2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

VQA-based Robotic State Recognition Optimized with Genetic Algorithm

Kento Kawaharazuka[1], Yoshiki Obinata[1], Naoaki Kanazawa[1], Kei Okada[1], and Masayuki Inaba[1]

***Abstract*— State recognition of objects and environment in robots has been conducted in various ways. In most cases, this is executed by processing point clouds, learning images with annotations, and using specialized sensors. In contrast, in this study, we propose a state recognition method that applies Visual Question Answering (VQA) in a Pre-Trained Vision-Language Model (PTVLM) trained from a large-scale dataset. By using VQA, it is possible to intuitively describe robotic state recognition in the spoken language. On the other hand, there are various possible ways to ask about the same event, and the performance of state recognition differs depending on the question. Therefore, in order to improve the performance of state recognition using VQA, we search for an appropriate combination of questions using a genetic algorithm. We show that our system can recognize not only the open/closed of a refrigerator door and the on/off of a display, but also the open/closed of a transparent door and the state of water, which have been difficult to recognize.**

I. INTRODUCTION

When a robot moves through space and performs a task, recognition of surrounding objects, tools, and environment is indispensable. For example, the robot needs to recognize the open/closed state of doors to rooms and elevators, the on/off of lights, etc. when moving around [1], [2]. In addition, it is necessary to recognize the open/closed state of refrigerator and cabinet doors, the on/off of displays, whether water is left running from the faucet, etc. when patrolling the area. So far, these state recognitions have been mainly based on human programming to extract features from raw images or point clouds [3], [4], human annotation of images and training with neural networks [5], or the attachment of appropriate sensors depending on the state to be recognized [6]. In other words, we have constructed a recognizer for each state that we want to recognize. On the other hand, with these methods, the number of recognizers and the data to be collected increases as the number of states to be recognized increases, which requires a large amount of time for system construction and makes resource management difficult. In addition, when humans walk around, they are constantly recognizing various states that cannot be easily programmed. These states include the open/closed of transparent doors, the state of water, etc. These states are not so easy to recognize for robots, and various studies have been conducted on the recognition alone [7].

Therefore, in this study, we propose a method of state recognition using the spoken language. We apply Visual Question Answering (VQA) [8] in a Pre-Trained Vision-Language Model (PTVLM) [9], [10] trained from a large-scale dataset. The state recognition is performed by asking a question about the current image and getting the answer in sentence form. By using the spoken language, it is possible to recognize even states that cannot be easily described by a program. On the other hand, preliminary experiments have shown that the recognition performance depends greatly on the question, especially on the form, articles, state expressions, and wording that are used. Since the states that are easy to recognize and the ones that are hard to recognize are different for each question, the states can only be partially recognized accurately with a single question. Therefore, in this study, we develop a method to search for appropriate question combinations using a genetic algorithm and construct a state recognizer with high performance. We show that our system can perform the state recognition of not only the open/closed of a refrigerator door and the on/off of a display, but also the open/closed of a transparent door and the state of water, which have been difficult to recognize. This study shows that an appropriate recognizer with high performance can be automatically generated by simply providing candidate questions in VQA, and that state

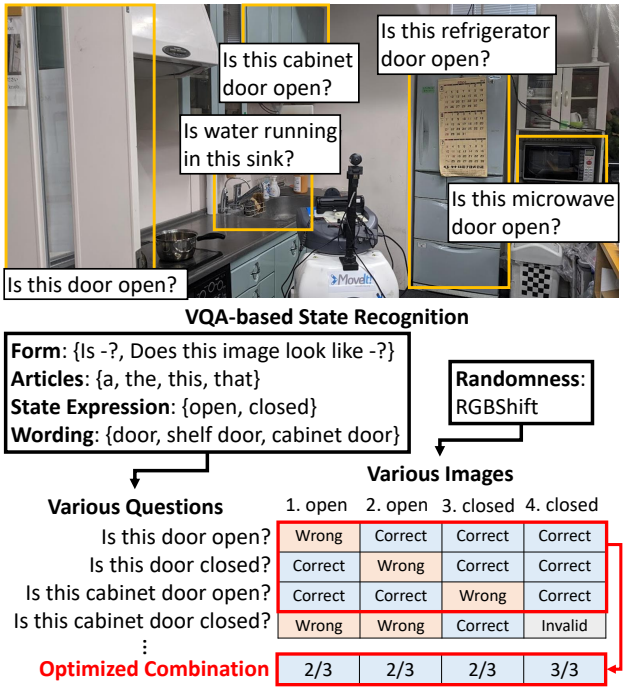

Fig. 1. The concept of this study. The combination of appropriate questions is optimized with genetic algorithm for binary state recognition using VQA.

