# Multimodal Object Recognition Using Bayesian Attractor Model for 2D and 3D data

Haruhito Ando*, Daichi Kominami*, Ryoga Seki*, Masayuki Murata*, Hideyuki Shimonishi†
* Graduate School of Information Science and Technology
Osaka University, Osaka, Japan
Email: {h-ando, d-kominami, r-seki, murata}@ist.osaka-u.ac.jp
† Cybermedia center, Osaka University, Osaka, Japan
Email: shimonishi.cmc@osaka-u.ac.jp

***Abstract*—The advent of beyond 5G/6G technologies is expected to revolutionize the development of digital twins. Recent development of sensors and deep learning techniques has significantly improved the precision of perception of surroundings, and multimodal object recognition approaches, which play an important role in the realization of the digital twin, have been studied. Previously, we developed a multimodal object recognition method using RGB and depth images. Although our method handled multimodal data sets, there are two problems with this method: Low accuracy of the depth image and dependence on the results of video recognition. The dependence is due to the fact that the recognition results of the RGB images are used to determine the distance between the object and the camera. In this paper, we propose advanced multimodal object recognition methods that expand on our previous method. The proposal includes an additional location-modal approach that utilizes PointNet for semantic segmentation and location estimation. We evaluate our method using our prepared data set consisting of RGB-D images and 3D point clouds, considering the situations where workers and robots collaborate in the warehouse. Our results show that multimodal recognition achieves a better precision score than only the video modality under uncertain measurements.**

***Index Terms*—Digital twin, Object recognition, Multimodality, Bayesian attractor model, Bayesian causal inference**

## I. Introduction

In next-generation mobile communication systems (beyond 5G/6G), the recognition of various objects in a real world by computers will form the essential foundation of the future digital world [1]. Massive advances in sensor technology, including visual, tactile, and LiDAR (Light Detection And Ranging) sensors, are poised to enhance the machine's ability to perceive environments with more precision and create more detailed digital twins [2], [3]. Furthermore, sensor data analysis methodologies are rapidly evolving, driven by the growth of deep learning technologies [4]. In the expected applications in 6G networks, for example warehouse robotics, where robots work with human workers, the need for precise sensing and operation is paramount for safety and efficiency [5].

However, as the use of various sensors becomes more prevalent for recognition tasks, a critical limitation arises when relying on a single sensor for comprehensive environment sensing. Single-sensor systems often provide incomplete understanding of environments influenced by various uncertainties [6]. For example, in video camera-based sensing, the way light reflects influences the recorded data clarity. Additionally, the distribution of the target's movement affects the instance analysis. Therefore, recognition that takes these uncertainties into account is becoming important.

Our previous study has successfully realized multimodal object recognition under uncertainties utilizing two models inspired by brain information processing, which make decisions under uncertain multimodal observations [7]. For instance of the advantages of multimodal recognition, at the decision-making level, when it is difficult to visually distinguish between a dog and a cat –it indicates that the information obtained from vision is uncertain–, the sound of barking heard by the ears can lead us to recognize that the object is a dog. We have adapted two models based on this brain system for object recognition, the Bayesian Attractor Model (BAM) [8] for the sensory level, and the Bayesian Causal Inference (BCI) [9], [10] for the decision-making level. The BAM simulates how a person makes decisions based on the observed information, using Bayesian estimation, and the BCI represents the process by which humans identify a signal source using various sensory modalities, such as vision and auditory.

Since we have shown the effectiveness of multimodal object recognition with brain architecture, several challenges remain. One is that recognition using 2D video information is difficult to grasp three-dimensional movements of objects. In addition, the depth acquired by stereo cameras, which many 2D cameras use, is not as precise as the distance that can be handled. Another is that the feature extraction techniques that we use have dependencies before the integration of modality in the decision-making level.

The purpose of this paper is to propose a stochastic object recognition method by expanding our previous methods [7] using multimodal input with a mutually independent feature extract network (Fig. 1). We use a modality that captures location information more directly than depth images. To do so, we propose a new object recognition method that leverages point-cloud data obtained from a LiDAR sensor, which can achieve a reliable detection range greater than the depth perception by a stereo camera. We use a semantic segmentation technique (PointNet [11]) for obtaining the location of an object. To evaluate our method, we also created a data set

