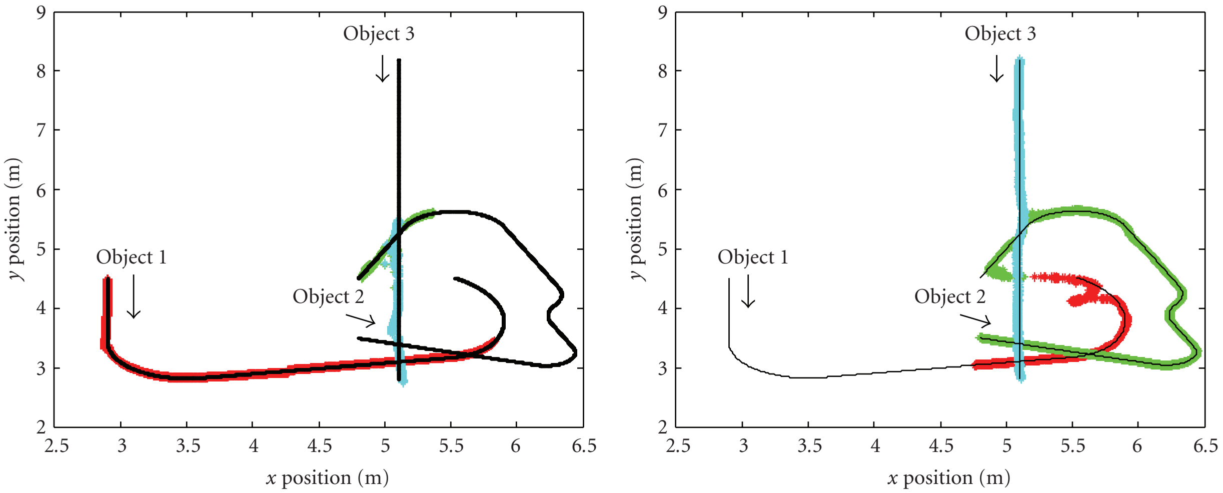
(c) Triggering probabilities without visual sensor cooperation in the sensor A_1

(d) Triggering probabilities without visual sensor cooperation in the sensor A_2

Figure 21: Triggering probabilities of two sensors between $1101T_s$ and $1900T_s$.

(16), and the 95% confidence true bearing range in (15) should be considered together.

3.4. Performance Evaluation with Simulation. In this subsection, the performance of the triggering-based visual sensor cooperation is evaluated with the comparison to the performance of the periodic visual sensor cooperation as well as nonvisual sensor cooperation (acoustic-only case). For the performance evaluation, the environment described in Figure 6(a) is considered with 200 acoustic sampling times, and the simulated bearing estimates are corrupted by noise variance 3. There are 100 trials to get the average results. Figure 13 shows the average RMS position errors corresponding to triggering-based visual sensor cooperation, periodic visual sensor cooperation, and nonvisual sensor cooperation (acoustic-only case). As shown in Figure 13(a), the average RMS position errors of object O^1 is 1.38 based on the triggering-based visual sensor cooperation and 7.54 without the visual sensor cooperation. Also, the average RMS position errors with the periodic visual sensor cooperation are shown according to different visual sensor's sampling time T_v: $1T_s$ to $100T_s$. In the triggering-based visual sensor cooperation, the average visual sensor's sampling time T_v is approximately $4.55T_s$. In the periodic visual sensor cooperation, on the other hand, the visual sensor's sampling time T_v corresponding to the average RMS position error 1.38 is approximately $4.16T_s$. It shows that the triggering-based visual sensor cooperation requires less visual sensor resources than the periodic visual sensor cooperation for the same tracking performance. Similarly, Figures 13(b) and 13(c) show the same pattern for objects O^2 and O^3. Furthermore, in the periodic visual sensor cooperation, the visual sensor's sampling time T_v corresponding to the average RMS position error 0.64 is approximately $4.21T_s$ on an average over the three objects.



(a) Final estimated position for acoustic sensor A_1

(b) Final estimated position for acoustic sensor A_2

FIGURE 22: Final estimated position for acoustic sensors A_1 and A_2.

TABLE 1: Performances of the triggering-based visual sensor cooperation, the periodic visual sensors cooperation and the nonvisual sensor cooperation (acoustic-only case).

(a) Average RMS position errors.

