Semantic Scene Difference Detection in Daily Life Patroling by Mobile Robots using Pre-Trained Large-Scale Vision-Language Model

Yoshiki Obinata[1], Kento Kawaharazuka[1], Naoaki Kanazawa[1], Naoya Yamaguchi[1], Naoto Tsukamoto[1], Iori Yanokura[1], Shingo Kitagawa[1], Koki Shinjo[1], Kei Okada[1] and Masayuki Inaba[1]

Abstract— It is important for daily life support robots to detect changes in their environment and perform tasks. In the field of anomaly detection in computer vision, probabilistic and deep learning methods have been used to calculate the image distance. These methods calculate distances by focusing on image pixels. In contrast, this study aims to detect semantic changes in the daily life environment using the current development of large-scale vision-language models. Using its Visual Question Answering (VQA) model, we propose a method to detect semantic changes by applying multiple questions to a reference image and a current image and obtaining answers in the form of sentences. Unlike deep learning-based methods in anomaly detection, this method does not require any training or fine-tuning, is not affected by noise, and is sensitive to semantic state changes in the real world. In our experiments, we demonstrated the effectiveness of this method by applying it to a patrol task in a real-life environment using a mobile robot, Fetch Mobile Manipulator. In the future, it may be possible to add explanatory power to changes in the daily life environment through spoken language.

I. INTRODUCTION

Robots are becoming capable of performing a variety of life-support behaviors. For robots to spontaneously perform these life-support actions, it is necessary to capture changes in the daily life environment in which the robot is operating.

In the field of computer vision, the task of capturing changes in images includes anomaly detection. Research on anomaly detection has been conducted for many years. Methods based on probabilistic models like Bayesian networks[1] and hidden Markov models[2] are used for it, and in recent years, deep learning methods have been used to deal with complex situations. For example, some use Variational Auto Encoder[3], some use generative models such as GAN[4]. These methods require collecting a large amount of data to train the model. In contrast, our study uses language to detect scene difference in daily life environment. Starting with Transformer[5], models such as BERT[6], T5[7] have demonstrated remarkable performance in language tasks. Language models are also beginning to be introduced into robotics fields[8][9][10][11][12][13][14]. In computer vision, large-scale vision-language models trained using these language models have been studied extensively, starting with VQA[15], and in recent years, trained models capable of performing various vision-language tasks, such as CLIP[16],
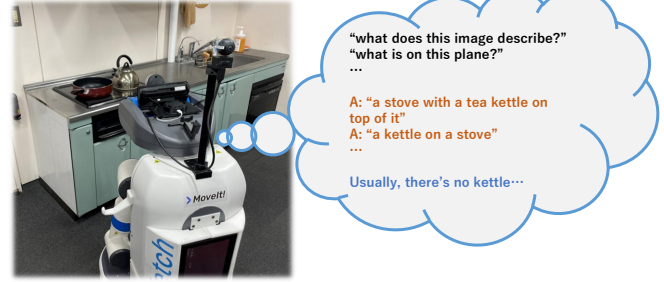

Fig. 1. Semantic scene difference detection using spoken language by the robot. The robot perceives its daily life environment in the space of spoken language and uses its semantic differential to detect changes.

GLIP[17][18], and OFA[19] have begun to be made available. Unlike image recognition models such as SSD[20] and VGG[21], these models are difficult to train locally on a small number of GPUs, but they are trained on a large amount of text and image datasets with rich computational resources and have knowledge about the language of human society and images. Therefore, it is possible to accurately extract features from a single image and translate them into language. There are some studies that use pre-trained models for anomaly detection in videos and datasets[22][23].

Based on this background, we propose a method to apply a large-scale vision-language model that has already been trained to calculate scene distance in daily life environment for a mobile robot. The large-scale model answers questions about the image captured by the mobile robot's camera and compares the sentences between reference and current images to detect how different the situation is in the same location. This sentence comparison is performed by preparing multiple questions in the VQA task and numerically comparing the answers between reference and current image. This method requires only one reference image for each location and does not require any model re-training.

