A Service Robot with Interactive Vision
- Object Recognition Using Dialog with User -

Masao Takizawa, Yasushi Makihara, Nobutaka Shimada, Jun Miura and Yoshiaki Shirai

Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
E-mail: shimada@eng.osaka-u.ac.jp

Abstract

This paper proposes a service robot which can bring user-specified objects from a refrigerator to a user. Speech interface is used not only for specifying objects but also for helping the robot vision. The focus is placed on how vision tasks are realized by interaction and how to interpret ambiguous outputs of a commercial speech recognition system. Experimental results with real environments are shown.

Key Words: Service robot, Object recognition, Interactive Vision, Speech recognition

1 Introduction

There is a growing necessity of welfare robots in this aging society. One of the important tasks for such a robot is to bring a user-specified object[1][2]. This paper deals with tasks of bringing cans, bottles, and PET bottles from a refrigerator.

The necessary functions for the task are understanding users commands, recognizing user-specified objects by vision, and manipulating objects, and so on. This paper focuses on how vision tasks are realized by interaction and how to interpret ambiguous outputs of a commercial speech recognition system.

It is not easy to automatically recognize objects in the presence of partial occlusion of objects or illumination change. In such difficult conditions, a service robot may be helped by the user who is more intelligent than the robot.

Among early works of connecting visual information and verbal information, some of them tried to generate scene explanations based on visual recognition results[9][10]. Watanabe et al.[11] proposed a system to recognize flowers and fruits in a botanical encyclopedia using explanation texts. There are also approaches to interactive vision: some of them search regions where image features are most consistent with user's advice [12][13]. However, no attempt was made to recover recognition errors by verbal interactions. Takahashi et al.[5] proposed verbal and gestural interaction to directly point the object position. They uses verbal information only for choosing an object from multiple candidates.

The service robot employs a speech recognition system "ViaVoice" [6] which is trained for long sentences found in newspapers. Usually the service robot has to understand short sentences which includes proper nouns such as names of beverages, where user's utterances are not correctly recognized. In such cases, the system has to estimate the meanings of unknown recognized words.

(a) Service robot (b) Interactive devices

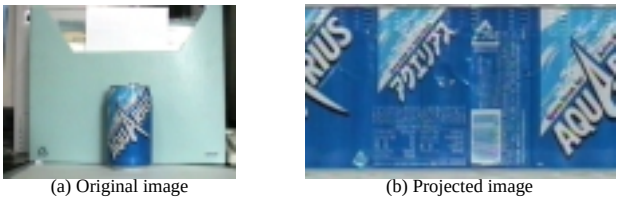
Figure 1: Our service robot system

Some attempts are made for dealing with unknown words. Attributes of unknown words such as the name of place were estimated by using grammars[3], or by using the discourse[4]. Damnati et al.[7] estimated a class (a part of speech) of unknown words. Nagata[8] also estimated a class of unknown words in a text (not in speech) considering the context.

In this paper, we recover the following utterances: the registered words which are recognized by mistake or synonyms of the registered words. We call them "unknown words". The system estimates which registered word they correspond to, considering the state, the context (i.e. the preceding and following words) and the pronunciation similarity between the unknown words and the registered words. Assuming both the case of one unknown word and that of two successive unknown words, the system calculates the estimation scores both as the registered words which are recognized by mistake and as synonyms of the registered words, and adopts the estimation with the highest score.

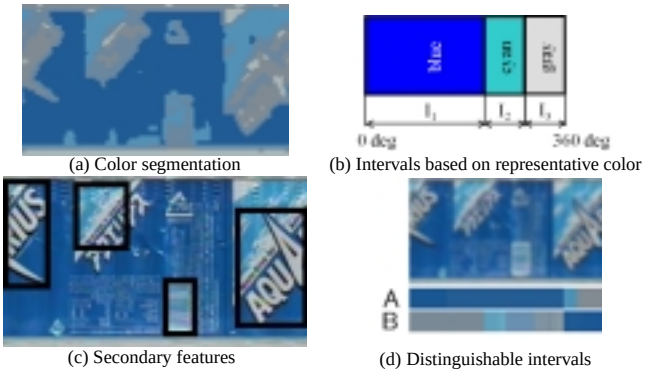
2 Object recognition

In order to make a model of an object, the system observes it from many directions and extracts its features at each direction. Features consist of the size of an object, representative colors, and secondary features such as the color, the position,



(a) Original image (b) Projected image

Figure 2: Construction of a projected image



(a) Color segmentation (b) Intervals based on representative color

(c) Secondary features (d) Distinguishable intervals

Figure 3: Extraction of features

and the size of uniform color region. We propose a strategy to register minimum number of features hierarchically for discriminating the object[16]. If there are no objects with the same representative colors, the system registers only the representative colors. Every time an object with the same feature is registered, the system adds distinguishable features.