[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. [kawaharazuka, obinata, kanazawa, k-okada, inaba]@jsk.t.u-tokyo.ac.jp

recognition can be performed simply and quickly using the spoken language and a single vision-language model, thus dramatically improving the recognition behavior of robots.

II. VQA-based Robotic State Recognition Optimized with Genetic Algorithm

A. VQA-based Robotic State Recognition Using Pre-Trained Vision-Language Models

First, we describe a state recognition method using VQA. VQA is a task to obtain an answer A by asking a question Q to an image V. In particular, we deal with binary state recognition, which can be easily used for action branching of robots. In other words, the task is to get a Yes or No answer by asking a question such as "Is -?" or "Does -?". On the other hand, such questions do not always result in Yes or No answers, but often in other phrases. Therefore, the obtained answers are classified into three categories: Correct, Wrong, and Invalid. Here, by augmenting the image or by changing the question, multiple answers can be obtained for the same image. The number of Correct, Wrong, and Invalid responses obtained is expressed as $N_{\{correct,wrong,invalid\}}$, and we define Correct Rate $R_{correct}$ as $N_{correct}/(N_{correct}+N_{wrong})$ and Invalid Rate $R_{invalid}$ as $N_{nvalid}/(N_{correct}+N_{wrong}+N_{invalid})$ ($N_{invalid}$ is not included in the denominator in $R_{correct}$). If $R_{correct}$ is greater than 0.5 for a certain image, then the image is correctly recognized.

Next, we describe how to augment images and generate various questions regarding the image. First, we use RGBShift as image augmentation, which adds a random value sampled from a uniform distribution in [-0.1, 0.1] to each RGB value. This allows us to obtain multiple responses to the same image and question and to take their average (in this study, we prepared six augmented images for each image-question pair). Of course, other augmentation such as Gaussian noise are also possible. Next, we generate various types of questions by combining four types of sentence variation: question form, article, state expression, and wording. The form refers to the way of asking a question, and in this study, the forms of "Is -?" such as "Is this door open?", and "Does -?" such as "Does this image look like this door is open?" are used. The article refers to the use of "a", "the", "this", "that" for the target object. The state expression refers to the use of antonyms, for example, "closed" for "open", "is not" for "is", and so on. The wording refers to, for example, the use of "display", "monitor", "TV", etc. for a monitor, and in this study, up to four types of wording are prepared for each state recognizer. Therefore, we will prepare a maximum of $2\times4\times2\times4$, that is, 64 questions. Here, there are at most $2^{64}-1$, or about 10^{19} combinations of questions. Of course, the number of question types is not limited to these, but can be increased in any number of ways, such as singular and plural, etc.

In this study, we use OFA [11] as a Pre-Trained Vision-Language Model that can be used for VQA. OFA is a model with high generalization ability by learning multiple tasks such as image captioning, visual grounding, and text-to-image generation at the same time.

B. Optimization of Appropriate Question Combination Using Genetic Algorithm

For the large number of questions mentioned in Section II-A, we search for the combination of questions that can achieve the best accuracy. In this study, a genetic algorithm is used for this purpose. The question combination is represented by a binary vector $\boldsymbol{s}$ ($\in\{0,1\}^{N_q}$, where N_q denotes the number of all questions) and is optimized. Questions with the value of 1 are used, while questions with the value of 0 are not used. As a dataset for optimization, we collect $N_{data}^{train}=20$ images. For example, in the case of the open/closed state recognition of doors, we prepare 10 images of the door in the open state and 10 images of the door in the closed state.

