# Multimodal Object Recognition Using Bayesian Attractor Model For 2D and 3D Data

H. Ando, D. Kominami, R. Seki, H. Shimonishi and M. Murata
Osaka Univ.

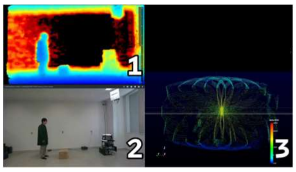
27th Conference on Innovation in Clouds, Internet and Networks
Workshop on 6G Network Use Cases and Verticals 6GN
Mar.11th-14th, 2024, Paris.

1

## Background



- **Developing 6G technologies**[1]
  - Connecting various types of sensors[2,3,4]
    - RGB Camera, LiDAR sensor, Infrared sensor
    - Analysis at edge and cloud
  - ➢ **Exploring 6G Use Case**
- **Robot control with digital twin**
  - Humans and robots cooperate in logistics warehouse
  - Stochastic robot control for safety and efficiency
  - ➢ **Stochastic recognition techniques are essential**



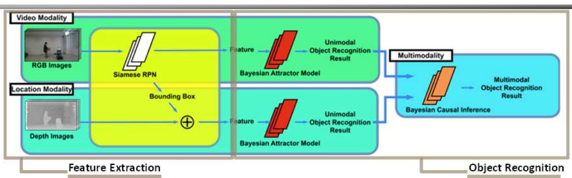

**Example of co-operative working**

[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] Wu, Hai, et al. "Virtual Sparse Convolution for Multimodal 3D Object Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[4]Jiao, Yang, et al. "MSMDfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

2

## Our Previous Work: Overview





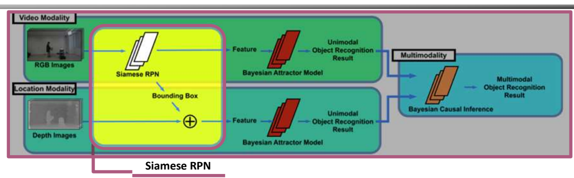

- **Object recognition method inspired by brain information processing mechanism**[5]
  - Calculates Posterior probability density (**Confidence**) about the observed object
  - Realizes **Uncertainty-aware object recognition** using confidence levels and enable risk assessment
- **Remaining Issue: Dependencies between modalities**
  - **Dependencies** arise between recognition results before integration **due to shared feature extraction parts**

[5]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.

3

## Feature Extraction by Siamese-RPN





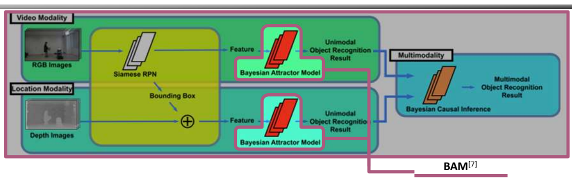

- **One-shot object detection method**[6]
  - **Video features** : Extract features **by CNN-based extraction** as the region (Bounding Box) where an object exists
  - **Location features** : Extract features **by combining position of Bounding Box and depth images**
    - **Center of Bounding Box (x, y) and depth (z)**
    - **Dependencies** arise in the assumptions of Siamese RPN feature extraction **before integration**

[6]B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking With Siamese Region ProposalNetwork," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

4

## Unimodal Object Recognition by Bayesian Attractor Model





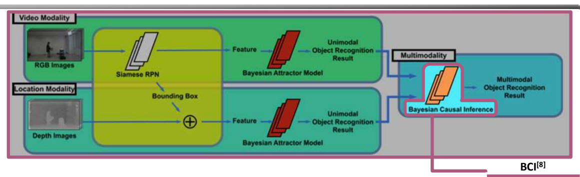

- **Unimodal Object Recognition using Brain-inspired model considering uncertainty**
  - Represents human decision state using state-space model that has fixed points called **attractors**
    - Attractors store the characteristics of each object
  - **Updates its state** according to the observation and recognize objects stochastically
    - Posterior probability density (**confidence**) on each attractor is derived as the certainty of recognition
  - ➢ **Construction of risk-aware systems** is made possible by recognition with confidence

[7] S. Bitzer, J. Bruineberg, and S.J. Kiebel, "A Bayesian Attractor Model for Perceptual Decision Making," PLOS Computational Biology, 2015

5

## Multimodal Object Recognition by Bayesian Causal Inference






- **Multimodal object recognition using BCI Model that integrates different modalities**
  - **Reduces recognition error** by leveraging different modalities
    - Assesses if the observation is focused on the same object and establish the weights accordingly
    - Calculates confidence by integrating the probability distributions obtained from each modality
  - ➢ Support decision-making by **considering uncertainties** associated with modalities

[8] K.P. K¨ording, U. Beierholm, W.J. Ma, S. Quartz, J.B. Tenenbaum, and L. Shams, "Causal Inference in Multisensory Perception," PLOS ONE, 2007.

