Realtime Object Recognition Method Inspired by Multimodal Information Processing in the Brain for Distributed Digital Twin Systems

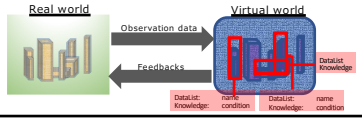
Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

Ryoga Seki

Background — Growing expectations of the Digital Twin

❖Digital Twins
- Represent virtually every aspect of our real world
- Feed back simulated results in digital world to the real world
- Have various applications: warehouse management, safety management, and autopilot

❖Research challenges
- Robust sensing data processing
- Accurate identification of objects and self-location information of sensor devices
- Real-time and wide-area object recognition



Purpose — Realizing the Edge-Cloud Digital Twin

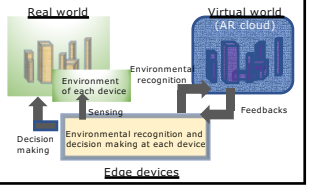
Overall purpose

❖Realizing the edge-cloud digital twin concept using the AR cloud
- Build a localized virtual world by recognizing the environment with sensing devices
- Integrate each device's information on the AR cloud to create a huge virtual world

In this paper

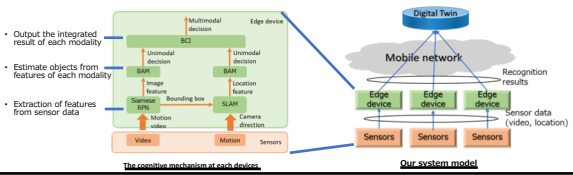
❖Problem: real-time update of virtual world from observed information
- Limitation of processing capability on the cloud
- Latency between devices and the cloud

❖Approach
- Learning from the superior cognitive mechanisms of the human brain to realize the edge-cloud digital twin
- we focus on the multimodal information processing mechanism of the brain



The constructure digital twin learned from human brain

❖Constructing a **probabilistic virtual world** with **the cognitive mechanisms of the brain**
- The cognitive mechanisms of the brain are hierarchically structured
- Define edge devices as cognitive mechanisms that have such functions
- Summarize cognitive results from multiple them to cloud



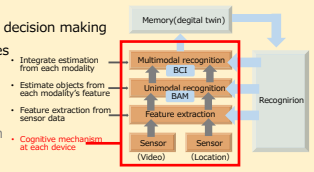
Brain-inspired object recognition methods

Purpose in this paper: Realize quickly and robust object estimation that can cope with real-world environmental changes

Approach

❖Propose an object estimation method inspired by the brain's cognitive mechanism
- Multimodal processing using uncertain information obtained from eyes, ears, and the semicircular canal
- Hierarchical, lightweight and robust decision making

❖Models of the brain's cognitive processes
- Unimodal Processing by "Yuragi Learning" of Brain and Organism **B**ayesian **A**ttractor **M**odel (BAM)
- Multimodal processing based on causal inference by brain perception **B**ayesian **C**ausal **I**nference (BCI)

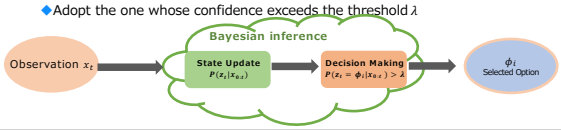


Bayesian Attractor Model (BAM)

❖Mathematical model of the brain's behavior in making decisions based on observations
- Estimate which of the pre-stored options matches the observation target
- Determine which of the previously learned options is closest

❖**State Update:** Internal decision state z_t is updated by receiving external observations x_t
→by Bayesian inference, z_t is a probability distribution $P(z_t)$ that reflects the uncertainty of the brain state.

❖**Decision Making** : One of the pre-learned options (attractors) is adopted
- ◆Prepare as many stable points that are named attractors
- ◆Derive a probability density (hereafter called the confidence level) where zt = φi
- ◆Adopt the one whose confidence exceeds the threshold λ



6

Extending the BAM with BCI

❖BCI is a mathematical model of human cognition based on multimodal perceptual stimuli

1. Uni-modal object estimation: Two input stimuli (features of an object) x_L, x_V calculated from 3D location modality and visual modality are input to the BAM
   - The BAM outputs cognitive results S_L and S_V for each modality
2. Multimodal object estimation: Outputs of the BAM is integrated by the BCI
   - Probability that the two stimuli are generated from the same object (or not) is calculated (Causal Inference)
   - Cognitive results are calculated according to the probability (Model Average)

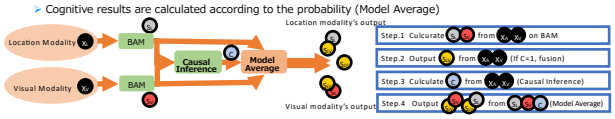


Figure. Extending the BAM with BCI

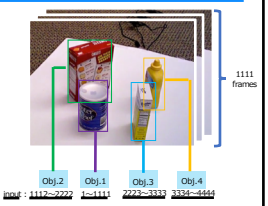
[1] K. P. Kording, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisen- sory perception," PLoS one, vol. 2, no. 9, p. e943, 2007.

7

Simulation evaluation

❖Feature extraction in visual and location modalities from video data[2]
- 4 objects in 1111 video frames
- Features are extracted from each bounding box of an object in a frame
- The BAM learns and stores features of each object in the first frame
- The BAM outputs which object is observed with confidence

❖Compare the estimation accuracy between unimodal and multimodal object estimation



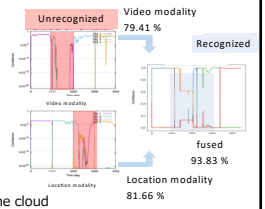
[2] Video data of public dataset(Yale-CMU-Berkeley (YCB) Object and Model set)

8

Conclusion

Results

❖I Proposed a method for object estimation based on multimodal uncertain observed information
- **Lightweight** enough to perform estimation on per-device, and **robust** by using multimodality
- Limitation: My proposed method is for a single object only



Future tasks

❖Propose a multi-object estimation method

❖Fusion of estimation results from multi-devices on the cloud

❖Feedback from the cloud to the device and its utilization

9

Further Future Challenges

① Detailed system design for single device
- Recognize multiple objects simultaneously

② Construction of an upper layer to store object recognition results from a device

③ Design a system to integrate the recognition results from multi-device

④ Build a method of cooperation between layers