For extracting features from images of many directions, we use one image made by projecting the surface texture of the object from the center axis to the virtual cylinder and by developing the cylinder into a rectangle plane. This is called "projected image" (Figure 2(b)).

2.1 Object model

Instead of registering feature values for every direction, we determine the intervals of the similar feature values. For example, the projected image in Figure 2(b) is divided into the interval I_1 of blue, I_2 of cyan and I_3 of gray. An interval is further divided into multiple intervals according to secondary features (Figure 3(c)) if distinct secondary features are extracted in the interval. We continue this process until all the intervals are distinguishable from other objects. For example, Supposing object 1 in Table 1(a) is already registered, object 2 with the same feature (blue, [white]) as object 1 is added. Then secondary features of object 1 and object 2 are added as shown in Table 1(b) to distinguish object 1 from object 2. The final interval segmentation for the object in Figure 2 is shown in Figure 3(d) where A and B respectively show the intervals of representative color and secondary features.

Table 1: Registration of object 2 ([interval index], representative color, [secondary feature])

(a) Before registering object B with (blue, [white])

1	([1], yellow), ([2], green), ([3], blue, [white])

(b) After registering

1	([1], yellow), ([2], green), ([3], blue, [white-top-large])
2	([1], blue, [white-middle]), ([2], blue, [white-top-small])

2.2 Automatic recognition

We consider the following two cases for object recognition: (1) recognize a specified object and (2) recognize all objects in the refrigerator. In both cases, object recognition proceeds in the following steps:

1. Extract candidate regions for the object.
2. Recognize object types (can, bottle, PET bottle, or unknown) and shapes for each region.
3. If the region is recognized as a known object type, classify the region into one of the model objects.

The recognition result is presented to a user by speech and using a display. Object recognition in the refrigerator often fails due to change of light condition, complicated textures of the objects and occlusion. In such cases, the system first presents already recognized result (it may be wrong) to the user. Then considering what information is required to recognize the object correctly, the system makes questions to promote the user's advice to recover the recognition failures.

2.3 Recovery from recognition failure using dialog

Results of the automatic recognition are classified into the following only four cases. For each case, the system makes utterances to obtain appropriate advices for failure recovery from the user.

1. One object is found.
   If the result is incorrect, (a) the user specifies the location of the target object, and if necessary, (b) the user can make the system learn to avoid the same mistake.
2. More than two objects are found.
   The system says, "I have found n objects", and selects one object among n objects and says, "I will bring this. Is it all right?" The user's answer is divided into the following four cases:
   (a) If it is correct, the user says, "Yes."
   (b) If the user wants to select another object, the user specifies it by the location or the type of the object. Because the system has already recognized the type and the location of each found object, it can determine the target object.

(c) If the target object is not found, the user specifies the location of the target object (see 1(a)).

(d) If some of the found objects are not correct, the user can make the system learn.

After the system begins to get the target object, the system asks whether unselected objects are the target objects. If the user tells another name, the system learns it to avoid the same mistake.

3. The target object is not found although the candidate region is found.
The system says, "I have no confidence. Is this all right?", in order to ask the user to look at the candidate region carefully. This case is further divided into the following three cases.

(a) If it is correct, the user says, "Yes." In order to avoid the same mistake, the system learns it.

(b) If the target object overlaps with another object of the same color, they may be regarded as the same object. In this case the user helps the system in one of the following ways:

(1) The user selects the front object. The system tries to recognize the target object considering the overlap.

(2) The user selects the back object. Because the system has to recognize the front object in order to recognize the back object, it asks what the front object is. Then the system recognizes the front object and tries to recognize the back object carefully.

(3) The user points out that two objects overlap. Then the system asks which object to bring. Depending on the user's response, the system acts just as (1) or (2).

(c) If the candidate region is something else (the background or an unregistered object), the user specifies the location of the target object (this case is handled in the same way as 1(a)) or gives the name of the unregistered object (in this case the system needs to registers it, but it is not dealt with in this paper).

