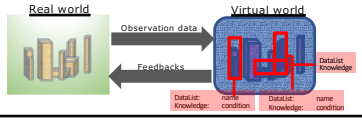
Multi-Object Recognition Method Inspired by Multimodal Information Processing in the Human Brain

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
- Constructing a wide-area digital twin in real time
- High-speed processing of video from many cameras
- Recognizing the surrounding environment with each sensor device



Purpose — Realizing the Edge-Cloud Digital Twin

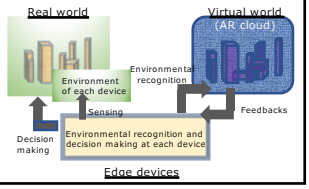
Overall purpose

❖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

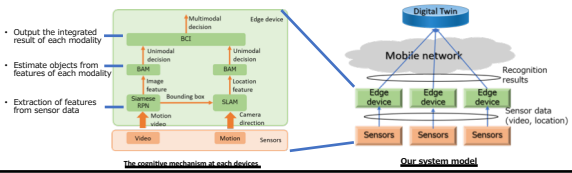
❖Problem with real-time update from observed information
- Limitation of processing capability on the cloud
- Latency between devices and the cloud

❖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



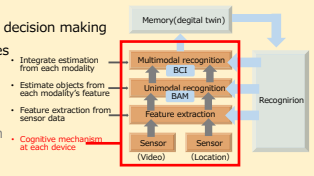
Brain-inspired object recognition methods

Purpose in this paper: Modeling the features of the brain's perception of spatial context and incorporate them into our method

Our existing method

❖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)

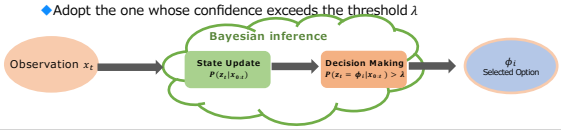


Bayesian Attractor Model (BAM)

❖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



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 two stimuli are generated from the same object (or not) is calculated (**Causal Inference**)
   - Cognitive results are calculated according to the probability (**Model Average**)

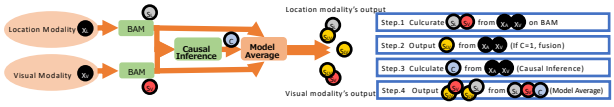
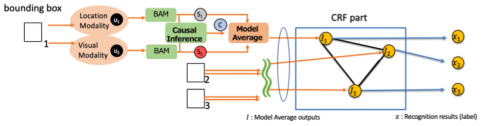


Figure. Extending the BAM with BCI

Modeling the brain's perception of spatial context

❖Recognizing multiple objects considering with **CRF**

- A number of features are extracted for each modality, which are processed by the BAM and the BCI.
- Building a CRF with outputs of the BCI as nodes, it decides recognition results as a label corresponding to each bounding box.
- Define an energy function on the labels of objects that can be minimized to obtain the optimal solution.



The spatial constraints we set

❖The input data is streamed without any loss.

❖**The input data has a one-to-one correspondence with the attractor.**

❖**Each attractor stores a different object.**

→They are **NOT** appropriate for real-world environments and need to be changed in our future studies.
(after building **a layer of memory** for the digital twin)

Simulation evaluation

❖Feature extraction in visual and location modalities from video data

- 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
- Then construct a CRF from the output confidence

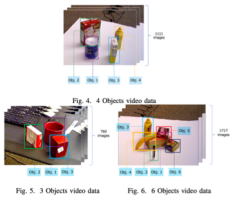


Fig. 4. 4 Objects video data

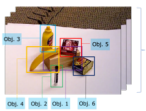
Fig. 5. 3 Objects video data

Fig. 6. 6 Objects video data

Result

Unimodal : Correct responses rate to **BAM alone**
Multimodal : Correct responses rate fused with confidence in **BCI**
CRF version : Correct recognition rate with **spatial context**

→ Without being pulled toward lower recognition rates



6 Objects correct response rate (%)

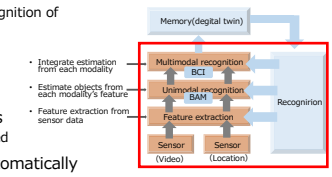
Modality	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Total
Unimodal video	85.3	99.4	95.9	91.7	96.9	73.7	90.5
Unimodal location	39.8	92.6	67.8	92.0	74.6	66.6	72.2
Multimodal video	75.2	99.4	96.5	94.8	96.9	83.0	91.0
Multimodal location	39.8	92.7	67.9	92.2	74.8	66.6	72.3
CRF version	94.5	83.8	83.2	92.4	86.6	86.8	87.9

Conclusion

Results

❖**Flexible System** that can compensate for the instability of one source of information with data from another source

❖Incorporated features of the brain's recognition of spatial context by CRF



Future tasks

❖Eliminate extra graph connections
- for when many objects are recognized

❖Constraints should be defined automatically

12

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