II. SCENE DIFFERENCE DETECTION USING PRE-TRAINED VISION LANGUAGE MODELS

A. Overview of semantic scene difference detection system using pre-trained vision-language model

The concept of this research is shown in Fig. 1. Robots can capture semantic changes by comparing situations in their daily life environment through language. It is important to quantify these changes to detect anomalies and initiate tasks in the daily life environment based on these results. Fig. 2

[1]The authors are with the Department of Mechano-Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan. [obinata, kawaharazuka, kanazawa, yamaguchi, tsukamoto, yanokura, s-kitagawa, shinjo, k-okada, inaba]@jsk.imi.i.u-tokyo.ac.jp
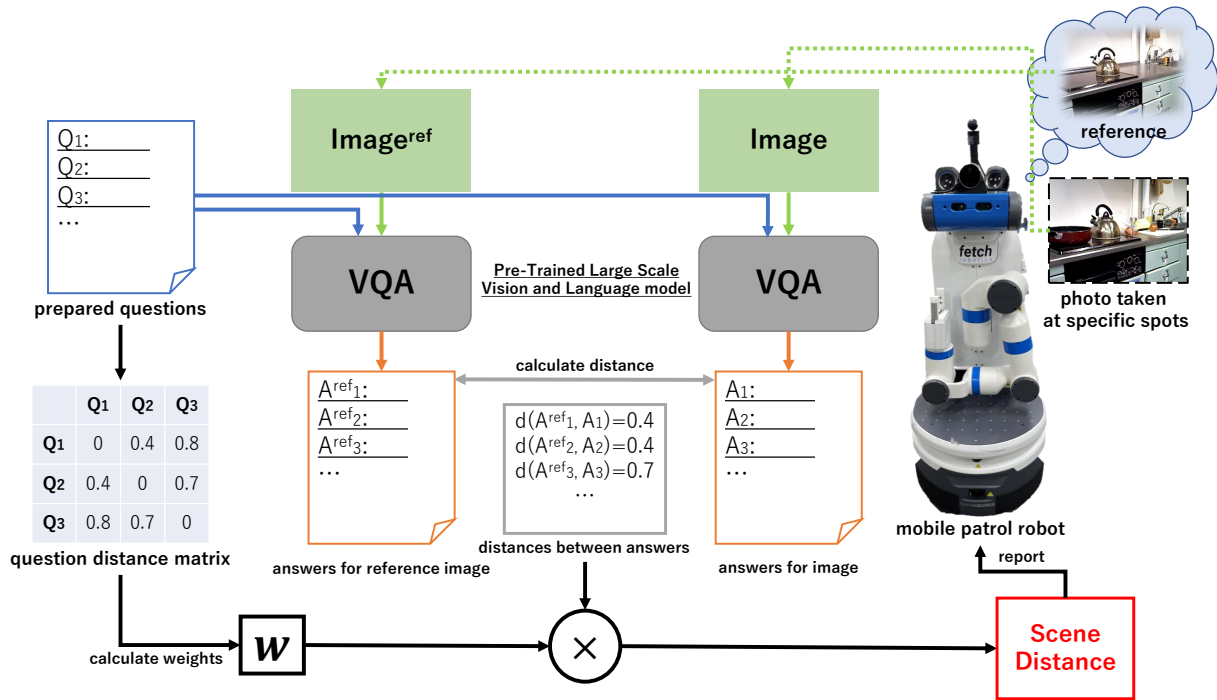
Fig. 2. Overall view of semantic scene difference quantifying system using pre-trained vision-language model. The robot uses a large-scale vision-language VQA model to compare reference and captured images, deriving their semantic distance by analyzing the answers to questions about the images. The distance between the answer sentences is calculated using doc2vec, resulting in two vectors whose distance determines the semantic distance between the images. Weights can be given to elements of the vectors to reduce the effect of question proximity.
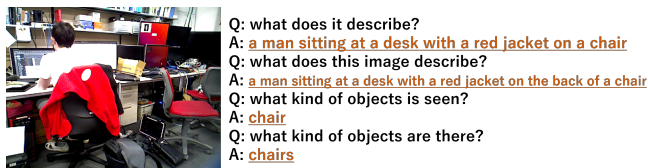

