Robust Classification Model with Multimodal Learning for Home Service Robots

Ikuya Matsumoto
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan

Daiju Kanaoka
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan

Hakaru Tamukoh
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan
E-mail:matsumoto.ikuya585@mail.kyutech.jp, kanaoka.daiju327@mail.kyutech.jp, tamukoh@brain.kyutech.ac.jp
https://www.lsse.kyutech.ac.jp/english/

Abstract

We propose an auxiliary data stream structure as a robust classification model. The model treats one modal as the main input and the other modals as supports. The model chooses how much of the sub-modal is used for classification. We experimented with two and three modal inputs. Moreover, we added pseudo-shadows to the visual information for the experiment with three modal inputs. In all experiments, our proposed model improves the accuracy and robustness to environmental disturbances by using multiple modals.

Keywords: Multimodal learning, Classification, Home service robot, Gate structure

1. Introduction

In recent years, service robots have attracted attention owing to the shortage of workers caused by the aging population[1]. Among them such robots, home service robots work in human living environments to assist people. We expect home service robots to perform domestic tasks in the home, such as cleaning up rooms[2][3][4][5][6]. Object classification is one of the functions required for home service robot operation. Currently, object classification uses visual information for home service robots. However, recognition accuracy decreases if the conditions for capturing images are poor (e.g., shadows, blurring). Multimodal learning is one of the solutions to this problem. The method combines several different types of information, such as image and tactile[7]. The application of multimodal learning to object classification will solve the abovementioned problem (e.g., operating in a poor visual environment). We propose a novel multimodal object classification model in which friction and pushing modals to assist with the visual modal.

2. Related Work

2.1. *Multimodal learning*

Multimodal learning is a learning method that combines multiple types of modals, such as images, text, and voice. This learning method produces more accurate output than unimodal learning and has been applied in various fields, such as speech recognition and disease prediction[8][9].

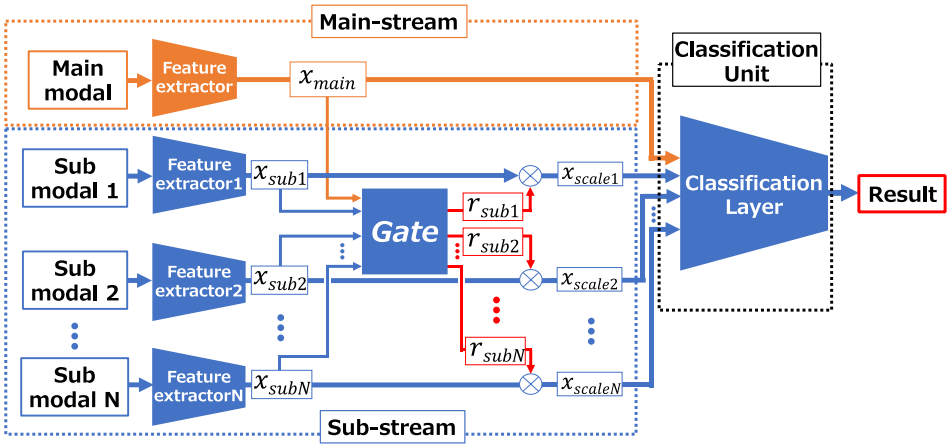

Fig. 1. Multimodal learning with auxiliary data stream model

2.2. *Multimodal learning without gate structure (Gao's model)*

Gao et al. conducted object classification experiments by using images and tactile information[10]. Gao's model was generally more accurate than the unimodal model. However, for some objects, multimodal was degraded accuracy than unimodal. We need to devise a structure for input rather than simply inputting multimodal.

2.3. *Deep gated multimodal learning (DGML)*

Anzai et al. proposed a gate structure for multimodal learning to estimate the pose of a grasping object using images and tactile inputs[11]. The gate structure calculates reliability values that adjust the effect of each modal on the overall recognition result. The sum of the reliability values for the image and haptic is one, and a modal with a higher reliability value is more effective for the estimation. Anzai et al. showed that gate structures are effective for multimodal learning. Therefore, gate structures can be applied to multimodal object classification to improve accuracy.

3. Proposed Model

We propose multimodal learning with an auxiliary data stream model shown in Figure 1. The proposed model is treated as one modal as the main input, whereas the other modals are used to support it. The proposed model can input multiple sub-modals, while maintaining the dominance of the main-modal.

