Master's Thesis

Title

Human Perception Inspired Multimodal Object Recognition by Considering Time-Dependent Reliance on Modalities

Supervisor

Professor Masayuki Murata

Author

Haruhito Ando

February 3rd, 2025

Department of Information Networking

Graduate School of Information Science and Technology

Osaka University

Master's Thesis

Human Perception Inspired Multimodal Object Recognition by Considering
Time-Dependent Reliance on Modalities

Haruhito Ando

Abstract

In recent years, smart manufacturing (SM) has gained significant attention in the manufacturing industry due to its potential to address labor shortages and optimize energy resources. One of the key elements of SM is the Collaborative Robot, which enhances operational efficiency and production speed by enabling human-robot collaboration. Since Collaborative Robots are designed to share workspaces with humans for efficiency, ensuring human safety is a critical challenge.

Current Automatic Guided Vehicles (AGVs) and Autonomous Mobile Robots (AMRs) are equipped with increasingly precise sensors, such as RGB cameras and LiDAR (Light Detection and Ranging), allowing them to acquire diverse environmental information. Given the limitations of individual sensors, research on multimodal object recognition, which integrates multiple sensors to enhance recognition accuracy, is being hot topic. While multimodal sensing has been widely applied in real-world robotic systems, existing approaches have not fully addressed the time-dependent variation in sensor reliability caused by environmental factors such as lighting conditions and moving obstacles. This limitation may lead to inaccurate perception of the surroundings in dynamic environments.

To address this issue, in this thesis we propose a multimodal object recognition method inspired by human perception, which dynamically adjusts the weighting of sensor inputs based on real-time environmental changes. The method employs MLE model-based reliability weighting approach, allowing the system to adaptively adjust sensor confidence scores over time. To evaluate how dynamic assessment of sensor reliability affects recognition, we conducted experiments where one modality was given while varying the other.

Specifically, we prepared recognition task results using real-world video data and combined them with simulated modality data, whose mean and variance changed over time, to comprehensively verify the effectiveness of the proposed method in suppressing false positives.

The evaluation results demonstrate that by incorporating dynamic sensor reliability assessment, our proposed approach significantly improved recognition performance compared to conventional methods, such as arithmetic averaging and maximum value selection. Notably, the method reduced fluctuations in unreliable scores and achieved a more stable score transition, leading to an improvement in precision and specificity over the baseline approach. In particularly, in one situation set up in this study, precision improved from 0.00 to 1 for the average method, and specificity improved from 0.89 to 1 for max selecting method. These findings suggest that dynamically adjusting sensor reliability can enhance recognition robustness in real-world robotic applications.

Keywords

Smart Manufacturing (SM)
Collaborative Robot
Human Perception
Autonomous Mobile Robot (AMR)
Multimodal Object Recognition

Contents

List of Figures

List of Tables

1 Introduction

The advent of technology, digitalization is progressing across various fields, including manufacturing, where Smart Manufacturing (SM) has attracted attention as a means to address labor shortages and optimize energy use [1–4]. SM incorporates multiple technologies, such as Cloud Computing and Augmented Reality, but one of the most actively researched areas is Collaborative Robots. Collaborative Robots are designed to operate in shared workspaces with humans, allowing for simultaneous collaboration. Unlike traditional industrial robots, which are confined to separate work areas for safety reasons, Collaborative Robots eliminate the need for protective barriers, making more efficient use of workspace.

Robots used in manufacturing environments are categorized based on their level of autonomy. Automatic Guided Vehicles (AGVs) operate along predefined routes, while Autonomous Mobile Robots (AMRs) navigate freely without fixed pathways. AMRs are typically equipped with multiple sensors, such as cameras, LiDAR (Light Detection and Ranging), and radar, enabling them to estimate their position, detect humans, and navigate their environment (Fig. 1).

By leveraging AMRs, manufacturers can offload repetitive and physically demanding tasks to robots while allowing human workers to focus on more expert or complex operations that are costly to automate. For example, in small-batch production, automating every step of the process is economically unfeasible, however, using AMRs to transport materials within a factory can significantly improve efficiency. Unlike traditional AGVs, which require extensive infrastructure modifications, AMRs can adapt to changes in factory layouts, making them a more flexible solution for automation.

Despite these advantages, ensuring human safety remains a critical requirement for AMRs. Since these robots operate in environments where humans frequently move, accidental collisions must be prevented. AMRs are also high-mass systems, often carrying power units and cargo, which makes collisions particularly hazardous. To enable safe and efficient operation, AMRs must incorporate robust environmental perception capabilities, allowing them to identify objects and humans using Computer Vision (CV) technologies [1–5].

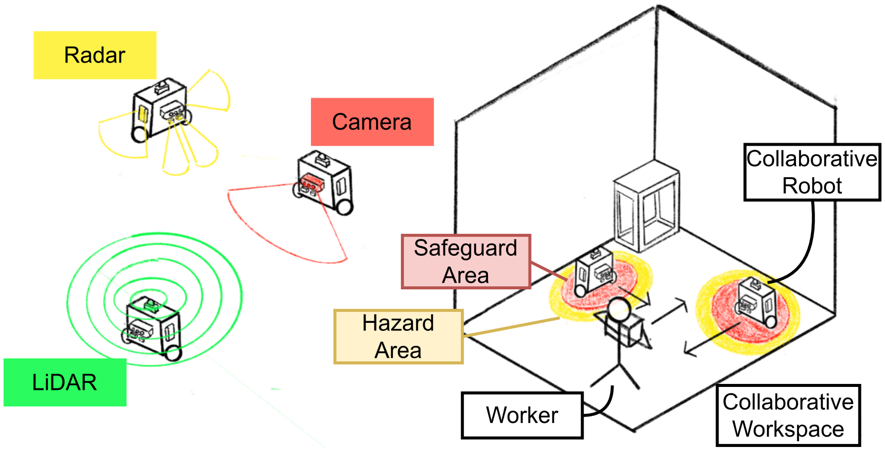


Figure 1: **Collaborative Robots example.** Collaborative Robots have multiple sensors, such as cameras, LiDAR, and radar to recognize their surroundings. They share workspace with human workers, making more efficient use of workspace. For human safety, they need to implement safety measures and stop all motion to prevent any potential injury.

To recognize their surroundings, AMRs integrate a variety of sensors that provide information for obstacle detection, path planning, and task execution [6–8]. Each sensor has its own strengths and limitations. For example:

- Cameras capture detailed visual information but may be affected by lighting conditions.
- LiDAR provides high-precision distance measurements but struggles on surfaces with low reflectivity.
- Radar measures object velocity but lack detailed spatial resolution.

In indoor environments such as factories and warehouses, lighting conditions vary, causing the appearance of objects to change. Similarly, some materials may not reflect light well, making LiDAR-based detection unreliable. Each unimodal sensor suffers from information loss, e.g., cameras lack depth information, while LiDAR and radar lack color and texture data [9–11]. To address these limitations, multimodal object recognition has been studied as a way to combine multiple sensors and compensate for missing information [8–12]. A common approach is to integrate data from cameras and LiDAR to combine visual and depth information for improved recognition performance.

Multimodal Object Recognition methods have been explored by many work, and they are classified into two main approaches based on the timing of integration. [8–12]:

- **Early Fusion (Feature-Level Fusion):** Integrates sensor data at the feature extraction stage, converting them into a common format for joint processing. This approach preserves sensor information but is highly sensitive to sensor failures.

- **Late Fusion (Decision-Level Fusion):** Processes sensor data separately and then combines the final outputs. This approach is more robust to sensor failures but may lose cross-modal relationships.

While these methods focus on when sensor data should be combined, they do not consider how sensor reliability changes over time. In real-world applications, the importance of each sensor fluctuates due to environmental changes. For example, in human detection tasks, sensor reliability shifts as:

- Distant objects are better detected by LiDAR, but cameras and radar become more reliable at closer distances.

- Changing lighting conditions degrade camera reliability, while low-reflectivity surfaces weaken LiDAR performance.

Despite these known challenges, existing approaches do not dynamically adjust sensor weighting based on changing reliability. Some studies have explored climate effects on sensing [8] and illumination effects on recognition accuracy [13], but few have incorporated real-time sensor reliability estimation into their methods. To improve environmental perception, it is essential to develop a method that can adapt to time-dependent sensor reliability changes.

To address this issue, we focus on human perception models, as humans make decisions based on uncertain and noisy sensory inputs [14–16]. Human recognition processes have been studied from multiple perspectives, including neurophysiology and psychophysics. Among these, psychophysical approaches suggest that humans integrate sensory inputs by weighting them based on modality-specific reliability, a concept that has been applied in our prior work on object recognition [17, 18]. In this prior research, we incorporated predefined reliability values for different sensor inputs, using these reliability estimates to enhance object recognition. However, this approach did not account for real-time fluctuations in sensor reliability, as the reliability values remained constant throughout the evaluation process. As a result, it was unable to adapt to changes in reliability over time.