4. Neither the target object nor candidate regions are found.
The system says, "I have not found it. Where is it?" While waiting for the user's response, the system tries to recognize all objects in the refrigerator. If the user specifies the location before the system finishes recognizing all objects, this case is handled in the same way as 1(a). Otherwise, the system shows the found objects so that the user may specify easily. The user can specify the location relative to the found object.

An example of hidden object recognition is shown in Figure 4, where a front object is recognized (a), the visible part of the target object is extracted (b), the target object is verified by the representative color and edges (left rectangle of (c)) and finally recognized (d).

(a) Recognition of the front object

(b) Visible part of the target

(c) Verification by the representative color and edges

(d) Recognition result

Figure 4: Recognition of the hidden object

3 Speech recognition

3.1 Conventional recognition engines

"ViaVoice" supports two recognition engines: a user-defined grammar engine and a dictation engine. The former can recognize only sentences matching to pre-registered context-free grammar. It requires a set of words to be recognized for the task. They are categorized and registered the words in advance (Table 2). The engine also requires acceptable grammars written by context-free grammar which defines the order of the word categories in a sentence (3). Because it accepts only several commands and object names, it has comparatively high recognition rate. However, it cannot recognize unregistered words and sentences.

The dictation engine can accept freely spoken sentences. It can transform the continuous wave pattern of an utterance consisting of several words into a series of recognized words. The engine retrieves the best word considering phonetic likelihoods and grammars. It is often mis-partitioned or incorrect for ambiguously pronounced words or noisy audio input. Table 4 is a recognition result for "*shiroi kan no ushiro* (behind the white can)" by the dictation engine. *shiroi* is ambiguously pronounced and mis-recognized.

3.2 Our approach

Our speech recognition system is shown in Figure 5. A user's utterance is first sent to the user-defined grammar engine. If the utterance is matched to one of the grammars, it is recognized. Otherwise, it is sent to the dictation engine. It gives a recognized result and several alternative candidates. If the registered words are found in the candidates, they are adopted

Table 2: Categories and examples of registered words

Category	Registered words (Japanese)
[brand name]	*nohohon-cha*(green tea), *koora*(coke), etc.
[shape]	*kan*(can), *bin*(bottle) etc.
[relative pos..	*migi*(right), *ushiro*(behind), *ue*(up) etc.
[absolute pos.]	*ue-no-tana*(upper shelf), *poketto*(door pocket)
[pronoun]	*kore*(this), *sore*(that), *sono*(its)
[kind of drink]	*juusu*(juice), *koohii*(coffee), etc.
[intransitive verb]	*arimasu*(exist), *haitte-imasu*(be in)
[transitive verb]	*totte*(take), *dokete*(move), etc.
[adjective]	*ookii*(big), *hosoi*(thin), etc.
[adverb]	*motto*(more), *ichiban*(most)
[interrogative]	*nani*(what), *ikutsu*(how many)
[numeral]	*hutatsu*(two), *san-banme*(third)
[yes/no]	*hai*(yes), *iie*(no)
[color]	*aka*(red), *kiiro*(yellow), etc.

Table 3: Examples of registered grammars

grammar	example sentence
[Yes/No]	*hai*(yes), *iie*(no)
[brand name][transitive verb]	*koora*(coke) *totte*(take)
[color][shape]	*akai*(red) *kan*(can) *aoi*(blue) *bin*(bottle)
[brand name]*no*[relative pos.]	*nohohon-cha no hidari* (left of green tea)
[absolute pos.]*kara*[numeral]	*migi kara san-banme* (the third one from the left)

as the estimates. If not found, the dictation result is tried to be matched to the registered words by considering the phonetic likelihood, the state of the task in operation and the context of the utterance.

3.3 Detection of unknown word

Our aim is to recover the following utterances: the registered words recognized incorrectly and synonyms of the registered words. They are called "unknown words" in this paper. Among the words obtained by the dictation engine, those which are not matched to any of the registered words are regarded as parts of unknown words.

