Local and Global Information Exchange for Enhancing Object Detection and Tracking

Jinseok Lee[1], Shung Han Cho[2], Seong-Jun Oh[3] and Sangjin Hong[2]
[1]Dept. of Biomedical Engineering,
Worcester Polytechnic Institute, Worcester, MA 01609 - USA
[2]Mobile Systems Design Laboratory, Dept. of Electrical and Computer Engineering,
Stony Brook University-SUNY, Stony Brook, NY 11794 - USA
[3]College of Information and Communications, Korea University,
Seoul, 136-713 - South Korea
[e-mail: jinseok@wpi.edu , {shcho, snjhong}@ece.sunysb.edu sjoh@korea.ac.kr]
*Corresponding author: Seong-Jun Oh

R*eceived October 30, 2011; revised January 17, 2012; accepted February 13, 2012; published May 25, 2012*

Abstract

Object detection and tracking using visual sensors is a critical component of surveillance systems, which presents many challenges. This paper addresses the enhancement of object detection and tracking via the combination of multiple visual sensors. The enhancement method we introduce compensates for missed object detection based on the partial detection of objects by multiple visual sensors. When one detects an object or more visual sensors, the detected object's local positions transformed into a global object position. Local and global information exchange allows a missed local object's position to recover. However, the exchange of the information may degrade the detection and tracking performance by incorrectly recovering the local object position, which propagated by false object detection. Furthermore, local object positions corresponding to an identical object can transformed into nonequivalent global object positions because of detection uncertainty such as shadows or other artifacts. We improved the performance by preventing the propagation of false object detection. In addition, we present an evaluation method for the final global object position. The proposed method analyzed and evaluated using case studies.

Keywords: visual sensor, localization, distributed detection system, sensor network, data combination, information exchange

This research was supported partly by the International Collaborative R&D Program of the Ministry of Knowledge Economy (MKE), the Korean government, as a result of Development of Security Threat Control System with Multi-Sensor Integration and Image Analysis Project, 2010-TD-300802-002.
This research was also supported partly by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency (NIPA-2012-(C1090-1221-0011)).

http://dx.doi.org/10.3837/tiis.2012.05.009

1. Introduction

Visual sensor-based surveillance systems are of great interest to a diversity of fields, and many researchers have tried to enhance the object detection, tracking, and localization performance [1][2][3][4]. In particular, object detection and tracking with a visual sensor is a critical component when evaluating a complete surveillance system and is also a challenging problem [5][6][7]. Difficulties occur with object detection and tracking because of abrupt object motion, variable lighting conditions, the changing appearance of an object and its background, non-rigid object structures, object-to-object occlusions, and object-to-background occlusions. Many approaches and algorithms have been proposed to overcome these problems. To address the lighting problem, [8] combined color and gradient features during quasi-illumination invariant background subtraction. [9] applied time-varying reflectance images and their corresponding illumination images to a sequence of images, which was followed by a normalization process. [10] used a probabilistic method for adaptive background subtraction, which produced a stable, real-time tracker that dealt with lighting changes reliably. Recently, [11] developed background subtraction algorithms, which particularly targeted rapid illumination changes. To address the changing appearance problem, [12] adaptively selected object features to discriminate an object from a background more effectively. [13] proposed a method with a training phase that learned an object's geometry and appearance using a randomized tree classifier. Recently, [14] proposed a method for tracking objects with a changing appearance based on a sparse, local feature-based object representation. In this context, a dynamic model was proposed to evolve a feature graph that was dependent on the appearance and structure changes by adding new stable features, as well as removing inactive features. To address the occlusion problem, [15] predicted occlusion by searching for the pairwise overlapping of bounding boxes in the predicted positions. [16] produced an occlusion map, where potentially occluding pixels were detected if they were part of the reference image. Recently, [17] developed a method that generated a probability density function for the depth of a scene at each pixel using a training set of detected blobs. Furthermore, to represent object characteristics more accurately, [1] used a library containing a set of counterexamples, while [2] and [3] used a more detailed construction that corresponded to an object.

However, single visual sensor-based object detection/tracking still has limitations with complex scenes such as underground stations and malls. [18] presented a comprehensive survey of object detection based on an object's motion and behavior. It has been suggested that the most promising and practical method for overcoming the shortcomings of visual sensor-based surveillance systems is to use multiple visual sensors. Multiple visual sensor-based object detection and tracking can be conducted as follow. If an object is present, information about the object and its position is generated locally by each visual sensor, before the local information is combined at a global information center. Based on information derived from multiple local visual sensors, which can be inconsistent among the local visual sensors, the global information center makes a single global decision about the existence of an object and it continues to track the object when it is detected. The advantage of using multiple visual sensors is that when an object is detected and tracked by one or more visual sensors, a missing local object position can be recovered by the global information center based on the detection of one or a few objects using local visual sensor(s). An object can also be tracked by a visual

sensor that did not detect it originally, because it can be recovered from the global information center.

The combination of a diverse range of local information, however, does not always provide a better performance. This is because the combination of incorrect information by the global information center may degrade the performance by propagating the detection of a false object. For example, when some visual sensors falsely detect an object, the falsely detected object's position is propagated to other non-detecting visual sensors. Furthermore, a detected local object's position has uncertainty, even if it represents a true identical object. In an outdoor environment, for example, a change detection algorithm may detect an object together with its shadow [19]. The local object positions corresponding to an identical object are then transformed into nonequivalent global object positions. The nonequivalent global object positions may be recognized as multiple objects.

In this paper, we propose a method for combining local information from multiple visual sensors. Initially, we formulate a true/false decision problem for object detection and we present an optimal decision fusion method, which uses quality information for each local visual sensor. The use of quality information can minimize the propagation of falsely detected objects. The global object position identified by the global information center is also transferred to all the local visual sensors, and each visual sensor can track the object better based on the global object position. Thus, the global information center tracks the detected object based on the weighted sum of the local positions reported by the visual sensors. The main concept is the exchange of local and global information between local visual sensors and the generation of quality information from local visual sensors.

The remainder of this paper is organized as follow. In Section 2, we introduce an application of object detection enhancement based on multiple sensors and the focal problems are described. In Section 3, we focus on a case where local information from multiple visual sensors is transferred to the global decision fusion center and the global decision is made based only on local information without information exchange. Local and global information exchange is discussed in Section 4, as well as the performance analysis. Finally, our conclusions are summarized in Section 5.

2. System Model

2.1 Application

Fig. 1 shows a model of an application for the enhancement of object detection, where multiple visual sensors share an overlapping viewable range. If a visual sensor detects an object, e.g., visual sensor 1, the local information detected is transformed into a global coordinate at the global position fusion center. The global coordinate is then re-transformed into local information for all the visual sensors and any missing object detections can be recovered, i.e., by visual sensors 2 and 3 in the example. Thus, local object tracking by visual sensors 2 and 3 can improve improve the overall object tracking performance, even if visual sensors 2 and 3 do not participate in tracking the object.

2.2 Problem Description

Denote V^j as the *j-th* visual sensor, where $j = 1, 2, \cdots, J$, where J is the number of visual sensors. If an object is detected by V^{j}, we denote the local position (information) of the detected object viewed by V^j as $\mathbf{l}^{\mathrm{j}}(n) = \left(l_x^j(n),\ l_y^j(n)\right)$, where n is the discrete-time index.

The local object position $\mathbf{l}^j(n)$ is transformed into a global object position in the global coordinates, which is denoted as $\mathbf{g}^j(n) = \left(g_x^j(n), g_y^j(n)\right)$. A perspective model is used for the surveillance system [20] during this local to global transformation. The global object position $\mathbf{g}^j(n)$ of object j at time index n, can be re-transformed into the local object position, which is denoted as $\tilde{\mathbf{l}}^j(n)$ or $\left(\tilde{l}_x^j(\mathrm{n}), \tilde{l}_y^j(n)\right)$. The local-global-local transformation using multiple visual sensors allows any missed local object position $\mathbf{l}^j(n)$ to be recovered by $\tilde{\mathbf{l}}^j(n)$, provided at least one visual sensor detects an object.