Fig. 3. Examples of differences in output sentences depending on the words used. Pronouns and singular and plural forms slightly change the output sentences.

shows an overall view of the method used to quantifying its semantic change. The patrol robot compares the reference and captured images and derives their semantic distance. A large-scale vision-language VQA model is given multiple question sentences, a reference image, and a captured image, and the answer sentences for each image are obtained. The distance between the answer sentences corresponding to the same question sentence in the two images is determined by doc2vec, and two vectors of the same dimension as the number of question sentences are derived. By finding the distance between these two vectors, the semantic distance between the two images can be obtained, but by giving appropriate weights to the elements of the vectors, the effect of the semantic proximity of the questions can be reduced.

B. Features of Large-Scale Vision-Language Models in Scene Difference Detection Tasks

In this study, we use OFA as a large-scale vision-language model. There are five types of tasks for generating rectangles and text from images: Visual Grounding, Grounded Captioning, Image Text Matching, Image Captioning, and Visual Question Answering. Image Captioning (IC) and Visual Question Answering (VQA) are suitable for explaining the situation, as they allow the user to ask questions and obtain the answers. VQA in OFA sometimes outputs strange sentences like "prototype prototype prototype of table", "messily messy office space" so we use IC.

The model outputs slightly different answers to questions with similar expressions, as shown in Fig. 3. Since the output of the model changes depending on the prepositions, pronouns, and verbs used, it makes sense to input a variety of expressions, even for the similarly worded questions.

C. Evaluation of Scene Difference

We describe a method for calculating the scene difference distance at a patrol location. As shown in Fig. 2, dozens of questions are prepared in advance to ask a large-scale model to describe a situation. Suppose that m questions are prepared, and m answers can be obtained by inputting a single image and m questions to OFA. The questions and answers are transformed into a 768-dimensional vector via all-mpnet-base-v2 [24], a pre-trained language model that further fine-tunes MPNet [25]. Define these as

$$q_1, q_2, ..., q_m$$
$$a_1, a_2, ..., a_m$$

respectively. Let

$$a_1^{ref}, a_2^{ref}, ..., a_m^{ref}$$

be the sentence vectors of the reference responses, and define the Scene Distance SD as

$$SD = \sum_{k=1}^{m} w_k D_c(\boldsymbol{a_k}, \boldsymbol{a_k^{ref}})$$

using the weights w_k. D_c is the cosine distance, defined by follows

$$D_c(\boldsymbol{p}, \boldsymbol{q}) = 1 - \frac{\boldsymbol{p} \cdot \boldsymbol{q}}{||\boldsymbol{p}||||\boldsymbol{q}||}$$

Normally, the weights w_k is calculated as follows

$$w_k = \frac{1}{m} \tag{1}$$

However, this method makes it difficult for questions of different types to be reflected in the scene distance. Therefore, we propose the following method to calculate the weights $\boldsymbol{w}$

$$
\begin{aligned}
M_{rel} &= \begin{pmatrix} D_c(\boldsymbol{q_1},\boldsymbol{q_1}) & D_c(\boldsymbol{q_1},\boldsymbol{q_2}) & \cdots & D_c(\boldsymbol{q_1},\boldsymbol{q_m}) \\ D_c(\boldsymbol{q_2},\boldsymbol{q_1}) & D_c(\boldsymbol{q_2},\boldsymbol{q_2}) & \cdots & D_c(\boldsymbol{q_2},\boldsymbol{q_m}) \\ \vdots & \vdots & \ddots & \vdots \\ D_c(\boldsymbol{q_m},\boldsymbol{q_1}) & D_c(\boldsymbol{q_m},\boldsymbol{q_2}) & \cdots & D_c(\boldsymbol{q_m},\boldsymbol{q_m}) \end{pmatrix} \\
&= \begin{pmatrix} 0 & D_c(\boldsymbol{q_1},\boldsymbol{q_2}) & \cdots & D_c(\boldsymbol{q_1},\boldsymbol{q_m}) \\ D_c(\boldsymbol{q_1},\boldsymbol{q_2}) & 0 & \cdots & D_c(\boldsymbol{q_2},\boldsymbol{q_m}) \\ \vdots & \vdots & \ddots & \vdots \\ D_c(\boldsymbol{q_1},\boldsymbol{q_m}) & D_c(\boldsymbol{q_2},\boldsymbol{q_m}) & \cdots & 0 \end{pmatrix} \\
\boldsymbol{v} &= \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} \\
\boldsymbol{w} &= \frac{M_{rel}\boldsymbol{v}}{||M_{rel}\boldsymbol{v}||}
\end{aligned}
\tag{2}
$$