Table.5 shows a detection result of unknown words for

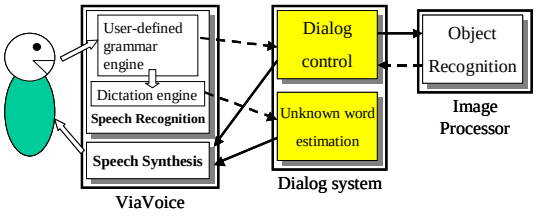


Figure 5: Speech recognition process

Table 4: Recognition result of dictation engine

Actual utterance	*shiroi* (white)	*kan* (can)	*no* postpos.	*ushiro* (behind)
Recognized result	*tsuyoi* (strong)	*kan* (can)	*no* postpos.	*ushiro* (behind)
Alternative 1	*seejika* (politician)	*kan* (can)	*wuo* (fish)	*shiryoo* (data)
Alternative 2	*seoi* (on back)	*kan* (can)	*na* aux.verb	*shunoo* (executives)
Alternative 3	*sugoi* (terrible)	*kan* (can)	*na* postpos.	*shirou* (proper noun)

Table 5: Unknown word detection for "*nohohon-cha no hidari*"

recognized word	alternatives
go	*no, doo, go, do, ko*
hon	*kon, oo-gon, hoo-mon, obon*
cha	*chi, cho, jo, ja, chan*
no	*na, nai, no, ga*
hideri	*higeki, hiraki, hidari, higaeri*

"*nohohon-cha no hidari* (left of green tea)". The underlined words are registered word and bold-type words are the parts of unknown words. It shows "nohohon-cha" is over-segmented and mis-interpreted as "go-hon-cha". Table.6 is another example of unknown word detection for "*koora totte* (take the coke)". "*koora*" is over-segmented mis-interpreted as "*oo-da* (it's the king)" and "*totte*" is mis-interpreted as "*to*(postposition)". Because a word sandwiched unknown words is highly possible to be a mis-interpreted word, such a word is regarded as also a part of unknown word.

Since an actual word may be over-segmented and/or successive words may be incorrectly connected into one unknown word, the system supposes that the detected unknown word part corresponds to one or two actual words. Then the system estimates the meanings of the unknown words by finding the plausible interpretation among the registered words. If not found, the system gives up the estimation and asks the user for more information.

3.4 Estimation of unknown word

The problem of the estimation of unknown words is formulated as finding the registered word(s) W (W_1 and W_2) with the

Table 6: unknown word detection for "*koora totte*"

recognized word	alternatives
oo	*wo,oo,o*
da	*ga,ha,no,wo,ro,he,me,ra*
to	*ten,too,ta*

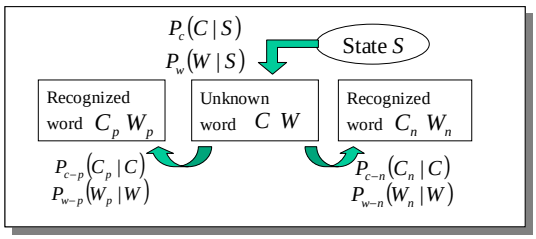


Figure 6: Estimation of category C and word W

maximum probability, given state S, context γ and pronunciation R recognized by the dictation engine (see Figure 6). In this paper, the following conditions are used as S.

1. waiting the user's first utterance
2. questioning upon the user's previous utterance: "What shall I bring?" or "What color is the object?"
3. questioning upon the object recognition result: "Not found. Where is it?" or "Found. May I bring it?"

The preceding and following words of the unknown word are used as γ. R is represented as a series of Japanese vowels (a, i, u, e, o) and consonants (k, s, t, n, h, m, y, r, w and etc.).

An unknown word is supposed to be either mis-recognized word or a synonym of a registered word. The most plausible word and its probability are estimated respectively in each of the two cases. Then the word having the larger probability is adopted as the estimate of the unknown word. When both probabilities are under a threshold, such a part of the utterance is ignored.

3.4.1 Word estimation as a synonym

Because the pronunciation of synonyms are not necessarily similar to that of a registered word, the probability that the unknown word is a synonym is calculated based on only S and γ without R. In order to reduce the amount of computation, we first estimate the probability $P(C|S,\gamma)$ of each word category (see Table 2). Then only for the category with high probability, we estimate the word. We find W having the best of the following score:

$$\begin{aligned} Score_s(W;S,\gamma) &= P(W|S,\gamma) \\ &= \sum_C \{P(C|S,\gamma)P(W|S,\gamma,C)\} \end{aligned} \quad (1)$$

where C is the category with the probability larger than a threshold. $P(C|S,\gamma)$ is computed as follows:

$$P(C|S,\gamma) \simeq P(C|C_p,C_n,S) = \frac{P(C_p,C,C_n|S)}{\sum_C P(C_p,C,C_n|S)} \quad (2)$$

$$P(C_p,C,C_n|S) \simeq P(C|S)P_{c-p}(C_p|C)P_{c-n}(C_n|C) \quad (3)$$

where

$P_{c-p}(C_p|C)$: $Prob(C_p$ is the preceding category of C under the condition of utterance of $C)$

$P_{c-n}(C_n|C)$: $Prob(C_n$ is the following category of C under the condition of utterance of $C)$.