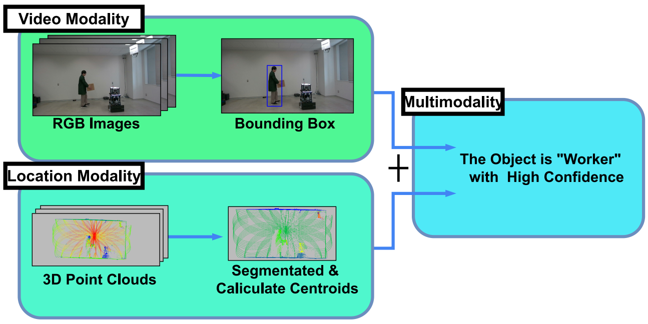


Fig. 1. **Flow of multimodal object recognition.** RGB images and 3D point clouds are utilized for object recognition in our experiments. These data sets are analyzed using separate networks and inferences will be generated for each objects with confidence independently. Finally, these results are integrated and the final decision is made taking into account their confidence.

using a LiDAR sensor and an RGB camera, combining a three-dimensional point cloud with an RGB image. In our model, we confirmed that a multimodal approach achieves a higher perception rate than a unimodal method that uses only the video modality. We also observed that this approach reduces the errors caused by the dependence on both modalities, which occurred in our previous method.

The key contributions of our work are as follows:

- Enhancing multimodal object recognition for location modalities
- Demonstrating the applicability of brain architecture in processing diverse multimodal inputs

Section II introduces our previous research and related studies. Our proposed methodology is described in Section III, and we verify its effectiveness using our new data set in Section IV. Section V concludes with a brief summary.

## II. Object Recognition Method Inspired by Brain Perception

In this section, we present our previous research that is strongly related to the proposal in this paper. First, we show a unimodal object recognition method that uses the *Bayesian Attractor Model (BAM)* [8]. The BAM computes subjective confidence in what is being observed. Details are described in Section II-A. Subsequently, we detail the modality integration model using the *Bayesian Causal Inference (BCI)* [9] in Section II-B. Finally, we present an overview of the multimodal object recognition method [7] in Section II-C. This paper shows an extension of this method.

### A. Bayesian attractor model

Bayesian attractor model [8] represents a human perceptual decision making by merging two prominent ideas into a new model. This model combines attractor dynamics and Bayesian inference. The model consists of an observed value $\mathbf{x_t}$, an internal state $\mathbf{z_t}$ for time $t$, and attractor $\phi_i$ for decision option $i$. The BAM makes a decision by updating $\mathbf{z_t}$ from the accumulated evidence for different decision options, as the activity of neurons is mutually inhibitory. Using Bayesian theory, the posterior distribution of $\mathbf{z_t}$, i.e., $P(\mathbf{z_t})$, is updated with observed values and a generative model.

The generative model of $\mathbf{z_t}$ is based on the Hopfield dynamics $f$, which has constant points in the internal state, called attractors. This is given by:

$$\mathbf{z}_t - \mathbf{z}_{t-\Delta t} = \Delta t f\left(\mathbf{z}_{t-\Delta t}\right) + \sqrt{\Delta t} w_t \tag{1}$$

where $\Delta t$ represents a small amount of time and $w_t$ represetns process noise. $w_t$ follows multivariate normal distribution $\mathcal{N}(0, (q^2/\Delta t)\mathbf{I})$ where $q$ means uncertainty of dynamics.

The present observed value is then forecasted on the basis of the internal state. This is given by:

$$\mathbf{x}_t = \mathbf{M}\boldsymbol{\sigma}(\mathbf{z}_t) + \mathbf{v}_t \tag{2}$$

where the $\mathbf{v}_t$ represents measurement noise at $t$, and $\mathbf{v}_t$ follows multivariate normal distribution $\mathcal{N}(0, r^2\mathbf{I})$ where $r$ means uncertainty of sensory. $r$ is the amount of noise expected in the observed data. $\sigma$ represents a sigmoid function, mapping the arguments to the range $[0, 1]$. The relationship between an attractor and its representative value of $\mathbf{x}_t$ is represented by the matrix $\mathbf{M} = [\mu_1 \ldots \mu_N]$. Since the authors of [8] define $\phi_i$ so that $\boldsymbol{\sigma}(\phi_i)$ is close to a vector with 1 in the $i$th dimension and 0 in all other dimensions, $\mathbf{x}_t \simeq \mu_\mathbf{i}$ when $\mathbf{z}_t = \phi_i$.

Using Eq. (1) and Eq. (2), Bayes' theorem allows us to estimate $P(\mathbf{z}_t|\mathbf{x}_t)$ from the actual observed values $\mathbf{x}_t$. When the BAM observes values associated with decision option $i$, $\mathbf{z}_t$ is close to $\phi_i$ and $P(\mathbf{z}_t = \phi_i|\mathbf{x}_t)$ takes a large value. We can consider that sufficient evidence has been collected to adopt option $i$ and therefore $P(\mathbf{z}_t = \phi_i|\mathbf{x}_t)$ is called the *confidence* level for decision option $i$.

