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

Ryoga Seki
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
r-seki@ist.osaka-u.ac.jp

Daichi Kominami
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
d-kominami@ist.osaka-u.ac.jp

Hideyuki Shimonishi
Graduate School of Information Science and Technology, Osaka University and NEC Corp.
Osaka, Japan
h-shimonishi@ist.osaka-u.ac.jp

Masayuki Murata
Graduate School of Information Science and Technology, Osaka University
Osaka, Japan
murata@ist.osaka-u.ac.jp

Masaya Fujiwaka
System Platform Research Labs, NEC Corporation
Kanagawa, Japan
fujiwaka@nec.com

***Abstract*—Recently, digital twins have been paid much attention as a major application towards Beyond 5G/6G network, and real-time object recognition methods are key technology to digitize the real world as a digital twin. However, it is challenging to make a fast and accurate decision on what the object is from real-time streaming information such as video because accurate object recognition algorithms require a huge computation. To satisfy delay requirement of digital twin applications, such computations have to be moved from cloud to edges or even small terminal devices, where computing capacity is very limited. Thus, recognition mechanisms have to be simplified for small devices but they would result in degraded accuracy. In this paper, we focus on the multimodal information processing mechanism of the brain, which makes decisions based on multiple types of uncertain observed information, to improve accuracy of simplified recognition mechanisms. We first propose a unimodal object recognition mechanism based on the Bayesian attractor model, which continuously recognizes objects from noisy streaming media data. Then, we extend the mechanism with Bayesian causal inference to fuse the results of unimodal media recognition. Through computer simulations, we show that our proposed method identifies an object accurately and quickly from uncertain observed information.**

***Index Terms*—Mobile AR, digital twin, multimodal recognition, Bayesian attractor model, Bayesian causal inference.**

I. Introduction

Beyond 5G/6G network is expected to be a major social infrastructure toward 2030, where humans and machines cooperates to solve various social problems by using AR applications or cyber physical systems (CPS). Digital twins have been paid much attention as a key technology to bridge real and digital worlds; humans and robots in real world are digitized via various sensor devices and represented in virtual world as a digital twin.

Beyond 5G/6G network would provide huge bandwidth and low latency communication, however, it would be still inefficient or unrealistic to send all the sensor data to a central cloud because the amount of data could be tremendous to digitalize whole world around us. Data processing should be distributed from a cloud to edges or even small terminal devices, where the computational capability is very limited, and only recognition results should be gathered at the cloud.

To understand and control the real world using digital twins, it is necessary to uniquely identify what kind of object exists in front of us, locate its position, and represent it with a digital twin. In recent years, technologies like convolutional neural networks (CNNs) have made remarkable progress, but it is still challenging to make a fast and accurate decision on what an object is by analyzing real-time streaming information at small devices because deep CNN models require a huge amount of computation. Another difficulty for deep CNN models used in real-time AR applications is their learning cost. While AR applications require so-called single- or few-shot learning when target objects are specified and learnt from a single or few images that AR users initially encounter in the field, deep CNN models need thousands of images to learn an object before being deployed for AR activities in the field.

Uncertainty in image-based object recognition is another challenge. Because the real world in front of a camera is continuously changing, uncertainty in real-time observation of the real world cannot be avoided owing to noise or instability in real-time streaming information, as well as unavoidable incompleteness of the observation itself. Moreover, when multiple objects having a similar shape and color are located in front of the camera, it is quite difficult to identify them using only video information. Therefore, multimodal decision making is expected to mitigate such incomplete information and improve the accuracy of unimodal recognition [1], [2], but the tradeoff between accuracy and computational complexity of deep CNN models has not been fully solved for edge devices performing real-time recognition of moving objects.

The information processing mechanism of a brain is a familiar example of a system that makes decisions from such uncertain observation. The brain uses uncertain information obtained from the eyes, ears, skin, and semicircular canals to infer the state of the surrounding environment and to make final decisions. In recent years, mathematical modeling of the brain's information processing mechanisms has been promoted, and such models include the Bayesian attractor model (BAM) [3] and Bayesian causal inference (BCI) [4]–[6]. As shown in [7], information processing of the brain can be modeled hierarchically; feature, unimodal, and multimodal levels. Our proposed model employs this model and for the unimodal and multimodal levels, we use the BAM and BCI, while basic media-specific data preprocessing is performed at the feature level.