The evaluation value for a given $\boldsymbol{s}$ is the sum of the following three values (a)–(c). First, it is most important to correctly recognize as many states of the images in the dataset as possible. That is, we use the number $N_{correct}^{img}$ of images for which the Correct Rate $R_{correct}^{i}$ exceeds 0.5 for each image i ($1\le i\le N_{data}^{train}$) in the dataset. This $R_{correct}^{i}$ is the rate of correct answers for all questions where the value of $\boldsymbol{s}$ is 1, and for the image i augmented by RGBShift. In practice, we use the ratio of the number of correctly recognized images among all images (a) $R_{correct}^{img}=N_{correct}^{img}/N_{data}^{train}$ as the evaluation value. Second, if $R_{correct}^{img}$ is the same, the Correct Rate should be higher. Therefore, as the evaluation value, we use the expected value of the Correct Rate, i.e. the average value (b) $R_{correct}^{ave}$ of $R_{correct}^{i}$ for all images in the dataset. For (a) and (b), we also define $R'_{correct}=N_{correct}/(N_{correct}+N_{wrong}+N_{invalid})$ and use $R_{correct}^{img'}$ and $R_{correct}^{ave'}$ as the evaluation values for comparison. This is expected to have the effect of treating Invalid answers as equivalent to Wrong answers, so that questions with many Invalid answers are not selected. Third, if $R_{correct}^{img}$ is the same, having a fewer questions reduces the computational complexity. On the other hand, a larger number of questions is more likely to provide stable state recognition based on a larger number of answers. Therefore, we use the ratio of questions used for the combination among all questions (c) $R_q=N_q^s/N_q$ (where N_q^s is the number of 1 in $\boldsymbol{s}$) as the evaluation value. In this study, we compare the results of optimization using the following four evaluation values E,

$$E_+=\alpha R_{correct}^{img}+R_{correct}^{ave}+\beta R_q \tag{1}$$

$$E'_+=\alpha R_{correct}^{img'}+R_{correct}^{ave'}+\beta R_q \tag{2}$$

$$E_-=\alpha R_{correct}^{img}+R_{correct}^{ave}-\beta R_q \tag{3}$$

$$E'_-=\alpha R_{correct}^{img'}+R_{correct}^{ave'}-\beta R_q \tag{4}$$

where α and β are weight coefficients, and in this study, $\alpha=100$ and $\beta=0.1$ to prioritize the evaluation values in the order of (a), (b) and (c). Likewise, we denote the question combinations obtained by using these evaluation values as $\boldsymbol{s}_+,\boldsymbol{s}'_+,\boldsymbol{s}_-,\boldsymbol{s}'_-$.

The detailed setup of a genetic algorithm is shown below. The library DEAP [12] is used, and the function cxTwoPoint is used for crossover with a probability of 50%, and the function mutFlipBit is used for mutation with a probability of