The BAM can adjust and refine its internal state and predictions despite any observational inaccuracies, so utilizing it can provide robustness for a decision-making task with data including noise.

### B. Bayesian causal inference

Bayesian causal inference [9] is a statistical method to estimate the probability of a causal relationship between variables using Bayes' theorem and has been used to explain multimodal information processing in the human brain.

In this section, we explain the integration of two modalities represented by $A$ $(audio)$ and $V$ $(visual)$. The BCI estimates whether the observations in each modality $(x_A, x_V)$ originate from the same source $(C = 1)$ or from different sources $(C = 2)$. The prior probability of $C = 1$ is $p_{common}$ and is usually set to 0.5. Ref. [9] gives models for the location distribution of a common source and the distribution of observed values in each modality when $C = 1$, and for the location distribution of two sources and the distribution of observed values in each modality when $C = 2$.

From these prior distributions it is possible to infer the posterior probability of the causal structure ($C = 1$ or $C = 2$) using Bayes' rule:

$$p(C|x_A, x_V) = \frac{p(x_A, x_V|C)p_{common}}{p(x_A, x_V)} \tag{3}$$

In the context of multisensory perception of the brain, model averaging can be used to combine the predictions of different causal models that explain how different sensory signals are related to each other. That is, the integrated estimate of $\hat{x}_A$ is derived from the following equation with the estimated value $\hat{x}_{A,C}$ at $C = 1$ and $C = 2$, with $P(C = 1|x_A, x_V)$ and $P(C = 2|x_A, x_V)$ as weights (The same applies to $\hat{x}_V$).

$$\hat{x}_A = P(C = 1|x_A, x_V)\hat{x}_{A,C=1} + P(C = 2|x_A, x_V)\hat{x}_{A,C=2} \quad (4)$$

By averaging the predictions of different models with causal structure, it can make more accurate and reliable perceptual decision making, taking into account the relative strengths and weaknesses of different models. The detailed derivation of $\hat{x}_{A,C}$ is omitted here as it depends on the generative model. The BCI-based model integration yields two results, $\hat{x}_A$ and $\hat{x}_V$, as described above. For example, $\hat{x}_V$ would be used for visual tasks and $\hat{x}_A$ for auditory tasks to make decisions.

### *C. Multimodal object recognition*

Using brain-based models, the BAM and the BCI, robust multimodal object recognition has been realized in our previous work [7]. In this method, the BAM performs object recognition for video and depth modalities. In each modality, the BAM associates the features of the observed targets with the attractors as training. Inference is made using the input observations and then outputs the similarity for each target with confidence. The results obtained from these modalities are integrated using the BCI to obtain the final result.

This method becomes possible to make more accurate decision-making by compensating for each other's recognition results in situations where a unimodal method cannot make a decision, such as when a video modality has low visibility due to filming in a dark room. However, our previous method used video-based information to obtain the distance between the camera and the objects. Therefore, a decrease in the accuracy of the observed video information was directly related to a decrease in the accuracy of the depth information.

In this paper, we use 3D point-cloud data obtained from a LiDAR to directly observe the position of objects and use it to solve the simultaneous loss of accuracy in the two modalities. Our method could recognize objects even if there is a loss of accuracy in one modality. A detailed description of the method is given in the next section.

## III. Method

In this section, we show the multimodal information processing method that integrates 2D-video modality and 3D-location modality from an RGB-video camera and a LiDAR.

### *A. Object recognition based on information processing model of brain*

Integration of information acquired from multiple modalities has been studied in our previous work [7]. As described in the previous section, the problem of inaccuracy in multiple modalities can arise at the same time in our previous method. There were two reasons for this. One was that the information obtained from one modality was used to extract features from the other modality. The other is that the integration model did not include which of the two modalities should be prioritized in the integration of information using BCI.

In this paper, we expand our previous method [7] by adding a new location modal object recognition approach (Fig. 2).

Our proposal also employs a video and location modality for object recognition. For the video modality, an RGB camera is the tool of choice, and for the location modality, a LiDAR is used. A LiDAR is useful for obtaining high-precision 3D point-cloud data [13]. We gather 2D images and 3D point clouds from each device, process them through Siamese RPN [12] and PointNet [11], and subsequently extract modality-specific features.