	Nonvisual sensor cooperation (acoustic-only case)	Triggering-based visual sensors cooperation
Object 1	7.54	1.38
Object 2	3.70	0.24
Object 3	5.54	0.32
Total average	5.59	0.64

(b) Equivalent visual sensor cooperation period.

	Nonvisual sensor cooperation (acoustic-only case)	Triggering-based visual sensors cooperation
Object 1	$4.16T_s$	$4.55T_s$
Object 2	$4.28T_s$	
Object 3	$4.19T_s$	
Total average	$4.21T_s$	

Table 1(a) summarizes the average RMS position errors with the triggering-based visual sensor cooperation and the nonvisual sensor cooperation (acoustic-only case). Table 1(b) summarizes the average triggered visual sensor's sampling time and the periodic visual sampling time corresponding to the performance level as same as the RMS position error in the triggering sensor-based cooperation. In practice, the optimal period of visual sensor cooperation is unknown since the triggering mechanism is dependent upon the system dynamics and the estimation performance. In any environment, triggering the visual sensor cooperation can adapt to the cooperation period, while periodic visual sensors cooperation may waste resources. In addition, under the cooperation period restriction due to network delay and image processing, the triggering mechanism may support the cooperation to the objects with the highest priority since it can recognize critical cases.

4. Experiment and Analysis

4.1. Experiment Setup. The visual sensor cooperation with the acoustic sensor-based estimation is considered in an indoor environment with size 14.63 m × 8.23 m illustrated in Figure 14. Object O^1 starts with initial velocity $(0\,\text{m/s}, -0.3\,\text{m/s})$ from position $(2.9\,\text{m}, 4.5\,\text{m})$ object O^2 starts with initial velocity $(0.3\,\text{m/s}, -0.1\,\text{m/s})$ from position $(4.8\,\text{m}, 3.5\,\text{m})$, object O^3 starts with initial velocity $(0\,\text{m/s}, 0\,\text{m/s})$ from position $(5.1\,\text{m}, 8.2\,\text{m})$. Two acoustic sensors A_1 and A_2 are deployed on the ceiling positioned at $(3.2\,\text{m}, 1.9\,\text{m})$ and $(7.6\,\text{m}, 6.8\,\text{m})$ each with 100 emulated samples per second. Each acoustic sensor receives the acoustic samples with variance 3 during 19 seconds and tracks the objects independently. Three visual sensors V_1, V_2, and V_3 are placed at positions $(1.9\,\text{m}, 6.3\,\text{m})$, $(13.4\,\text{m}, 5.0\,\text{m})$, and $(5.5\,\text{m}, 0.3\,\text{m})$ each with 6 samples per second.

4.2. Object Tracking with an Acoustic Sensor. The acoustic sensor is used with a micropower gradient flow acoustic localizer, where four microphones measure interaural time differences of an object [3]. The microphones are connected to the National Instruments USB-9162 to obtain the acoustic signals. By scaling the speed of wave propagation and the unit dimensions of the microphones array, the direction of azimuth and elevation angles are derived. Figure 15 shows the microphones for the micropower gradient flow acoustic localizer deployed in the sensor network.

4.3. Object Tracking with a Visual Sensor. In our application, once an acoustic sensor triggers for visual sensor cooperation, the visual sensor performs the object localization and supports the acoustic sensor with the localized position. The visual sensor localizes the object positions with the parallel projection model which supports zooming, panning, and tilting of the visual sensor [33, 34] and simplifies the computational complexity in determining the object positions with automatically focusing on the objects. As the visual sensor cooperation is triggered, a pair of visual sensors simultaneously detect, identify, and localize the multiple objects as shown in Figure 16. The objects are detected with motion analysis and color information as shown in [35, 36]. We assume that the viewable range of the visual sensors and the measurable range of the acoustic sensor are overlapped so that the visual sensors support the localized positions of the objects moving within the measurable range of acoustic sensors.