In this work, we propose a sensor integration method that dynamically evaluates the reliability of modality-specific prediction scores over time. Unlike prior approaches that used fixed reliability values, our method continuously updates the weighting of each modality in response to environmental variations. The proposed method is based on a MLE model framework [19, 20], which allows for adaptive integration of multimodal sensor data while considering time-dependent reliability changes. To evaluate the effectiveness of this approach, we conducted experiments where the recognition result of one modality was fixed while that of the other was varied, allowing us to observe how the integration process adapted to different conditions. For the fixed modality, we used real-world video-based recognition results and combined them with a simulated modality whose mean and variance changed over time. We then compared the proposed method to conventional approaches, such as arithmetic averaging and maximum value selection, to assess its impact on recognition performance and error reduction. The results demonstrate that our proposed method reduces the influence of unstable scores and achieves more consistent recognition performance, even in environments with high uncertainty.

The key contributions of our work are as follows:

- Proposal of a reliability-weighted multimodal object recognition method that adjusts sensor weights based on time-dependent reliability changes.
- Experimental validation of dynamic sensor reliability assessment using real-world video-based recognition results combined with simulated modality scores that change over time.
- Demonstration of improved recognition stability, showing that the proposed method reduces the impact of unstable sensor scores compared to arithmetic averaging and maximum value selection methods.

The remaining of the thesis is organized as follows. Section 2 introduces our previous research and related studies. Our proposed methodology is described in Section 3, and we verify its effectiveness using our prepared results of processed real-world video data combined with a generated modality data in Section 4 . Section 5 concludes with a brief summary.

2 Human Perception-Inspired Multimodal Object Recognition

Multimodal object recognition has been extensively studied to improve perception in real-world environments. Many approaches integrate RGB images and point cloud data, aiming to compensate for the limitations of individual modalities. However, most existing methods assume that sensor reliability remains constant over time, failing to account for environmental variations such as changes in lighting conditions or object occlusions. To address these issues, our study explores an adaptive multimodal recognition framework inspired by human perception, dynamically adjusting sensor contributions based on real-time changes in reliability.

This section first outlines conventional object recognition methods (Section 2.1), discussing both unimodal and multimodal approaches. Then, we examine prior research that incorporates human perceptual models into multimodal recognition (Section 2.2) and highlight the key distinctions between existing methods and our proposed framework.

2.1 Object Recognition

2.1.1 Unimodal Object Recognition

Unimodal object recognition relies on a single type of sensor input, typically RGB images or point cloud data [8–12,21,22]. In RGB-based recognition, deep learning techniques such as YOLO Series [23–25] and DETR [26] are widely used, leveraging large-scale datasets like COCO [27] to detect objects in complex scenes. RGB images provide rich texture and color information, making them effective for identifying object categories, but they lack depth and spatial structure.

In contrast, point cloud-based recognition processes data from sensors like LiDAR and Radar, which capture 3D spatial information essential for estimating object position and shape. Deep learning models such as PointNet [28] and PointPillars [29] are commonly used in this domain. Datasets like KITTI [30] and nuScenes [31] offer benchmark evaluations for point cloud-based recognition. While point cloud data provides precise distance measurements, it lacks color and texture, making it difficult to distinguish objects with similar shapes but different appearances.

2.1.2 Multimodal Object Recognition

To overcome the limitations of unimodal approaches, multimodal object recognition integrates multiple sensor inputs, combining their strengths to enhance perception [8–12]. These methods are broadly categorized into two main types:

- **Early Fusion:** Sensor data is combined at the feature extraction stage, converting all inputs into a common representation before recognition. For example, Frustum PointNet [32] projects RGB-based bounding boxes onto point clouds for 3D object localization, while PointPainting [33] uses semantic segmentation from RGB images to enrich point cloud data.

- **Late Fusion:** Each sensor's data is processed separately, and the recognition results are combined at a later stage. Methods like MV3D [13] convert point clouds into a bird's-eye view before fusion, while CLOCs [34] aligns bounding boxes detected from different modalities based on their spatial overlap.

Despite the progress in multimodal fusion, existing approaches typically assign fixed weights to each sensor, assuming that all modalities contribute equally throughout the recognition process. However, in real-world environments, sensor reliability fluctuates over time, for instance, distant objects are better detected by LiDAR, but RGB images and radar become more reliable as objects move closer. Moreover, lighting conditions significantly affect RGB-based recognition, while low-reflectivity surfaces degrade Point Cloud-based recognition. These variations highlight the need for a dynamic sensor weighting mechanism that adapts to environmental changes.

2.2 Multimodal Integration in Human Perception

Since sensor inputs contain noise and uncertainty, researchers have explored how human perception integrates multimodal sensory information [14–16]. Humans do not rely on a single sense but instead adjust sensory reliance dynamically, prioritizing the most reliable source based on context. This concept has been formalized using Bayesian Causal Inference (BCI) [19, 20, 35–38], which models how humans determine whether sensory signals originate from a common or independent source. We have been applied this concept in our prior work on object recognition [17,18].

2.2.1 Generative Model

BCI models sensory integration using causal structures (C) that define the relationship between sensory observations and their sources. Given two sensory inputs (e.g., RGB images and LiDAR data), the model estimates whether they originate from the same source ($C = 1$) or different sources ($C = 2$). The prior probability of $C = 1$, denoted as p_{common}, represents the likelihood that the two modalities should be integrated.

2.2.2 Causal Inference

Using Bayes' theorem, the probability of the causal structure given the observed data is computed as Eq. ((1)):

$$p(C|x_a, x_v) = \frac{p(x_a, x_v|C)p_{common}}{p(x_a, x_v)} \tag{1}$$

where x_a and x_v are the observations from different modalities. Based on this, the integrated estimation is performed as:

$$\hat{x}_a = P(C = 1)\hat{x}_{a,C=1} + P(C = 2)\hat{x}_{a,C=2} \tag{2}$$

where $\hat{x}_{a,C=1}$ and $\hat{x}_{a,C=2}$ represent the estimated values depending on the causal condition C.

2.2.3 Estimation

Because both prior distributions and likelihoods are assumed to be Gaussian, the posterior distribution remains Gaussian, leading to the following estimation equations:

$$\hat{x}_{a,c=1} = \hat{x}_{v,c=1} = \frac{x_a/\sigma_a^2 + x_v/\sigma_v^2 + \mu_p/\sigma_p^2}{1/\sigma_a^2 + 1/\sigma_v^2 + 1/\sigma_p^2} \tag{3}$$

$$\hat{x}_{a,c=2} = \frac{r_a/\sigma_a^2 + \mu_p/\sigma_p^2}{1/\sigma_a^2 + 1/\sigma_p^2} \tag{4}$$

where μ_p, σ_p represent the prior mean and variance of the observed data, and σ_a, σ_v indicate the observation noise of each modality.

2.2.4 Human Perception-inspired multimodal integration model

By incorporating Bayesian inference, existing research has shown that dynamically adjusting sensory reliance improves object recognition robustness [17,18]. However, previous studies used fixed reliability values, meaning they did not account for real-time fluctuations. This limitation makes them less adaptable to changing environments. To address this, in this thesis we propose an adaptive multimodal recognition framework that continuously evaluates sensor reliability over time. Unlike prior work, our method dynamically adjusts modality weighting based on environmental conditions, enabling more reliable recognition in uncertain and dynamic settings.

3 Proposal of Reliability-Weighted Multimodal Object Recognition

This section details the proposed reliability-weighted multimodal object recognition method, which dynamically adjusts the weight of each sensor based on real-time changes in reliability. Unlike conventional fusion methods that apply fixed sensor weighting, our approach continuously re-evaluates sensor confidence using historical detection performance over a time window.

Section 3.1 provides an overview of the proposed framework, Section 3.2 discusses the reliability-weighted multimodal integration approach, and Section 3.3 describes how the computed scores are used for decision-making.

3.1 Overview

The proposed method takes inputs from two modalities, processes detection results from each, and integrates them to produce a confidence-weighted recognition score. Unlike conventional multimodal fusion, which treats all sensors as equally reliable, our method monitors temporal fluctuations in sensor performance and adjusts their contributions accordingly. Figure 2 illustrates an overview of the proposed method. The key steps are as follows:

- **Input Processing:** Object recognition is performed independently for each modality.
- **Temporal Reliability Estimation:** The system tracks the stability of each modality's detection score over a defined time window.
- **Reliability-Weighted Fusion:** Sensor scores are dynamically weighted based on their estimated reliability.
- **Final Decision:** A threshold-based approach determines whether an object is detected.