Siamese RPN [12] takes a template image and a detection image as input and outputs the position of the detected similarities as a bounding box. It is composed of a SiameseNetwork and a Region Proposal Network. The SiameseNetwork extracts the features of the images, while the Region Proposal Network uses the extracted features to calculate the similarities between the template image and the detected image. In [7], we use this extracted feature for the input of BAM.

PointNet [11] is a neural network architecture designed to process and analyze point cloud data, which is a type of spatial data representation commonly used in computer vision and 3D perception tasks. Point clouds are sets of data points in a 3D coordinate system, where each point represents a specific position in space.

When dealing with the location modality with PointNet, it produces labeled point clouds of detected objects. This output is then converted to each centroid of the object to reduce the data dimension fed to the BAM, since we prevent the BAM from increasing the calculation time when handling high-dimensional features.

For the BCI, we use the probability distribution of each object sourced from the BAM and then merge the results across the modalities. Through this approach, we can achieve multi-modal object recognition leveraging different modalities originating from different devices.

### *B. Location-modal object recognition*

We developed a new object recognition method using PointNet to incorporate the location modality into our previous method. At first, we feed the inputted 3D point clouds into PointNet, which can perform semantic scene segmentation. Subsequently, we aggregated the semantic segmented data and calculated the centroids of each object. The aggregation is essential in this context, as the BAM tends to exhibit unstable behavior with high-dimensional input. When the data are provided to the BAM, probabilistic analysis using confidence in location modality becomes possible.

### *C. Weighted object fusion using BCI*

Following the object estimation process in the BAM for every modality, the BCI performs causal inference. The BCI

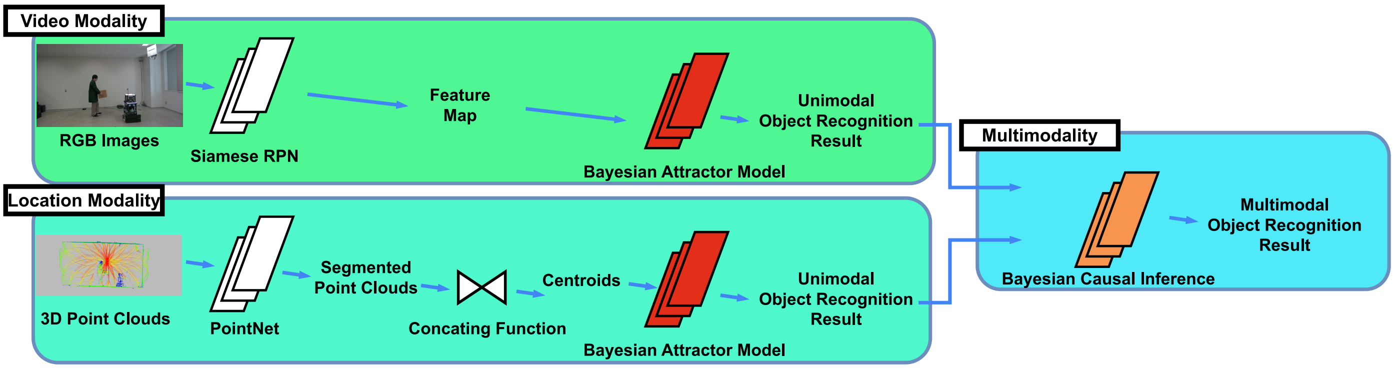
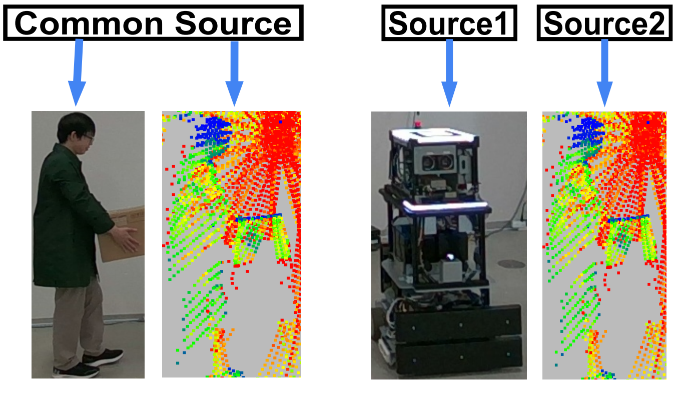


Fig. 2. **Object recognition system using 2D and 3D data based on brain system.** The data fusion method is constructed based on our previous research [7]. The diagram illustrates the process when using an RGB camera for the video modality and a LiDAR for the location modality. Our proposed approach employs Siamese RPN [12] for video modality analysis and PointNet [11] for location modality. The video-modal input for the BAM consists of a feature extracted from the image in the Bounding Box that wraps around each object for the video modality. The location-modal input for BAM consists of the center position of the objects. Using BCI for the fusion of the results of two modalities, the model allows for probabilistic recognition of objects.