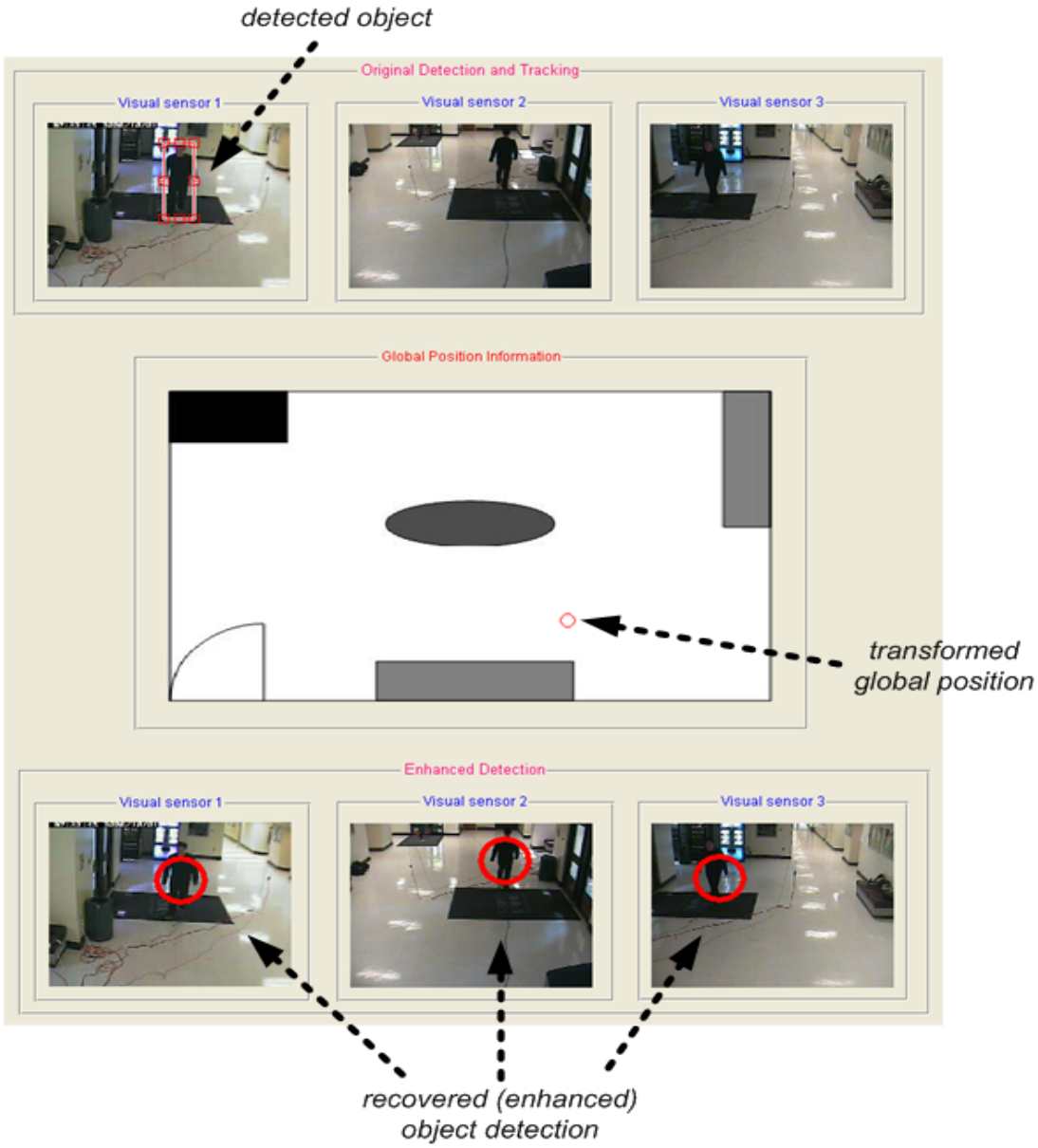


Fig. 1. Application model for the enhancement of an object's detection and tracking.

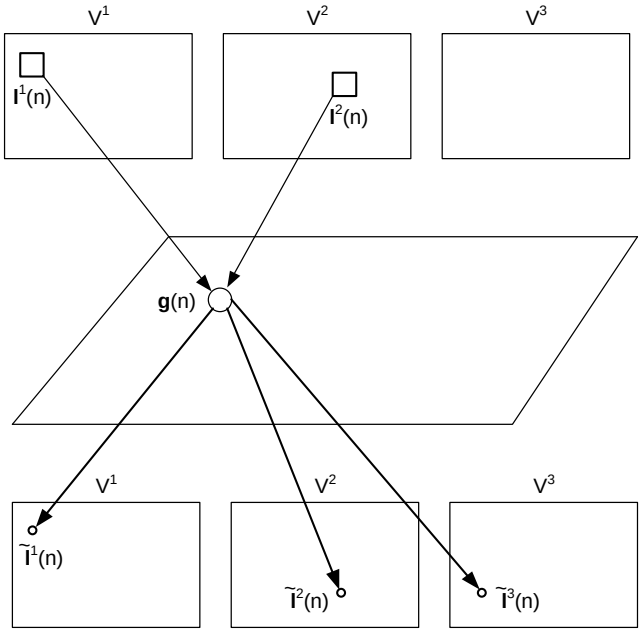


Fig. 2. If $\mathbf{V^1}$ detects an object but $\mathbf{V^2}$ and $\mathbf{V^3}$ do not detect an object, $\tilde{\mathbf{I}}^{\mathbf{2}}(\mathbf{n})$ and $\tilde{\mathbf{I}}^{\mathbf{3}}(\mathbf{n})$ are recovered based on the global information.

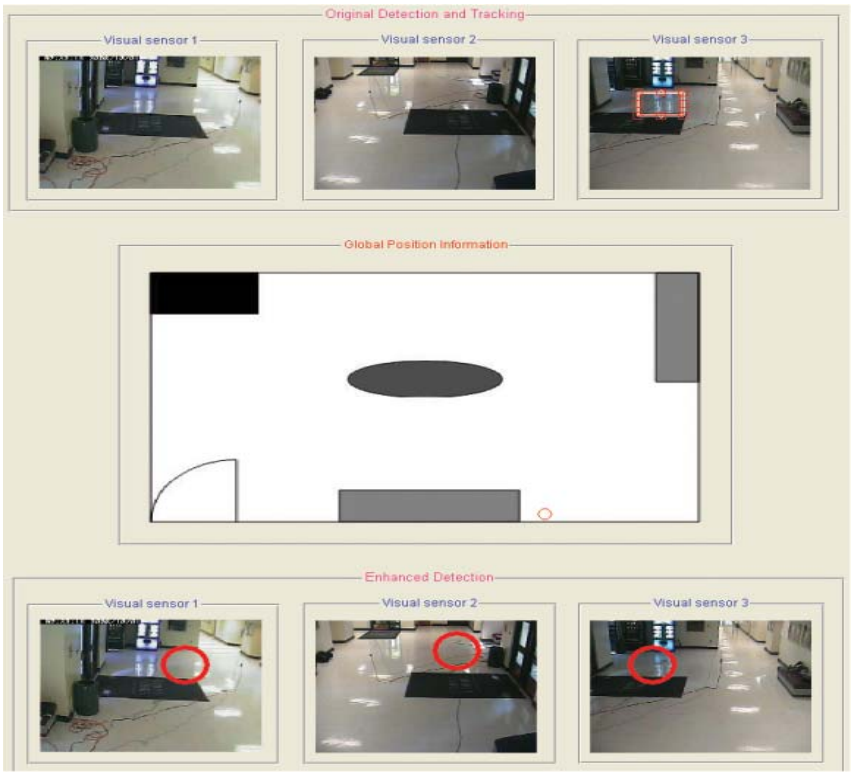


Fig. 3. False object detection propagation during local and global information exchange.

Local and global information exchange can support missed detection recovery, but there is a risk of increased false detection propagated by a false local object position. **Fig. 2** shows that V^1 detects an object, whereas V^2 and V^3 do not. Therefore, the only detected local object position $\mathbf{l}^1(n)$ is transformed into $\mathbf{g}^1(n)$, which is re-transformed into the local object positions $\tilde{\mathbf{l}}^1(\mathrm{n}), \tilde{\mathbf{l}}^2(\mathrm{n})$ and $\tilde{\mathbf{l}}^3(\mathrm{n})$. Given the condition that an object is present, the missed object detections by V^2 and V^3 are recovered using $\tilde{\mathbf{l}}^2(\mathrm{n})$ and $\tilde{\mathbf{l}}^3(\mathrm{n})$. However, when no actual object (i.e. $\mathrm{l}^1(n)$ is obtained from a false detection), $\tilde{\mathbf{l}}^2(\mathrm{n})$ and $\tilde{\mathbf{l}}^3(\mathrm{n})$ are falsely (unnecessarily) recovered. Thus, the information exchange increases the number of false objects detected as shown in **Fig. 3**. We also need to consider minimizing the propagation of false object detection.