6

## Contribution in Present Work



- **Problem：Dependencies before multimodal integration**
  - Use the information extracted from RGB images by Siamese RPN when processing depth images
    - **Dependencies** arise between recognition results before the integration of multimodal data
  - ➢ The reduction in recognition in the video modality results in decreased recognition in the location modality
    - ➢ Deterioration of dependability
- **Contribution：Resolve dependencies in the feature extraction**
  - **Employ point clouds** as a modality capable of standalone analysis, independent from video modality
    - **Adopt PointNet**[9] for semantic segmentation
  - ➢ Construct to merge video and location modalities following their separate recognition
    - ➢ Improvement of dependability

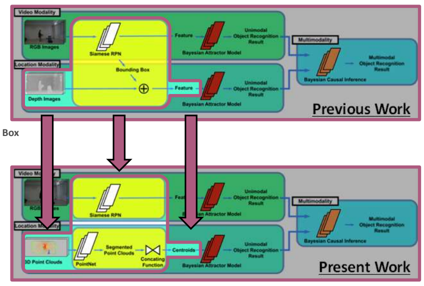
[9] C.R. Qi, H. Su, K. Mo, and L.J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, 2017



## Difference Between Previous and Present Work



- **Previous Work**
  - Video Modality
    - RGB images
    - Siamese RPN[6]
  - **Location Modality**
    - **Depth images**
    - **Combining with Bounding Box**
- **Present Work**
  - Video Modality
    - RGB images
    - Siamese RPN[6]
  - **Location Modality**
    - **3D Point Clouds**
    - **PointNet**[9]

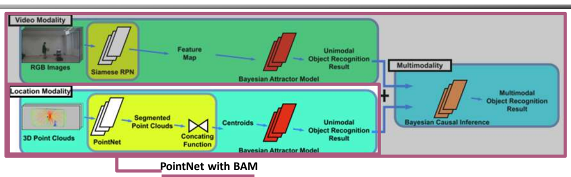



**Comparing Two Methods**



## Present Work: Unimodal Object Recognition by PointNet






**Our present work using Bayesian Attractor Model for RGB and Point Clouds**

- **Recognition model combining BAM [7] and PointNet [9]**
  - Utilize **3D centroids** of objects instead of location based on depth images and Bounding Box
    1. **Performs semantic segmentation** of point clouds using PointNet
    2. Acquires labelled point clouds and calculates centroids of objects
  - Recognition with 3D centroids using BAM
    - Memorize the centroids at 1st frame in BAM, compare observation with them, and recognize an object with confidence

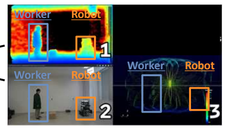


## Evaluation Settings



- **Create dataset using an RGB-D camera and a LiDAR sensor**
  - Time-series data consisting of **RGB-D images** and **3D point clouds** with labels
  - A scene where **a human and a robot** are performing cooperative work indoors
- **Tasks for object recognition from multimodal sources**
  - Targets : Worker / Robot
  - Task : Estimate what given feature corresponding to each object represents

**Intel® RealSense™ Depth Camera D455**[10]




**Task**

**Livox Avia**[11]

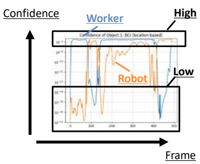
[10] https://www.intelrealsense.com/depth-camera-d455/
[11]https://www.livoxtech.com/jp/avia



## Evaluation Metrics



- **Inputs and outputs**
  - **Input (video)** : image features representing each object per frame
  - **Input (location)** : position features representing each object per frame
  - **Output** : confidence for each object
    - Example of output where **worker features** are given as input
    - **Horizontal axis**: Time [frames], **Vertical axis**: Confidence
    - **Blue:** Worker confidence, **Orange:** Robot confidence
- **Quantitative and qualitative evaluation**
  - **Quantitative evaluation : Transition of confidence**
    - **How recognize** the objects in each modality and integrated one
  - **Qualitative evaluation : Precision**
    - Percentage of objects that are truly correct when an object is recognized
    - Calculated by assuming that the object with **the highest confidence as recognized**