(a) A common source (b) Separate sources

Fig. 3. **Causal inference model.** In the left case, representing the forced-fusion model, both RGB and PC inputs come from one common source, suggesting they identify the same subject: a worker. However, in the right case, showcasing the task-relevant model, RGB and PC are derived from separate sources, indicating different objects.

takes the confidence of the BAM as its observed value. Using causal inference, it assesses whether both the video and location modalities identify the same object. This assessment averages the results from two sensory input models: the forced-fusion model (indicating a common source) and the task-relevant unisensory segregation model (indicating separate sources). A typical example of a causal structure with a common case is shown in Fig. 3. Following this approach, model averaging (Eq. (4)) is employed to produce the object estimation from the results.

To perform BCI, we define the generative models $P(c_V, c_L|C = 1)$ and $P(c_V, c_L|C = 2)$ where $c_V$ and $c_L$ are the confidence obtained from the video and location modalities, respectively.

$$P(c_V, c_L|C = 1) = \begin{cases} 0.5 + \sigma_{c=1} & \text{if } L_V = L_L, \\ \sigma_{c=1} & \text{otherwise.} \end{cases} \tag{5}$$

$$P(c_V, c_L|C = 2) = \begin{cases} 1 & \text{if } L_V \neq L_L \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\sigma_{c=1} = 0.5\sigma(\max(c_V)) + 0.5 * \sigma(\max(c_L))$, $\sigma = 1/(1 + \exp(-x - \lambda_{bci}))$, $L_V = \mathrm{argmax}(\mathrm{c_V})$, and $L_L = \mathrm{argmax}(\mathrm{c_L})$. The generative model assumes that the confidence obtained from the BAM increases in at least one of the modalities when observing the same object. It also assumes that the recognition results in the two modalities differ when observing separate objects. The sigmoid function is used to indicate whether the confidence obtained exceeds the threshold value ($\lambda_{bci}$).

## IV. Evaluation and Results

### A. Evaluation metrics

We adopt *Precision* as the primary evaluation metric. When provided with features representing "worker", we check how much our proposal confidently recognizes the observed object as "worker." When the stakes of false positives are high or the objective is to minimize erroneous detection, precision is a key factor. It quantifies the percentage of correct positive identifications made. For instance, in applications where robots are employed in tasks involving human interaction, precision is especially crucial to avoid false positives, which are instances where an object is mistakenly identified as something else, potentially leading to significant consequences.

### B. Data set description

We need to prepare images and 3D point clouds in a time series for evaluation that meet the following requirements:

- Scenarios envisioned within a warehouse
- Captured using actual, physical equipment
- Moving entities including both workers and robots

Therefore, we created a data set with images and point clouds using the Intel RealSense D455 and Livox AVIA. We shot 60 sequences for training and 495 sequences for testing. We annotated 3D point clouds to be manageable inside the

TABLE I
KEY PARAMETERS

| Name | Value |
|---|---|
| Dynamics uncertainty($q$) | $1.0 \times 10^0$ |
| Video-modal sensory uncertainty ($r_v$) | $1.0 \times 10^1$ |
| Depth-modal sensory uncertainty($r_d$) | $1.0 \times 10^0$ |
| Location-modal sensory uncertainty ($r_l$) | $3.0 \times 10^{-1}$ |
| Confidence threshold ($\lambda_{bci}$) | $1.0 \times 10^{-5}$ |

program with 7 types of objects, "worker," "robot," "ceiling," "floor," "wall," "container," and "clutter." An example of the shots is shown in Fig. 1. For the analysis of the video image, the model learned in previous studies [7] was used. siameseRPN is a one-shot learning method and can be applied to the current study.

### C. Setting of model parameters

We set several parameters that affect results (Table I). The dynamics uncertainty and the sensory uncertainty are the main factor in the BAM, and the confidence threshold plays a role in determining the outcome of the BCI. These values should be determined on the basis of the characteristics of the input data. Inspecting the data set collected, we adjusted the values to match those presented in Table I. We also set the features obtained from the 1st frame as attractor in the BAM.