In order to compute the above formula, $P_{c-s}(C_i|S)$, $P_{c-p}(C_i|C_j)$ and $P_{c-n}(C_i|C_j)$ should be given as a language model. These model parameters are computed based on histograms $N(C_i,C_j)$ and $N(S,C)$. In our experiment, these histograms were automatically made by synthesizing all the possible sentences from the registered word and grammars for each state.

For each word in the category C having $P(C|S,\gamma)$ larger than a threshold, $P(W|S,\gamma,C)$ is computed as follows:

$$\begin{aligned} P(W|S,\gamma,C) &\simeq P(W|W_p,W_n,S,C) \\ &= \frac{P(W_p,W,W_n|S,C)}{\sum_W P(W_p,W,W_n|S,C)} \end{aligned} \quad (4)$$

$$P(W_p,W,W_n|S,C) \simeq \quad (5)$$

$$P(W|S,C)P_{w-p}(W_p|W)P_{w-n}(W_n|W) \quad (6)$$

where

$P_{w-p}(W_p|W)$: $Prob(W_p$ is uttered just before W under the condition of utterance of $W)$

$P_{w-n}(W_n|W)$: $Prob(W_p$ is uttered just after W under the condition of utterance of $W)$.

$P_{w-s}(W_i|S,C_i)$, $P_{w-p}(W_i|W_j)$ and $P_{w-n}(W_i|W_j)$ is also given as the parameters of the language model. Finally, the word having the best $Score_s(W;S,\gamma)$ is determined as the estimated synonym.

3.4.2 Word estimation as a mis-recognized word

The probability that the unknown word is a mis-recognized word of a registered word depends on the pronunciation R recognized by the dictation engine, in addition to S and γ. We find W having the best of the following score:

$$\begin{aligned} Score_m(W;S,\gamma,R) &= P(W|S,\gamma,R) \\ &= \frac{\sum_C \{P(C|S,\gamma)P(W|S,\gamma,C)\}P(R|W)}{\sum_W P(W,R|S,\gamma)} \end{aligned} \quad (7)$$

where $\sum_C$ is done for the category C with the probability larger than a threshold and $\sum_W$ is done for W belonging to the categories. Strictly, $P(R|W)$, called *pronunciation similarity*, should be represented as $P(R|W,S,\gamma)$ by employment of Bayesian rule. Considering that the recognized pronunciation depends on W rather than S and γ, we can take an assumption $P(R|W,S,\gamma) \simeq P(R|W)$.

In computation of Eq.7, $P(C|S,\gamma)$ and $P(W|S,\gamma,C)$ are obtained by Eq.4. For calculation of $P(R|W)$, all the combinations of the recognized pronunciation and its alternative interpretations given by the dictation engine are used as R. In the

case of Table 5, for example, “go-hon-cha”, “go-kon-cha”, “no-hon-cha” and so on are considered. For each R, $P(R|W)$ is calculated by Ristad’s method [15] as follows.

$$\begin{aligned} P(R|W) &= P(r_1,r_2,\cdots,r_m|w_1,w_2,\cdots,w_n) && (8)\\ &= P(\varepsilon,r_1,r_2,\cdots,r_m|w_1,w_2,\varepsilon,\cdots,w_n) \\ &\quad +P(r_1,\varepsilon,r_2,\cdots,r_m|w_1,w_2,\varepsilon,\cdots,w_n)+\cdots && (9)\\ &\simeq P(\varepsilon|w_1)P(r_1|w_2)P(r_2|\varepsilon)\cdots P(r_m|w_n) \\ &\quad +P(r_1|w_1)P(\varepsilon|w_2)P(r_2|\varepsilon)\cdots P(r_m|w_n)+\cdots && (10)\end{aligned}$$

where r_i and w_j respectively denote the ith and the jth vowels and consonants consisting of R and W. ε denotes a blank symbol. The calculation of Eq.10 can be rapidly done by a DP-like method (see Figure 7). If $P(R|W)/P(R) > threshold$, such a W is rejected because of less confidence where

$$\begin{aligned} P(R) &= P(r_1,r_2,\cdots,r_m) \\ &\simeq P(r_1)P(r_2)\cdots P(r_m) \\ &= \sum_w P(r_1,w)\sum_w P(r_2,w)\times\cdots\times\sum_w P(r_m,w). && (11)\end{aligned}$$

Each $P(r_i|w_j)$ and $P(r_i,w_j)$ can be given as a mis-recognition model of the dictation engine by making a histogram $N(r_i,w_j)$ of mis-recognition counts repeating recognition tests in many times.