If multiple visual sensors simultaneously detect an object, nonequivalent global object positions can be obtained, as shown in **Fig. 4**. More specifically, the local object positions, $\mathbf{l}^j(n)$'s, corresponding to a single object are transformed differently due to detection uncertainty (i.e., $\mathbf{l}^1(n)$ and $\mathbf{l}^3(n)$ are transformed into $\mathbf{g}^1(n)$ and $\mathbf{g}^3(n)$). Thus, a false local object position such as $\mathbf{l}^2(n)$, increases the confusion by finding a final global object position. Therefore, the final global object position denoted as $\mathbf{g}(n)$ or $\left(g_x(n), g_y(n)\right)$ should be evaluated based on $\mathbf{g}^j(n)$ for all V^j's that detect an object. This can be considered as an assignment problem given that some local object positions represent a true object whereas the others represent a false object. Furthermore, although $\mathbf{l}^i(n)$ and $\mathbf{l}^j(n)$ represent an identical true object, $\mathbf{g}^i(n)$ and $\mathbf{g}^j(n)$ do not necessarily coincide for $i \neq j$.

Throughout this paper, we minimize the false detection based on the recovery of incorrect local object position by filtering out the false detection. In addition, we find a single global object position given a mixture of true and/or false detections, so the local object positions $\tilde{\mathbf{l}}^j(n)$ are enhanced in comparison with the original local object positions $\mathbf{l}^j(n)$. We make use of the quality information during detection, where each visual sensor's coordinates are considered to have a confidence level, such that a global coordinate with a higher confidence level is attributed a higher emphasis. Note that in the rest of this paper, exchangeability refers to each local and global object position such as $\mathbf{l}^j(n)$ or $(l_x^j(n), l_y^j(n))$, $\mathbf{l}^j(n)$ or $(\tilde{l}_x^j(n), \tilde{l}_y^j(n))$, $\mathbf{g}^j(n)$ or $(g_x^j(n), g_y^j(n))$, and $\mathbf{g}(n)$ or $(g_x(n), g_y(n))$.

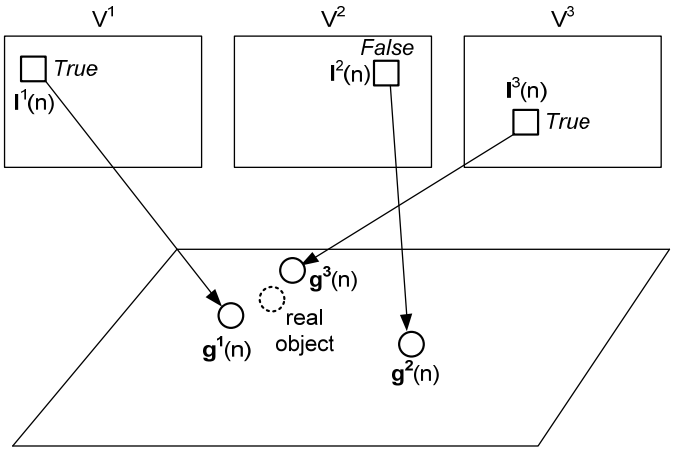


Fig. 4. Illustration of local and global information exchange based on a mixture of true and/or false object detection with nonequivalent local to global transformation.

3. Theory of Global Decisions based on Local Information

3.1 Binary Decision Problem based on Binary Local Information

Fig. 5 shows the distributed detection system and the decision fusion center, where information from local visual sensors is fused to make a global decision. For an object, all signal processing is conducted locally at each sensor V^j and each local decision is available as E^j, where $E^j \in \{0,1\}$ for $j = 1,2, \cdots, J$. More specifically, a binary hypothesis testing problem is formulated as

$$H_0: an\ object\ is\ absent \quad \text{and} \quad H_1: an\ object\ is\ present,$$

while E^j is the local decision described as follow,

$$E^j = 0 \quad if\ H_0\ is\ declared\ by\ V^j \text{ and } E^j = 1 \quad if\ H_1\ is\ declared\ by\ V^j.$$

The *apriori* probabilities of the two hypotheses are denoted as $P(H_0) = P_0$ and $P(H_1) = P_1$. We assume that J observations from J sensors are conditionally independent, where the conditional probability is denoted by $P(E^j|H)$. More specifically, $P(E^j = 0|H_1)$ represent the probability that an object is missed, while $P(E^j = 1|H_0)$ represents the probability that a false source is detected. In addition, $P(E^j = 1|H_1)$ represents the probability that an object is detected when there is actually an object, while $P(E^j = 0|H_0)$ represents the probability that a false source is not detected. We denote the false and missing detection probabilities as P_M^j and P_F^j, where the probabilities are equivalent to the following conditional probabilities:

$$P_M^j = P(E^j = 0|H_1) \quad and \quad P_F^j = P(E^j = 1|H_0). \tag{1}$$

After processing the source detections locally, a global decision is available in the decision fusion center as E by collecting all E^j's, where $E \in \{0,1\}$. The global decision E is decided from $D(E^1, E^2, \cdots, E^J)$ where $E = 0(1)$ if $H_0(H_1)$ is declared in the decision fusion center. The optimal decision rule begins with the maximum *a posteriori* probability (MAP) decision rule, which is equivalent to a minimum probability error (MPE) decision rule. The corresponding likelihood ratio test (LRT) is formulated as

$$D(E^1, E^2, \cdots, E^J) = \frac{P(E^1, E^2, \cdots, E^J|H_1)}{P(E^1, E^2, \cdots, E^J|H_0)} \underset{E=0}{\overset{E=1}{\gtrless}} \frac{P_0}{P_1}. \tag{2}$$

Based on this specification, the optimal decision fusion rule is that shown in (3) from [21], when only binary decisions from local sensors are available.

$$D(E^1, E^2, \cdots, E^J) = \log\frac{P_1}{P_0} + \sum_{S_{E_P}} \log\frac{1-P_M^j}{P_F^j} - \sum_{S_{E_N}} \log\frac{1-P_F^j}{P_M^j} \underset{E=0}{\overset{E=1}{\gtrless}} 0, \tag{3}$$

where S_{E_P} is the set of all j such that $E^j = 1$ and S_{E_N} is the set of all j where $E^j = 0$. Given the following conditional probabilities

$$P_M \simeq P_M^j \simeq P_M^i \; and \; P_F \simeq P_F^j \simeq P_F^i, where \; i \neq j, \; i,j = 1,\ldots J, \tag{4}$$

the optimal decision is simplified as

$$E \simeq D(E^1,\ldots,E^J) \simeq \log\frac{P_1}{P_0} + N(S_{E_P}) \cdot \log\frac{1-P_M}{P_F} - N(S_{E_N}) \cdot \log\frac{1-P_F}{P_M}, \tag{5}$$

where $N(S_{E_P})$ is the element number of a set S_{E_P} while $N(S_{E_N})$ is the element number of a set S_{E_N}. In addition, when P_M is approximately equal to P_F, the optimal decision is

$$E \simeq D(E^1,\ldots,E^J) \simeq \log\frac{P_1}{P_0} + \left(N(S_{E_1}) - N(S_{E_0})\right) \cdot \log\frac{1-P_M}{P_F}. \tag{6}$$

Note that an implication of (4) is that identical performances are derived from the individual sensors and, given that assumption, it is natural that the global decision is based on the number of sensors that declare an object's detection.

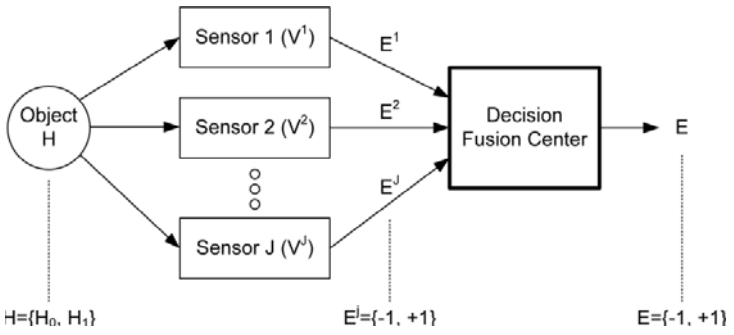


Fig. 5. Distributed detection system with a decision fusion center.

3.2 Binary Decision Problem based on Quality Local Information

In a case where the local visual sensors only report their binary decisions, the decision is mainly dependent on the difference between $N(S_{E_P})$ and $N(S_{E_N})$, if the performance of the multiple sensors is identical. If $N(S_{E_P}) > N(S_{E_N})$, the decision fusion center is biased into making a decision, $E = 1$, regardless of how the information quality corresponding to each local decision E^j, where $j \in S_{E_P}$, is degraded. Similarly, if $N(S_{E_N}) > N(S_{E_P})$, the decision fusion center is biased to make the opposite decision, $E = 0$, regardless of whether the quality information corresponding to each local decision E^j, where $j \in S_{E_P}$, is enough to indicate the existence of an object. The overall decision can be improved if each visual sensor reports the local decision and its confidence level to the decision fusion center, or the confidence levels of each visual sensor are computed by the decision fusion center.