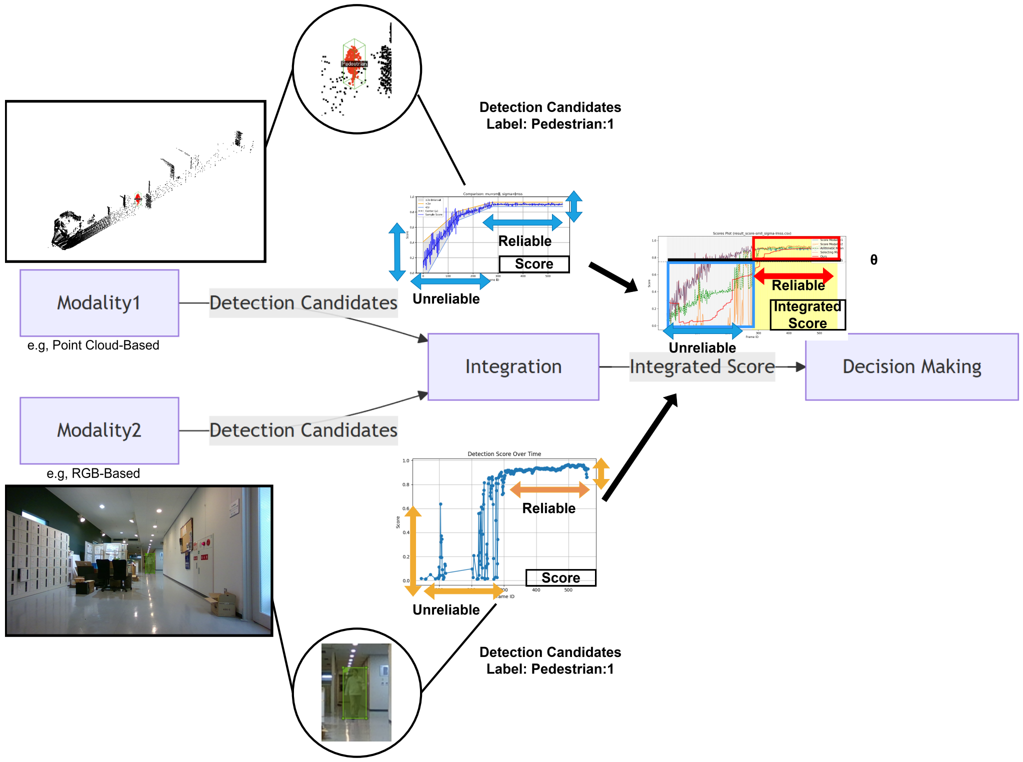


Figure 2: **Overview of the proposal object recognition.** The final decision is made using the resulting integration score and a set threshold value, which is weighted according to the reliability of the candidate detections from the two modalities.

3.2 Reliability-weighted Multimodal Integration

3.2.1 Dynamic Reliability Estimation

To account for time-dependent variations in sensor reliability, we introduce a MLE Model-based approach [19, 20]. This model continuously evaluates the stability of each sensor's detection score over recent frames, adjusting its contribution accordingly. For example, in environments the point cloud data of small objects become sparse, its reliability decreases, and RGB images degrade due to lighting changes, their reliability decreases. Rather than relying solely on the instantaneous detection score from each modality, we incorporate historical performance trends to ensure more stable recognition. Specifically, we define mean ($\mu_{(T)}$) and variance ($\sigma^2_{(T)}$) at time T over a time window W as follows:

$$\mu_{(T)} = \frac{1}{W} \sum_{i=1}^{W} x_{(T-i)}, \tag{5}$$

$$\sigma^2_{(T)} = \frac{1}{W-1} \sum_{i=1}^{W} \left[x_{(T-i)} - \mu_{(T)}\right]^2 + \epsilon. \tag{6}$$

where $x_{(T-i)}$ represents the detection score of a given modality at frame $T - i$. W is the window size and ϵ is a small constant ensuring numerical stability ($\epsilon << 1$). A high variance ($\sigma^2_{(T)}$) indicates inconsistent detection scores, suggesting that the modality is unreliable. Thus, when performing multimodal fusion, sensors with lower variance are assigned higher weights.

3.2.2 Reliability-Weighted Score Integration

To compute the final integrated recognition score, we assign weights based on each sensor's inverse variance:

$$I_{(t=T)} = \frac{\mu_{A(T)}/\sigma^2_{A(T)} + \mu_{V(T)}/\sigma^2_{V(T)} + \mu_{p(T)}/\sigma^2_{p(T)}}{1/\sigma^2_{A(T)} + 1/\sigma^2_{V(T)} + 1/\sigma^2_{p(T)}}. \tag{7}$$

where A and V refer to Modality A (e.g., RGB images) and Modality V(e.g., Point Cloud). $\mu_{A(T)}$ and $\mu_{V(T)}$ are the mean of recent detection scores, and $\sigma^2_{A(T)}$ and $\sigma^2_{V(T)}$ are the variance of them. $\mu_{p(T)}$ and $\sigma_{p(T)}$ denotes prior knowledge of reliability, which is updated each time the score is calculated for each $W/$ This allows us to retain the information measured at the previous time. If a modality lacks detection results at a given frame (e.g., missing data from RGB-based method due to blackout), integration is performed using only available inputs:

$$I_{(t=T)} = \frac{\mu_{A(T)}/\sigma^2_{A(T)} + \mu_{p(T)}/\sigma^2_{p(T)}}{1/\sigma^2_{A(T)} + 1/\sigma^2_{p(T)}}. \text{ if Modality } V \text{ score is missing} \tag{8}$$

$$I_{(t=T)} = \frac{\mu_{V(T)}/\sigma^2_{V(T)} + \mu_{p(T)}/\sigma^2_{p(T)}}{1/\sigma^2_{V(T)} + 1/\sigma^2_{p(T)}}. \text{ if Modality } A \text{ score is missing} \tag{9}$$

If both modalities are unavailable, only prior knowledge is used for estimation.

3.2.3 Updating Prior Distributions

To ensure adaptability to changing environmental conditions, the prior distribution is updated dynamically:

$$\mu_{p(t=T+1)} = I_{(t=T)}, \tag{10}$$

$$\sigma^2_{p(t=T+1)} = (\frac{1}{\sigma^2_{A(T)}} + \frac{1}{\sigma^2_{V(T)}} + \frac{1}{\sigma^2_{p(T)}})^{-1} \tag{11}$$

By iteratively refining prior knowledge, the model gradually adapts to shifting sensor reliability trends.

3.3 Decision Making

3.3.1 Risk Assessment Using Integrated Scores

After computing the reliability-weighted score $I_{(T)}$, the final decision is made based on a predefined threshold θ.

$$R(T) = \begin{cases} 1, & \text{if } I_{(T)} \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The threshold θ is empirically determined based on experimental results, ensuring a balance between false positives and false negatives. By employing θ, the proposed method provides flexible response, such as lowering the threshold to increase safety or raising the threshold to avoid unnecessary detection.

3.3.2 Advantages of the Proposed Method

Unlike conventional multimodal fusion approaches that assign static sensor weights, our method provides several advantages:

- **Adaptability to Dynamic Environments:** Automatically adjusts to changes in lighting, occlusions, and sensor noise.
- **Reduction of Erratic Detections:** Unstable modalities contribute less to the final decision.
- **Improved Recognition Robustness:** Avoids over-reliance on a single sensor.

By incorporating historical reliability trends, our method ensures more consistent and accurate recognition in uncertain environments.

4 Evaluation

This section presents the experimental setup and evaluation results of the proposed reliability-weighted multimodal object recognition method. The objective is to assess how dynamically adjusting sensor reliability influences recognition performance under varying environmental conditions.

Section 4.1 describes the experimental setup, including the dataset, task definitions, evaluation metrics, and parameter settings. Section 4.2 presents quantitative and qualitative results, followed by a discussion in Section 4.3, and the contribution of this work in Section 4.4.

4.1 Setting

4.1.1 Dataset

To evaluate the proposed method, we conducted experiments using a combination of real-world video data and simulated modality data. The real-world data consists of RGB-based recognition results shown in Fig. 3, obtained using a unimodal object recognition method (StreamYOLO [25]). The dataset contains 560 frames collected in an indoor environment, where a person walks down a hallway (Fig. 4). In addition to real data, we generated a simulated modality to model the effects of environmental variations over time. This synthetic data was designed such that its mean detection confidence (μ_s) and variance(σ_s^2) change across different time segments, simulating conditions where sensor reliability fluctuates. These values were determined based on an analysis of the RGB-based recognition trends, ensuring that the simulated scores exhibit realistic behavior (Table. 1).