4.4. Objects Dynamic Characteristics with Acoustic Sensing Range/Capability. Figure 17 shows the three objects movement by switching three dynamic models: the constant velocity with $\mathbf{F}^{(1)}$ (CV), the clockwise coordinated turn with $\mathbf{F}^{(2)}$ (CT), and the anticlockwise coordinated turn with $\mathbf{F}^{(3)}$ (ACT) in (12). Also, the three people (objects) were instructed to make a constant-strength-sound when they move. Object O^1 starts with the CV model for 3.6 seconds. Between 3.6 seconds and 7.1 seconds, the object moves with the ACT model with $\alpha = 0.15\,\mathrm{m/s^2}$. Between 7.1 seconds and 13.5 seconds, the object moves with the CV model. Between 13.5 seconds and 15.0 seconds, the object moves with ACT model with $\alpha = 0.20\,\mathrm{m/s^2}$. Finally, between 15.0 seconds and 19.0 seconds, the object moves with the CV model. In the mean time, Object O^1 was additionally instructed not to make a sound between 5.0 seconds and 7.0 seconds (for two seconds). Object O^2 starts with the CV model for 4.5 seconds. Between 4.5 seconds and 6.1 seconds, the object moves with the ACT model with $\alpha = 0.45\,\mathrm{m/s^2}$. Between 6.1 seconds and 7.5 seconds, the object moves with the CV model. Between 7.5 seconds and 8.5 seconds, the object moves with the ACT model with $\alpha = 0.30\,\mathrm{m/s^2}$. Between 8.5 seconds and 9.5 seconds, the object moves with the CT model with $\alpha = 0.30\,\mathrm{m/s^2}$. Between 9.5 seconds and 13.0 seconds, the object moves with CV the model. Between 13.0 seconds and 16.0 seconds, the object moves with the CT model with $\alpha = 0.25\,\mathrm{m/s^2}$. Finally, between 16.0 seconds and 19.0 seconds, the object moves with the CV model. Object O^3 initially does not move without making any sound for 13.0 seconds, and starts to move with the CV model and sound between 13.0 seconds and 19.0 seconds.

In addition, given the acoustic sensors A_1 and A_2 shown in Figure 17, if the measurement is received by only acoustic sensor A_1, a circle is marked ("o"). If the measurement is received by only acoustic sensor A_2, a square is marked ("□"). If the measurement is received by both sensors A_1 and A_2, a diamond is marked ("◇"). If the measurement is not received by any of two sensors, a star is marked ("$*$"). Note even though the three people were instructed to make a constant-strength-sound, we observed that acoustic source strength changes according to an orientation and a distance between an acoustic sensor and a source (human). The acoustic sensor has a preprocess stage, which makes a decision whether the incoming acoustic source is from a real one or not (environment noise only). The decision is based on acoustic source strength with a threshold value. With the sound strength larger than threshold value, an acoustic sensor makes a decision that the source is from a real one. Otherwise, an acoustic sensor decides that the source is absent (environment noise only).

Figure 18(a) arranges non-measurement, new object appearance and movement out of sensing capability/range with respect to each sensor. Sensor A_1 initially receives one measurement from object O^1 and does not receive measurements between 5.0 seconds and 7.0 seconds. At time 15.0 seconds, sensor A_1 starts to receive new measurement from object O^2, but starts to miss the measurement from object O^1 since object O^1 moves out the sensing range/capability. At time 16.0 seconds, sensor A_2 starts to receive another new measurement from object O^3. Sensor A_2 initially receives one measurement from object O^2. At time 11 seconds, sensor A_2 starts to receive new measurement from object O^1. At time 13.0 seconds, sensor A_1 starts to receive another new measurement from object O^3. Figure 18(b) shows that the measured objects from each sensor A_1 and A_2.

4.5. Visual Sensor Cooperation with Triggering Timing Analysis. There are a few factors making the consistent sensing of the acoustic sources difficult—the background noise and the variations in the strength and the orientation of the incoming sound waves. With those, it is even more difficult to repeat the same movements of three people for the experiment. For these reasons, we use one set of real bearing data from the movement of three people, but we added 100 sets of noise with variance 3. For the triggering timing analysis, the triggering timings are considered as the triggering probabilities, and they are compared for the two cases. The case one is where the visual sensor supports the localized positions to the acoustic sensor estimator when they are triggered. On the other hand, the visual sensor does not support in the second case.

From time $1T_s$ to $500T_s$, acoustic sensor A_1 receives measurements from object O^1, and acoustic sensor A_2 receives measurements from object O^2. Since each sensor estimates different objects' state, it is considered as the single object estimation with a single sensor. Then, the triggering timing is obtained from the estimation performance only. Figures 19(a) and 19(b) show the triggering probabilities with the visual sensor cooperation in sensors A_1 and A_2, respectively, between $1T_s$ and $500T_s$. For comparison, Figures 19(c) and 19(d) show the triggering probabilities without visual sensor cooperation in sensors A_1 and A_2, respectively.