The confidence level of the decision, the quality information, indicates the degree of confidence when $E^j = 1$, which is denoted by W^j, where $j \in S_{E_P}$ and $0 \leq W^j \leq 1$. Note that W^j does not exist when $E^j = 0$. When W^j is close to zero, E^j has *less confidence*. However, when W^j is close to one, E^j has *more confidence*. **Fig. 6** shows two possible optimal decision flows, which are classified based on the quality information availability. **Fig. 6(a)** shows that each sensor makes a local decision, before evaluating the corresponding quality information together. However, **Fig. 6(b)** shows that each sensor makes a local decision only, while the

quality evaluator supports the quality information.

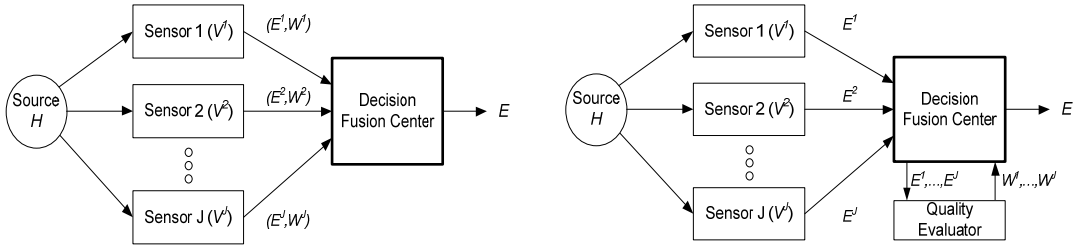


Fig. 6. Distributed detection system using an information fusion center.

If we suppose that the quality information W^j, $j = 1,2,\cdots,J$, is available, an optimum decision can also be made also using the MAP rule as follows:

$$P(H_1|E^1,\ldots,E^J,W^1,\ldots,W^J) \underset{H_0}{\overset{H_1}{\gtrless}} P(H_0|E^1,\ldots,E^J,W^1,\ldots,W^J), \tag{7}$$

where W^j indicates the degree of confidence for $E^j = 1$, where E^j is weighted by each corresponding quality information W^j as follows:

$$P(H_1|E^1 \cdot W^1,\ldots,E^J \cdot W^J) \underset{H_0}{\overset{H_1}{\gtrless}} P(H_0|E^1 \cdot W^1,\ldots,E^J \cdot W^J). \tag{8}$$

Based on Bayes' rule, the LRT is described as follow:

$$\frac{P(E^1 \cdot W^1,\ldots,E^J \cdot W^J|H_1)}{P(E^1 \cdot W^1,\ldots,E^J \cdot W^J|H_0)} = \begin{cases} > \frac{P_0}{P_1} \Rightarrow E = 1 \\ < \frac{P_0}{P_1} \Rightarrow E = 0. \end{cases} \tag{9}$$

For simplicity, we denote $E^j \cdot W^j$ by E_w^j, and the collection of the quality information $\{E_w^1,\ldots,E_w^J\}$ by $E_w^{1:J}$, while the left-hand side of (10) is simplified and decomposed as

$$\frac{P(E_w^{1:J}|H_1)}{P(E_w^{1:J}|H_0)} = \underbrace{\prod_{S_{E_N}} \frac{P(E_w^j|H_1)}{P(E_w^j|H_0)}}_{A} \cdot \underbrace{\prod_{S_{E_P}} \frac{P(E_w^j|H_1)}{P(E_w^j|H_0)}}_{B}, \tag{10}$$

where **A** only relates to E^j given the condition $E^j = 0$, while **B** relates to both E^j and W^j given the condition $E^j = 1$. Thus, the **A** for (11) is

$$\prod_{S_{E_0}} \frac{P(E_w^j|H_1)}{P(E_w^j|H_0)} = \prod_{S_{E_0}} \frac{P(E^j=0|H_1)}{P(E^j=0|H_0)} = \prod_{S_{E_0}} \frac{P_M^j}{1-P_F^j}, \tag{11}$$

while **B** for (11) is

$$\prod_{S_{E_1}} \frac{P(E_w^j|H_1)}{P(E_w^j|H_0)} = \prod_{S_{E_1}} \frac{P(E^j=1|H_1)\cdot W^j}{P(E^j=1|H_0)\cdot(1-W^j)} = \prod_{S_{E_1}} \frac{\left(1-P_F^j\right)\cdot W^j}{P_F^j\cdot\left(1-W^j\right)}. \quad (12)$$

By substituting (12) and (13) and into (11), the log-LRT is

$$\log D(E_w^{1:J}) = \log\frac{P_1}{P_0} - \sum_{S_{E_0}} \log\frac{1-P_F^j}{P_M^j} + \sum_{S_{E_P}} \log\frac{1-P_F^j}{P_F^j} + \sum_{S_{E_P}} \log\frac{W^j}{1-W^j} = \begin{cases} > 0 \Rightarrow E = 1 \\ < 0 \Rightarrow E = 0 \end{cases}. \quad (13)$$

By considering the quality information W^j corresponding to the local decision E^j, we reduce the decision bias as to the difference between $N(S_{E_P})$ and $N(S_{E_N})$.

3.3 Data Combination Problem based on Quality Local Information

We focus on making a better decision for E using the quality information W^j. However, the *apriori* probabilities of the two hypotheses, P_1 and P_0, remain unsolved in (16). Furthermore, we should consider evaluating the global object position $\mathbf{g}$ to support missing detection recovery, as shown in **Fig. 1**. Thus, $\mathbf{g}$ is transformed into $\tilde{\mathbf{l}}^j$, i.e., the missing detection recovery. However, based on multiple visual sensors, the multiple detected local object positions $\mathbf{l}^j$ are transformed into multiple nonequivalent global object positions $\mathbf{g}^j$, $j \in N(S_{E_P})$. Note that $\mathbf{l}^j$, where $j \in S_{E_N}$, does not exist.

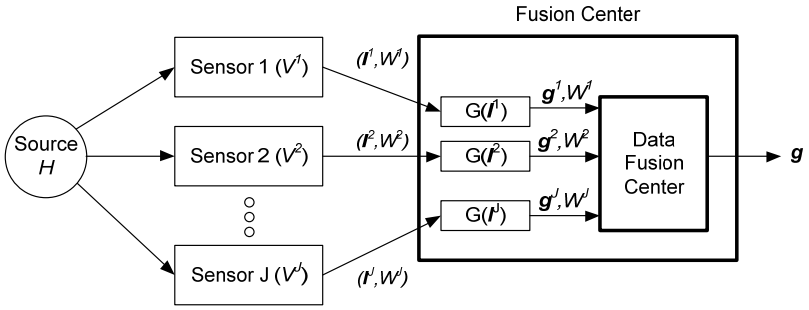


Fig. 7. Distributed detection system based on the data fusion center.

Fig. 7 shows that each collected and transformed global object position $\mathbf{g}^j$, where $j \in S_{E_P}$, is combined to produce the final result, $\mathbf{g}$, using the data fusion center. The role of the data fusion center is to find $\mathbf{g}$ among $\mathbf{g}^j$'s based on the quality information W^j as

$$\mathbf{g} = U(\mathbf{g}^{1:J}, W^{1:J}), \quad (14)$$

where $\mathbf{g}^{1:J}$ represents the set $\{\mathbf{g}^1, \ldots, \mathbf{g}^j, \ldots \mathbf{g}^J\}$, $j \in S_{E_P}$. The final result, $\mathbf{g}$, is obtained by weighting the quality information W^j on $\mathbf{g}^j$, as follow:

$$\mathbf{g} = U(\mathbf{g}^{1:J}, W^{1:J}) = \frac{\sum_{j\in S_{E_P}} W^j\cdot\mathbf{g}^j}{\sum_{k\in S_{E_P}} W^k}. \quad (15)$$

4. Local and Global Information Exchange Algorithm

In order to obtain the quality information, we first define a global object state $\hat{\mathbf{g}}(n)$ as

$$\hat{\mathbf{g}}(n) = \left[\hat{g}_x(n)\hat{g}_{vx}(n)\hat{g}_y(n)\hat{g}_{vy}(n)\right]^T \tag{16}$$

where $\left[\hat{g}_x(n)\hat{g}_y(n)\right]$ and $\left[\hat{g}_{vx}(n)\hat{g}_{vy}(n)\right]$ are the true global object positions and velocities at time n. Given the global object state $\hat{g}(n)$ at a discrete time instant $n \in \{1,2,\cdots\}$, this evolves according to

$$\hat{\mathbf{g}}(n) = \mathbf{F}(n-1) \cdot \hat{\mathbf{g}}(n-1) + \mathbf{Q} \tag{17}$$

where $\mathbf{Q}$ includes the Gaussian noise for an object position described as

$$\mathbf{Q} = [N(0,\sigma^2) \quad 0 \quad N(0,\sigma^2) \quad 0]^T \tag{18}$$

where $\mathbf{F}(n-1)$ is a dynamic transition function for $\hat{\mathbf{g}}(n-1)$. More dynamic transition functions are introduced in [22].