4.1.2 Task Setting

To evaluate the effectiveness of the proposed reliability-weighted fusion method, we designed a task in which recognition scores from two modalities (RGB-based and simulated) are integrated, with the simulated modality exhibiting time-dependent reliability variations. The simulated modality undergoes four distinct phases (Phase 1, Phase 2, Phase 3,

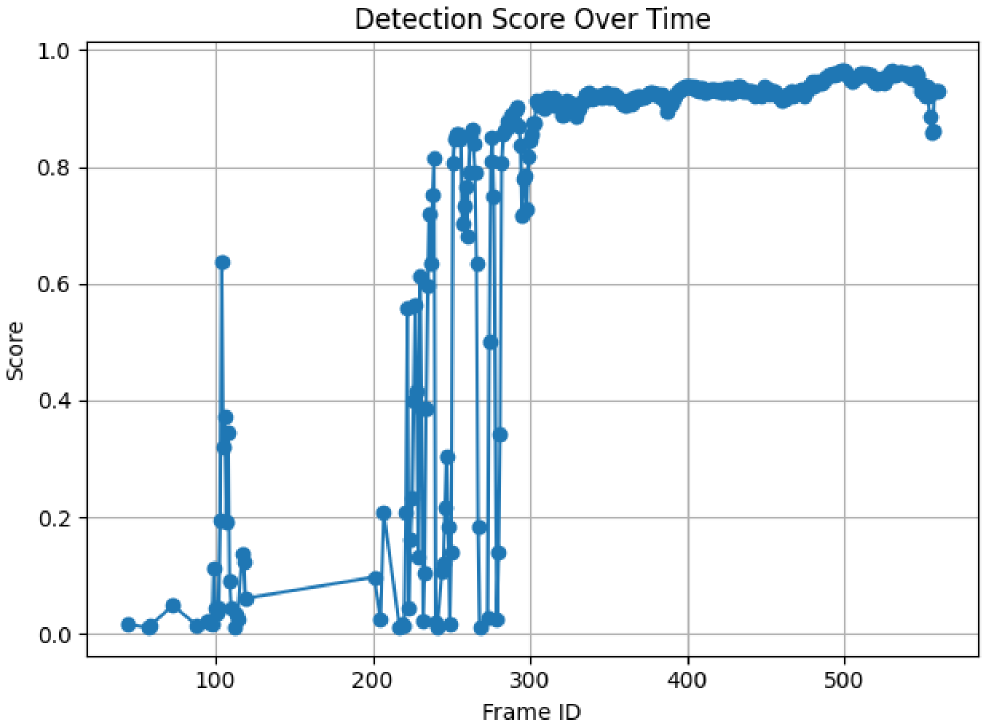


Figure 3: **Result of RGB-based unimodal object recognition.** Recognition results of the RGB-based recognition system prepared with the actual environmental data. The first half of the video was undetectable, and the score became unstable in the middle of the frames, while the second half of the video was stable and recorded a high score.

Phase 4), each characterized by different values for mean (μ_s) and variance (σ_s^2). These values are controlled using three predefined levels:

- **h (high):** For mean (μ_h), this represents high confidence in detection. For variance (σ_h^2), this represents unstable confidence (i.e., high fluctuation)
- **m (medium):** For mean (μ_m), this represents moderate confidence. For variance (σ_m^2) , this represents moderate fluctuation.
- **l (low):** For mean (μ_l), this represents low detection confidence. For variance (σ_l^2), this represents stable confidence (i.e., low fluctuation)

In the experiment, modalities changes their reliability as follows:

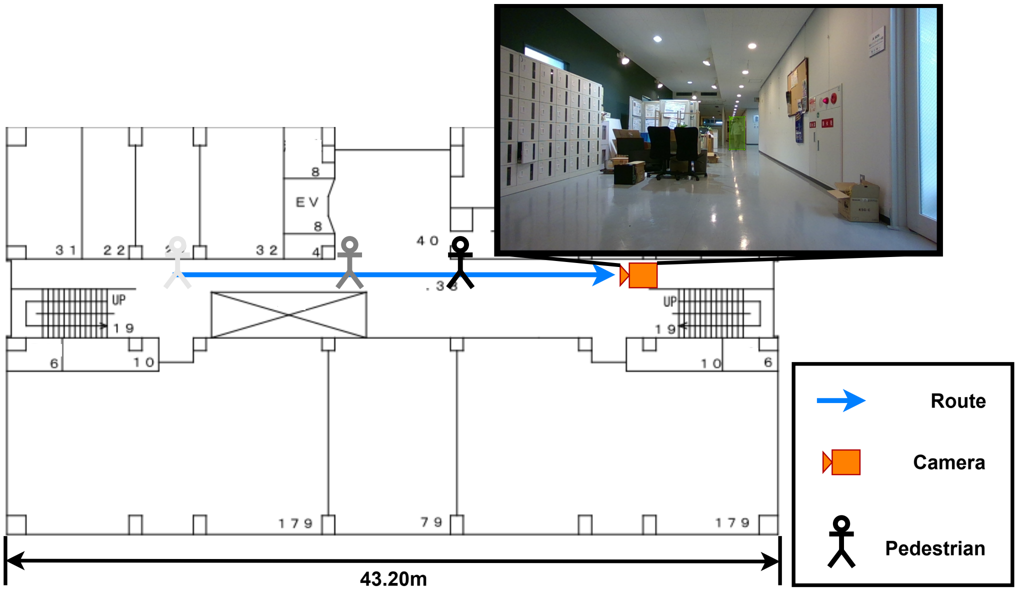


Figure 4: **Test scene description.** The person coming down the hallway towards the front of the camera.

- **Simulated Modality:** The simulated modality moves through different reliability states over time.
- **RGB Modality:** The RGB-based recognition scores become more stable after frame 300, reflecting improved tracking accuracy.

This experimental design evaluates whether the proposed method can dynamically adjust sensor weighting based on real-time reliability assessment.

In this experiment, the performance of the method is evaluated using whether the output data satisfy the following conditions with respect to a set threshold θ.

- Conditions in which a score above a set threshold θ should be output ($I_{(T)} \geq \theta$):
  - only when at least one modality is reliably detecting an object.
  - Example: If the RGB modality is stable and exhibits high confidence, the integrated score should follow it closely and exceed θ.
- Conditions in which a score below a set threshold θ should be output ($I_{(T)} < \theta$):

  - if neither modality provides a reliable detection.
  - Example: If both RGB and the simulated modality exhibit low confidence and high variance, the integrated score should not be falsely elevated.

By assessing whether the integrated score aligns with sensor reliability trends, this experiment evaluates the effectiveness of the proposed dynamic sensor fusion method. For example, the simulated modality parameters set as ($p = 1$, $\mu_s = llhh$, $\sigma_s = hhll$), the integrated score should remain below θ at Phase 1 because of the unstable detection period ($\sigma_s = h$) and it should exceed θ at Phase 3 because of the stable period and high-confidence ($\mu_s = h, \sigma_s = l$). By setting these conditions, the experiment evaluates whether the proposed fusion method can effectively differentiate between reliable and unreliable sensor inputs.

4.1.3 Metrics

To assess performance quantitatively, we use the confusion matrix-based evaluation:

- **True Positive (TP):** The integrated score correctly surpasses the threshold when a reliable modality's score is also above the threshold. (i.e., when at least one modality with low variance and high confidence exceeds θ, the integrated score does too.)
- **False Negative (FN):** The integrated score incorrectly remains below the threshold when a reliable modality's score is above the threshold. (i.e., a high-confidence, low-variance modality detects an object, but the integrated score fails to follow it.)
- **False Positive (FP):** The integrated score incorrectly surpasses the threshold when no reliable modality's score is above the threshold. (i.e., neither modality is reliable, but the integrated score falsely detects an object.)
- **True Negative (TN):** The integrated score correctly remains below the threshold when no reliable modality's score exceeds the threshold (i.e., both modalities indicate low confidence, and the integrated score appropriately does not detect an object)

Using these, we compute the following performance metrics:

Table 1: Parameters

Parameter	Value
Total Frames (N)	560
Detection Threshold (θ)	0.65, 0.75 ,0.85
Time Window (W)	5 , 10
Prior Distribution (μ_p, σ_p)	0.5, 0.05
Simulated Modality Pattern (p)	1,2
Simulated Modality Average Score ($\mu_l, \mu_m.\mu_h$)	$p = 1$, (0.15, 0.75, 0.90)
	$p = 2$, (0.25, 0.50, 0.80)
Simulated Modality Standard Deviation ($\sigma_l, \sigma_m.\sigma_h$)	$p = 1$, (0.015, 0.030, 0.13)
	$p = 2$, (0.030, 0.050, 0.15)

- **Accuracy (ACC):** Measures the overall correctness of object classification.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{13}$$

- **Precision (Prec):** Evaluates how many positive predictions are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{14}$$

- **Recall (Rec):** Assesses how many actual positives were correctly detected.

