LETTER *Special Section on Deep Learning Technologies: Architecture, Optimization, Techniques, and Applications*

Effectiveness of Feature Extraction System for Multimodal Sensor Information Based on VRAE and Its Application to Object Recognition

Kazuki HAYASHI[†], ***Nonmember and*** **Daisuke TANAKA**[†a)], ***Member***

SUMMARY To achieve object recognition, it is necessary to find the unique features of the objects to be recognized. Results in prior research suggest that methods that use multiple modalities information are effective to find the unique features. In this paper, the overview of the system that can extract the features of the objects to be recognized by integrating visual, tactile, and auditory information as multimodal sensor information with VRAE is shown. Furthermore, a discussion about changing the combination of modalities information is also shown.

key words: *dimensionality reduction, auto-encoder, object recognition*

1. Introduction

For a robot to operate autonomously in an environment coexisting with humans, it is required to accurately recognize the surrounding environment. Object recognition by these robots is one of required skills. The recognition methods that use multiple sensors have proposed the methods that integrate visual and tactile information [1] and auditory information with robot behavioral information [2], [3] in prior research. And, to achieve object recognition, it is necessary to find the unique features of the objects to be recognized.

In our previous study, a system that can extract the features of the objects to be recognized by integrating visual, tactile, and auditory information as multimodal sensor information is developed [4]. To verify the effectiveness of the system, the recognition problem of 8 colored balls are carried out. The extracted feature are well-decoupled based on its color. However, other features such as softness, size, and inner structure of the balls are not exploited in the obtained feature. Hence, the effectiveness of the integration of multimodal sensor information seems to be not sufficiently shown in the previous work.

In this paper, the results of feature extraction and the comparison experiments by changing the combination of modalities information is shown to demonstrate how information is utilized for feature extraction.

2. Problem Settings

In this study, we consider the recognition problem as shown in Fig. 1 using information about the objects. We have prepared different-type felt balls which include 8 different colors with different sizes, softness, and structures (with or without bell) inside the objects. Not only visual information but also tactile or auditory information would be useful to recognize these objects.

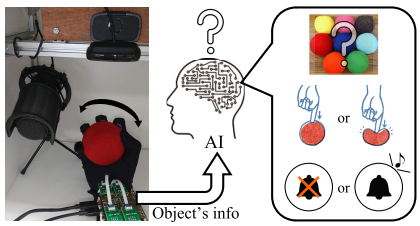

Fig. 1 The problem to be considered in this study. Using information about the objects, such as color, size, softness, and inner structure, the AI model recognizes them.

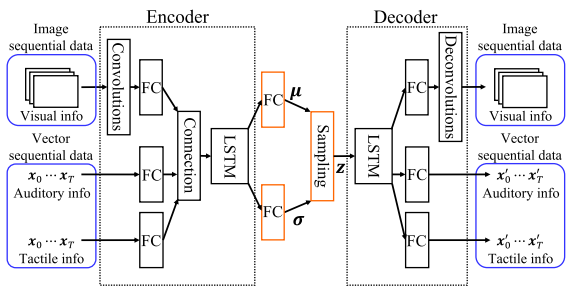

Fig. 2 VRAE model for multimodal sensor information.

3. Proposed Method

In this study, we consider the use of neural networks shown in Fig. 2, as a system for extracting object features based on visual, auditory, and tactile information as used in [4].

In this system, the VRAE (Variational Recurrent Auto-Encoder) used in [5], [6] is adopted, considering that demonstrated in [7] that deep auto-encoder outperforms PCA (Principal Component Analysis) in compressive feature acquisition. The VRAE is a Sequence-to-Sequence VAE (Variational Auto-Encoder) model that has been extended to handle time-series data and is capable of mapping a latent vector into the time-series data.

Images are captured from a camera as time-series data while grasping the objects, and these are input to the LSTM

Manuscript received March 31, 2022.
Manuscript revised September 14, 2022.
Manuscript publicized January 12, 2023.