As described in the following section, the BAM represents the behavior of a person's decision-making process based on observed information by using Bayesian estimation, and thus, it is expected to identify objects with high accuracy from uncertain time-varying information. BCI is a mathematical model of the process by which humans recognize perceptual objects using multiple modalities (e.g., vision and audition). In this cognitive model, humans infer whether two input stimuli originate from the same stimulus source, and then integrate

each input stimulus to make a final cognitive decision.

In this research, we propose a multimodal object recognition method using video and location sensors for edge devices. We use the BAM for unimodal object recognition and integrate the output obtained from each modality with BCI. We have promptly reported an early result [8], and in this paper, we provide more detailed explanation of the proposed method, concise investigation of the behavior of BAM and BCI modules, and results from various test data to show effectiveness of the proposed method. The image features of the reference image data are associated with attractors defined in the BAM. The reference image can be prepared beforehand, but in typical AR applications, the images are acquired in the field. The BAM is given a series of input data for each frame and determines which reference data correspond to the input data based on which attractor the internal state of the brain falls into. To deal with uncertainty in object identification with the video modality, we also introduce another modality, namely the location modality, of the object, which is inferred from the position of the camera and its direction. This also helps to distinguish between objects having similar shape and color within the camera view. In our method, another BAM is used to infer the object using noisy location input data from sensors.

The results of the recognition in each modality are integrated by BCI to make the final decision, thereby improving the accuracy. The BCI model infers the certainty indicating whether these two modalities are measuring the same object or different ones, and then integrates two decisions weighted by the certainty, where the highly integrated results are given a higher certainty, and rather separated results are given a lower certainty. This avoids incorrect integration when multiple objects are in the field of view and mismatched measurements are given to different modalities.

The remainder of this paper is organized as follows. In Section II, we describe the mathematical basis of the BAM and BCI. In Section III, we describe a method for applying the results of extracting the features of each modality from video and inputting them into the BAM and then the BCI model. In Section IV, we evaluate the proposed method. Section V gives the conclusion of this paper.

II. Related Work

A. Bayesian Attractor Model

The BAM estimates which of the pre-stored options matches the observation target. The BAM has an internal decision state $\mathbf{z}_t$, and it updates this state when receiving external observations $\mathbf{x}_t$. By updating the state based on Bayesian inference, $\mathbf{z}_t$ is not treated as a single point, but is represented as a probability distribution $P(\mathbf{z}_t)$ that reflects the uncertainty of the observation and brain state.

We prepare as many stable points (attractors) $_1; \ldots; _n$ in the state space where $\mathbf{z}$ exists as the number of choices (n), and make the decision to take the i-th choice when $\mathbf{z}$ is sufficiently close to . Because $\mathbf{z}_t$ is represented as a probability distribution, we derive a probability density (hereafter called the *confidence* level) where $\mathbf{z}_t =$ $_i$ and make a decision using this confidence level. The details of state updating and decision making are described below.

1) State Update: The state update is performed by finding the posterior distribution $P(\mathbf{z}_t|\mathbf{x}_t)$ of the decision state $\mathbf{z}_t$ by Bayesian estimation when the observed value $\mathbf{x}_t$ is obtained. The following generative model is assumed for $\mathbf{x}_t$ and $\mathbf{z}_t$:

$$\mathbf{z}_t \quad \mathbf{z}_t \quad _t = \quad tf(\mathbf{z}_t \quad _t) + \sqrt{t}\mathbf{w}_t; \tag{1}$$

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

where $f(\mathbf{z})$ represents the dynamics of a Hopfield network, one of the attractor models, and the dynamics has multiple attractors. When n is the number of options to be stored in the Bayesian attractor model, we design it so that f has n attractors $_1; \ldots; _n$. In the above equations, M is a feature matrix for each option, with $M = [\ _1; \ldots; \ _n]$ where $_i$ is a feature vector of the i-th option; is a multidimensional sigmoid function with a value range of 0 to 1; and $\mathbf{w}_t$ and $\mathbf{v}_t$ are random numbers following normal distributions $\mathbf{w}_t$ $N(0; \frac{q^2}{t}I)$ and $\mathbf{v}_t$ $N(0; r^2 I)$, respectively. N denotes a normal distribution. I is a unit matrix, and when t is 1, the respective standard deviations are q and r. These deviations determine the noise level of the dynamics and observation in the generative model, so q is called the dynamics uncertainty and r is called the sensory uncertainty.