3.1. *Auxiliary data stream model*

The proposed model consists of a mainstream, a substream, and a classification unit. The mainstream and a substream extract features using a feature extractor that is appropriate for the input modal. The extracted main-modal features (x_{main}) and sub-modal features ($x_{sub1}, x_{sub2}, \ldots, x_{subN}$) are input to the gate, which calculates the sub-modal reliability value ($r_{sub1}, r_{sub2}, \ldots, r_{subN}$). The substream scales the sub-modal features by multiplying the calculated reliability values by the sub-modal features. Finally, the main-modal features and the scaled sub-modal feature (x_{scale}) input to the classification unit output the results.

3.2. *Gate structure*

The gate structure used in the proposed model is shown in Figure 2. The fully connected (FC) layers ($FC0, FC1, \ldots, FCN$) reduce the number of feature dimensions extracted from each modal to one ($x_{\mathrm{FC1}}, x_{\mathrm{FC2}}, \ldots, x_{\mathrm{FCN}}$). The features were concatenated and input into FC. The FC layer and the sigmoid function calculated the reliability values of the sub-modals. The reliability value was obtained through learning.

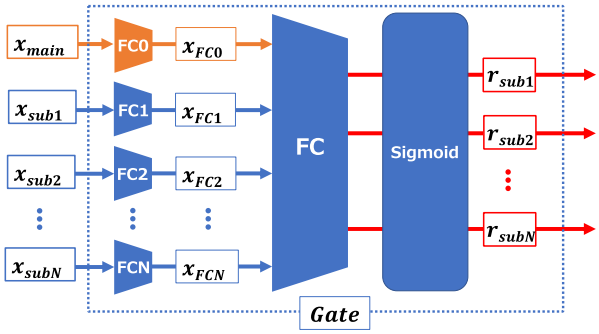

Fig. 2. Gate structure. The reliability value can be in the range of [0, 1]. The larger value, the more effective the modal is for classification.

4. Experiments

4.1. *Dataset*

We used the LMT haptic texture database for this experiment[12]. The dataset contains 10 training and 10 test data per class. We used visual, friction, and pushing modals in the experiment. The details of each modal are presented in Table 1. The visual modal is RGB images taken without flash. The friction modal is time series data of three axes (x, y, z) acquired by sliding the tactile sensor, and the pushing modal is time series data of one axis (z) acquired by pushing the tactile sensor.

Table 1. Data details

Modal	Shape
Visual	480 × 320 [px]
Friction	3 axes (x, y, z) × 48000 [steps]
Pushing	1 axis (z) × 1601 [steps]

4.1.1. *Preprocess*

We preprocessed each modal using the following procedure. The visual modal compressed the image size. The friction and pushing modals were downsampled to remove unstable data.

- Visual
  (i) Resized images to 50 × 50 [px].
  (ii) Normalized the images to [0 , 1].
- Friction
  (i) Skipped 3000 [steps] at the beginning as the data are unstable.
  (ii) Normalized data to [0, 1] for each axis.
  (iii) Down sampled 45000 [steps] to 300 [steps] using an anti-aliasing filter.
- Pushing
  (i) Skipped 101 [steps] at the beginning as the data are unstable.
  (ii) Normalized data to [0, 1] for each axis.
  (iii) Down sampled 1500 [steps] to 300 [steps] using an anti-aliasing filter.

4.2. *Evaluation experiment with two modals*

4.2.1. *Overview*

We conducted 108 class classification experiments using visual and friction as modal inputs. In this experiment, visual is the main-modal, and friction is the sub-modal. Figure 3 shows the model used in this experiment. The visual feature extractor was a CNN, and the friction feature extractor was a 3D CNN.

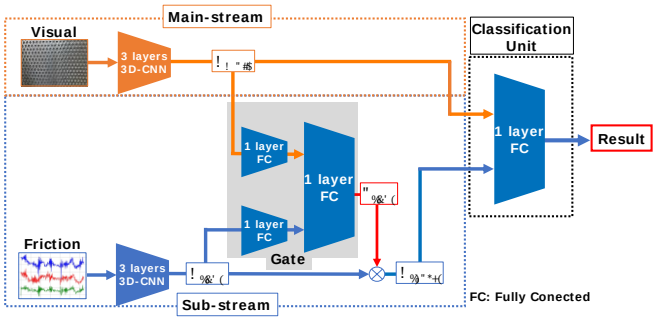

Fig. 3. Proposed model for two modals

4.2.2. *Result*

The experimental results are shown in Table 2. The proposed model has a gated structure and main and sub-stream, which allowed it to achieve an accuracy higher than that of other models. The proposed model is more accurate than DGML, and the structure that gives the main and sub-functions to the modal is effective.