### D. Results

To begin with, we tested unimodal object recognition using the BAM in each modality; the video modality, the depth modality, and the location modality (Figs. 4(a)–4(f)). The blue line indicates the confidence of the "worker," and the orange line indicates that of the "robot." In the video modality, recognition appears to be unstable at certain points when detecting the "worker," possibly because the two observed objects have a similar coloration and area after the worker crouching (Fig. 4(a)). As mentioned above, since depth modality utilizes the results of the video modality analysis, errors in the video modality can lead to mistakes in the depth modality analysis as well in Fig. 4(b). In contrast, for the location modality, the graph clearly indicates continuous attention to the same object when detecting the "worker" (Fig. 4(c)). This is because there are only two categories to identify and they are sufficiently distanced from each other in the first frame.

Subsequently, we combined the findings from the modalities, the video and the depth modality declined in our previous research [7], video and location modality we proposed, to conduct object recognition(Figs. 5(a)–6(d)). In the previous research method, even when multiple modalities are integrated, there is a tendency to have a higher confidence in incorrect options and a lower confidence in correct ones in areas where errors are present (Fig. 5(a)). However, in our method, when the primary modality is incorrect, it is possible to maintain high confidence in the correct object in our proposed method (Fig. 6(c)). For example, if the video modality errs in recognition, but the location modality is correct with high confidence, the result of fusion will still favor the video modal with greater confidence in the incorrect object, though the confidence in the other remains high (Fig. 6(a)).

TABLE II
RECOGNITION PRECISION

| Modality | Worker | Robot |
|---|---|---|
| Single modality (video) | 1.000 | 0.594 |
| Single modality (depth) | 1.000 | 0.806 |
| Single modality (location) | 0.788 | 1.000 |
| Multi-modalities (video based with depth) | 1.000 | 0.784 |
| Multi-modalities (video based with location) | 0.936 | 0.869 |
| Multi-modalities (depth based with video) | 1.000 | 0.786 |
| Multi-modalities (location based with video) | 0.741 | 1.000 |

Finally, we present the precision for each object using different methods (only video modality, only location modality, primarily integrating the video modality and primarily integrating the location modality, Table II). We calculate the precision according to the following (7):

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$TP$ represents the total number of frames where the confidence in the correct object is higher, while $FP$ refers total number of frames that is lower.

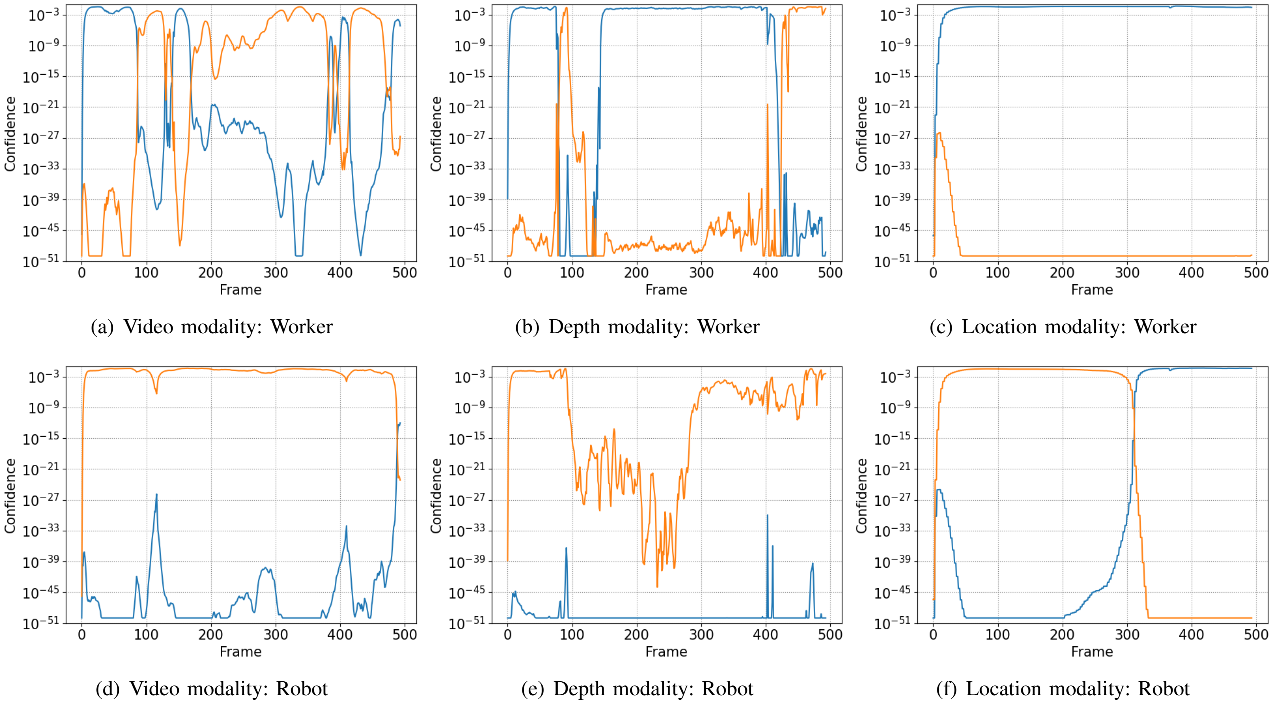
The results of multimodal integration both yield relatively high Precision for both "worker" and "robot." The precision in the video modality increases in robot recognition by integrating with the location modality. However, we also observe that if one modality is highly confident but incorrect, its influence can lead to a decrease in precision, in location modality-based fusion.