The weights $\boldsymbol{w}$ in (2) have the function of decreasing the contribution of similar meaning questions and increasing the contribution of questions with different meanings from others.

The further apart the meanings of the questions are, the more scene difference can be captured. We define Quality of Questions QoQ, a measure of the goodness of the questions being chosen, as

$$QoQ = \sum_{i,j} M_{rel\,ij}$$

This is the sum of all the elements of M_{rel}. The more different the meanings of the questions are from each other, the larger the sum becomes, so the larger QoQ is, the better the questions are selected.

III. EXPERIMENTS

Scene difference detection experiments were conducted in the environment of the 73B2 laboratory in Engineering Bldg. 2 at the University of Tokyo. As shown in Fig. 4, this facility is a laboratory for robots, but it is also equipped with a kitchen, dining table, and work desk, making it suitable for evaluating the proposed method in multiple scenes. We use
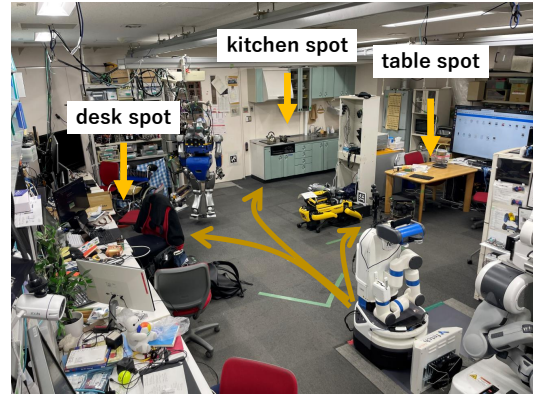

Fig. 4. Experimental environment and name of the spot. Fetch Mobile Manipulator navigates to each spot for scene difference detection and points the head camera to the spot to take pictures. Locations where the robot would be expected to perform life support tasks were selected: dining table (table spot), kitchen (kitchen spot), and office desk (desk spot).

TABLE I

NAVIGATION COORDINATES FOR EACH SPOT

Spot Name	x[m]	y[m]	yaw[rad]
table spot	4.036	7.344	1.753
kitchen spot	1.559	7.231	2.296
desk spot	4.319	6.108	-2.231

Fetch Mobile Manipulator[26] from Fetch Robotics to move to the table spot, kitchen spot, and desk spot where demand for life support behaviors is likely to be high in the figure, move the head camera to a specific position, and capture the scene. The coordinates for capturing these images are shown in Table I, and the angles of the torso and head are shown in Table II. In this experiment, the robot navigated to these spots and moved the height of the torso and the angle of the head camera to the target position and angle each time, so there were slight differences in position and posture each time. We performed this imaging four times a day for two months and collected 145 images at the table spot, 144 at the kitchen spot, and 141 at the desk spot. The images used below are representative images extracted from this data.

We prepared 67 questions to express the situation. The questions include

- what does this image describe?
- what is being done?
- what objects are seen?
- what is on it?
- how many people?

A. Scene Distance Measurement Using All Questions

The spot shown in Fig. 4 was imaged and the scene distance was quantified with the proposed method using all 67 questions. The reference images at the table spot, kitchen spot, and desk spot and the images taken at other times are shown in Fig. 5. The graphs of the scene distance at that time are shown in Fig. 6, Fig. 7 and Fig. 8, respectively.