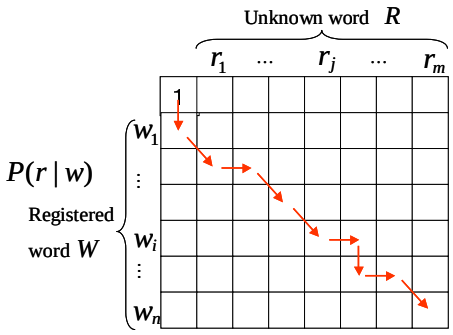
Finally the obtained $P(R|W)$ is substituted to Eq.7 and then the word W having the best $Score_m$ is picked up as the estimated word as *a mis-recognized word*.

If two successive unknown words (R_1,R_2) are detected, the system also considers that the two words are actually one mis-recognized word W. Because each of successive unknown words have multiple alternative interpretations and all the combinations of the alternatives of R_1 and R_2 should be considered, it takes $o(\#\text{ of } R_1 \times \#\text{ of } R_2)$ computations. In the case of two successive unknown words, the system first carries out DP-like calculation of $P(R_1|W)$ forward from the head and $P(R_2|W)$ backward from the tail (see Figure 8). Then total scores are obtained by combining them. Its computation order is $o(\#\text{ of } R_1 + \#\text{ of } R_2)$.

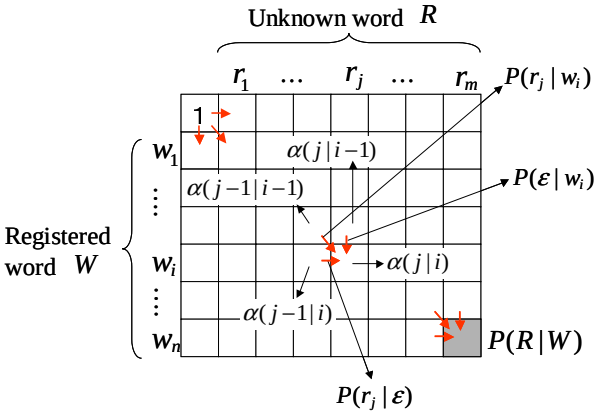
For cases that one unknown word R actually corresponds to two or more mis-recognized words, the similar computation is employed for repartitioning the word. First choose C_1 having $P(C_1,C_2|S,\gamma)$ larger than a threshold. For each W belonging to C_1, generate the DP matrix (Figure 7) made in calculation of $P(R|W)$. Because the value of the jth column in the lowest row of the matrix represents $P(r_1,\cdots,r_j|W)$, find j_{max} having the maximum value in the lowest row. If the maximum value is larger than a threshold, repartition the unknown word just after the j_{max}th cell and regard the W as the former mis-recognized word W_1 (see Figure 9(a)). The latter mis-recognized word W_2 can be found by generating the matrix backward from the tail of R (see Figure 9(b)).

3.5 Interactive learning of unknown word

Automatically estimated unknown words should be confirmed by the user. Then the system can obtain the correct word for mis-recognition or the corresponding word for a synonym. The correctly estimated word is added to the registered words in the system. For mis-recognized words, the pronunciation which the system first recognized is registered as an alternative pronunciation of the correct word.



(a) Computation of a certain path



(b) Summation of all the paths

Figure 7: Calculation of pronunciation similarity

4 Experimental Results

We implemented an interactive vision platform which can recognize beverages by analyzing a captured image inside of a refrigerator and interact with a user via a microphone and a

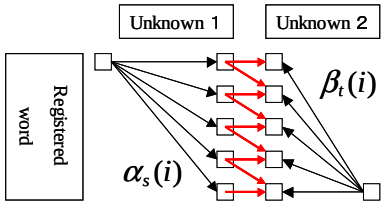
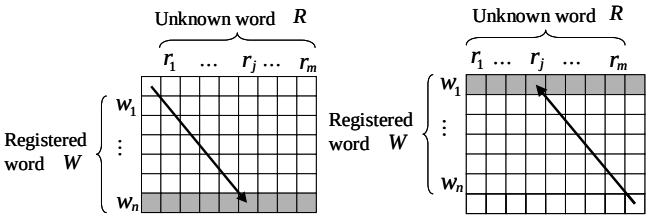


Figure 8: Calculation of $P(R|W)$ for multiple unknown words



(a) partitioning the first word (b) partitioning the second word

Figure 9: Repartitioning of one unknown word

(a) Image 1
State: (You've requested "pokari-suetto"(after-sport drink).)
System's utterance: "It's found. May I bring it?"