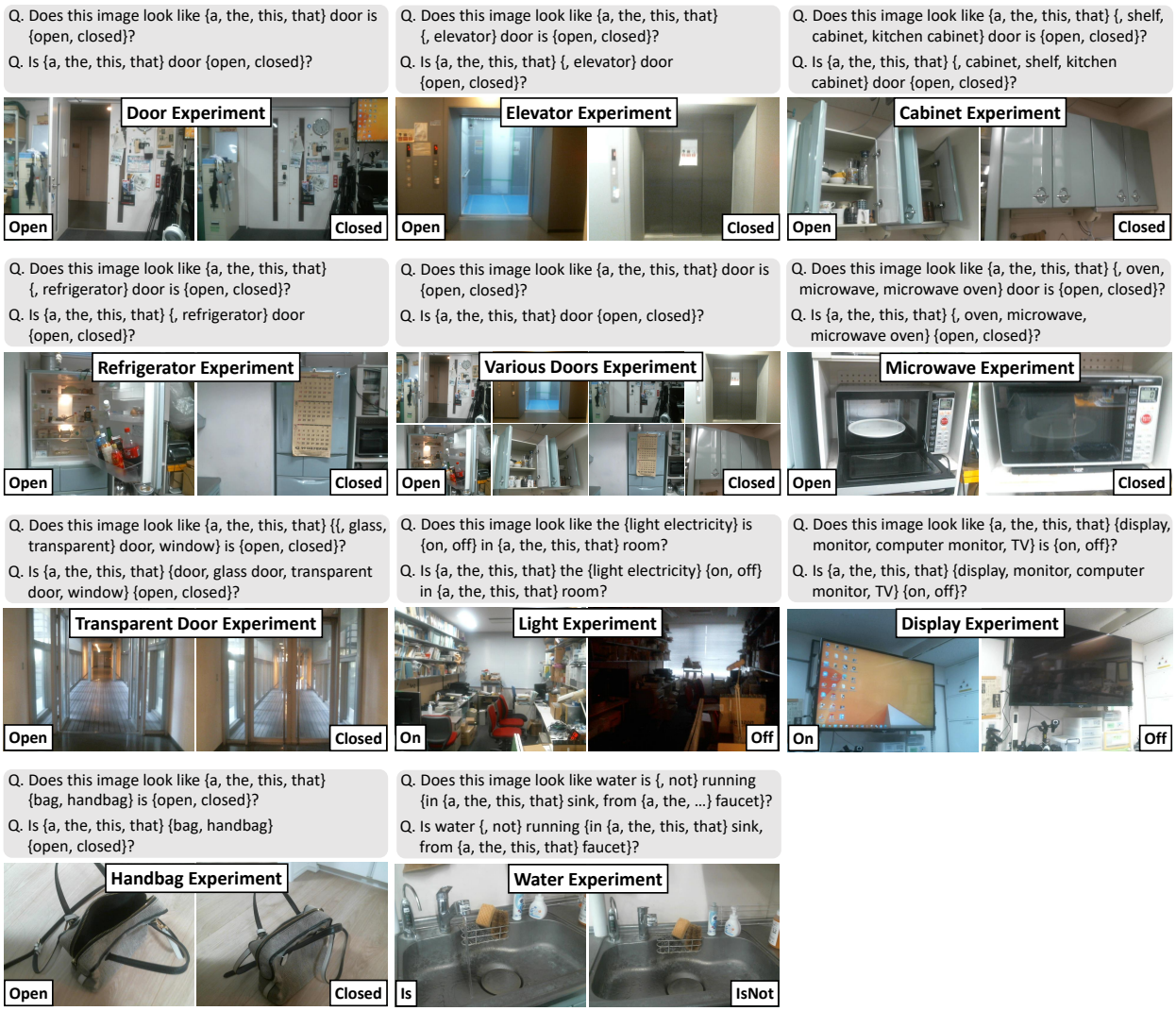

Fig. 2. The set of questions and representative images for the door, elevator, cabinet, refrigerator, various doors, microwave, transparent door, light, display, handbag, and water experiment.

20%. Individuals are selected by the function selTournament, and the tournament size is set to 5. The number of individuals and generations are set to 1000 and 200, respectively.

In addition, in this study, we prepare $\boldsymbol{s}_{\{does,is\}}$ where all the questions of the form "Does -?" or "Is -?" is used. We also denote $\boldsymbol{s}$ where all values of $\boldsymbol{s}$ are 1 as $\boldsymbol{s}_{all}$. After optimizing the question combination using the training dataset, we prepare a test dataset different from the training dataset and compare $N^{img}_{correct}$, $R^{ave}_{correct}$, $R_{invalid}$ and N^{s}_{q} ($R_{invalid}$ is the ratio of Invalid responses among all responses). The number of images in a test dataset is the same as that in a training dataset.

III. Experiments

The set of questions and representative images used in the experiments is shown in Fig. 2. The experiments are conducted to recognize whether a basic door, an elevator door, a cabinet door, a refrigerator door, various doors, a microwave door, and a transparent door are open or closed, whether a light or a display is on or off, whether a handbag is open or closed, and whether water is running or not.

A. Door Experiment

Table I shows the recognition results of the open/closed state of a standard door. Although only 15/20 of $\boldsymbol{s}_{does}$ were correctly recognized, $\boldsymbol{s}_{is}$, $\boldsymbol{s}_{all}$ and the optimization results show that all images were correctly recognized. While $\boldsymbol{s}_{+}$ and $\boldsymbol{s}'_{+}$ use a large number of questions, $\boldsymbol{s}_{-}$ and $\boldsymbol{s}'_{-}$ use only the best question. When comparing $\boldsymbol{s}_{+}$ and $\boldsymbol{s}'_{+}$, $\boldsymbol{s}'_{+}$ has less $R_{invalid}$ and uses fewer questions. This is considered to be because $\boldsymbol{s}'_{+}$ treats Invalid as the same as Wrong, so the choices are limited to those that are less likely to be Invalid.