Table 2. Results of the with two modals

Model	Accuracy [%]
Gao's	93.0
DGML	94.0
Ours	94.9

Table 3. Network design for 3 modals experiments

	Layer	In	Out	Filter Size	Activation Function
Visual Feature Extractor	1st 2D Conv.	3	10	(5, 5)	ReLU
	2nd 2D Conv.	10	30	(5, 5)	ReLU
	3rd 2D Conv.	30	108	(3, 3)	ReLU
	Adaptive Avg.Pool2D	108	108	-	-
Friction Feature Extractor	1st 3D Conv.	3	10	(1, 1, 6)	ReLU
	2nd 3D Conv.	10	30	(1, 1, 3)	ReLU
	3rd 3D Conv.	30	108	(1, 1, 3)	ReLU
	Adaptive Avg.Pool3D	108	108	-	-
Pushing Feature Extractor	1st 3D Conv.	1	10	(1, 1, 6)	ReLU
	2nd 3D Conv.	10	30	(1, 1, 3)	ReLU
	3rd 3D Conv.	30	108	(1, 1, 3)	ReLU
	Adaptive Avg.Pool3D	108	108	-	-
Gate	1st FC_{visual}	108	1	-	-
	1st $FC_{friction}$	108	1	-	-
	1st $FC_{pushing}$	108	1	-	-
	2nd FC	3	2	-	Sigmoid
Classifier	FC	324	108	-	-

4.3. *Evaluation experiment with three modals*

4.3.1. Overview

We conducted 108 class classification experiments using visual, friction, and pushing as the modal inputs. In this experiment, visual was the main-modal, and the others were sub-modal. The model used in this experiment is shown in Figure 4, and the network design is presented in Table 3. The feature extractor for pushing was a 3D CNN. A classification experiment was also conducted by adding pseudo-shadows to the visuals to simulate a real environment.

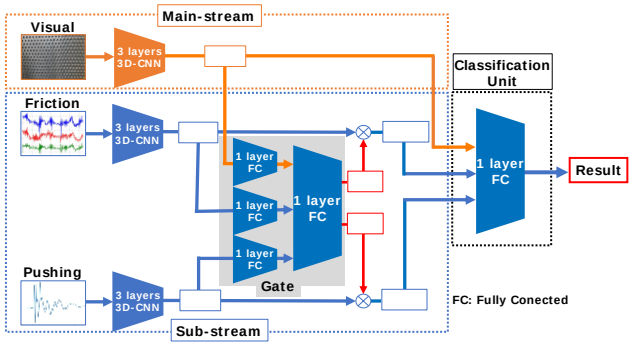

Fig. 4. Proposed model for three modals

4.3.2. *Result*

Table 4 summarizes the experimental results. The proposed model achieved higher recognition accuracy than Gao's model. Our model has a smaller loss of accuracy than Gao's model because the useful modals can be combined in pseudo-shadowed visual experiments.

Table 4. Results of the with three modals

Model	Accuracy [%] (w/o shadows)	Accuracy [%] (w/ shadows)	Amount of decrease
Gao's	97.7	97.0	−0.7
Ours	98.1	97.6	−0.5

5. Conclusion

We proposed an auxiliary data stream model as a robust object classification model and conducted validation experiments in this study. The proposed model achieved higher accuracy than the conventional model, and the experiments simulated real-world conditions also showed high accuracy and robustness of the model.
However, to apply the proposed model to a home service robot, it is necessary to validate the model using actual

data acquired by the robot. In addition, although we focused on shadows in the real environment to verify the robustness of the proposed model, it is necessary to verify the model against other disturbances, such as blurring and backlighting. In future works, we will attempt to implement and evaluate the proposed model on a home service robot.