2) Decision Making: Which attractor the internal state of the brain is close to can be determined based on the magnitude of the confidence level and therefore, decision making is done when the confidence level for one of the options exceeds the threshold value. The posterior probability distribution of $\mathbf{z}_t$, $P(\mathbf{z}_t|\mathbf{x}_t)$, is obtained by the state update. Here, approximate calculations are performed using the unscented Kalman filter (UKF) to account for the non-linearity of the generative model, as given in [3]. Finally, the confidence level is obtained for each pre-stored option, $\mathbf{c} = fc_1; \ldots; c_n g$, and $c_i = P(\mathbf{z}_t = \ _i|\mathbf{x}_t)$. Note that since the Kalman filter is used, estimated $P(\mathbf{z}_t|\mathbf{x}_t)$ is probability density function of a multivariate normal distribution.

B. Bayesian Causal Inference

BCI is a mathematical model of human cognition based on multimodal perceptual stimuli. For example, when we see something on the left side of our field of vision and hear a sound from the same direction, we may judge the direction of arrival by integrating our visual and auditory senses, assuming that these stimuli originate from the same source, or we may judge the direction of arrival separately, assuming that they come from different directions. In [5], a mathematical formulation for the following process is given. When an object is presented, both modalities (visual and auditory) perceive the location of the object and probabilistically infer whether both modalities observe the same object (causal inference), which is used to integrate the observations from each modality to recognize the object's location (the *model average*). The result is a weighted sum of the integrated recognition result and segregated recognition results. If the causal inference obtains a higher confidence that both modalities observe the same object, the model average results place a higher weight on the integrated result; otherwise, the model average results place a higher weight on the segregated results.

III. Method

Figure 1 shows our system mode. Small edge devices equipped with multiple sensors, such as video camera, depth camera, LiDAR, motion sensor, etc. are distributed over a network. As described earlier, object recognition tasks are not performed at a cloud but distributed to edge devices, which read sensor data and recognition tasks described below, and send recognition results to the cloud. At the cloud, recognition results from various devices are integrated and maintained as a digital twin. At each edge device, by using the BAM to estimate objects individually from the features of the video modality and the location modality obtained from the sensor devices, and then integrating them with BCI to make decisions, as shown in Fig. reffig:model. In the following sections, we describe a method to apply the decision model of the BAM to object estimation, and a method of integrating the confidence level of the BAM results with BCI.

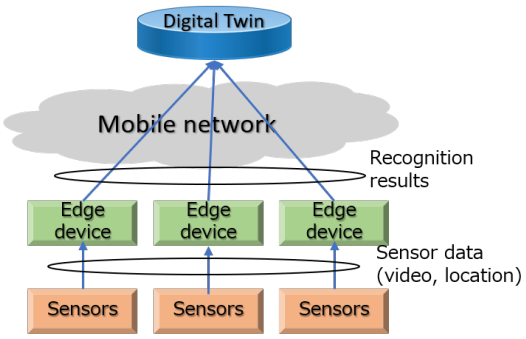


Fig. 1. System model

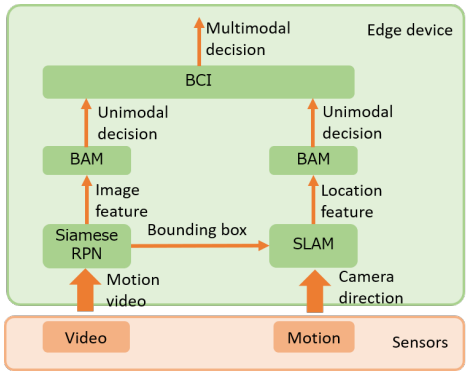


Fig. 2. Proposed method

A. Application of the BAM to Object Estimation

To apply the BAM to object estimation, it is necessary to determine a) the features to be observed by the BAM in each modality, b) the reference data to be stored in the attractors, and c) the values of the uncertainty parameters q and r. For c), the observation noise can be determined by transforming the features so that the variance of the features is set to a certain value. We discuss each of the features, transformations, and attractors below.