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{15}$$

- **Specificity(Spec):** Measures the method's ability to avoid false positives.

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{16}$$

These metrics allow us to evaluate the proposed method's ability to maintain high performance while reducing false positives and false negatives.

4.1.4 Parameter Settings

Table 1 lists the experimental parameters used in the evaluation. To assess the sensitivity of the proposed reliability-weighted fusion method under varying conditions, multiple detection thresholds (θ) and simulated modality patterns (p) were tested. The detection threshold (θ) was evaluated at three different values: 0.65, 0.75, and 0.85 allowing us to analyze how the trade-off between detection sensitivity and specificity affects performance. The simulated modality follows two distinct patterns ($p = 1$, $p = 2$), designed to resemble different sensor characteristics:

- **$p = 1$:** Simulates a modality similar to RGB-based recognition with higher confidence for high-reliability conditions.
- **$p = 2$:** Represents a modality similar to Point Cloud-based recognition, with more balanced confidence levels across different reliability states

For each simulated modality pattern, the mean detection confidence (μ_l, μ_m, μ_h) and standard deviation (σ_l, σ_m, σ_h) are predefined.

4.2 Experimental Results

This section presents the evaluation results of the proposed reliability-weighted multimodal object recognition method. The experiments examine how dynamically adjusting sensor reliability influences detection performance under varying conditions.

- **Averaging Method:** Computes the mean of detection scores from both modalities. This method is less affected by outliers but is influenced by modalities with low reliability. By comparing this method, we can evaluate the difference between integration approaches that incorporate dynamic reliability and those that do not.
- **Max Selection Method:** Selects the highest score from either modality. It is effective in scenarios where certain sensors function strongly but carries the risk of over-relying on unstable sensors. By comparing it with other approaches, we can assess its robustness against environmental changes.

Each method was tested under different reliability conditions for simulated modality that were characterized by variations in mean detection confidence (μ_s) and detection score variance (σ_s^2). The analysis focuses on four key metrics as we mentioned above: accuracy (ACC), precision (Prec), recall (Rec). and specificity (Spec).

4.2.1 Quantitative Results

Table 2 summarizes the performance comparison between the proposed method and the baseline approaches. The results show that the proposed method consistently achieves higher precision and specificity, particularly in high-variance conditions ($\sigma_s = h$), where sensor reliability fluctuates. In stable conditions ($\sigma_s = l$), all three methods achieve similar accuracy, indicating that when sensor reliability is high, fusion complexity is minimal. However, in high-variance scenarios, the proposed method consistently outperforms the baselines, achieving higher specificity while reducing false positive. For instance, in the ($p = 1, \mu_s = llll, \sigma_s = hhhh$) condition at $\theta = 0.75$, the proposed method achieves $Spec = 1.00$, while the averaging method only reaches $Spec = 1.00$, but with significantly lower accuracy, confirming that it avoids being influenced by unreliable sensor inputs.

The effect of threshold variations is also evident. Comparing ($p = 2, \mu_s = lmhh, hmll$) at $\theta = 0.75$ and $\theta = 0.85$ shows that increasing the threshold improves precision (remains at 1.00) but reduces recall (0.85 $\rightarrow$ 0.73). This confirms that raising θ filters out uncertain detections, making recognition more selective at the cost of missing some objects. Conversely, in ($p = 2, \mu_s = llmm, \sigma_s = mmmm$) at $\theta = 0.65$ and $\theta = 0.75$, lowering the threshold improves recall (0.81 $\rightarrow$ 0.85) without significantly specificity (remains at 1.00). This suggests that in moderate-reliability conditions, reducing θ increases detection rates without increasing false positive. Another comparison between ($p = 1, \mu_s = mhhh, \sigma_s = mlll$) and ($p = 1, \mu_s = mhhh, \sigma_s = mmmm$) at $\theta = 0.75$ highlights the role of variance. When variance increases from low to medium, recall drops slightly (0.94 $\rightarrow$ 0.80) while specificity increases (0.82 $\rightarrow$ 0.95). This confirms that higher variance reduces overall detection rates but ensures that recognized objects are more reliable, improving confidence filtering. These findings demonstrate that the proposed method effectively adapts to varying sensor reliability, maintaining robust detection performance while minimizing the impact of unreliable sensor inputs.

4.2.2 Qualitative Results

To further analyze the performance of the proposed method, we examine score transitions over time under various conditions. Figures 5–7 provide a qualitative comparison

Table 2: **Performance comparison of object recognition.** We conducted repeated experiments while varying factors such as the properties of the simulated modality, the threshold, the prior distribution of scores, and the values of the time window.

Cond.			**Thresh**	**Prior Dist.**	**W**	**Ours**				**Average**				**Sampling Max**			
p	μ_s	σ_s	θ	μ_p, σ_p	W	Acc	Prec	Rec	Spec	Acc	Prec	Rec	Spec	Acc	Prec	Rec	Spec
1	mhhh	mmmm	0.75	0.5, 0.05	5	0.93	0.94	0.87	0.95	0.92	0.86	1.00	0.86	0.49	0.47	1.00	0.04
1	mhhh	mlll	0.75	0.5, 0.05	5	0.90	0.94	0.94	0.82	0.78	1.00	0.72	0.99	0.75	0.76	1.00	0.04
1	lmhh	hmml	0.75	0.5, 0.05	5	0.96	1.00	0.91	1.00	0.96	0.93	0.99	0.92	0.75	0.67	1.00	0.50
1	lmhh	hmll	0.75	0.5, 0.05	5	0.79	1.00	0.71	1.00	0.80	1.00	0.72	1.00	0.97	0.98	1.00	0.95
1	llhh	hhll	0.75	0.5, 0.05	5	0.91	1.00	0.82	1.00	0.95	0.94	0.99	0.94	0.94	0.90	1.00	0.89
1	lllh	hhhl	0.75	0.5, 0.05	5	0.93	1.00	0.85	1.00	0.89	1.00	0.76	1.00	0.94	0.89	1.00	0.89
1	llll	hhhh	0.75	0.5, 0.05	5	0.92	1.00	0.83	1.00	0.54	0.00	0.00	1.00	0.94	0.89	1.00	0.89
2	mhhh	hhhh	0.75	0.5, 0.05	5	0.93	1.00	0.85	1.00	0.93	0.88	0.97	0.89	0.82	0.63	1.00	0.50
2	lmhh	hmmm	0.75	0.5, 0.05	5	0.93	1.00	0.85	1.00	0.93	0.88	1.00	0.89	0.91	0.83	1.00	0.82
2	llmm	mmmm	0.75	0.5, 0.05	5	0.91	1.00	0.81	1.00	0.54	1.00	0.03	1.00	0.94	0.89	1.00	0.89
2	lllh	hhhl	0.75	0.5, 0.05	5	0.93	1.00	0.85	1.00	0.88	1.00	0.73	1.00	0.94	0.89	1.00	0.89
2	lmhh	hmll	0.85	0.2, 0.05	5	0.87	1.00	0.73	1.00	0.94	0.98	0.90	0.98	0.96	0.94	1.00	0.94
2	mhhh	mlll	0.65	0.6, 0.05	5	0.86	1.00	0.81	1.00	0.78	1.00	0.73	0.99	0.87	0.86	1.00	0.51
2	llmm	mmmm	0.65	0.6, 0.05	5	0.93	1.00	0.85	1.00	0.97	0.93	1.00	0.94	0.93	0.87	1.00	0.87
1	mhhh	mmmm	0.75	0.5, 0.05	10	0.82	0.94	0.58	0.97	0.92	0.86	1.00	0.86	0.49	0.47	1.00	0.04
1	mhhh	mlll	0.75	0.5, 0.05	10	0.48	0.91	0.48	0.86	0.79	1.00	0.72	0.99	0.75	0.76	1.00	0.04
1	llhh	hhll	0.75	0.5, 0.05	10	0.84	1.00	0.68	1.00	0.94	0.94	0.99	0.94	0.94	0.90	1.00	0.89
2	lmhh	hmmm	0.75	0.5, 0.05	10	0.80	1.00	0.58	1.00	0.94	0.88	1.00	0.89	0.91	0.83	1.00	0.82
2	llmm	mmmm	0.75	0.5, 0.05	10	0.80	1.00	0.58	1.00	0.54	1.00	0.03	1.00	0.94	0.89	1.00	0.89
2	llll	hhhh	0.75	0.5, 0.05	10	0.80	1.00	0.58	1.00	0.54	0.00	0.00	1.00	0.94	0.89	1.00	0.89

between the proposed reliability-weighted fusion method and baseline approaches. The figures illustrate how each method responds to fluctuating sensor reliability. The colors represent different data sources: Blue lines represent simulated modality, orange lines represent RGB-based video modality, green lines represent averaging method, dotted red lines represent max selection method, and red lines represent proposed method.