TABLE I
The Result of Door Experiment

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$\boldsymbol{s}_{+}$	20 / 20	0.981	0.035	11 / 16
$\boldsymbol{s}'_{+}$	20 / 20	0.977	0.001	8 / 16
$\boldsymbol{s}_{-}, \boldsymbol{s}'_{-}$	20 / 20	1.000	0.000	1 / 16
$\boldsymbol{s}_{does}$	15 / 20	0.800	0.000	8 / 16
$\boldsymbol{s}_{is}$	20 / 20	0.904	0.056	8 / 16
$\boldsymbol{s}_{all}$	20 / 20	0.852	0.028	16 / 16

B. Elevator Experiment

Table II shows the recognition results of the open/closed state of an elevator door. Here, two kinds of words, "door" and "elevator door", are used as the wording. While $s_{\{does,is,all\}}$ makes recognition mistakes for some images, $s_{\{+,-\}}$ and $s'_{\{\{+,-\}}$ could recognize all images correctly. Moreover, the optimization results of $s_{\{+,-\}}$ and $s'_{\{+,-\}}$ are all consistent, suggesting that using only the best question is significantly more accurate than using the other questions. The best question was "Does this image look like the door is open?". In addition, when comparing s_{does} and s_{is}, s_{does} has lower $R^{ave}_{correct}$ and smaller $R_{invalid}$ than s_{is}, which is the same tendency as in Section III-A.

TABLE II
THE RESULT OF ELEVATOR EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$s_{\{+,-\}}, s'_{\{+,-\}}$	20 / 20	0.992	0.000	1 / 32
s_{does}	17 / 20	0.750	0.000	16 / 32
s_{is}	17 / 20	0.796	0.048	16 / 32
s_{all}	18 / 20	0.773	0.024	32 / 32

C. Cabinet Experiment

Table III shows the recognition results of the open/closed state of a cabinet door. Here, four words, "door", "cabinet door", "kitchen cabinet door", and "shelf door", are used as the wording. For all question combinations, all states are successfully recognized. The characteristics are the same as above experiments, s_+ has larger $R_{invalid}$ and N^s_q than s'_+. Also, s_{does} has lower $R^{ave}_{correct}$ and smaller $R_{invalid}$ than s_{is}.

TABLE III
THE RESULT OF CABINET EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
s_+	20 / 20	0.963	0.082	26 / 64
s'_+	20 / 20	0.971	0.000	17 / 64
s_-, s'_-	20 / 20	1.000	0.000	1 / 64
s_{does}	20 / 20	0.865	0.000	32 / 64
s_{is}	20 / 20	0.878	0.125	32 / 64
s_{all}	20 / 20	0.873	0.062	64 / 64

D. Refrigerator Experiment

Table IV shows the recognition results of the open/closed state of a refrigerator door. The optimized question combinations outperform $s_{\{does,is,all\}}$ in accuracy. Regarding all optimization results, $N^{img}_{correct}$ is 20/20 for the training dataset used in the optimization, while it is 19/20 in some cases for the test dataset. Other characteristics are similar to those in the above experiments.

E. Various Doors Experiment

Table V shows the recognition results of the open/closed states of all the four doors described so far. Only "door" is used as the wording, but the dataset given includes the standard door, the elevator door, the cabinet door, and the refrigerator door. 80 door images were used for the training and test datasets, respectively. The optimized question combination outperforms $s_{\{does,is,all\}}$ in accuracy. In other words, if the state to be recognized has similar characteristics, the state recognition is possible without changing the questions, and there is no need to construct each recognizer individually. Also, the number of images that can be recognized with only one question is 72/80, which is lower than 76/80 for optimization results, indicating that it is possible to achieve better recognition accuracy by using appropriate question combinations.

TABLE IV
THE RESULT OF REFRIGERATOR EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
s_+, s'_+	19 / 20	0.964	0.000	3 / 32
s_-	19 / 20	0.975	0.000	1 / 32
s'_-	20 / 20	0.983	0.000	1 / 32
s_{does}	15 / 20	0.727	0.000	16 / 32
s_{is}	17 / 20	0.737	0.025	16 / 32
s_{all}	17 / 20	0.732	0.013	32 / 32

TABLE V
THE RESULT OF VARIOUS DOORS EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
s_+	76 / 80	0.884	0.075	7 / 16
s_-	76 / 80	0.890	0.041	3 / 16
$s'_{\{+,-\}}$	76 / 80	0.890	0.004	4 / 16
s_{does}	64 / 80	0.789	0.000	8 / 16
s_{is}	73 / 80	0.857	0.092	8 / 16
s_{all}	72 / 80	0.823	0.046	16 / 16