a) Features:

Video modality: The features are extracted using a Siamese region proposal network (RPN) [9]. The Siamese RPN takes a template image and a detection image as input, detects similarities between these images, and outputs the location as a bounding box. The Siamese RPN consists of a Siamese neural network and an RPN. The former extracts the features of the image, and the latter uses the extracted features to calculate the similarities to the template image from the detected image. In the literature [9], existing CNNs such as AlexNet [10] are used as the Siamese network, but in the proposed method, a simple CNN consisting of four layers is used to greatly reduce the computational cost. Because the recognition decision is made by the BAM, the role of the CNN is just to extract feature values, and an encoder like a shallow CNN is suitable for this purpose. The feature values are 128-dimensional data from the output of the Siamese network (image features) corresponding to the bounding box output by the RPN.

Location modality: The feature values of the location modality are the 3D world coordinate system data calculated from the camera direction vector and the depth information integrated from multiple frames. Thus, the feature values to be fed to the BAM are 3D data, such as the x-y-z location in a world coordinate system.

b) Transformation:

The observed values obtained from sensor devices do not have a fixed value range, and the magnitude of noise is unknown. In addition, in consideration of the case where the scale of the numerical value is different for each dimension of the feature value, which is a multidimensional variable, r must be adjusted for each dimension of the feature value. Therefore, instead of using the feature data as is, we perform a conversion process to make the data easier for the BAM to process. The function S to transform each element of the feature vector $\mathsf{x}(t) = [x_1(t); x_2(t); \quad ; x_n(t)]$ to be observed at time t is defined as follows:

$$S[x_i(t)] = \frac{x_i(t) \quad _i(\mathsf{X})}{s_i(\mathsf{X})}; \tag{3}$$

where X is a set of vectors of the same size as $\mathsf{x}(t)$, $_i(\mathsf{X}); s_i(\mathsf{X})$ returns the mean and variance of the i-th element of each vector in X, and X is the set of feature vectors obtained in advance for each identification target. By choosing X well, we can calculate Eq. (3) and obtain the mean of $S[x_i(t)]$ as 0 with a variance of 1. By setting the mean to 0, the value range includes both positive and negative values. There are two reasons for this process. One is that the estimation accuracy of the BAM is better when the value range of the observed value includes both positive and negative values. This is illustrated using an example of identifying a 1D observable when all the values are positive. Now, two values are assumed to be stored in the BAM. When the decision state is near a certain attractor and the observed value increases, BAM calculates z_t by back-calculating Eq. (2), but because x_t is increasing, z_t is also increasing. Because this change in z_t is in the direction away from all attractors, the confidence level will only decrease and z_t will not approach another attractor. On the contrary, when the observed value decreases, it could approach another attractor. This means that if the sign of the observed value is only positive (or negative), when z_t is in the vicinity of an attractor, the increase (or decrease) of the observed value cannot be taken as a movement of z to another attractor in the space of z. In this case, the values stored in the attractor cannot be identified accurately. When the value range of the observed value contains both positive and negative values, the same thing happens if the two values stored in the attractor are both of the same sign, but if the values stored in the attractor have different signs, the above problem does not occur. Of course, the same problem can occur when storing three of more one-dimensional values.

Second, it provides a guideline for setting an appropriate r value: if X contains information on all of the target objects that the BAM may observe, the variance of the observed values will be close to 1. Therefore, the variance when continuously observing a particular object will be smaller than 1, and r will be chosen in the range of $0 < r < 1$.

c) Attractors:

The attractor stores the feature values of a reference image/location of the objects to be estimated. This is equivalent to assigning the feature values of each object to each of vectors $_1; \ldots; \ _n$, which are elements of the feature matrix M. For the video modality, to enable one-shot learning, that is, to use just one representative image and eliminate pre-training, only the first image of an object in the video where each object is first seen is used to calculate the feature values. For the location modality, the initial object location stored in memory is used as a feature value. We assume that objects are stationary during scene acquisition, and tracking for a moving object is for future study.