Firstly, we tested the effectiveness of the method by fixing the values of θ, μ_p, and σ_p (Fig. 5). The proposed method maintains smoother score transitions than both the averaging and max selection methods. Specifically, in Fig. 5a, the max selection method exhibits sudden spikes, whereas the proposed method stabilizes detection. It suggests that the proposed method filters out noisy detections and maintains robust integration.

Secondly, in the case of Pattern 2, we checked how the transition would change if the mean and standard deviation of the hypothetical were changed. In Fig. 6a, max selection frequently exceeds the threshold due to unreliable peaks, while the proposed method remains stable. In contrast, the proposed fusion approach effectively prevents unreliable detections by prioritizing stable modality inputs.

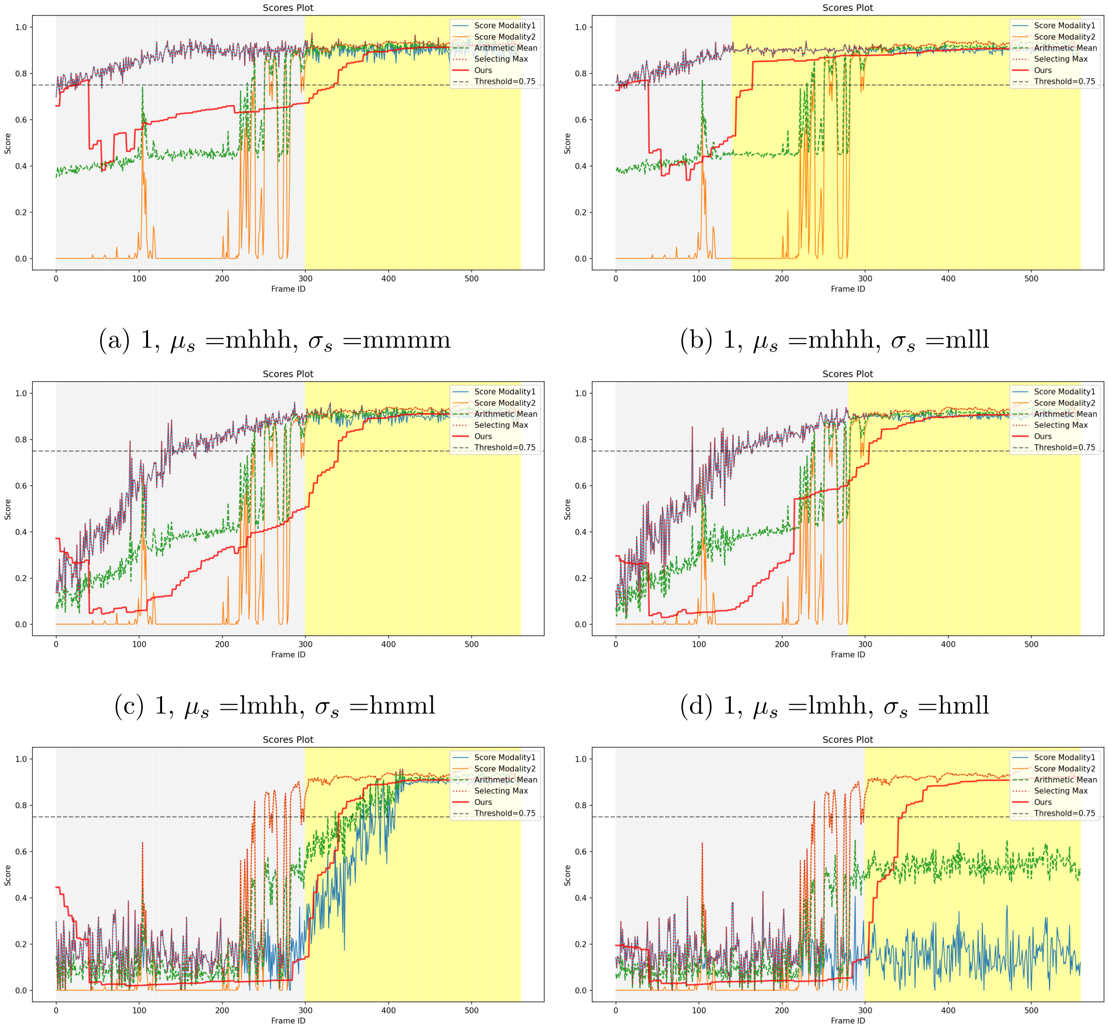
Thirdly, we compared the cases with fixed patterns and varying thresholds. In Fig. 7, increasing the threshold reduces false positives but may also miss some valid detections. A moderate threshold ($\theta = 0.75$) balances precision and recall, while lower thresholds increase sensitivity and higher threshold improve recall.

Finally, we tested the effect of Time Window W. A longer W helps to filter out short-term noise, stabilizing detection decisions. A shorter W enables quicker adaptation, which can be beneficial when sensor reliability changes rapidly but may also introduce higher sensitivity to noise.

4.3 Discussion

The experimental results show that the proposed reliability-weighted fusion method effectively reduces false positives (FP) while maintaining stable detection performance. In particular, the results confirm that the method prevents unreliable high-variance scores from disproportionately influencing the final integrated detection score.

While the results demonstrate the effectiveness of the proposed method, the current evaluation setup has several limitations that should be addressed in future research.



(a) 1, μ_s =mhhh, σ_s =mmmm

(b) 1, μ_s =mhhh, σ_s =mlll

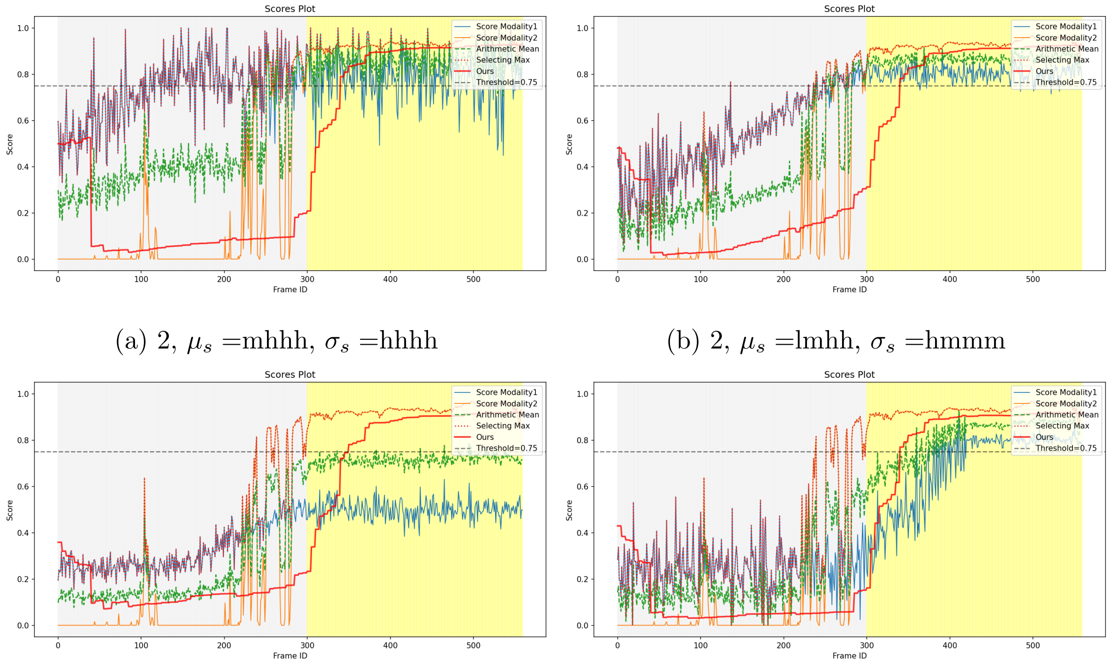
(c) 1, μ_s =lmhh, σ_s =hmml

(d) 1, μ_s =lmhh, σ_s =hmll

(e) 1, μ_s =lllh, σ_s =hhhl

(f) 1, μ_s =llll, σ_s =hhhh

Figure 5: **Score plots of pattern ($p = 1$, $\theta = 7.5$, $\mu_p = 0.5$, $\sigma_p = 0.05$, $W = 5$).** This figure presents the score transitions over time for Pattern 1 with detection threshold $\theta = 0.75$ and prior distribution ($\mu_p = 0.5$, $\sigma_p = 0.05$). μ_s and σ_s are simulated modality's mean and standard deviation as mentioned at Section 4.1.2. (a, b) Even when the simulated modality's mean score is high, the proposed method avoids being influenced when variance is large (highlighted yellow regions). (c, d) When the simulated modality increases earlier than the video modality, similar behavior is observed. (e) If the simulated modality score increases later, a different adaptation trend is seen. (f) When the simulated modality score remains consistently low and fluctuates, the proposed method successfully avoids using unreliable sensor inputs, indicating that this sensor is unsuitable for the task.