F. Microwave Experiment

Table VI shows the recognition results of the open/closed state of a microwave door. Here, four words, "door", "microwave door", "oven door", and "microwave oven door", are used as the wording. The optimized $s_{\{+,-\}}$ in particular outperforms $s_{\{does,is,all\}}$ in accuracy. On the other hand, $N^{img}_{correct}$ for $s'_{\{+,-\}}$ is smaller than that for $s_{\{+,-\}}$. This is considered to be because $s'_{\{+,-\}}$ treats Invalid responses as Wrong, and thus questions with many Invalid responses but with accurate state recognition are lost as choices. The value of $R_{invalid}$ for $s_{\{+,-\}}$ is larger than that for $s_{\{does,is,all\}}$, indicating that $s_{\{+,-\}}$ actively uses questions that output Invalid responses. Note that the microwave door is not included in the various doors experiment because the simple wording of "door" makes it difficult to recognize the state of the microwave door.

TABLE VI
THE RESULT OF MICROWAVE EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
s_+	17 / 20	0.703	0.315	6 / 64
s_-	16 / 20	0.689	0.210	5 / 64
s'_+	12 / 20	0.546	0.008	47 / 64
s'_-	12 / 20	0.596	0.015	8 / 64
s_{does}	12 / 20	0.506	0.000	32 / 64
s_{is}	15 / 20	0.621	0.084	32 / 64
s_{all}	14 / 20	0.561	0.042	64 / 64

G. Transparent Door Experiment

Table VII shows the recognition results of the open/closed state of a transparent door. Here, four words, "door", "transparent door", "glass door", and "window", are used as the wording. The optimized question combination significantly outperforms $\boldsymbol{s}_{\{does,is,all\}}$ in accuracy. The accuracy of $\boldsymbol{s}_{\{does,is,all\}}$ is 10/20 for $N^{img}_{correct}$, and this is due to the fact that the door is transparent, which leads to the incorrect recognition that the door is always open. On the other hand, some questions can correctly recognize the state even if the door is transparent, such as when asking the two questions "Does this image look like a window is closed?" and "Is this window open?" used for $\boldsymbol{s}'_-$. The questions obtained from other optimization results also suggest that the wording of "window" is very important.

TABLE VII

THE RESULT OF TRANSPARENT DOOR EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^s_q
$\boldsymbol{s}_{\{+,-\}}, \boldsymbol{s}'_+$	15 / 20	0.692	0.000	5 / 64
$\boldsymbol{s}'_-$	16 / 20	0.729	0.000	2 / 64
$\boldsymbol{s}_{does}$	10 / 20	0.574	0.000	32 / 64
$\boldsymbol{s}_{is}$	10 / 20	0.465	0.116	32 / 64
$\boldsymbol{s}_{all}$	10 / 20	0.519	0.058	64 / 64

H. Light Experiment

Next, we will show that it is possible to recognize not only the open/closed state of the door as we have seen so far, but also various other states. Table VIII shows the recognition results of the on/off state of a room light. Here, "light" and "electricity" are used as the wording. The optimized question combination significantly outperforms $\boldsymbol{s}_{\{does,is,all\}}$ in accuracy. $\boldsymbol{s}_{\{+,-\}}$ outperforms $\boldsymbol{s}'_{\{+,-\}}$ in $N^{img}_{correct}$, although $R^{ave}_{correct}$ is smaller. This is considered to be because $\boldsymbol{s}'_{\{+,-\}}$ treats Invalid responses as Wrong, and thus questions with many Invalid responses but with accurate state recognition are lost as choices. Note that the word used in all the optimized question combinations is "electricity". Other characteristics are the same as in the above experiments.

TABLE VIII

THE RESULT OF LIGHT EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^s_q
$\boldsymbol{s}_{\{+,-\}}$	20 / 20	0.799	0.094	3 / 32
$\boldsymbol{s}'_{\{+,-\}}$	17 / 20	0.856	0.000	3 / 32
$\boldsymbol{s}_{does}$	10 / 20	0.598	0.000	16 / 32
$\boldsymbol{s}_{is}$	10 / 20	0.573	0.069	16 / 32
$\boldsymbol{s}_{all}$	11 / 20	0.586	0.034	32 / 32

I. Display Experiment

Table IX shows the recognition results of the on/off state of a display. Here, four words, "display", "monitor", "computer monitor", and "TV", are used as the wording. The optimized question combination significantly outperforms $\boldsymbol{s}_{\{does,is,all\}}$ in accuracy. $\boldsymbol{s}_-$ and $\boldsymbol{s}'_-$ are identical, and only one question "Does this image look like this display is off?" is used. On the other hand, the combination of many questions like $\boldsymbol{s}_+$ or $\boldsymbol{s}'_+$ is somewhat more accurate than $\boldsymbol{s}_-$ or $\boldsymbol{s}'_-$ with $N^{img}_{correct}$ of 20/20.