### E. Discussion

In this study, we have shown that robust and scalable multimodal object recognition can be achieved by extracting different features from independent modalities using different networks. However, we must consider the limitations of our experience, the simplicity of the task, and how to combine the results of independent recognition.

First, the simplicity of the task can be identified as one of the limitations. Presently, our tests are conducted with a binary classification task, where each class corresponds to a single object. For practical applications, especially in a real-word environment, there is a need to employ multi-class and multi-object data sets for evaluation. This limitation occurs because the feature extraction network in use is tailored for one-class-one-object recognition. Since the component can be replaced with other networks, we believe that adopting this approach will lead to a successful resolution.

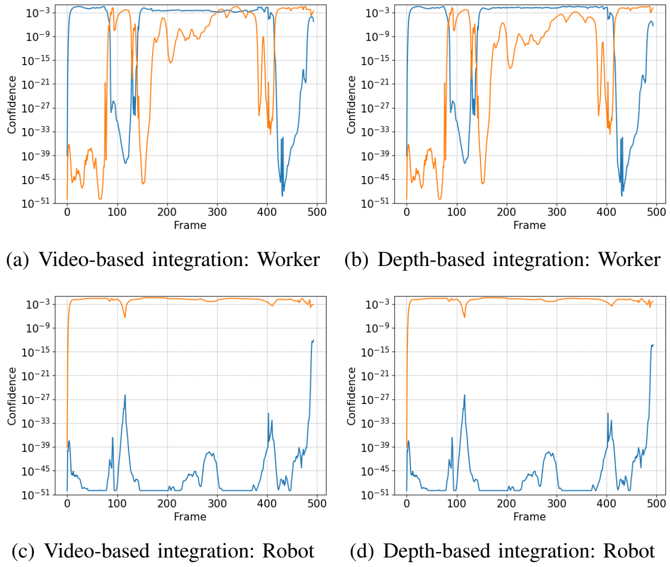
Secondly, the integration of independent modal recognition results is highlighted as a limitation. There are video and location modalities, and in both modalities, observation results are produced for both workers and robots. The premise here is that, in multimodal integration, each modality provides recognition results for "workers" and "robots" individually. Currently, the system processes integration separately for the output results observed for "workers" and those for "robots". The design



(a) Video modality: Worker (b) Depth modality: Worker (c) Location modality: Worker

(d) Video modality: Robot (e) Depth modality: Robot (f) Location modality: Robot

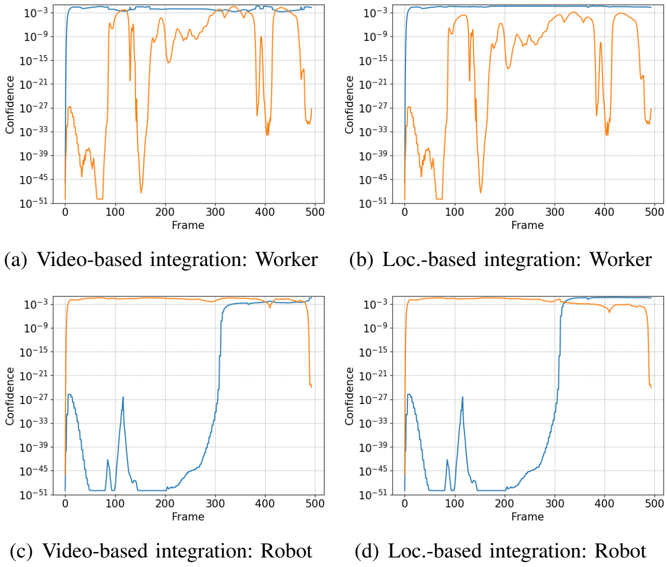
Fig. 4. **Object recognition using BAM (unimodality).** The results for the video modality are shown in Fig. 4(a) and Fig. 4(d), and those for the depth modality are shown in Fig. 4(b) and Fig. 4(e). The location modal results, obtained from the method we proposed in this paper, are shown in Fig. 4(c) and Fig. 4(f). In these figures, the vertical axis signifies the logarithmic confidence that the supplied features identify the object, and the horizontal axis signifies the temporal sequence of the input frames, each frame being updated every 33.3 milliseconds. The blue line indicates the confidence of identifying "worker," and the orange line indicates that of "robot." In the video modality and the depth modality, the probabilities show a fluctuating pattern, while in the location modality, the estimation remains constant for the same object.