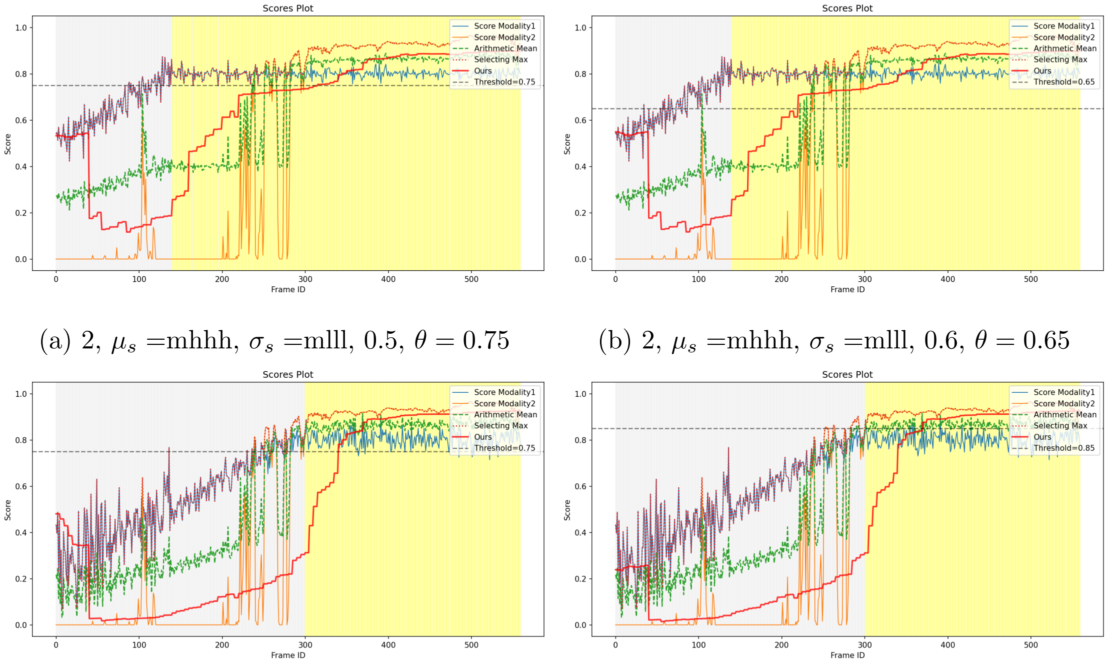
(a) 2, μ_s =mhhh, σ_s =hhhh

(b) 2, μ_s =lmhh, σ_s =hmmm

(c) 2, μ_s =llmm, σ_s =mmmm

(d) 2, μ_s =lllh, σ_s =hhhl

Figure 6: **Score plots of pattern (**$p = 2$**,** $\theta = 7.5$**,** $\mu_p = 0.5$**,** $\sigma_p = 0.05$**,** $W = 5$**).** Score transition plot for Pattern 2 under the same conditions as Figure 6 ($\theta = 0.75$, $\mu_p = 0.5$, $\sigma_p = 0.05$). (a, b) Compared to Pattern 1, the effect of high variance on reliability assessment is reduced, and the proposed method follows stable detections more accurately. (c, d) In early rising simulated modality scores, the proposed method again follows the stable modality rather than the fluctuating unreliable one.



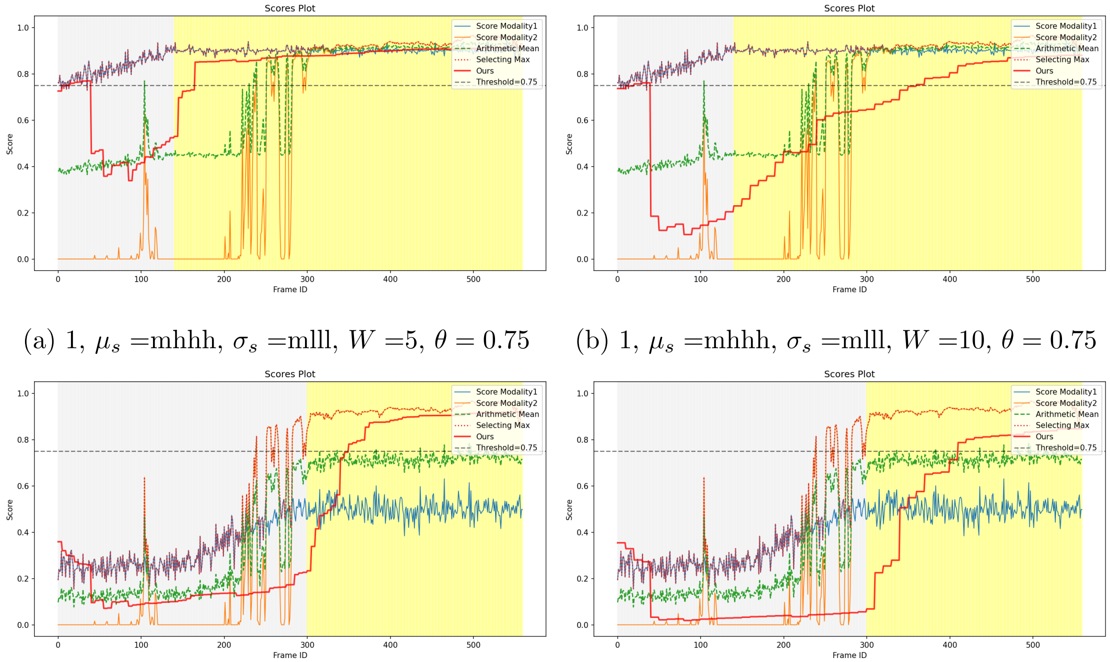
(a) 2, μ_s =mhhh, σ_s =mlll, 0.5, $\theta = 0.75$

(b) 2, μ_s =mhhh, σ_s =mlll, 0.6, $\theta = 0.65$

(c) 2, μ_s =lmhh, σ_s =hmmm, 0.5, $\theta = 0.75$

(d) 2, μ_s =lmhh, σ_s =hmmm, 0.2, $\theta = 0.85$

Figure 7: **Score plots with varying thresholds ($\theta = 0.65,\ 0.75,\ 0.85$).** This figure compares the effect of different detection thresholds (θ) on score transitions, demonstrating how varying the threshold influences detection behavior. (a, c) At a moderate threshold threshold ($\theta = 0.85$), recall improves, but some unreliable detections persist. (b) At a low threshold ($\theta = 0.65$), the balance between precision and recall is maintained, reducing false positives while still detecting. (d) At a high threshold ($\theta = 0.85$), precision improves significantly, but some valid detections are missed.



(a) 1, μ_s =mhhh, σ_s =mlll, W =5, $\theta = 0.75$

(b) 1, μ_s =mhhh, σ_s =mlll, W =10, $\theta = 0.75$

(c) 2, μ_s =llmm, σ_s =mmmm, W =5, $\theta = 0.75$

(d) 2, μ_s =llmm, σ_s =mmmm, W =10, $\theta = 0.75$

Figure 8: **Score plots with varying time window (W = 5, W = 10) .** The figure illustrates how varying the time window affects detection stability and responsiveness: (a, b) When the simulated modality's score rises earlier, the proposed method adapts efficiently, preventing abrupt changes despite variance fluctuations. A longer time window results in smoother score transitions, while a shorter window allows faster adjustments but increases fluctuations. (c, d) When the simulated modality's score increases later, the proposed method maintains robustness without being overly affected by early low-confidence scores. A shorter time window reacts more quickly, while a longer time window prevents unnecessary oscillations.

The current method assumes that the recognition results provided from multiple modalities correspond to the same object. In real-world applications, object recognition errors or misalignment between modalities could introduce inconsistencies in fusion. Future work should explore robust matching techniques to ensure that recognition results are reliably associated across modalities.

The experiments were conducted using a simplified simulation environment, where sensor reliability variations were modeled in a controlled manner. While this allows for a systematic analysis of reliability-based fusion, real-world sensor data may exhibit more complex variations, such as occlusions, dynamic environmental changes, or noise patterns that differ from the simulated conditions. Future work should include real-world experiments to validate the effectiveness of the proposed approach in practical scenarios.