As X in Eq. (3), we give $^{1st}_1; \ldots; \ ^{1st}_n$, which are the feature vectors calculated from the first frame of the video where

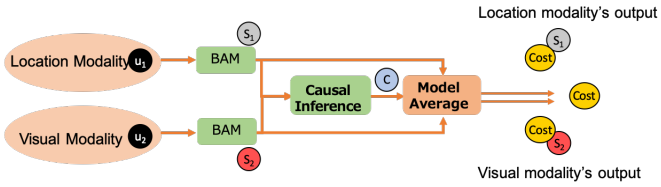


Fig. 3. Extending the BAM with BCI

each of the n objects was first seen, or the initial location of an object in memory. For the feature vector, which is an element of M, we give the value of $^{1st}_{1};:::;\ ^{1st}_{n}$ transformed by the function S, i.e., $M = [S(\ ^{1st}_{1});:::;S(\ ^{1st}_{n})]$.

B. Extension of the BAM with BCI

After the BAM performs object estimation in each modality, the BCI model performs causal inference as shown in Figure 3. Here, the confidence level of the BAM, c, is input to the BCI model as the observed value, causal inference is performed to infer whether the same object is observed in the video modality and the location modality, and the result is used for multimodal integration by the model average algorithm to output the final object. However, while the conventional BCI model targets continuous values (for location estimation), this method targets discrete values (for object identification), so it needs to be extended slightly. We explain how this extension is performed below.

1) Causal Inference: The BCI model in [5] performs causal inference according to Bayes' theorem, as shown in the following equation:

$$p(Cju_1;u_2) = \frac{p(u_1;u_2jC)p(C)}{p(u_1;u_2)} = \frac{p(u_1;u_2jC)p(C)}{p(u_1;u_2jC=1)p(C=1)+p(u_1;u_2jC=0)p(C=0)}; \quad (4)$$

where $p(C)$ is the probability of observing the same object in both modalities; C takes two values, 0 (observing separate objects) or 1 (observing the same object); and u_1 and u_2 are the observed values of each modality, respectively, in this case the confidence value of each BAM. Reference [5] defines $p(u_1;u_2jC)$ in a continuous manner, and we redefine it in a discrete manner as follows:

$$p(u_1;u_2jC=1) = \int p(u_1;u_2js)p(s)ds = \sum_{k=1}^{K} p(u_1;u_2jO_k)p(O_k); \quad (5)$$

$$p(u_1;u_2jC=0) = \int p(u_1js)p(s)ds \int p(u_2js)p(s)ds = \sum_{k=1}^{K} p(u_1jO_k)p(O_k) \sum_{k=1}^{K} p(u_2jO_k)p(O_k); \quad (6)$$

where s is the position of the object, $p(s)$ is its distribution, and $p(ujs)$ is the probability that the position of an object is observed as u in a certain modality when it is at s.

2) Model Average: Based on the results of causal inference, multimodal integration is performed to output the final object estimation results. Here, a cost function weighted by the results of causal inference is calculated as in the equation below, and the objects $O^{\emptyset}_m$ that minimize it are used as the final object estimation results for modality m (1 m 2). Here, if $C = 1$, $O^{\emptyset}_1$ and $O^{\emptyset}_2$ output the same object, and if $C = 0$, the estimation result of each modality is output as it is.

$$Cost_m(O_m) = p(C=1)\sum_{k=1}^{K} jO_m \quad O_kj^2 p(O_kju_1;u_2) + p(C=0)\sum_{k=1}^{K} jO_m \quad O_kj^2 p(O_kju_m); \quad (7)$$

where $jO_m \quad O_kj$ is 0 if $O_m = O_k$, and otherwise it is 1. Although the distance error of the estimated position of the object are used in the literature [5], because the distance cannot be defined when performing object estimation, it is assumed that the calculation is based only on whether the modalities agree or disagree.