From the perspective of Bayesian estimation, the posterior probability density function (PDF) for $\mathbf{g}(n)$ is estimated by propagating the PDF over time [1]:

$$p(\mathbf{g}(n)|Z(1:n)) \propto p(Z(n)|\mathbf{g}(n)) \cdot p(\mathbf{g}(n)|Z(1:n-1)), \tag{19}$$

where $Z(n)$ represents a measurement at time n, and $Z(1:n)$ represents the history of measurements up to time n. Note that we use the time notation n in this section. Generally, the measurement term depends on the type of sensor and the application (i.e., TDE, signal strength and/or bearings from acoustic sensors [24][25][26]). In this paper, the measurement $Z(n)$ is replaced by $\bar{\mathbf{g}}(n)$, which is obtained as

$$\bar{\mathbf{g}}(n) = \mathbf{F}(n-1) \cdot \mathbf{g}(n-1), \tag{20}$$

where $\mathbf{g}(n-1)$ is the final global object state at time $n-1$, and $\bar{\mathbf{g}}(n)$ represents $\left[\bar{g}_x(n)\bar{g}_{vx}(n)\bar{g}_y(n)\bar{g}_{vy}(n)\right]^T$. Given that an object follows the dynamic model $\mathbf{F}(n-1)$, the weight $w^j(n)$ corresponding to $\mathbf{g}^j(n)$ is evaluated when $\mathbf{g}^j(n)$ is close to $\bar{\mathbf{g}}(n)$ as

$$w^j(n) = \exp\left(-\left\{\frac{(g_x^j(n)-\bar{g}_x(n))^2}{2\sigma^2} + \frac{(g_y^j(n)-\bar{g}_y(n))^2}{2\sigma^2}\right\}\right), \tag{21}$$

where $w^j(n)$ is obtained based on the 2-D Gaussian distribution function, which denotes the probability that an object corresponding to $\mathbf{g}^j(n)$ and $\mathbf{l}^j(n)\big|$ follow the dynamic model $\mathbf{F}(n-1)$. Given that an object moves according to a given dynamic model, $w^j(n)$ represents the quality information for $\mathbf{g}^j(n)$ and $\mathbf{l}^j(n)$: $W^j(n) = w^j(n)$.

The associated set $\{\mathbf{g}^j(n), w^j(n)\}$ approximates the posterior pdf $p(\mathbf{g}(n)|\bar{\mathbf{g}}(1:n))$ as [1]

$$p\left(\mathbf{g}(n)\middle|\bar{\mathbf{g}}(1:n)\right); \sum_{j=1}^{N(S_{E_1})} \frac{w^j(n)}{\sum_{k=1}^{N(S_{E_1})} w^k(n)} \cdot \delta\left(\bar{\mathbf{g}}(n) - \mathbf{g}^j(n)\right). \tag{22}$$

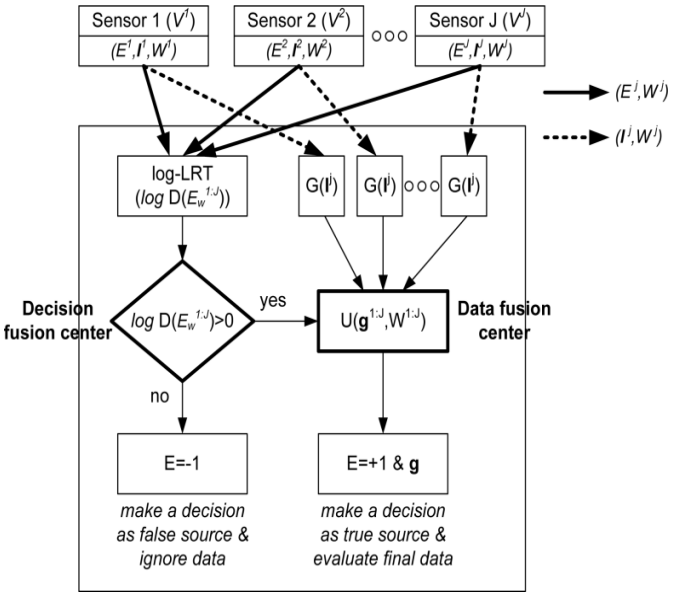


Fig. 8. Illustration of the overall collaborative data flow for the object detection enhancement.

Each component of $\mathbf{g}^j(n)$ is weighted and averaged for the final global object state $\mathbf{g}(n)$, which is based on a probability data association method (PDA) [23]. Thus, $g_x^j(n)$ and $g_y^j(n)$ contribute to the final global object position for each corresponding $w^j(n)$ as

$$g_x(n) = \frac{\sum_{S_{E_1}} g_x^j(n) w^j(n)}{\sum_{S_{E_1}} w^j(n)}, \quad g_y(n) = \frac{\sum_{S_{E_1}} g_y^j(n) w^j(n)}{\sum_{S_{E_1}} w^j(n)}. \tag{23}$$

When the final global object position $(g_x(n), g_y(n))$ is obtained, the position is evaluated, where $(g_x(n), g_y(n))$ is close to $(\bar{g}_x(n), \bar{g}_y(n))$ as

$$w(n) = \exp\left(-\left\{\frac{(g_x(n)-\bar{g}_x(n))^2}{2\sigma^2} + \frac{(g_y(n)-\bar{g}_y(n))^2}{2\sigma^2}\right\}\right). \tag{24}$$

The evaluated $w(n)$ also denotes the probability that an object follows the dynamic model $\mathbf{F}(n-1)$. If $w(n)$ is close to zero, the object is totally deviated from the position based on the dynamic model. However, if $w(n)$ is close to one, the object follows the dynamic model completely. Given a condition where an object moves according to a dynamic model, $w(n)$ represents the *apriori* probability $P_1(n)$. Thus, the *apriori* probabilities are

$$P_1(n) = w(n) \quad and \quad P_0(n) = 1 - w(n). \tag{25}$$

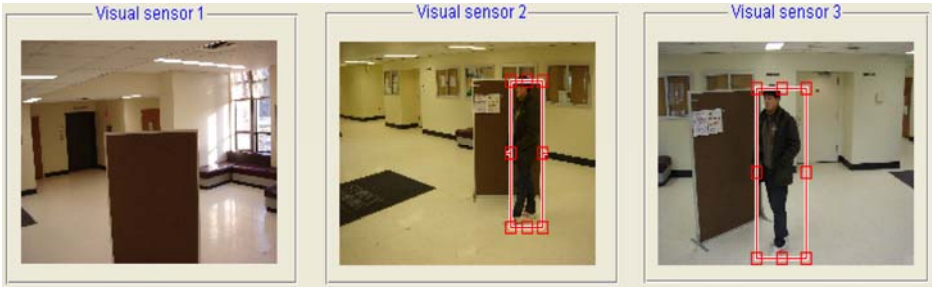
The decision/data flow for object detection enhancement is summarized in **Fig. 8**. When the global information center collects the detected local object positions $\mathrm{l}^j(n)$, each corresponding $w^j(n)$ and the *apriori* probabilities, $P_0(n)$ and $P_1(n)$, are obtained given $(\bar{g}_x(n), \bar{g}_y(n))$ and $(g_x^j(n), g_y^j(n))$. The evaluated probabilities for $W^j(n)$, $P_0(n)$ and $P_1(n)$, are delivered to the decision center, where $E(n)$ is declared. If $E(n) = 1$, the final estimated global object position $(g_x(n), g_y(n))$ is transformed into local object positions $(\tilde{l}_x^j(n), \tilde{l}_y^j(n))$

for all visual sensors V^j. Simultaneously, $g_{vx}(n)$ and $g_{vy}(n)$ are derived from $(g_x(n), g_y(n))$ and $(g_x(n-1), g_y(n-1))$, while the global object state $\mathrm{g}(n)$ is applied recursively for $(\bar{g}_x(n+1), \bar{g}_y(n+1))$ at the time $n+1$ in (20). If $E(n) = 0$, however, the global object position $(g_x(n), g_y(n))$ and all detected local object positions $(l_x^j(n), l_y^j(n))$ are eliminated.