4.4 Contribution of This Work

The proposed method enables multimodal recognition based on reliability, thereby allowing for flexible decision-making in robot control. This section discusses the contributions of the proposed approach through its application in real-world robotic systems.

Beyond controlling the movement of collaborative robots, this method is expected to be applicable to systems requiring decision-making in complex environments. By evaluating variations in scores obtained from multiple sensors in real time, robots can perform adaptive braking operations. For instance, if the integrated score is 0.3, the system considers the possibility of an object's presence and begins to decelerate. When the score exceeds 0.7, the robot recognizes a potential hazard and applies stronger braking. Decision-making based on reliability helps reduce unnecessary sudden braking and incorrect avoidance maneuvers, ultimately enhancing safety. Furthermore, by incorporating a mechanism that allows the system to learn the energy required for restarting after braking and the additional effort involved in route adjustments, overall energy efficiency is expected to improve.

The findings of this study contribute not only to the advancement of autonomous systems but also serve as a significant step toward achieving safer and more efficient robot control.

5 Conclusion

In this thesis, we proposed a reliability-weighted multimodal object recognition method inspired by human perception. The method dynamically adjusts sensor contributions based on time-dependent variations in reliability, addressing limitations in conventional multimodal fusion approaches that assume constant sensor reliability.

The key contributions of this research can be summarized as follows:

- Proposal of a reliability-weighted multimodal integration method
  - We introduced MLE model-based framework that dynamically adjusts sensor reliability over time, preventing unreliable high-variance scores from dominating recognition decisions.
- Experimental validation with real-world video data and simulated modality inputs
  - We conducted experiments using RGB-based recognition results combined with a simulated modality, where mean confidence and variance varied over time to simulate environmental changes.
- Improvement in recognition robustness and false positive suppression
  - The proposed method significantly reduced false positives (FP) while maintaining stable score transitions, as demonstrated in Sections 4.2.
  - Compared to baseline methods (arithmetic averaging and max selection), the method achieved higher precision and specificity, particularly under high-variance conditions.

Moving forward, future research will focus on improving cross-modal feature matching, validating the method in real-world environments, and expanding its applicability to multi-object recognition. Furthermore, integrating this approach with more advanced perception frameworks will be essential to enhance its robustness and practicality in complex, real-world applications.

Acknowledgments

I would like to express my deepest gratitude to Professor Masayuki Murata of the Graduate School of Information Science and Technology, Osaka University, for his invaluable guidance and support throughout my research. I would like to express my sincere appreciation to Professor Hideyuki Shimonishi of d3 center for his suggestions for my research. I am also sincerely grateful to Associate Professor Shin'ichi Arakawa of the Graduate School of Information Science and Technology, Osaka University, Associate Professor Yuichi Ohsita of D3 Center, Osaka University, for their objective reviews. Furthermore, I would also like to express my sincere appreciation to Associate Professor Daichi Kominami of the Graduate School of Information Science and Technology, Osaka University, research initiatives for his continuous advice and encouragement, which enabled me to advance my research. Finally, I extend my heartfelt appreciation to all members of the Murata Laboratory at the Graduate School of Information Science and Technology, Osaka University, for their insightful feedback, valuable assistance, and generous support in both conducting this research and preparing this thesis.

References

[1] L. D. Evjemo, T. Gjerstad, E. I. Grøtli, and G. Sziebig, "Trends in smart manufacturing: Role of humans and industrial robots in smart factories," *Current Robotics Reports*, vol. 1, no. 2, pp. 35–41, 2020. [Online]. Available: https://doi.org/10.1007/s43154-020-00006-5

[2] S. Phuyal, D. Bista, and R. Bista, "Challenges, opportunities and future directions of smart manufacturing: A state of art review," *Sustainable Futures*, vol. 2, p. 100023, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666188820300162

[3] B. Wang, F. Tao, X. Fang, C. Liu, Y. Liu, and T. Freiheit, "Smart manufacturing and intelligent manufacturing: A comparative review," *Engineering*, vol. 7, no. 6, pp. 738–757, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809920302502

[4] X. Ding, J. Guo, Z. Ren, and P. Deng, "State-of-the-art in perception technologies for collaborative robots," *IEEE Sensors Journal*, vol. 22, no. 18, pp. 17 635–17 645, 2022.

[5] L. Zhou, L. Zhang, and N. Konz, "Computer vision techniques in manufacturing," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 1, pp. 105–117, 2023.

[6] Y. Liu, S. Wang, Y. Xie, T. Xiong, and M. Wu, "A review of sensing technologies for indoor autonomous mobile robots," *Sensors*, vol. 24, no. 4, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/4/1222

[7] R. Roriz, J. Cabral, and T. Gomes, "Automotive lidar technology: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6282–6297, 2022.

[8] X. Yu and M. Marinov, "A study on recent developments and issues with obstacle detection systems for automated vehicles," *Sustainability*, vol. 12, no. 8, 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/8/3281

[9] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, and L. Wang, "Robustness-aware 3D object detection in autonomous driving: A review and outlook," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 15 407–15 436, 2024.

[10] D. Wu, F. Yang, B. Xu, P. Liao, and B. Liu, "A survey of deep learning based radar and vision fusion for 3D object detection in autonomous driving," 2024. [Online]. Available: https://arxiv.org/abs/2406.00714

[11] Z. Wei, F. Zhang, S. Chang, Y. Liu, H. Wu, and Z. Feng, "Mmwave radar and vision fusion for object detection in autonomous driving: A review," *Sensors*, vol. 22, no. 7, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2542

[12] X. Wang, K. Li, and A. Chehri, "Multi-sensor fusion technology for 3d object detection in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1148–1165, 2024.

[13] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] C. R. Fetsch, G. C. DeAngelis, and D. E. Angelaki, "Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons," *Nature Reviews Neuroscience*, vol. 14, no. 6, pp. 429–442, 06 2013. [Online]. Available: https://doi.org/10.1038/nrn3503

[15] H. Colonius and A. Diederich, "Formal models and quantitative measures of multisensory integration: a selective overview," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1161–1178, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.13813

[16] L. Li, R. Rehr, P. Bruns, T. Gerkmann, and B. Röder, "A survey on probabilistic models in human perception and machines," *Frontiers in Robotics and AI*, vol. 7, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2020.00085

[17] R. Seki, D. Kominami, H. Shimonishi, M. Murata, and M. Fujiwaka, "Multi-object recognition method inspired by multimodal information processing in the human brain," in *Proceedings of IEEE Global Communications Conference, Workshops*, 2022, pp. 569–574.

[18] Haruhito Ando, Daichi Kominami, Ryoga Seki, Masayuki Murata and Hideyuki Shimonishi, "Multimodal object recognition using bayesian attractor model for 2D and 3D data," to appear in *Proceedings of 27th Conference on Innovation in Clouds, Internet and Networks (ICIN 2024)*, March 2024.

[19] Y. Cao, C. Summerfield, H. Park, B. L. Giordano, and C. Kayser, "Causal inference in the multisensory brain," *Neuron*, vol. 102, no. 5, pp. 1076–1087.e8, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627319303320

[20] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *Trends in Cognitive Sciences*, vol. 8, no. 4, pp. 162–169, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661304000385

[21] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[22] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: exceeding YOLO series in 2021," *CoRR*, vol. abs/2107.08430, 2021. [Online]. Available: https://arxiv.org/abs/2107.08430

[25] J. Yang, S. Liu, Z. Li, X. Li, and J. Sun, "Real-time object detection for streaming perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5385–5395.

[26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of European conference on computer vision.* Springer, 2020, pp. 213–229.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision– ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[28] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 689–12 697.

[30] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. [Online]. Available: https://doi.org/10.1177/0278364913491297

[31] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, 2019.

[32] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 386–10 393.

[35] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisensory perception," *PLOS ONE*, vol. 2, no. 9, pp. 1–10, 09 2007. [Online]. Available: https://doi.org/10.1371/journal.pone.0000943

[36] T. Rohe and U. Noppeney, "Sensory reliability shapes perceptual inference via two mechanisms," *Journal of Vision*, vol. 15, no. 5, pp. 22–22, 04 2015. [Online]. Available: https://doi.org/10.1167/15.5.22

[37] T. Rohe, A.-C. Ehlis, and U. Noppeney, "The neural dynamics of hierarchical bayesian causal inference in multisensory perception," *Nature communications*, vol. 10, no. 1, p. 1907, 2019.

[38] D. C. Knill, "Mixture models and the probabilistic structure of depth cues," *Vision Research*, vol. 43, no. 7, pp. 831–854, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0042698903000038