(b) Image 2
State: (You can choose any drink you want.)
System's utterance: (waiting)

Figure 10: State of the dialog: the user specified object is pointed by an arrow and the object selected by the system is surrounded by a box line.

headphone. In the speech recognition module, 104 words, 21 categories and 245 grammars are pre-registered. "Enrolling" (speaker adaptation) was not used for evaluation of robustness of our speech recognition method. We showed 10 different images inside of the refrigerator to 5 users and let them interact with our system. Figure 10 shows two of the images shown to the users, the state at that time and the preceding system's utterance. The user specified a preferred drink and then the system uttered to the user in order to obtain more information to progress the task (bringing a beverage). The recognition results of the user's utterances are shown as follows.

Table 7 shows an example for the recovery of a mis-recognized word. The system first detected the user-specified drink (a blue can), displayed it to the user (Figure 10(a)) and then said "It's found. May I bring it?" to the user. Because the detected object is another drink with blue package but in PET-bottle, the user corrected it by saying "*aoi petto-botoru* (the blue PET-bottle)". The dictation engine mistook "*aoi* (blue)" for two words "*ha omoni*". Because no registered word found in their alternative interpretations, it was verified that they were mis-recognized words. Then it gave the highest probability that the two unknown words corresponded to one registered word "*aoi*". Finally the unknown words were correctly recovered as *aoi*.

Table 7: Estimation of a mis-recognized word (for image 1)

utterance	auto recognition	recovered
aoi petto-botoru (blue pet-bottle)	(unknown) *petto-botoru* (unknown pet-bottle)	*aoi petto-botoru* (blue pet-bottle)

recognized result	alternatives
ha	*a*, *aa*, *ka*
omoni(mainly)	*on*(sound), *oni*(demon), *omoi*(heavy)
petto-botoru(pet-bottle)	—

	class	estimation	score	
one word	mis-recognized	*aoi*(blue)	1.000	*
	synonym	*aoi*(blue)	1.000	
two words	mis-recognized	—	—	
	synonym	—	—	

Table 8: Estimation of a synonym (for image 2)

utterance	auto recognition	recovered
dakara choodai bring "dakara"(drink name)	*dakara* (unknown) "dakara" unknown	*dakara totte* bring "dakara"

primary recognition	alternatives
dakara	*takara*(treasure)
choodai	*choodai*(bring), *choozai*(mixing), *choonai*(neighborhood)

	class	estimation	score	
one word	mis-recognized	*totte*(bring)	0.997	
	synonym	*totte*(bring)	0.998	*
two words	mis-recognized	—	—	
	synonym	*no-migi*(right of)	0.012	

Table 8 shows an example for the recovery of a synonym. The user said "*dakara choodai* (bring 'dakara (drink name)')" viewing Figure 10(b). Although *choodai* was correctly recognized by the dictation engine, no registered word is found among its alternative interpretations. As a result of verification, *choodai* was regarded as a synonym of *totte* (bring). Because "t" sound is often mis-recognized as "d" and "ch" sound in learning of $N(r,w)$, the probability of "*totte*" as the mis-recognized word is also high accidentally.

Table 9 shows an example for repartitioning of mis-connected words viewing Figure 10(b). The user said "*dakara totte* (bring 'dakara')" in two words. The dictation engines, however, mis-connected these words as "*dakara-toitte* (but)". Since this word is not registered, the registered categories and words were checked based on the current state and then "*wanda* (coffee name)" as one mis-recognized word, "*nohohon-cha* (green tea)" as a synonym, "*dakara totte* (bring 'dakara')" as two mis-recognized word and two synonyms were respectively estimated. Finally "*dakara totte*" as two mis-recognized word was adopted.