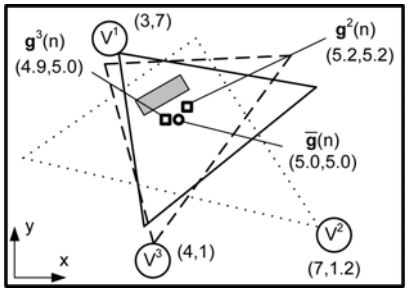
5. Simulation and Experiments

5.1 Simulation Results

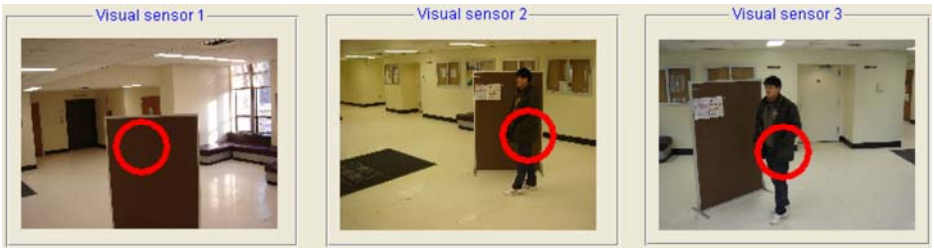
We investigated our proposed combined system using three visual sensors for each different scenario, i.e., the occlusion problem, shadow problem, and false detection problem. To achieve object detection enhancement, we assume that $P_M^j = P_F^j = 0.2$ and $W^j(n)$ is investigated with the Gaussian variance $\sigma^2 = 1$.



(a) Original detection



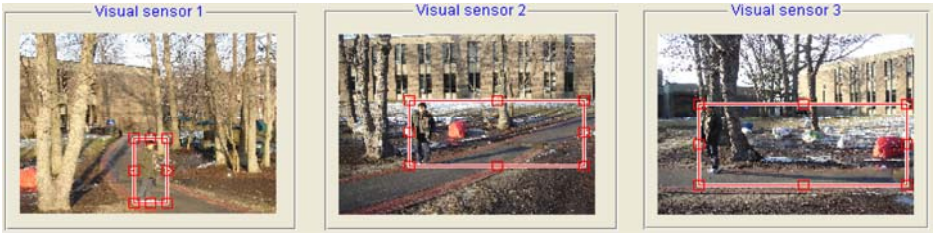
(b) Global map: $(\tilde{g}_x(n), \tilde{g}_y(n) = (5.0, 5.0)$



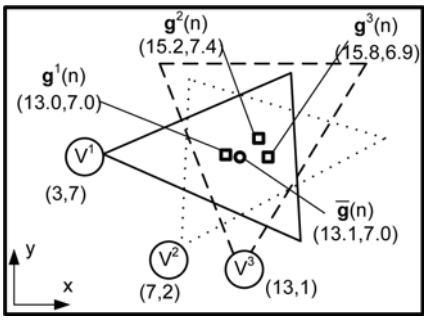
(c) Enhanced detection

Fig. 9. Original and enhanced detection (case 1).

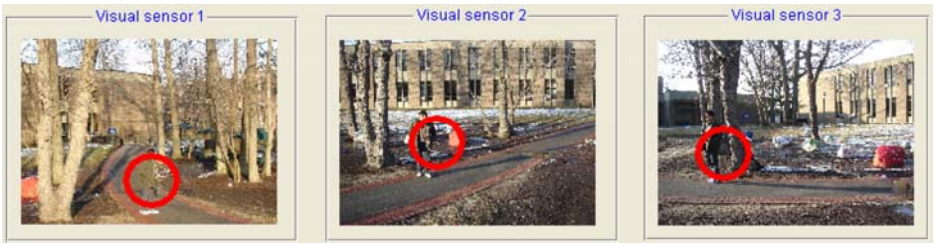
In **Fig. 9(a)**, visual sensor 1 did not detect an object due to an occlusion whereas the other two visual sensors correctly detected the object. Both detected local object positions were transformed into the global object position candidates as $\mathbf{g}^2(n) = (5.2, 5.2)$ and $\mathbf{g}^3(n) = (4.9, 5.0)$. **Fig. 9(b)** shows the surveillance environment that uses the global coordinates, including the visual sensor positions, viewable ranges, and the positions $\mathbf{g}^2(n)$, $\mathbf{g}^3(n)$ and $\bar{\mathbf{g}}(n)$. Given the position $\bar{\mathbf{g}}(n)$ as (5.0, 5.0), the quality information $W^2(n)$ and $W^3(n)$ corresponding to $\mathbf{g}^2(n)$ and $\mathbf{g}^3(n)$ are 0.9608 and 0.9950, respectively. The final global object position is (5.05, 5.10), while the corresponding $W(n)$ is 0.9938. Thus, the value for the overall decision is 2.78. Based on the true object declaration, the global information assisted enhanced object detection is shown in **Fig. 9(c)**, where the missed local object position from visual sensor 1 is recovered.



(a) Original detection



(b) Global map: $(\tilde{g}_x(n), \tilde{g}_y(n) = (13.1, 7.0)$

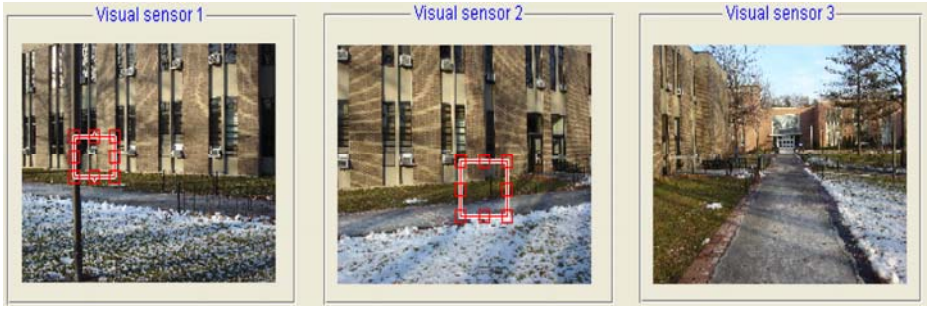


(c) Enhanced detection

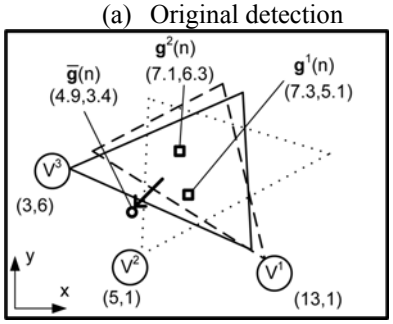
Fig. 10. Original and enhanced detection (case 2).

In **Fig. 10(a)**, visual sensor 1 correctly detects an object whereas visual sensors 2 and 3 detect an object with a deviated local object position that is merged with a shadow. The three detected local object positions are transformed into the global object position candidate as $\mathbf{g}^1(n) = (13.0, 7.0)$, $\mathbf{g}^2(n) = (15.2, 7.4)$, and $\mathbf{g}^3(n) = (15.8, 6.9)$. **Fig. 10(b)** also shows the surveillance environment global coordinates including the visual sensor positions, the

viewable ranges, and the positions $\mathbf{g}^1(n)$, $\mathbf{g}^2(n)$, $\mathbf{g}^3(n)$, and $\bar{\mathbf{g}}(n)$. Given that the position $\bar{\mathbf{g}}(n)$ is (13.1, 7.0), the quality information $W^1(n)$, $W^2(n)$, and $W^3(n)$ corresponding to $\mathbf{g}^1(n)$, $\mathbf{g}^2(n)$, and $\mathbf{g}^3(n)$ are 0.9950, 0.1018, and 0.026, respectively. The final global object position is (13.55, 7.18), while the corresponding $W(n)$ is 0.8892. The result for the overall decision is 0.13. Based on the true object declaration, the global information assisted object detection enhancement is shown in **Fig. 10(c)**, where the deviated local object positions from visual sensors 2 and 3 are correctly recovered.



(a) Original detection



(b) Global map: $(\tilde{g}_x(n), \tilde{g}_y(n) = (4.9,\ 3.4)$ (out of viewable range)

(c) Enhanced detection

(d) based on (3) without *apriori* probabilities

Fig. 11. Original and enhanced detection (case 3).

In **Fig. 11(a)**, visual sensors 1 and 2 detect a false object given that an object is positioned outside the viewable ranges. The detected local object position is transformed into the global

object position candidate as $\mathbf{g}^1(n) = (7.3, 5.1)$ and $\mathbf{g}^2(n) = (7.1, 6.3)$. **Fig. 11(b)** also shows the surveillance environment with global coordinates including the visual sensor positions, the viewable ranges, and the positions $\mathbf{g}^1(n)$, $\mathbf{g}^2(n)$, and $\bar{\mathbf{g}}(n)$. Given that position $\bar{\mathbf{g}}(n)$ is (4.9, 3.4), the quality information $W^1(n)$ corresponding to $\mathbf{g}^1(n)$ and $\mathbf{g}^2(n)$ are 0.0132 and 0.0013, respectively. The final global object position is (7.28, 5.21), while the corresponding $W(n)$ is 0.0114. The overall decision value is -6.101. Based on the false object declaration, the global information assisted object detection enhancement is shown in **Fig. 11(c)**, where all of the detected local object positions from visual sensors 2 and 3 are eliminated. As a performance comparison, **Fig. 11(d)** shows the result based on (3), which were considered using only E^j without *apriori* probabilities (i.e. $P_1(n) = P_0(n) = 0.5$). The overall decision value is 0.4, so a true object is declared.