[†]The authors are with the National Institute of Technology (KOSEN), Niihama College, Niihama-shi, 792–8580 Japan.
a) E-mail: d.tanaka@niihama-nct.ac.jp

DOI: 10.1587/transinf.2022DLL0008

(Long Short-Term Memory) layer via CNN (Convolutional Neural Network). Then, the auditory and tactile information are input to the LSTM layer via the DNN (Deep Neural Network) and concatenated with the output vectors from the CNN. And, the latent variables based on the normal distribution $\mathcal{N}(\mu, \sigma^2)$ are obtained from the sampling layer. Furthermore, in order to visualize the features, the dimensions of the latent variables that are input from the LSTM layer are lowered to three dimensions. When we perform multi-class classification inputting this latent vector, we can obtain high recognition accuracy. The sampled $\mathbf{z}$ was used to reconstruct and the network is trained to restore to the original input data. Note that the output of the LSTM layer in the decoder part was split and input to each FC through the reverse operation of the one performed in the encoder part.

4. Experiments

As compared to our previous work [4], we limit the object colors to verify the effectiveness of the other sensor modality. The characteristics of the balls used in this experiment are shown in Table 1. In addition, smaller (6 cm) hard-type blue-colored ball without bell is used. Consequently, 9 balls are used in this experiment.

4.1 Training Data and Network Settings

All data are obtained from the same system used in our previous study [4]. Here, we show the outline of the system.

We have prepared 1000 data for each object as training data. Using the measurement environment shown in Fig. 3, we acquire visual and auditory information. This environment consists of a webcam and a condenser microphone, and the resolution of the webcam is set to 320×640 pixels, and the sampling frequency of the condenser microphone is set to 48000 Hz.

Table 1 The characteristics of the balls used in the experiment. A ball has a combination of the characteristics (for example, a hard blue-colored ball with bell).

Characteristic	Availability
Color	Blue (BL) / Red (RD)
Softness	Hard / Soft
Inner structure	with Bell / without Bell

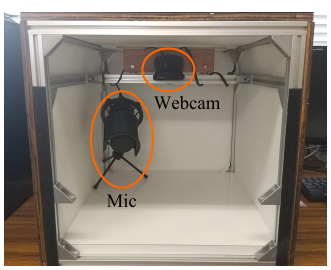

Fig. 3 Measurement environment with webcam and condenser microphone.

In this environment, objects are grasped using a glove device. Four tactile sensors (Shokac Chip by Touchence Corp.) are attached to the glove device referring to [8], and the object's states are captured using webcam, microphone, and tactile sensors while the objects are grasped.

To speed up the processing of VRAE, the resolution of the acquired image is reduced by a factor of 10, and resized image is used as visual information. In addition, the recorded sound is preprocessed by a BPF (Band Pass Filter) to reduce noises. The passband edge frequency is set to 1500–4500 Hz and the stopband edge frequency is set to 500–6000 Hz. Then, the denoised sound is performed a short-time Fourier transform with an overlap rate of 50%, using a Hanning window, so that the interval between each step is 30 Hz. Since we have already known the frequency of bell inside the objects, the information of 400 dimensions was extracted from the obtained frequency spectrogram and input to the VRAE as auditory information. The tactile information is obtained by the sensor on the glove device. Since each tactile sensor can acquire 6 dimensions (pressure in three axes and moment of force in each axis), the 24-dimensional data obtained by this glove device is input to the VRAE as tactile information.

After training the VRAE model, another model (called the classification model) to classify 9 objects using obtained three-dimensional latent values is trained. Additional 1000 data for each ball to train classification model are prepared. The hyperparameters for all model are determined using validation data prepared by dividing the training data.