TABLE IX

THE RESULT OF DISPLAY EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^s_q
$\boldsymbol{s}_+$	20 / 20	0.950	0.557	20 / 64
$\boldsymbol{s}'_+$	20 / 20	0.957	0.000	6 / 64
$\boldsymbol{s}_-, \boldsymbol{s}'_-$	19 / 20	0.925	0.000	1 / 64
$\boldsymbol{s}_{does}$	12 / 20	0.716	0.000	32 / 64
$\boldsymbol{s}_{is}$	14 / 20	0.693	0.491	32 / 64
$\boldsymbol{s}_{all}$	14 / 20	0.705	0.245	64 / 64

J. Handbag Experiment

Table X shows the recognition results of the open/closed state of a handbag. Here, "bag" and "handbag" are used as the wording. Compared to $\boldsymbol{s}_{\{does,is,all\}}$, the optimized question combination is much more accurate. Other characteristics are the same as in previous experiments.

TABLE X

THE RESULT OF HANDBAG EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^s_q
$\boldsymbol{s}_+$	18 / 20	0.815	0.497	8 / 32
$\boldsymbol{s}_-$	18 / 20	0.815	0.195	5 / 32
$\boldsymbol{s}'_{\{+,-\}}$	18 / 20	0.750	0.000	3 / 32
$\boldsymbol{s}_{does}$	15 / 20	0.670	0.002	16 / 32
$\boldsymbol{s}_{is}$	11 / 20	0.530	0.461	16 / 32
$\boldsymbol{s}_{all}$	14 / 20	0.622	0.232	32 / 32

K. Water Experiment

Table XI shows the recognition results of whether water is running or not. Here, after "water is running", two modifications are added: "in a, the, this, that sink" or "from a, the, this, that faucet". The optimized question combination significantly outperforms $\boldsymbol{s}_{\{does,is,all\}}$ in accuracy. Similarly to Section III-H, $\boldsymbol{s}_{\{+,-\}}$ outperforms $\boldsymbol{s}'_{\{+,-\}}$ in $N^{img}_{correct}$, indicating the importance of questions with many Invalid responses. Note that $N^{img}_{correct}$ is 19/20 for all the data used in the optimization, indicating that the accuracy is somewhat lower for the test dataset. Also, all the questions used in the optimization results were the ones with the modification "from a, the, this, that faucet".

TABLE XI

THE RESULT OF WATER EXPERIMENT

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^s_q
$\boldsymbol{s}_+$	17 / 20	0.797	0.179	6 / 32
$\boldsymbol{s}_-$	17 / 20	0.796	0.033	5 / 32
$\boldsymbol{s}'_{\{+,-\}}$	15 / 20	0.735	0.033	5 / 32
$\boldsymbol{s}_{does}$	12 / 20	0.580	0.000	16 / 32
$\boldsymbol{s}_{is}$	13 / 20	0.667	0.114	16 / 32
$\boldsymbol{s}_{all}$	12 / 20	0.622	0.057	32 / 32

IV. Discussion

The following is a summary of the important points from the obtained experimental results.

- The accuracy of using many questions such as $\boldsymbol{s}_{\{does,is,all\}}$ at once is not high, and it is necessary to use only appropriate questions.
- The easier the state recognition is, the higher the accuracy is even if there is only one question.
- In many cases, one question alone is not accurate enough, and it is possible to construct a recognizer with better accuracy by appropriately combining multiple questions.
- Invalid can be treated the same as Wrong to reduce the number of Invalid answers, but this does not necessarily lead to better accuracy. Rather, some important questions are often left out of the choices and the accuracy is often decreased.
- In some state recognition, having a large number of questions can achieve higher accuracy than having fewer questions, even for a dataset different from the one used in the optimization.

Taking all of the above into consideration, we should basically use the questions obtained by $\boldsymbol{s}_{\{+,-\}}$. $\boldsymbol{s}_{+}$ is recommended for accurate recognition even if it takes longer, while $\boldsymbol{s}_{-}$ is recommended for accurate recognition in a shorter time span.