(a) Video-based integration: Worker (b) Depth-based integration: Worker

(c) Video-based integration: Robot (d) Depth-based integration: Robot

Fig. 5. **Multimodal Object Recognition Using BCI (Previous Research).** Figs. 5(a) and 5(c) show the results when primarily using the video modality for the fusion, while Figs. 5(b) and 5(d) show the results when primarily using the depth modality. The blue line indicates the confidence of identifying the "worker," and the orange line indicates that of "robot." Our findings indicate that it can be observed that the benefit of acquiring diverse information through multimodal integration is not fully utilized.



(a) Video-based integration: Worker (b) Loc.-based integration: Worker

(c) Video-based integration: Robot (d) Loc.-based integration: Robot

Fig. 6. **Multimodal Object Recognition Using BCI (This Research).** Figs. 6(a) and 6(c) show the results when primarily using the video modality for the fusion, while Figs. 6(b) and 6(d) show the results when primarily using the location modality. The blue line indicates the confidence in identifying "worker," and the orange line indicates that of "robot." Our findings indicate that fusion in decision level enables the system to continue to recognize with high confidence, even when the main modality's recognition is incorrect.

should evolve to automatically handle the combination of recognition results, which are the product of the number of modalities and the number of objects. To address this, our team is conducting research to solve this issue by incorporating additional models.

## V. Conclusion

The fusion method inspired by brain information processing is applicable to object recognition using data from different sensors. Previously, the efficacy of the method was verified only in experiments using data obtained from a single sensor. Therefore, we focused on whether it could be applied to multimodal data obtained from different sensors by expanding the prior research using the location modality. With the expansion, we develop a new object recognition method that employs PointNet to handle point clouds and calculate the object locations, and integrate this with our previous methodologies. Our experiment indicates that precision improves when the video modal analysis is adjusted with the location modality in robot recognition. Furthermore, the fact that our approach can maintain high confidence in the correct candidate, even when recognition errors occur, shows its effectiveness in representing scenarios probabilistically.

We primarily aimed to prove our system's flexibility in using different sensors, so processing efficiency was not part of our scope. While existing methods that depend on a single sensor did not have issues identifying objects, our system has not yet considered several practical challenges, like how to deal with combinations of different modalities' recognition. These challenges will need to be addressed for real-world application.

## Acknowledgements

This work was partially supported by Grant No. JPJ012368C 00701 from the National Institute of Information and Communications Technology (NICT) in Japan.

## References

[1] H. Viswanathan and P. E. Mogensen, "Communications in the 6g era," *IEEE Access*, vol. 8, pp. 57 063–57 074, 2020.

[2] 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.

[3] M. Attaran and B. G. Celik, "Digital twin: Benefits, use cases, challenges, and opportunities," *Decision Analytics Journal*, vol. 6, p. 100165, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S277266222300005X

[4] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, "Multi-modal 3d object detection in autonomous driving a survey," *International Journal of Computer Vision*, pp. 1–31, 2023.

[5] S. Yasuda, T. Kumagai, and H. Yoshida, "Cooperative transportation robot system using risk-sensitive stochastic control," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 5981–5988.

[6] D. Liu, Y. Cui, Z. Cao, and Y. Chen, "Indoor navigation for mobile agents: A multimodal vision fusion model," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8.

[7] 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 *2022 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2022, pp. 569–574.

[8] S. Bitzer, J. Bruineberg, and S. J. Kiebel, "A Bayesian Attractor Model for Perceptual Decision Making," *PLOS Computational Biology*, vol. 11, no. 8, pp. 1–35, august 2015. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1004442

[9] 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

[10] 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.

[11] 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*, 2017, pp. 652–660.

[12] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking With Siamese Region Proposal Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 8971–8980.

[13] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020. [Online]. Available: http://dx.doi.org/10.1561/0600000079