3) Object Identification Method: To use an extended version of BCI for object estimation, in the above equation, the probability that the object O_k is observed is $p(O_k)$. In our proposal, the initial value of $p(O_k)$ is $1=N_{obs}$, where N_{obs} is the number of objects and $p(O_k)$ is updated by Bayesian inference after every observation. For object identification, we substitute c for u. The probability that the BAM confidence value is c when the object O_k is observed is defined as $p(cjO_k)$. Then, finally, we obtain the object label that minimizes Eq. (7). Note that because the confidence level may take very small values, all values below the threshold are taken as the same value as the threshold as input to the causal inference model (the threshold in this case is 10^{50}).

IV. Results

To confirm the effectiveness of the object estimation method applying the BAM and BCI, as described above, we conducted a simulation-based evaluation.

A. Simulation Environment

For the video dataset, we used a real measurement public dataset of various objects (Yale-CMU-Berkeley Object and Model set) [11]. Figure 4 shows the video data used for input. The video was shot while moving an object, and the number of frames was 1,111. For each frame and each of four objects in the video data, we extracted the features of the video modality and location modality using the method in Section III-A a). As for the two parameters of the BAM, (q, r), we set (0.3, 2.5) for the video modality and (1.5, 0.04) for the location modality. In the following evaluation, for each object, we determined whether the object was correctly identified when the video modality and location modality features were observed. For this purpose, the BAM stored the feature values of each of the video and location modalities for the four objects. In addition, we concatenated the time series data of feature values for 1,111 frames for each object and used a total of 4,444 frames as input to the proposed object recognition method.

For the video modality in this evaluation, we assumed one-shot learning for AR applications, and thus only the first image of the object in the video was used as training data. As a result, we cannot directly compare our results and existing CNN-based object classification methods. We can, however, compare our BAM output of the video modality and the output from the classification branch of the Siamese RPN. In the latter case, recognition results are given for each video

frame independently without considering other frames in the sequence, so this approach would suffer from recognizing an object from motion video data in comparison with our method using the BAM. Therefore, in this evaluation, we compared our method with Oracle value.

Since the confidence of the BAM is defined as the probability density of n-dimensional normal distribution, when $\mathbf{z}_t$ is consistent with the correct attractor ($_c$), it is expressed as $P(\mathbf{z}_t = \ _c j \mathbf{x}_t) = 1=((0{:}5n\ )^2\ \sqrt{\ j\ j})$, where $j\ j$ is the determinant of the variance-covariance matrix of $\mathbf{z}$. We define the confidence level at this time as the Oracle. Since $P(\mathbf{z}j\mathbf{x})$ is the distribution obtained as a result of the Bayesian estimation, $j\ j$ is also obtained after the estimation. Therefore, we assume that the variance-covariance matrix of $\mathbf{z}$ is consistent with that of the generative model, and then $j\ j = q^2$. Thus the values of Oracle are 0.0169 for location modality and 0.0844 for video modality.