Some other notable properties are:

- There are more Invalid responses for questions in $\boldsymbol{s}_{is}$ than for questions in $\boldsymbol{s}_{does}$.
- The recognition accuracy is better for questions in $\boldsymbol{s}_{is}$ than for questions in $\boldsymbol{s}_{does}$.
- All the questions obtained by optimization use the same wording.

These characteristics are considered to vary depending on the model used for VQA, and are not universally applicable knowledge. However, by using this study, it is possible to obtain an appropriate combination of questions without worrying about the characteristics that vary from model to model, by simply providing some dataset.

The important contribution of this study is that the recognizer can be constructed intuitively by the spoken language, without the need for retraining of the network or individual programming. In addition, only a single pre-trained model is required, and the resource is very easy to manage. On the other hand, since the choice of the question sentences is heuristic, we have added an optimization mechanism to solve this problem. This makes it possible to construct a state recognizer very easily, in which the only part that humans have to consider is to increase the variety of questions. In the future, it is necessary to incorporate this study into actual tasks and to identify problems that may arise in the real world. In particular, we would like to examine state recognition in more cluttered situations, and recognition performance when various other objects, such as humans and robot bodies, appear in the image. In addition, we believe that anomaly detection and object detection will become possible in the same way as in this study, and we would like to deal with a wider range of state recognition in the near future.

V. Conclusion

In this study, we proposed a method of binary state recognition for robots using Visual Question Answering (VQA) in a Pre-Trained Vision-Language Model (PTVLM). In VQA, for multiple augmented images, we integrate the answers from various questions with different forms, articles, state expressions, and wording. Since there are states that are easy to recognize and states that are difficult to recognition for each question, by using an appropriate combination of questions, it is possible to construct a recognizer that can accurately recognize any state. By optimizing this combination with a genetic algorithm, it is now possible to recognize the states of transparent doors, water, etc. that have been difficult to recognize so far. We believe that this method will revolutionize the recognition strategies of robots, since it does not require any retraining of the network or programming, and a recognizer of complex states can be easily constructed by simply providing a set of questions to a single model.

References

[1] M. Saito, H. Chen, K. Okada, M. Inaba, L. Kunze, and M. Beetz, "Semantic Object Search in Large-scale Indoor Environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, Workshop on Active Semantic Perception and Object Search in the Real World*, 2011.

[2] K. Okada, M. Kojima, Y. Sagawa, T. Ichino, K. Sato, and M. Inaba, "Vision based behavior verification system of humanoid robot for daily environment tasks," in *Proceedings of the 2006 IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 7–12.

[3] R. T. Chin and C. R. Dyer, "Model-Based Recognition in Robot Vision," *ACM Computing Surveys*, vol. 18, no. 1, pp. 67–108, 1986.

[4] S. M. Z. Borgsen, M. Schöpfer, L. Ziegler, and S. Wachsmuth, "Automated Door Detection with a 3D-Sensor," in *Proceedings of the 2014 Canadian Conference on Computer and Robot Vision*, 2014, pp. 276–282.

[5] X. Li, M. Tian, S. Kong, L. Wu, and J. Yu, "A modified YOLOv3 detection method for vision-based water surface garbage capture robot," *International Journal of Advanced Robotic Systems*, vol. 17, no. 3, 2020.

[6] T. Takahata, "Coaxiality Evaluation of Coaxial Imaging System with Concentric SiliconGlass Hybrid Lens for Thermal and Color Imaging," *Sensors*, vol. 20, 2020.

[7] M. P. Khaing and M. Masayuki, "Transparent Object Detection Using Convolutional Neural Network," in *Proceedings of the International Conference on Big Data Analysis and Deep Learning Applications*, 2018.

[8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *Proceedings of the 2015 IEEE/CVF International Conference on Computer Vision*, 2015, pp. 2425–2433.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," arXiv preprint arXiv:2103.00020, 2021.

[10] F. Li, H. Zhang, Y. Zhang, S. Liu, J. Guo, L. M. Ni, P. Zhang, and L. Zhang, "Vision-Language Intelligence: Tasks, Representation Learning, and Large Models," arXiv preprint arXiv:2203.01922, 2022.

[11] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework," arXiv preprint arXiv:2202.03052, 2022.

[12] F. Fortin, F. D. Rainville, M. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary Algorithms Made Easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, 2012.