In table spot A, the placement of the remote control and keyboard is slightly different but almost the same as in
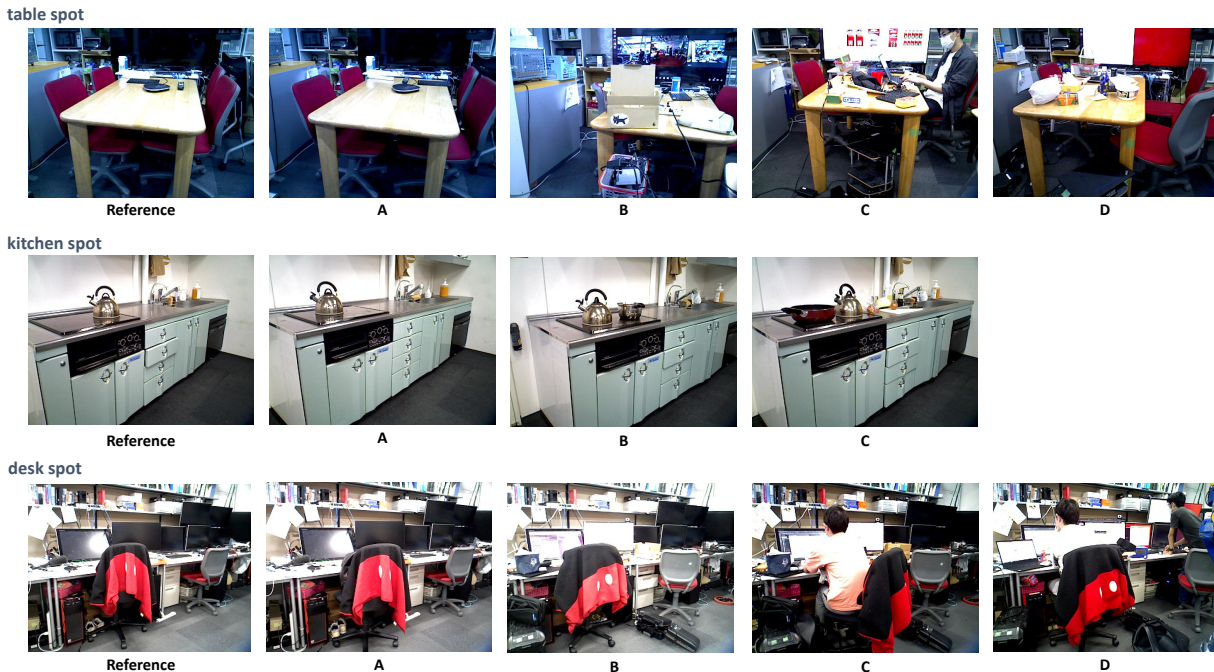
Fig. 5. Images of table spot, kitchen spot, and desk spot taken by Fetch Mobile Manipulator. Only one image of the reference state exists for each spot. Images A, B, C, and D taken at different times are compared with this reference image.

TABLE II

ROBOT POSTURE FOR TAKING A PICTURE AT EACH SPOT

Spot Name	torso[mm]	neck yaw[deg]	neck pitch[deg]
table spot	21.57	-2.170	19.57
kitchen spot	21.56	3.178	16.32
desk spot	21.58	-1.116	10.83

Normal; in B, a cardboard box and a bag are placed; in C, a person is working on a PC; and in D, food containers are scattered around. C had the highest scene distance.

In kitchen spot, A has the kettle on the induction stove shifted to the next position; B has a pot; C has food and cooking utensils on the sink and a frying pan on the induction stove. C had the highest scene distance.

In the desk spot, A is in the same state as Normal; B has a laptop PC with a monitor; C has a person working; D has two people working. C had the highest scene distance.