Table 9: Estimation of a mis-recognized word by repartitioning (for image 2)

utterance	auto recognition	recovered
dakara totte bring "dakara"(drink name)	(unknown) unknown	*dakara totte* bring "dakara"

primary recognition	alternatives
dakaratoitte (but)	—

	class	estimation	score	
one word	mis-recog.	*wanda* (coffee name)	0.612	
	synonym	*nohohon-cha* (green tea)	0.317	
two words	mis-recog.	*dakara totte* (bring "dakara")	1.000	*
	synonym	*dakara totte* (bring "dakara")	0.006	

5 Conclusion

This paper proposed a dialog system which interprets user's utterances for a service robot bringing user-specified objects from a refrigerator. The system estimates meanings of unknown recognized words in case of speech recognition failure or unexpected utterance. In order to improve speech recognition rate, acceptable words and grammars are registered in the system beforehand. When user's utterance is not consistent with the registered grammars, the system detects unknown words in the utterance and recovers mis-recognized words and synonyms of the registered words considering the state, the context and the pronunciation similarity. Experimental results with real environments are shown.

For future works, the following problems are to be solved. When the number of the registered word is much larger, a number of registered words might be mis-matched to the alternative interpretations. The system should choose an appropriate word by evaluating the reliability of the matched registered word. In addition, If the estimated word of the unknown word is incorrect, the current system can only repeat the previous question. It cannot respond to user's utterances during the image processing. The utterance ability of the system should be further improved so that the user naturally help the system perform the tasks.

References

[1] Y. Takahashi, S. Nakanishi, Y. Kuno, and Y. Shirai, "Human-Robot Interface by verbal and Nonverbal Communication", Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 924-929, 1998.

[2] T. Takahashi, T. Komeda, T. Uchida, M. Miyagi, and H. Koyama, "Development of the mobile robot system to aid the daily life for physically handicapped", Proc. of ICMA2000, pp. 549-554, 2000.

[3] K. Itou and S. Hayami and H. Tanaka, "Treatment of Unknown Words on Speech Dialog System" JSAI SIG-SLUD-9201-1, pp. 1-9, 1992 (in Japanese).

[4] Y. Takahashi, K. Douzaka and K. Aikawa, "Attribute Estimation of Unknown Words on Speech Dialog Based on Discourse", IEICE, NLC2001-35, pp.101-106, 2001 (in Japanese).

[5] T. Takahashi, S. Nakanishi, Y. Kuno and Y. Shirai, "Helping Computer Vision by Verbal and NonVerbal communication", Proc. of 14th Int. Conf. on Pattern Recogntion, pp.1216-1218, 1998.

[6] "Voice Land", http://www-6.ibm.com/jp/voiceland/.

[7] G. Damnati, F. Panaget, "Adding New Words in a Spoken Dialogue System Vocabulary Using Conceptual Information and Derived Class-based LM", Proc. of Workshop on Automatic Speech Recognition and Understanding, 1999.

[8] M. Nagata, "A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context", 37th Annual Meeting of the Association for Computational Linguistics, pp.277-284, 1999.

[9] K. Fujii and K. Sugiyama, "A Method of Generating a Spot-Guidance for Human Navigation", Trans. of IEICE D-II Vol. J82-DII No. 11, pp. 2026-2034, 1999 (in Japanese).

[10] M. Iwata and T. Onisawa, "Linguistic Expressions of Picture Information Considering Connection between Pictures", Trans. of IEICE D-II Vol. J84-DII No. 2, pp. 337-350, 2001 (in Japanese).

[11] Y. Watanabe, M. Nagato, and Y. Okada, "Image Analysis Using Natural Language Information Extracted from Explanation Text", Proc. of MIRU'96 Vol. 2, pp. 271-276, 1996 (in Japanese).

[12] S. Wachsmuth and G. Sagarer, "Connecting Concepts from Vision and Speech Processing", Workshop on Integration of Speech and Image Understanding, 1999.

[13] U. Ahlrichs, J. Fischer, J. Denzler, C. Drexler, H. Niemann, E. Noth, and D. Paulus, "Knowledge Based Image and Speech Analysis for Service Robots", Workshop on Integration of Speech and Image Understanding, 1999.

[14] Y. Shirai "Three-Dimensional Computer Vision", Springer-Verlag, pp. 62-68, 1987.

[15] E. S. Ristad, P. N. Peter, "Learning String-Edit Distance", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 5, pp. 522-532, 1998.

[16] Y. Makihara, M. Takizawa, Y. Shirai, J. Miura and N. Shimada, "Object Recognition by Supported by User Interaction for Service Robots", Proc. of Asian Conf. on Computer Vision, pp.719-724, 2002.