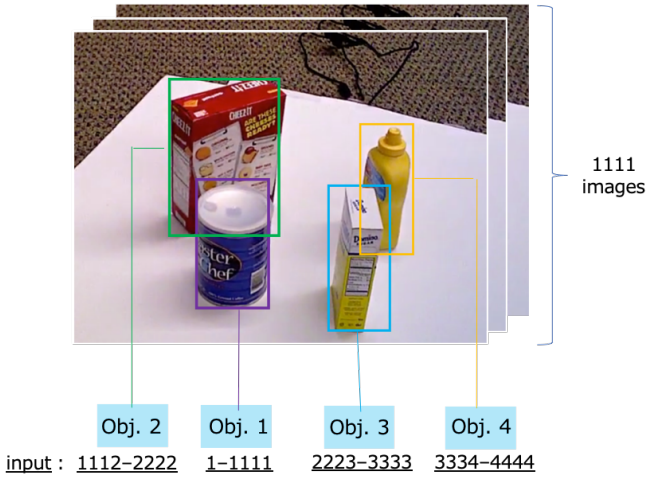
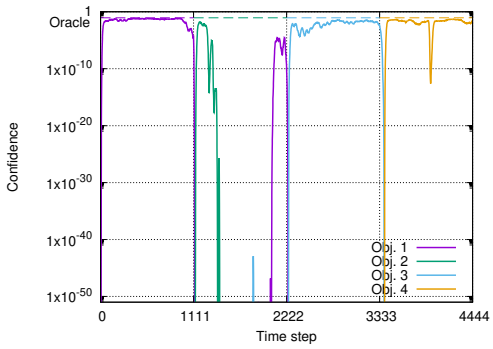
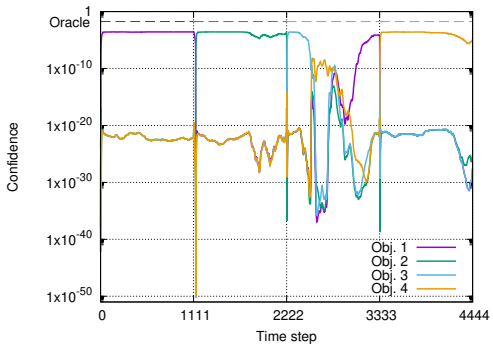


Fig. 4. Video data



(a) Video modality



(b) Location modality

Fig. 5. Confidence level results for the BAM

TABLE I
CORRECT RESPONSE RATE (%)

Modality	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Total
Unimodal video	100.0	25.2	97.8	94.6	79.4
Unimodal location	99.6	97.7	29.6	99.7	81.7
Multimodal video	99.6	82.1	98.5	98.5	93.8
Multimodal location	99.6	97.7	36.6	99.7	83.4

B. Evaluation Results

1) Unimodal Object Estimation: Figure 5 shows the confidence outputs of the BAM. When the object with the highest confidence level is used as the identification result, the correct response rate of the BAM with the video modality only was 79.41%, and that with the location modality only was 81.66%, where the correct response rate is defined as the percentage of identification results output per input for 4,444 inputs that match the observed object. In Fig. 5, the vertical axis is the confidence level expressed as a normal logarithm, which is the result of a decision about which object is currently being observed, and the horizontal axis is the time step. In the first 1 to 1,111 time steps, the features of the first object (Obj. 1) are input to the BAM; in the next 1,112 to 2,222 time steps, the features of the second object (Obj. 2) are input; and so on up to 4,444 time steps. By the 4,444th time step, the features of four objects are input to the BAM. Figure 5(a) shows the results for the video modality. The confidence level is seen to drop in the middle of recognizing the second object, making it unrecognizable, but the observed object is correctly recognized in almost all other frames. Figure 5(b) shows the results in the location modality. The confidence level shows that the third object is not recognized from around 2,500 time steps, but the other objects are correctly recognized as observed. The failure of object recognition in each modality is due to the fact that the video is acquired while the camera is moving, and the extracted object features change significantly over time. In such a case, it is difficult to perform accurate object estimation with a unimodal approach, no matter what video analysis method is used.

2) Multimodal Object Estimation: The confidence levels in the video and location modalities, obtained in Section IV-B1, were given to the BCI model to evaluate the results of integrating causal inference and discrimination results. Table I shows the percentage of correct answers for recognition in each modality, and Fig. 6 shows the simulation results. In Table I, each "Unimodal" modality shows the results of decision making based only on the confidence output of each BAM, and each "Multimodal" modality shows the percentage of correct responses resulting from calculating the model average for each modality . In the model average algorithm, results are output from Eq. (7), and the output is based on one modality while the other modality complements the recognition result. We do not discuss in this paper which modality should be treated as primary in the multimodal approach, and consider it as a future issue, but all of the results in this study show an improvement in the correct answer rate when all objects are estimated compared to the unimodal results.

Figure 6(a) shows the results of causal inference. Basically, causal inference indicates that different modalities can observe the same object if two of the modalities agree, even though one of them has lower confidence, and indicates that they observe different objects if they disagree. For time steps 1 to 2,222, the causal inference result successfully shows that these modalities observed the same object. For the location modality