References

1. New Energy and Industrial Technology Development Organization (NEDO), Future Robot Market Forecasts Released, (Accessed 2022-01-23).
2. Tomohiro Ono, Daiju Kanaoka, Tomoya Shiba, Shoshi Tokuno, Yuga Yano, Akinobu Mizutani, Ikuya Matsumoto, Hayato Amano, and Hakaru Tamukoh, "Solution of World Robot Challenge 2020 Partner Robot Challenge (Real Space)," Advanced Robotics, Vol. 36, Issue 17-18, pp. 870-889, 2022
3. Yuma Yoshimoto, Hakaru Tamukoh, "FPGA Implementation of a Binarized Dual Stream Convolutional Neural Network for Service Robots," Journal of Robotics and Mechatronics, Vol. 33, No. 2, pp. 386-399, 2021.
4. Yuichiro Tanaka, Takashi Morie, Hakaru Tamukoh, "An amygdala-inspired classical conditioning model on FPGA for home service robots," IEEE Access, Vol. 8, pp. 212066-212078, November 2020.
5. Yutaro Ishida, Takashi Morie, Hakaru Tamukoh, "A hardware intelligent processing accelerator for domestic service robots," Advanced Robotics, Vol. 34, Issue 14, pp. 947-957, June 2020.
6. Yutaro Ishida, Hakaru Tamukoh, "Semi-Automatic Dataset Generation for Object Detection and Recognition and its Evaluation on Domestic Service Robots," Journal of Robotics and Mechatronics, Vol. 32, No. 1, pp. 245-253, 2020.
7. Denis Ivanko, Alexey Karpov, Dmitrii Fedotov, Irina Kipyatkova, Dmitry Ryumin, Dmitriy Ivanko, Wolfgang Minker, and Milos Zelezny, "Multimodal Speech Recognition: Increasing Accuracy Using High Speed Video Data." Journal on Multimodal User Interfaces, Vol. 12, No. 4, pp. 319-328, 2018.
8. Hiroshi Kawasaki, "Clinical Multimodal Data Analysis (in Japanese), " Japanese Journal of Allergology, Vol. 69 No.8. pp. 708-709, September 2020.
9. Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep Multimodal Learning for Audio-Visual Speech Recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2130–2134, 2015.
10. Yang Gao, Lisa Anne Hendricks, Katherine J. Kuchenbecker, and Trevor Darrell. Deep Learning for Tactile Understanding From Visual and Haptic Data. IEEE International Conference on Robotics and Automation, Vol. 2016-June, pp. 536–543, 2016.
11. Tomoki Anzai and Kuniyuki Takahashi. Deep Gated Multi-modal Learning: In-hand Object Pose Changes Estimation using Tactile and Image Data. IEEE International Conference on Intelligent Robots and Systems, pp. 9361–9368, 2020.
12. Chair of Media Technology Technical University of Munich. LMT Haptic Texture Database. (Accessed 2021-10-26).

Authors Introduction

Mr. Ikuya Matsumoto

He received the B.Eng. degree from Kyushu Institute of Technology, Japan, in 2022. He is currently in a Master's degree student the graduate school of Life Science and Systems Engineering, Kyushu Institute of Technology. His research interest includes multimodal learning and domestic service robots.

Mr. Daiju Kanaoka

He received the B.Eng. and the M.Eng. degree from Kyushu Institute of Technology, Japan, in 2020 and 2022, respectively. He is currently in a Ph.D. student in the graduate school of Life Science and Systems Engineering, Kyushu Institute of Technology. His research interest includes object recognition, multimodal learning, and domestic service robots. He is a student member of IEEE.

Prof. Hakaru Tamukoh

He received the B.Eng. degree from Miyazaki University, Japan, in 2001. He received the M.Eng and the Ph.D. degree from Kyushu Institute of Technology, Japan, in 2003 and 2006, respectively. He was a postdoctoral research fellow of 21st century center of excellent program at Kyushu Institute of Technology, from April 2006 to September 2007. He was an assistant professor of Tokyo University of Agriculture and Technology, from October 2007 to January 2013. He is currently an associate professor in the graduate school of Life Science and System Engineering, Kyushu Institute of Technology, Japan. His research interest includes hardware/software complex system, digital hardware design, neural networks, soft-computing and home service robots. He is a member of IEICE, SOFT, JNNS,IEEE, JSAI and RSJ..