B. Evaluation of Variance Values for Two Different Weighting Methods

We evaluated the weighting method proposed in Sec.II-C for calculating the scene distance at the largest scene distance for each spot. 10 questions are randomly selected from 67 questions, and 10,000 sets are created. For each set of questions, QoQ of the questions is calculated. The horizontal axis represents the QoQ, and the vertical axis represents the variance of the scene distance for each separation bin, as plotted in Fig. 9, Fig. 10 and Fig. 11. Both the methods Eq.(1) and Eq.(2) in Sec.II-C are plotted. The variance of both methods decreases as QoQ increases and indicates that
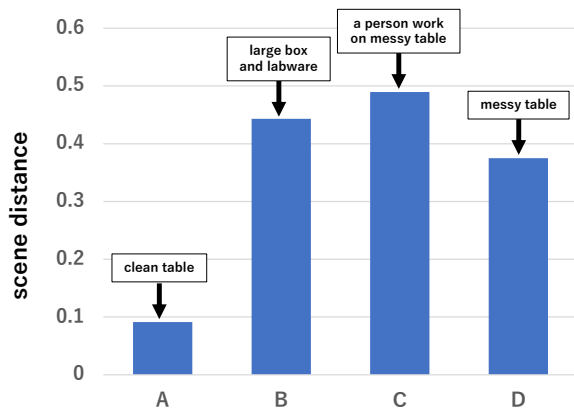

Fig. 6. Scene distance and feature of four types of scenes at table spot. The scene distance of image A, which is almost the same situation as the reference, is small, and the scene distances of the other images change with the degree of semantic difference in the situation.

the variance of the proposed method is smaller in regions where QoQ is small.

IV. DISCUSSIONS

A. Qualitative Evaluation of Scene Distance

We discuss the results of Fig. 6, Fig. 7 and Fig. 8. In the table spot A and the kitchen spot A, the scene distances were less than 0.2 for the movement of the remote controller and the movement of the kettle, and the scene distances were not high for slight movement of the objects. The scene distance of A in the desk spot is about 0.14, and the change in the scene distance is considered to be small in relation to the robot's navigation and head posture deviation. The scene
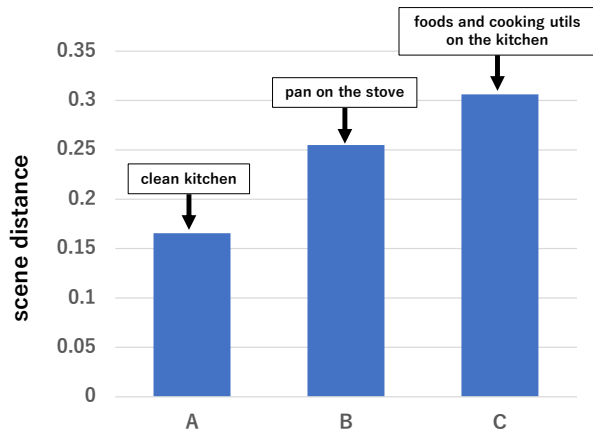
Fig. 7. Scene distance and feature of three types of scenes at kitchen spot. The scene distance of image A, which is almost the same situation as the reference, is small, and the scene distances of the other images change with the degree of semantic difference in the situation.

Fig. 9. Comparison of variance of scene distance in C in table spot for same weights and suggested methods. Our method has a smaller variance regardless of QoQ.
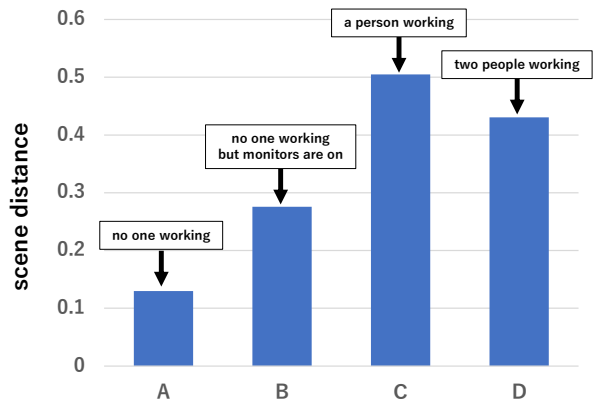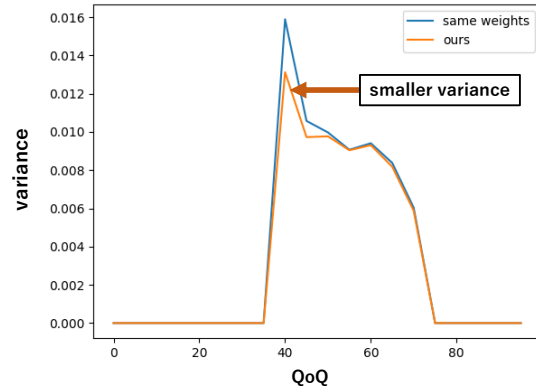