5.2 Experimental Results

In this experimental subsection, we verified our proposed combination system in an indoor environment. In the experiment, two cameras monitored the movements of three people in an indoor environment measuring *5.22* x *4.31* x *2.96m* as shown in **Fig. 12**. Cameras 1 and 2 were placed at the positions *(2.34 m, 0.12 m, 2.45 m)* and *(0.12 m, 2.32 m, 2.45 m)*, respectively, to monitor the movement of people from different viewing angles.

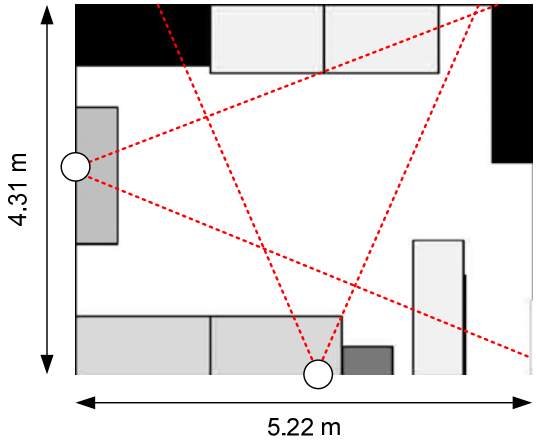


Fig. 12. Illustration of the experimental environment, which measured *5.22m* x *4.31m* x *2.96m* and contained two cameras

In the experiment, we detected the face and the body, which corresponded to local information. The body and face were treated as a single moving object. We can reduce the likelihood of missing detection by using face and body during detection. In addition, each detected face and body was tracked using the dynamic transition function given in (17). The face was detected based on skin color information, whereas the body was detected based on motion [27][28]. From a global coordinates perspective, we used the multiple dynamic models given in (22) and (23), with α=0.05. Given the local information detected, our proposed combination system exchanged the local and global information between the two cameras, which corrected the missing and false detection, as shown in **Fig. 13**. In frame #1, Camera 2 detected three people separately whereas Camera 1 had an overlapping view, which resulted in

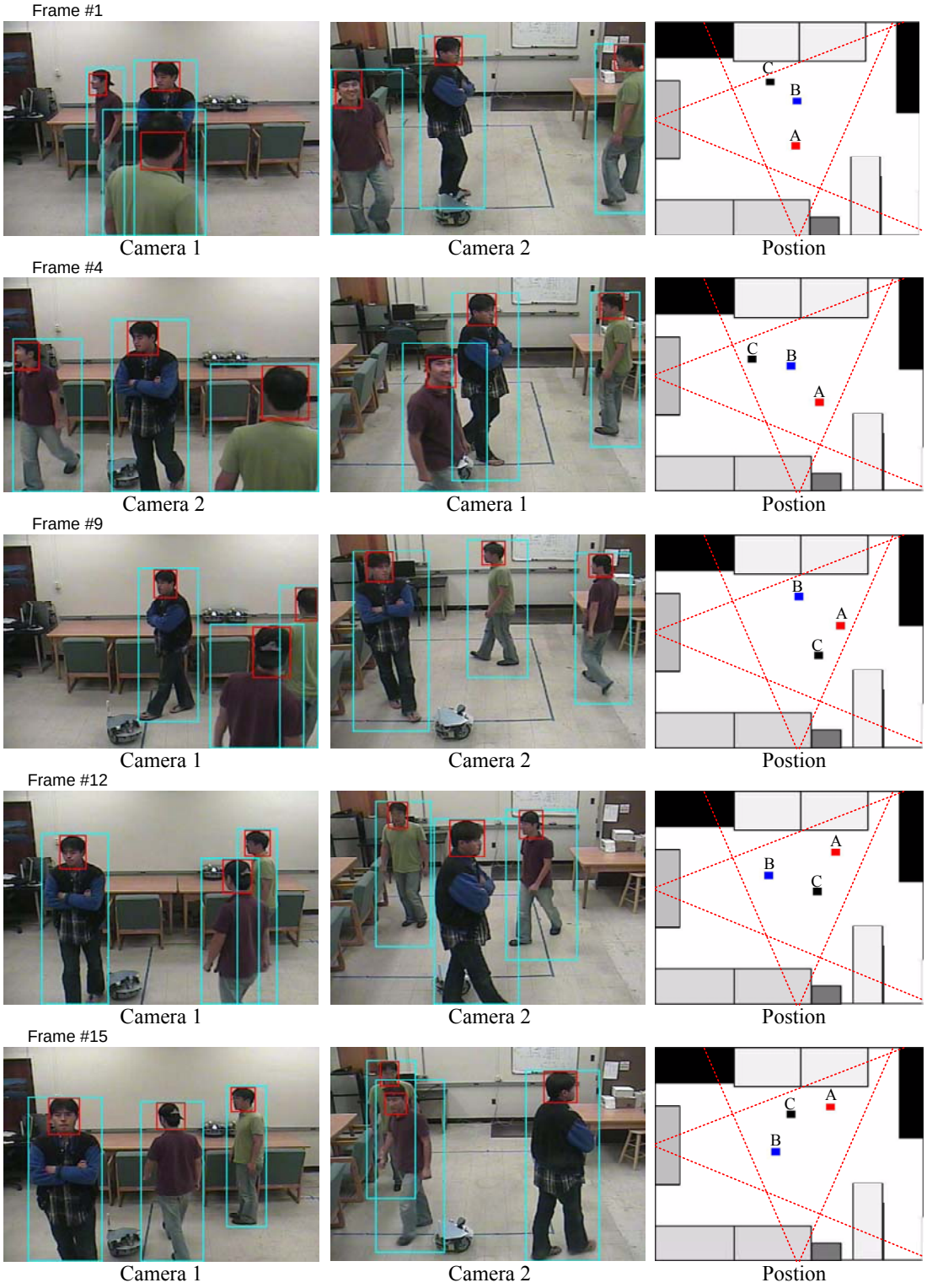


Fig. 13. Results of indoor environment experiment

the detection of a single human. The detected local information was transformed into global information, which was re-transformed into the local information, and this was separated correctly into three people with Camera 1. In frame #4, Camera 1 initially detected three people correctly whereas Camera 2 had an overlapping view of two people. After local and global information exchange, Camera 2 separated three people correctly. Similarly, in frames #9, #12, and #15, the correct local information recovered incorrect local information via local and global exchange.

5. Discussion and Conclusions

In this paper, we proposed object detection enhancement based on the combination of multiple visual sensors. Given the estimated mean position obtained from a dynamic object model, each transformed global object position candidate was evaluated, which contributed to a final global object position. We also analyzed the dynamic object model and the object detection accuracy based on object detection enhancement. Finally, the performance was evaluated for occlusion, deviated detection, false detection, and overlapping scenarios. The root mean square error (RMSE) was 0.183 (m) in the global coordinates. With the perspective model, transformation errors occurred with different camera heights, tilt angles, and distances between a camera and an object. We expect that the error could be reduced by identifying appropriate camera parameters in future research. During this study, we limited the combinations to a constrained space that provided an overlapped viewable range for the visual sensors, but our future goal is to extend the capability of the proposed system so it can be applicable to a non-overlapped viewable range in a large space. We are also currently exploring the effect of incorrect local information. The propagation of incorrect local information is also important for the maintenance of the combined system, so our future goal is to find the method that minimizes the incorrect local information. Furthermore, we need to consider an association issue [29][30] in more complicated areas such as places where people move in and out, which would assign consistent identifications to each person detected. This association issue is now being considered for the combined system in future work.

References

[1] R. Feraud, O.J. Bernier, J.-E. Viallet and M. Collobert, "A fast and accurate face detection based on neural network," *IEEE Trans. on Signal Processing*, vol.50, no.2, pp.174-188, Feb.2002. Article (CrossRef Link).

[2] R. Hsu, M. Mottaleb and A. Jain, "Face detection in color images," *IEEE Trans. on Patterns on Analysis and Machine Intelligence*, vol.24, no.5, pp.696-706, May.2002. Article (CrossRef Link).

[3] B. Heisele, T. Serre, M. Pontil and T. Poggio, "Component-based face detection," in *Proc. of the 2001 IEEE Conf. on Computer Vision and Pattern Recognition*, pp.657-662, 2001. Article (CrossRef Link).

[4] F. Xu, X. Liu and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Trans. on Intelligent Transportation Systems*, vol.6, no.1, pp.63-71, Mar.2005. Article (CrossRef Link).

[5] R.J. Radke, S. Andra, O. A. Kofahi and B. Roysam, "Image change detection algorithms: A survey," *IEEE Trans. on Image Processing*, vol.14, no.3, pp.294-307, Mar.2005. Article (CrossRef Link).