4.2 Visualization of Extracted Feature

Using the VRAE model, the three-dimensional latent variables of the objects are obtained as shown in Fig. 4. Figure 4 shows that features are successfully extracted based on information of each modality because the features are sepa-

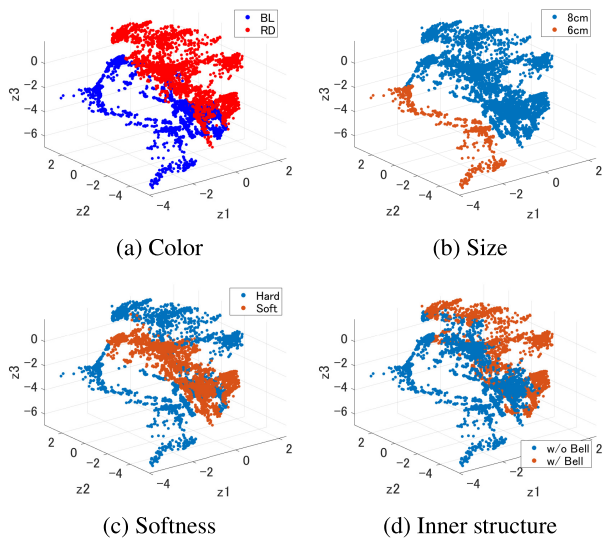

Fig. 4 The 3D feature values extracted from our VRAE model are colored by each feature.

Table 2 Accuracy comparison with sensor information used. The marks ∘ and – stand for "used," and "not used" respectively.

Input Data			Accuracy
Visual	Tactile	Auditory	
∘	–	–	56.8%
–	∘	–	65.4%
–	–	∘	55.2%
∘	∘	–	92.7%
∘	–	∘	90.7%
–	∘	∘	87.8%
∘	∘	∘	**97.7%**

rated in terms of the conditions of each modality. In addition, the features seem to be widely distributed so that they are separable based on each information. These results indicate that the VRAE model proposed in this study can extract features contained in the input information as features based on the information about the modality used.

4.3 Effectiveness of Integrating Multimodal Information

We conduct a comparison experiment by changing the combination of information from each sensor to discuss the effectiveness of integrating multimodal information. The obtained results are shown in Table 2.

Table 2 shows that using more sensory information performs better. Furthermore, it also shows that adding visual information to tactile and auditory information rapidly increases the accuracy. As seen in the results, the effectiveness of our system with VRAE has been shown, and it is also shown that the recognition for objects with multiple features would be improved with multimodal sensor information.

In addition, the model's accuracy using only information about tactile is obtained the highest accuracy of any model using a single modality, with higher classification accuracy for size as well as softness. These results indicate two things. One is that the pressure information is likely to include information on the size of the objects. The other is that size-related features are more likely to be expressed in the pressure information than image information. Therefore, when extracting features related to the size of an object, a method that uses not only visual but also tactile information may be effective.

5. Conclusion

In this study, it is shown that our system with VRAE can be used to integrate multimodal sensor information and extract the features of the objects to be recognized. In addition, in order to obtain a high accuracy of the classification model, the multimodal sensor information can be used to extract multiple features of an object and it is effective to apply them to object recognition. As our future work, the method would be able to apply to the actual robot.

Acknowledgments

This work is supported by a research grant from The Mazda Foundation.

References

[1] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-tactile fusion for object recognition," IEEE Trans. Autom. Sci. Eng., vol.14, no.2, pp.996–1008, 2017.

[2] J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," Robotics and Autonomous Systems, vol.62, no.5, pp.632–645, 2014.

[3] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," Robotics and Autonomous Systems, vol.62, no.6, pp.721–736, 2014.

[4] K. Hayashi and D. Tanaka, "Integration of multimodal sensor information using VRAE and application to object recognition," RISP International Workshop on Nonlinear circuits, Communications and Signal Processing, pp.33–36, Feb. 2022.

[5] O. Fabius and J.R. van Amersfoort, "Variational recurrent auto-encoders," arXiv:1412.6581, 2014.

[6] S.R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," Proc. 20th SIGNLL Conference on Computational Natural Language Learning, pp.10–21, 2016.

[7] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol.313, pp.504–507, 2006.

[8] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," Nature, vol.569, pp.698–702, 2019.