Fig. 8. Scene distance and feature of four types of scenes at desk spot. The scene distance of image A, which is almost the same situation as the reference, is small, and the scene distances of the other images change with the degree of semantic difference in the situation.
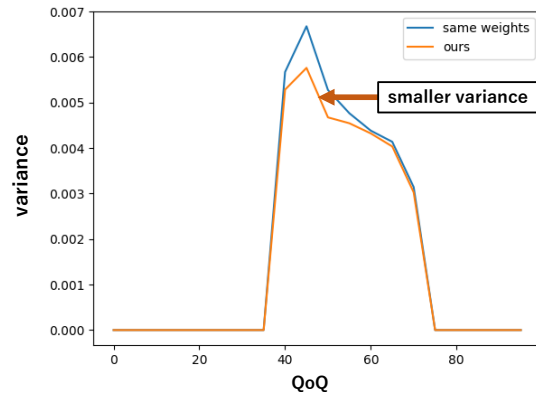

Fig. 10. Comparison of variance of scene distance in C in kitchen spot for same weights and suggested methods. Our method has a smaller variance regardless of QoQ.

distance of C in the kitchen spot is larger than that of B, indicating that the scene distance can describe the amount of things and the complexity of the situation. The scene distance of C in the desk spot is higher than that of B. It is clear that the scene distance increases when the complexity of the situation where the objects in the image remain unchanged and a person starts working is added. The above results suggest that, with appropriate thresholding, it is possible to detect semantic anomaly.

B. Variation in Scene Distance due to Question Bias

We discuss the results of Fig. 9, Fig. 10 and Fig. 11. Depending on how the questions are selected, the scene distance will have a wide range of values: for a small QoQ, the variance tends to be large, and for a larger QoQ, the variance tends to be small. This is because when QoQ is small, either all the questions are close in meaning, or there are a small number of questions with different meanings mixed in with the close meaning questions, and the output similarity is affected by the close meaning questions. The

proposed method achieves a smaller variance when QoQ is small than when the same weights are used. This is because the contribution of a small number of questions with different meanings is increased and the influence of many questions with the same meaning is decreased. The effect of this variance suppression varies depending on the scene conditions. This is thought to be due to the influence of the similarity of the output responses, and it is necessary to examine the contribution of these responses to each other.

C. Variation in Scene Distance due to Question Content

Scene Distance varies depending on the questions used. In the experiment, we prepared many variations of the questions to capture a wide range of semantic differences. The questions used can be customized for each task. For example, if the user wants to classify the state of a person, it is possible to input questions focusing on the presence or movement of the person. In the future, when performing clustering and anomaly detection in daily life environments, the proposed system will allow users to specify clusters they
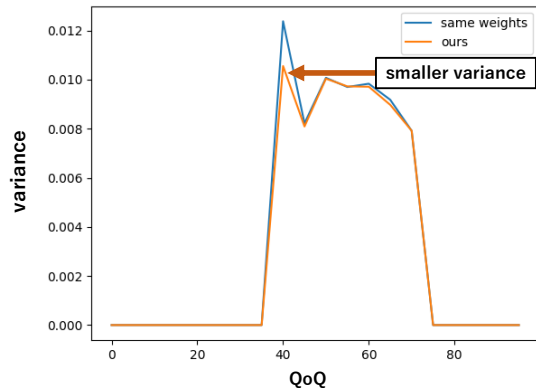
Fig. 11. Comparison of variance of scene distance in C in desk spot for same weights and suggested methods. Our method has a smaller variance regardless of QoQ.

wish to classify and anomalies they wish to detect using natural language questions.