From time $501T_s$ to $700T_s$, object O^1 does not transmit sound wave. Due to the non-measurement, the acoustic sensor A_1 triggers visual sensor cooperation: the number of objects and the number of measurements are different. Figure 20 continually shows the triggering probabilities of the two sensors between $1T_s$ and $1{,}100T_s$ through 100 times trial. Figures 20(a) and 20(b) show the triggering probabilities with visual sensor cooperation in sensors A_1 and A_2, respectively. Also, Figures 20(c) and 20(d) show the triggering probabilities without the visual sensor cooperation in sensors A_1 and A_2, respectively.

At time $1{,}101T_s$, acoustic sensor A_2 receives additional new measurement from object O^2. At time $1{,}300T_s$, acoustic sensor A_2 receives additional new measurement from object O^3. At time $1{,}500T_s$, acoustic sensor A_1 receives additional new measurement from object O^2, but the measurement from object O^1 is not received simultaneously. At time $1{,}600T_s$, acoustic sensor A_1 receives additional new measurement from object O^3. Figure 21 shows the triggering probabilities of the two sensors between $1{,}100T_s$ and $1{,}900T_s$. Figures 21(a) and 21(b) show triggering probabilities with the visual sensor cooperation in sensors A_1 and A_2, respectively. Also, for the comparison, Figures 21(c) and 21(d) show the triggering probabilities without the visual sensor cooperation in sensors A_1 and A_2, respectively,

Finally, Figure 22 shows the estimated final position of the three objects in each sensor. Figure 22(a) shows the final estimated position with acoustic sensor A_1, and Figure 22(b) shows the final estimated position with acoustic sensor A_2.

5. Conclusion and Final Remarks

In this paper, the acoustic-visual sensor cooperation method for multiple object tracking was presented. Since the visual sensor-based object localization requires much higher computational complexity than the acoustic sensor-based estimation, minimized visual sensor cooperation is adopted throughout this paper. The visual sensor cooperation method was proposed based on the analysis of the limitation in the acoustic sensor-based estimation. In order to alleviate the limitation, the visual sensor is triggered for the cooperation. For comparison, the proposed acoustic-visual sensor cooperation method was evaluated with a periodic visual sensor cooperation and the no cooperation case. Finally, the cooperation method was verified in a real environment.

As a future work, the cooperation method can be extended to a large-scale environment. Since an acoustic sensor has a limited coverage as well as a limited capacity in measuring the sound wave, it is required to deploy multiple acoustic sensors to cover a large area. We investigate the effects of interaction among multiple acoustic sensors. In addition, we analyze the effect of visual sensor cooperation delay time since the visual sensor and the acoustic sensor receive measurements with different sampling rates, and there are synchronization issues between the two sensors.

Acknowledgments

This work was in part supported by the Mid-career Researcher Program of Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2010-0000487).

References

[1] Y. Bar-Shalom and X. R. Li, *Estimation and Tracking: Principles, Techniques and Software*, Artech House, Norwood, Mass, USA, 1993.

[2] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filter for Tracking Application*, Artech House, Boston, Mass, USA, 2002.

[3] M. Stanaćević and G. Cauwenberghs, "Micropower gradient flow acoustic localizer," *IEEE Transactions on Circuits and Systems I*, vol. 52, no. 10, pp. 2148–2157, 2005.

[4] R. J. Kozick and B. M. Sadler, "Source Localization With Distributed Sensor Arrays and Partial Spatial Coherence," *IEEE Transactions on Signal Processing*, vol. 52, no. 3, pp. 601–616, 2004.

[5] G. Jacovith and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 525–533, 1993.

[6] D. D. Feldman and L. J. Griffiths, "Projection approach for robust adaptive beamforming," *IEEE Transactions on Signal Processing*, vol. 42, no. 4, pp. 867–876, 1994.

[7] T. Kirubarajan, Y. Bar-Shalom, and D. Lerro, "Bearings-only tracking of maneuvering targets using a batch-recursive estimator," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 3, pp. 770–780, 2001.

[8] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[9] J. Lee, S. H. Cho, S. Hong, and W. D. Cho, "Multitarget tracking (MTT) in 3-D using 2-D particle filters with single passive sensor," in *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems (MWSCAS '07)*, pp. 389–392, August 2007.