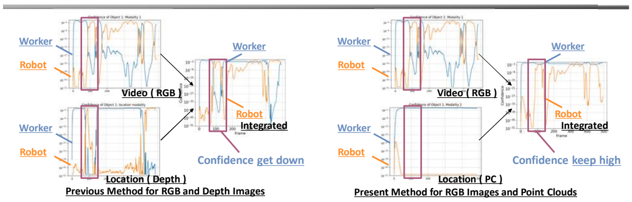



**Example of output when observing Worker**



## Quantitative Evaluation: Misprediction of Bounding Box






**Transition of Confidence (Target: Worker)**

- **Integration when one modality's recognition is high and the other's is low.**
  - **A decrease in recognition in the video modality leads to a decline in recognition in the location modality as well**. previous method **does not improve** the recognition by integration
  - When the confidence of the worker is **low** in the video modality and **high** in the location modality, present method **keeps the confidence of the worker in integrated result high**, even if robots are also recognized

## Qualitative Evaluation: Multimodal Object Recognition



Result on 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 |
| Video-based Multi-modalities (with depth) | 1.000 | 0.784 |
| Video-based Multi-modalities (with location) | 0.936 | 0.869 |

- **The precision for both targets are improved by multimodal integration**
  - **Single modality** reduces the certainty of one side
  - **Multimodal integration** results in high precision score for both targets



## Conclusion



- **Proposed method**
  - Multimodal object recognition method for 6G use case
    - Develop unimodal object recognition using point clouds with confidence
    - Improve our previous multimodal object recognition method using brain mechanisms
- **Experimental results**
  - Capture and test the environment for our use case using an RGB-D camera and a LiDAR sensor
  - Observe improvement of recognition by integrating recognition results
- **Future work**
  - Expand recognition targets
    - 2-class classification and one object in each class, assuming both exist in the frame in current experiments
    - For practical use, it is necessary to support multi-class multi-objects.
  - Propose a method to associate results from different modalities
    - Association of recognition results is given In current work



## Appendix







## Bayesian Attractor Model (BAM)[7]



- **A model combining Attractor model with Bayesian theory**
  - A model inspired by the process through which the brain makes decisions based on sensory organ information
    - Fusion of the attractor model for memory and the Bayesian estimation model for state updating
  - Produce recognition outcomes along with confidence based on continuous observations of what is being observed
- **Attractor model**
  - A model of recognition grounded in the brain's mechanisms of memory and cognition
    - Place fixed points (attractors) on the state space that correspond one-to-one with the objects of observation
    - Internal state changes in response to input
  - If inputs for the same target are sustained, it converges to the attractor that represents that target
- **Bayesian theory**
  - Calculate the posterior probability using prior probabilities and observations
  - Compile information continually upon receiving new inputs to revise decisions



## Bayesian Causal Inference (BCI)[8]

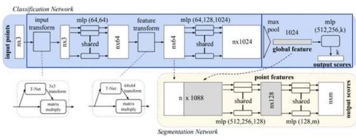


- **A statistical model combining Bayesian theory and causal inference**
  - A model inspired by the process through which the brain makes decisions based on multimodal information
    - Infer whether objects recognized by two different modalities are the same and integrate
  - Combine modalities with different advantages to provide more accurate predictions
    - Evaluate the likelihood that inputs from various modalities stem from a single external occurrence
    - Based on evaluated causal links, determine if sensory inputs should be merged or regarded independently
- **Formula for generating confidence**
  - $c_V, c_L$ : Confidence in video and location modalities
  - $P(c_V, c_L \mid C = 1) = \begin{cases} 0.5 + \sigma_{c=1} & \text{if } L_V = L_L, \\ \sigma_{c=1} & \text{otherwise,} \end{cases}$
  - $P(c_V, c_L \mid C = 2) = \begin{cases} 1 & \text{if } L_V \neq L_L \\ 0 & \text{otherwise} \end{cases}$
  - $\sigma_{c=1} = 0.5\sigma(\max(c_V)) + 0.5\sigma(\max(c_L))$
  - $\sigma = 1/(1 + \exp(-x - \lambda_{bci}))$
  - $L_V = \mathrm{argmax}(c_V)$
  - $L_L = \mathrm{argmax}(c_L)$