[6] Y. Wu and T.S. Huang, "Recent development in human motion analysis," *Pattern Recognition*, vol.36, no.3, pp.585-601, Mar.2003. Article (CrossRef Link).

[7] R. Bodor, B. Jackson and N. Papanikolopoulos, "Vision-based human tracking and activity recognition," in *Proc. of the 11th Mediterranean Conference on Control and Automation*,

pp.18-20, Jun.2003. Article (CrossRef Link).
[8] O. Javed, K. Shafique and M. Shah, "Hierarchical approach to robust background subtraction using color and gradient information," in *Proc. of the IEEE Workshop on Motion and Video Computing*, pp.22-27, Dec.2002. Article (CrossRef Link).
[9] Y. Matsushita, K. Nishino, K. Ikeuchi and M. Sakauchi, "Illumination normalization with time-dependent intrinsic images for video surveillance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.26, no.10, pp.1336-1347, Oct.2004. Article (CrossRef Link).
[10] C. Stauffer and W. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.22, no.8, pp.747-757, Aug.2000. Article (CrossRef Link).
[11] M. Shah, O. Javed and K. Shafique, "Automated visual surveillance in realistic scenarios," *IEEE Multimedia*, vol.14, no.1, pp.30-39, Mar.2007. Article (CrossRef Link).
[12] R. Collins, Y. Liu and M. Leordeanu, "On-line selection of discriminative tracking features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.27, no.10, pp.1631-1643, Oct.2005. Article (CrossRef Link).
[13] S. Avidan, "Ensemble Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.29, no.2, pp.261-271, Feb.2007. Article (CrossRef Link).
[14] F. Tang and H. Tao, "Probabilistic object tracking with dynamic attributed relational feature graph," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.18, no.8, pp.1064-1074, Aug.2008. Article (CrossRef Link).
[15] T. Yang, S. Li, Q. Pan and J. Li, "Real-time multiple objects tracking with occlusion handling in dynamic scenes," in *Proc. of the 2005 IEEE Conf. on Computer Vision and Pattern Recognition*, pp.970-975, Jun.2005. Article (CrossRef Link).
[16] A. Senior, "Tracking people with probabilistic appearance models," in *Proc. of the IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, pp.48-55, 2002. Article (CrossRef Link).
[17] D. Greenhill, J. Renno, J. Orwell and G.A. Jones, "Occlusion analysis: Learning and utilizing depth maps in object tracking," in *Proc. of Annual British Machine Vision Conf.*, vol.26, no.3, pp.430-441, Mar.2008. Article (CrossRef Link).
[18] W. Hu, T. Tan, L. Wang and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Information Systems, Man, And Cybernetics - Part C: Applications and Reviews*, vol.34, no 3, pp.334-352, Aug.2004. Article (CrossRef Link).
[19] S. Chien, Y. Huang, B. Hsieh, S. Ma and L. Chen, "Fast video segmentation algorithm with shadow cancellation, global motion compensation and adaptive threshold techniques," *IEEE Transactions on Multimedia*, vol.6, no.5, pp.732-748, Oct.2004. Article (CrossRef Link).
[20] M. Pollefeys, R. Koch and L. V. Gool, "Self-calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters," in *Proc. of the Sixth International Conference on Computer Vision*, pp.90-95, Jan.1998. Article (CrossRef Link).
[21] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol.22, no.1, pp.98-101, Jan.1986. Article (CrossRef Link).
[22] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking-Part I," *IEEE Trans. on Aerospace and Electronic Systems*, vol.39, no.4, pp.1333-1364, Oct.2003. Article (CrossRef Link).
[23] Y. B. Shalom and W. D. Blair, *Multitarget-Multisensor Tracking: Applications and Advances*, Artech House Publishers, 2000.
[24] D. B. Ward, E. A. Lehmann and R. C. Williamson "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Processing*, vol.11, no 6, pp.826-836, Nov.2003. Article (CrossRef Link).
[25] J. Lim, J. Lee, S. Hong and P. Park, "Algorithm for detection with localization of multi-targets in wireless acoustic sensor networks," in *Proc. of the 18th IEEE Int. Conference on Tools with Artificial Intelligence*, pp.547-554, Nov.2006. Article (CrossRef Link).
[26] 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.1, pp.2351-2365, Jan.2004. Article (CrossRef Link).
[27] Z. Jin, Z Lou, J. Yang and Q. Sun, "Face detection using template matching and skin-color information", *Neurocomputing*, vol.70, no.4-6, pp.794-800, Jan.2007. Article (CrossRef Link).
[28] K. Park, J. Lee, M. Stanacevic, S. Hong and W. Cho, "Iterative object localization algorithm using visual images with a reference coordinate", *EURASIP Journal on Image Video Processing*, Article ID 256896, vol.2008. Article (CrossRef Link).
[29] S. H. Cho, Y. Nam, S. Hong and W. D. Cho, "Locally initiating line-based object association in large scale multiple cameras environment," *KSII Transactions on Internet and Information Systems*, vol.4, no.3, pp.358-379, Jun.2010. Article (CrossRef Link).
[30] S. H. Cho, Y. Nam and S. Hong, "Multiple camera collaboration strategies for dynamic object association," *KSII Transactions on Internet and Information Systems*, vol.4, no.6, pp.1169-1193, Dec.2010. Article (CrossRef Link).

Jinseok Lee received the dual B.S. degree in electrical engineering from both Stony Brook University – State University of New York and Ajou University – Korea; and the Ph.D. degree in Electrical engineering from Stony Brook University. Currently, he is a Postdoctoral Associate of Biomedical Engineering at Worcester Polytechnic Institute. His current research interests include medical instrumentation, biomedical signal processing, and modeling of physiological systems.

Shung Han Cho received B.E. degree (Summa Cum Laude) with specialization in Telecommunications from both the department of Electronics Engineering at Ajou University, Korea and the department of Electrical and Computer Engineering at Stony Brook University - SUNY, NY in 2006. He was a recipient of Award for Academic Excellence in Electrical Engineering by College of Engineering and Applied Sciences at Stony Brook University. He received M.S. degree with Award of Honor in Recognition of Outstanding Achievement and Dedication and Ph.D. degree in Electrical and Computer Engineering from Stony Brook University in 2008 and 2010 respectively. He is currently a post-doctoral researcher at Stony Brook University. He was a recipient for International Academic Exchange Program supported by Korea Research Foundation (KRF) in 2005. He was a member of Sensor Consortium for Security and Medical Sensor Systems sponsored by NSF Partnerships for Innovation from 2005 to 2006. His research interests include collaborative heterogeneous signal processing, distributed digital image processing and communication, networked robot navigation and communication, heterogeneous system modeling.

Seong-Jun Oh (S'98-M'01-SM'10) is an Associate Professor at the Department of Computer and Communications Engineering, Korea University, Seoul, Korea. Before joining Korea University in September 2007, he was with Ericsson Wireless Communication, San Diego, CA, USA as a senior Engineer from September 2000 to March 2003 and with Qualcomm CDMA Technologies (QCT), San Diego, CA, USA as a Staff Engineer from September 2003 to August 2007. He received his B.S. (magna cum laude) and M.S. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1991 and 1995, respectively, and received Ph.D. at the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor in September 2000. He served for Korean Army during 1993-1994. His current research interests are in the area of wireless/mobile networks with emphasis on the next-generation (4G) cellular networks, resource allocation, and physical-layer modem implementation. While he was with Ericsson Wireless Communication, he has been an Ericsson representative for WG3 (physical layer) of 3GPP2 standard meeting. While in QCT, he has developed CDMA modems in ASIC for base station and mobile station. He was a Vice-Chair of TTA PG 707, the Korean evaluation group registered in ITU-R, where he was in charge of performance evaluations of LTE-Advanced and IEEE 802.16m systems, submitted as IMT-Advanced technologies in ITU-R WP-5D. He has received the Seoktop Teaching Award from the College of Information and Communication, Korea University for outstanding lectures in fall semester of 2007 and spring semester of 2010. He was a recipient of the Korea Foundation for Advanced Studies (KFAS) Scholarship 1997-2000.

Sangjin Hong received the B.S and M.S degrees in EECS from the University of California, Berkeley. He received his Ph.D in EECS from the University of Michigan, Ann Arbor. He is currently with the department of Electrical and Computer Engineering at Stony Brook University. Before joining Stony Brook University, he has worked at Ford Aerospace Corp. Computer Systems Division as a systems engineer. He also worked at Samsung Electronics in Korea as a technical consultant. His current research interests are in the areas of multimedia wireless communications and digital signal processing systems, reconfigurable VLSI Systems and optimization. Prof. Hong is a Senior Member of IEEE and a member of EURASIP journal editorial board. Prof. Hong served on numerous Technical Program Committees for IEEE conferences.