[10] J. Lim, J. Lee, S. Hong, and P. Park, "Algorithm for detection with localization of multi-targets in wireless acoustic sensor networks," in *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '06)*, pp. 547–554, November 2006.

[11] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1154–1164, 2002.

[12] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 38, no. 3, pp. 799–807, 2008.

[13] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Person tracking with audio-visual cues using the iterative decoding framework," in *Proceedings of the IEEE 5th International Conference on Advanced Video and Signal Based Surveillanc (AVSS '08)*, pp. 260–267, September 2008.

[14] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[15] V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 715–726, 2007.

[16] K.-Y. Chow, K.-S. Lui, and E. Y. Lam, "Efficient on-demand image transmission in visual sensor networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 95076, 11 pages, 2007.

[17] D. K. Park, H. S. Yoon, and C. Sun Won, "Fast object tracking in digital video," *IEEE Transactions on Consumer Electronics*, vol. 46, no. 3, pp. 785–790, 2000.

[18] M. Kushwaha, S. Oh, I. Amundson, X. Koutsoukos, and A. Ledeczi, "Target tracking in heterogeneous sensor networks using audio and video sensor fusion," in *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI '08)*, pp. 14–19, August 2008.

[19] A. O'Donovan, R. Duraiswami, and D. Zotkin, "Imaging concert hall acoustics using visual and audio cameras," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 5284–5287, April 2008.

[20] W. R. Gilks and C. Berzuini, "Following a moving target—Monte Carlo inference for dynamic Bayesian models," *Journal of the Royal Statistical Society B*, vol. 63, no. 1, pp. 127–146, 2001.

[21] P. M. Djurić, J. H. Kotecha, J. Zhang et al., "Particle filtering," *IEEE Signal Processing Magazine*, vol. 20, no. 5, pp. 19–38, 2003.

[22] J. Carpenter and P. Clifford, "Improved particle filter for nonlinear problems," *IEE Proceedings: Radar, Sonar and Navigation*, vol. 146, no. 1, pp. 2–7, 1999.

[23] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[24] M. S. Arulampalam, B. Ristic, N. Gordon, and T. Mansell, "Bearings-only tracking of manoeuvring targets using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 15, pp. 2351–2365, 2004.

[25] A. Doucet, N. J. Gordon, and V. Krishnamurthy, "Particle filters for state estimation of jump Markov linear systems," *IEEE Transactions on Signal Processing*, vol. 49, no. 3, pp. 613–624, 2001.

[26] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part I. Dynamic models," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, 2003.

[27] M. Bolić, P. M. Djurić, and S. Hong, "Resampling algorithms and architectures for distributed particle filters," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2442–2450, 2005.

[28] Y. Boers and J. N. Driessen, "Interacting multiple model particle filter," *IEE Proceedings: Radar, Sonar and Navigation*, vol. 150, no. 5, pp. 344–349, 2003.

[29] R. Velmurugan, S. Subramanian, V. Cevher et al., "On low-power analog implementation of particle filters for target tracking," in *Proceedings of the 14th European Signal Processing Conference (EUSIPCO '06)*, September 2006.

[30] M. Yeddanapudi, Y. Bar-Shalom, and K. R. Pattipati, "IMM estimation for multitarget-multisensor air traffic surveillance," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 80–94, 1997.

[31] G. Peremans, K. Audenaert, and J. M. Van Campenhout, "High-resolution sensor based on tri-aural perception," *IEEE Transactions on Robotics and Automation*, vol. 9, no. 1, pp. 36–48, 1993.

[32] C. Hue, J.-P. Le Cadre, and P. Pérez, "Tracking multiple objects with particle filtering," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 791–812, 2002.

[33] K.-S. Park, J. Lee, M. Stanaćević, S. Hong, and W.-D. Cho, "Iterative object localization algorithm using visual images with a reference coordinate," *EURASIP Journal on Image and Video Processing*, vol. 2008, Article ID 256896, 16 pages, 2008.

[34] J. Lee, K.-S. Park, S. Hong, and W.-D. Cho, "Object tracking based on rfid coverage visual compensation in wireless sensor network," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '07)*, pp. 1597–1600, May 2007.

[35] M. Han, A. Sethi, W. Hua, and Y. Gong, "A detection-based multiple object tracking method," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, vol. 5, pp. 3065–3068, October 2004.

[36] E. Hjelmås and B. K. Low, "Face detection: a survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.