V. CONCLUSIONS

In this study, we proposed a scene difference detection method using a pre-trained large-scale vision-language model for mobile robots to detect changes in our daily life environment. Multiple questions are prepared in advance, and reference images and current images are input to the model together with the question sentences, and the current difference is calculated by quantifying the difference between each answer sentence. In this process, the similarity of the meanings of the questions is taken into account, and the scene distance is calculated by weighting the value of the distance between the answers. Experimental results using these methods showed that, first, the semantic scene distance of the situation can be quantified and that the contribution of weak changes in the object, robot navigation, and posture errors to the scene distance is small. Second, the proposed weighting method was found to reduce the variance of the scene distance due to differences in the meaning of the selected questions. This method eliminates training costs in the local environment and may accelerate the spread of robots that can detect any semantic change in the living environment. The proposed method can be directly applied to a system in which a robot spontaneously decides whether or not to help a person by using questions about the person's actions or decides whether or not to clean up a room by using questions about the state of the objects in the room based on the normal state of its environment.

As a prospect of this research, it has the potential not only to quantify scene distances from the obtained answer sentences but also to explain changes using a spoken language with a large-scale language model such as GPT-3[27]. In addition, constructing a system that clusters environmental conditions based on the obtained scene distance and unsupervised learning may also allow the robot to perform accurate tasks based on the conditions.

REFERENCES

[1] J. Pearl. Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pp. 15–17, 1985.
[2] L. E Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, Vol. 37, No. 6, pp. 1554–1563, 1966.
[3] D. P Kingma, et al. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
[4] I. Goodfellow, et al. Generative adversarial networks. *Communications of the ACM*, Vol. 63, No. 11, pp. 139–144, 2020.
[5] A. Vaswani, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, Vol. 30, , 2017.
[6] J. Devlin, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
[7] C. Raffel, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arxiv. *arXiv preprint arXiv:1910.10683*, 2019.
[8] P. Anderson, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683. IEEE, 2018.
[9] M Ahn, et al. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.
[10] K. Kawaharazuka, et al. Vqa-based robotic state recognition optimized with genetic algorithm. In *International Conference on Robotics and Automation*, pp. 8306–8311. IEEE, 2023.
[11] K. Kawaharazuka, et al. Robotic applications of pre-trained vision-language models to various recognition behaviors. *arXiv preprint arXiv:2303.05674*, 2023.
[12] A. Das, et al. Neural modular control for embodied question answering. In *The 2nd Conference on Robot Learning*, pp. 53–62. PMLR, 2018.
[13] D. Shah, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *The 6th Conference on Robot Learning*, pp. 492–504. PMLR, 2023.
[14] C. Huang, et al. Visual language maps for robot navigation. In *International Conference on Robotics and Automation*, pp. 10608–10615. IEEE, 2023.
[15] S. Antol, et al. Vqa: Visual question answering. In *International Conference on Computer Vision*, pp. 2425–2433. IEEE, 2015.
[16] A. Radford, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.
[17] L. Li, et al. Grounded language-image pre-training. In *CVPR*, 2022.
[18] H. Zhang, et al. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022.
[19] P. Wang, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.
[20] W. Liu, et al. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pp. 21–37. Springer, 2016.
[21] K. Simonyan, et al. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[22] C. Wu, et al. An explainable and efficient deep learning framework for video anomaly detection. *Cluster Computing*, Vol. 25, No. 4, pp. 2715–2737, 2022.
[23] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 6568–6576, 2022.
[24] sentence-transformers/all-mpnet-base-v2. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`. [Online; accessed 17-September-2022].
[25] K. Song, et al. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 16857–16867, 2020.
[26] M. Wise, et al. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*, pp. 1–6, 2016.
[27] T. Brown, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.