## PointNet[9]

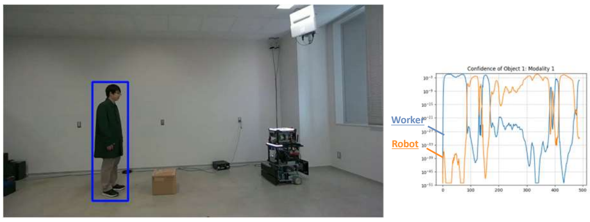





PointNet Architecture[9]

- **Deep learning framework for processing 3D point clouds**
  - Apply the same **Multilayer Perceptron (MLP)** to each point, followed by the application of **Max-Pooling**
    - Produces the same output, **independent of the order in which the point cloud is inputted**
  - Utilize a subnet called T-Net
    - Produces an affine transformation matrix for the input, allowing for adjustments such as rotation and shifting
  - **Utilizing global features**
    - **Integrate local and global features** and apply MLP for the realization of **semantic segmentation**

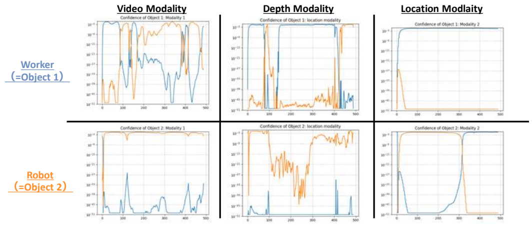
## Example of Bounding Box Detection Employing Siamese RPN 19




19

## Evaluation Results: Unimodal Object Recognition 20

- **Unimodal Object Recognition**
  - Use the term "**depth modality**" to distinguish the location modality using point clouds that employs depth images
  - Observe the correlation between video and depth modalities, which is a challenge in previous work

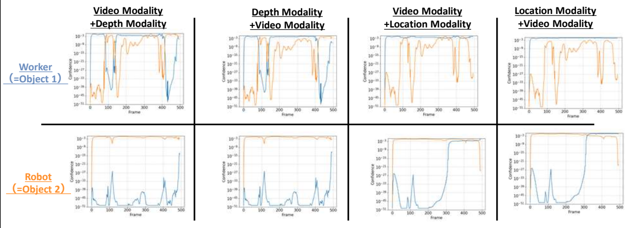



Evaluation results of unimodal object recognition

20

## Evaluation Results: Multimodal Object Recognition 21

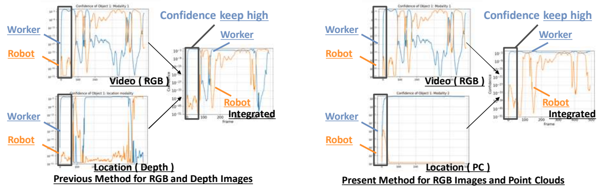
- **Multimodal Object Recognition**
  - When using **depth images**, the parts where correlation was observed have not been corrected
  - When using **point clouds**, the problem has been resolved




Evaluation results of multimodal object recognition

21

## Quantitative Evaluation 22

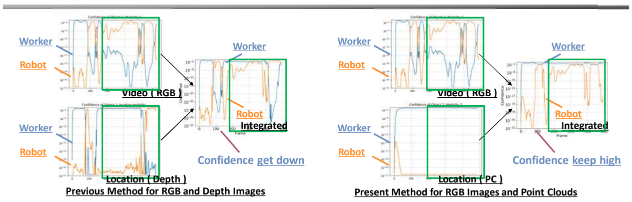



Transition of Confidence (Target: Worker)

- **Integration when both modalities produce recognition results with the same trend**
  - When the confidence of the worker in the video and location modalities are **high**, the confidence of the worker in integrated result is also **high**
  - When the confidence of the robot in the video and location modalities are **low**, the confidence of the worker in integrated result is also **low**

22

## Quantitative Evaluation: Misrecognition in Previous Method 23




Transition of Confidence (Target: Worker)

- **Integration when both modalities confidently provide different recognition results**
  - Bounding Box is right in the video modality, but the perception **continues to be wrong**
  - **Misrecognition occur** in the depth modality where they pass each other in the latter part of the scene
  - It is **clearly recognizable** by using centroids in location modality

23