Object Estimation Method for Edge Devices Inspired by Multimodal Information Processing in the Brain

Ryoga Seki
Osaka University
Osaka, Japan
r-seki@ist.osaka-u.ac.jp

Daichi Kominami
Osaka University
Osaka, Japan
d-kominami@
ist.osaka-u.ac.jp

Hideyuki Shimonishi
Osaka University
Osaka, Japan
and NEC Corp.
h-shimonishi@
ist.osaka-u.ac.jp

Masayuki Murata
Osaka University
Osaka, Japan
murata@ist.osaka-u.ac.jp

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

***Abstract*—To realize real-time mobile augmented reality applications, various objects in the real world need to be instantly identified, located, and represented as a digital twin through sensor devices and edge IoT systems. However, it is challenging to make a fast and accurate decision on what the object is from real-time noisy streaming information. Multimodal decision making has been expected to mitigate such incomplete information and improve the accuracy of simplified recognition algorithms tuned for edge devices. In this paper, we propose an object estimation method inspired from the multimodal information processing mechanism of the brain, which makes decisions based on multiple types of uncertain observed information. 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

Real-time mobile augmented reality (mobile AR) applications, or more generally cyber physical systems (CPS), are promising use cases for 5G mobile network and edge computing systems. For these uses, digital representations of physical objects, or so-called digital twins, play a critical role in representing the whole 3D space. To understand and control the real world using digital twins, it is necessary to instantly understand various objects in the real world through sensor devices. In other words, 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 from real-time streaming information such as video because deep CNN models require a huge amount of computation.

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], 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.

In this paper, we propose a multimodal object recognition method inspired by the superior function of the human brain because 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 sensory organs 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) [2] and Bayesian causal inference (BCI) [3], [4].

II. Method

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, we can achieve robust object recognition based on uncertain observation information.

a) Object Estimation by the BAM: The BAM outputs values called *confidence*, which indicate what is observed corresponds to one of the options in memory, based on the information acquired by sensory organs. To apply the BAM to object estimation, it is necessary to determine the features to be observed by the BAM in each modality, and the reference data to be memorized in the BAM.

The feature values of the video modality are extracted using a Siamese region proposal network (RPN) [5]. In our proposal, 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 corresponding to the bounding box output by the RPN.

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.

The BAM stores the feature values of a reference image/location of the objects to be estimated. 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.

b) Extension of the BAM with BCI: After the BAM performs object estimation in each modality, the BCI model performs causal inference. Here, the *confidence* of the BAM, $\mathbf{c}$, 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.

The BCI model in [3] performs causal inference according to Bayes' theorem, as shown in the following equation: $p(C|u_1, u_2) = p(u_1, u_2|C)p(C)/p(u_1, u_2)$, 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 [3] defines $p(u_1, u_2|C)$ in a continuous manner, and we redefine it in a discrete manner as follows: $p(u_1, u_2|C = 1) = \sum_{k=1}^{K} p(u_1, u_2|O_k)p(O_k)$ and $p(u_1, u_2|C = 0) = \sum_{k=1}^{K} p(u_1|O_k)p(O_k) \sum_{k=1}^{K} p(u_2|O_k)p(O_k)$.

c) 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'_m that minimize it are used as the final object estimation results for modality m ($1 \leq m \leq 2$). Here, if $C = 1$, O'_1 and O'_2 output the same object, and if $C = 0$, the estimation result of each modality is output as it is. $Cost_m(O_m) = p(C = 1) \sum_{k=1}^{K} |O_m - O_k| p(O_k|u_1, u_2) + p(C = 0) \sum_{k=1}^{K} |O_m - O_k| p(O_k|u_m)$, where $|O_m - O_k|$ 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 [3], 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.

d) Object Identification Method: To use the BAM and 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 $\mathbf{c}$ for u. The probability that the BAM confidence value is $\mathbf{c}$ when the object O_k is observed is defined as $p(\mathbf{c}|O_k)$, which is used for the calculation of $p(O_k|\mathbf{c})$. Then, finally, we obtain the object label that minimizes the cost function. 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}).

III. Results

To confirm the effectiveness of the object estimation method applying the BAM and BCI, as described above, we conducted a simulation-based evaluation. For the video dataset, we used a real measurement public dataset of various objects (Yale-CMU-Berkeley Object and Model set) [6]. For each frame and each of four objects in the video data, we extracted the features of the video modality and 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.

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

a) Accuracy: Table I shows the percentage of correct answers for recognition in each modality. In the table, each "Unimodal" modality shows the results of decision making based only on the maximum 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, 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.

b) Calculation Time: As for the computational time of the proposed method, the most computationally intensive operation is when 128-dimensional video features are input to the BAM, $\mathbf{z}_t$ is estimated, and the confidence level is output. The actual computation time on a desktop computer (CPU: Core-i7 8700, RAM: 16.0 GB) for this operation was 1.18 ms per frame input, which can be applied to 30-fps and 60-fps videos in this evaluation environment.

IV. Conclusion

In this paper, we proposed a method for object estimation from noisy observed information based on multiple types of uncertain observation information. We introduce a mechanism for making appropriate decisions by processing and combining incomplete observation information from two modalities. Computer simulations showed that the proposed method can combine the parts that are recognized with high confidence in each modality and can make decisions with higher accuracy than that of a unimodal method. Future work includes investigating how to learn a new attractor when an object is recognized that has not been learned beforehand and how to estimate multiple objects by observing them simultaneously.

References

[1] M. Subedar, et al., "Uncertainty-aware audiovisual activity recognition using deep Bayesian variational inference," in *Proc. IEEE ICCV*, Oct. 2019, pp. 6301–6310.

[2] S. Bitzer, et al., "A Bayesian attractor model for perceptual decision making," *PLoS Comput. Biol.*, vol. 11, no. 8, p. e1004442, 2015.

[3] K. P. Körding, et al., "Causal inference in multisensory perception," *PLoS one*, vol. 2, no. 9, p. e943, 2007.

[4] T. Rohe and U. Noppeney, "Cortical hierarchies perform bayesian causal inference in multisensory perception," *PLoS Biol.*, vol. 13, no. 2, p. e1002073, 2015.

[5] B. Li, et al., "High performance visual tracking with siamese region proposal network," in *Proc. IEEE CVPR*, Jun. 2018, pp. 8971–8980.

[6] "YCB benchmarks–object and model set," available at http://www.ycbbenchmarks.com/.