IEICE TRANS. INF. & SYST., VOL.E98–D, NO.3 MARCH 2015

PAPER

Discriminating Unknown Objects from Known Objects Using Image and Speech Information

Yuko OZASA[†a)], ***Student Member***, **Mikio NAKANO**[††b)], ***Member***, **Yasuo ARIKI**[†c)], ***Fellow***, ***and*** **Naoto IWAHASHI**[†††d)], ***Member***

SUMMARY This paper deals with a problem where a robot identifies an object that a human asks it to bring by voice when there is a set of objects that the human and the robot can see. When the robot knows the requested object, it must identify the object and when it does not know the object, it must say it does not. This paper presents a new method for discriminating unknown objects from known objects using object images and human speech. It uses a confidence measure that integrates image recognition confidences and speech recognition confidences based on logistic regression.

key words: *multimodality, unknown object discrimination, object recognition, information integration*

1. Introduction

When a household robot works with a human in a home environment, the robot needs to understand and ground the human language to the physical world. The grounding between language and the physical world requires the representation of real objects in the physical world. The real objects are represented by multiple modalities. Roy et al. [1] presented an implemented computational model of word acquisition which learns directly from raw multimodal sensory input. Specifically, in object-mediated communication, the multiple modalities are needed. Examples of such modalities are the language information, human voice, and physically observed information of the objects. So far methods for learning language and its meaning using several modalities such as voice and the object image have been proposed [1]–[7].

We learn knowledge not only from books but also from conversation and interaction with others. It is more natural for robots to learn knowledge through mutual interaction with humans. Several researchers have proposed this learning method for robots through interaction [8]–[12]. There are two approaches. One approach imitates the learning of children [8]. Its purpose is to make the robot learn and ground the language and concepts in the same way as children. The other approach deals with leaning while executing tasks [9]–[12]. It deals with the grounding problem in the task. For example, let us consider the task where a robot brings an object requested by a human voice. The grounding between the human speech of the object name and the image of the object is required in order to achieve the task. Our research belongs to the latter approach. The purpose of our work is to make robot learn unknown objects through the natural interaction between human and robots. For the interaction, we consider an object manipulation task. The task assumes that there are several objects which are known or unknown on a table, and a human tells the robot "bring me ⟨object name⟩." as shown in Fig. 1. Although there have been several pieces of work that deal with the object manipulation task in the situation where all the objects on the table are known, there has been no research dealing with the task in the situation where there are objects some of which are unknown to the robot on the table.

People sometimes refer to an object that the other does not know. They can discriminate whether the object is known or unknown to themselves. When the object is unknown, they learn the object at that time. In this paper, we mean by "an unknown object" an object whose name and image model the hearer does not have. So its name is out of vocabulary of the hearer.

In this paper, we propose an object recognition method using integrated confidence measures of image and speech, and an unknown object discrimination method by extending the object recognition method as the first step of unknown object learning through the interaction between a human and robots. Under the assumption that the spoken object name is the name of an object on the table, the image feature of the objects on the table and human speech are integrated so that the robot can detect the indicated object.

Fig. 1 Autonomous discrimination of unknown objects and their names by a robot.

Manuscript received July 28, 2014.
Manuscript revised November 10, 2014.
Manuscript publicized December 16, 2014.

[†]The authors are with the Graduate School of System Informatics, Kobe University, Kobe-shi, 657–8501 Japan.
[††]The author is with Honda Research Institute Japan Co., Ltd., Wako-shi, 351–0188 Japan.
[†††]The author is with the Graduate School of Computer Science and Systems Engineering, Okayama Prefectural University, Sojashi, 719–1197 Japan.
a) E-mail: y_ozasa@stu.kobe-u.ac.jp
b) E-mail: nakano@jp.honda-ri.com
c) E-mail: ariki@kobe-u.ac.jp
d) E-mail: naoto.iwahashi@gmail.com
DOI: 10.1587/transinf.2014EDP7260

The achievement of the task requires speech and image recognition. Then, there are four types of pairs of speech and image, a speech of a known name and an image of known object, a speech of a known name and an image of an unknown object, a speech of an unknown name and an image of a known object, and a speech of an unknown name and an image of an unknown object. The robot needs to discriminate these four types of pairs. To consider the task where a robot selects the object requested by a human voice from the multiple objects on the table, the task can be achieved by discriminating these four types of pairs. The discrimination enables the robot to select the object if it does not know in some cases.

The rest of the paper is organized as follows. Section 2 gives the details of the object manipulation task in this paper. Section 3 describes the proposed method for unknown object discrimination. The experimental methodology and results are presented in Sect. 4. Finally, Sect. 5 concludes the paper.

2. Task Settings

The task this paper deals with is to select an object requested by a human voice among the objects including the unknown objects on the table. It is different from an object grasping task which many robotics researchers deal with. As far as we know, it has not been dealt with in previous studies although this task is important for domestic robots that assist humans' daily lives.

In more detail, the task is described as follows:

- There are several objects on a table - Some or all objects may be unknown to the robot.
- A human tells the robot "bring me ⟨object name⟩ on the table", and the robot behaves as requested.

Two types of behaviors are prepared in this task. Ideally, the robot is expected to respond as follows (Fig. 2):

1. When the robot can select the object requested to bring, it says "Here you are" and brings the object to the user.
2. When the robot cannot select the object, it says "I don't know.", without doing any actions.

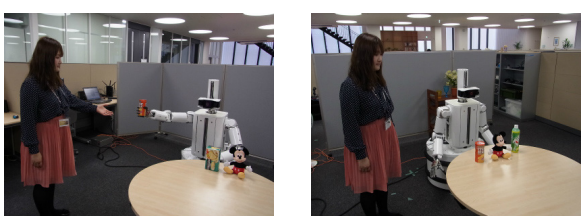
Let us consider the interactions between humans in the case that there are multiple objects on the table and one of the objects is unknown and other objects are known, and a human requests the other human to bring the unknown object. The human can bring the unknown object since he knows the sets of pairs of names (speech) and images of objects except for the unknown object.

The robot can select the object in this case since the integrated information of speech and image is used in the proposed method. The method using only speech and image cannot be applied to this case. Through this interaction, the robot can learn unknown objects in a natural way.

There can be the following three cases when multiple objects are on the table.

- The input speech is the name of a known object that is on the table.
- There are multiple objects on the table, the input speech is the name of an unknown object, only one object is unknown, and the remaining objects are known.
- There are multiple objects on the table, the input speech is the name of an unknown object, and there are multiple unknown objects on the table.

In the first and second cases, the behavior of the robot is (a), and in the third case, the behavior of the robot is (b) in Fig. 2.



(a) Bring the object and say " Here you are.".

(b) Say " I don't know.".

Fig. 2 Variation of robot behaviors.

3. Proposed System

The object grasping task requires the robot to grasp an object in a certain way, but our task requires the robot to discriminate the known and unknown objects and recognize the objects.

The proposed system is composed of two parts, estimating confidence and detecting unknown objects. The proposed system diagram is shown in Fig. 3. The unknown object discrimination algorithm is as follows:

The Unknown Object Discrimination Algorithm

Input: C_s, C_o
Output: "*Known*/*Unknown*", *Object name*

if $\max_i(F(C_s(s;\Lambda_i), C_o(o;G_i))) < \delta$ **then**
  Output: "*Unknown*", *Object name* of i
else

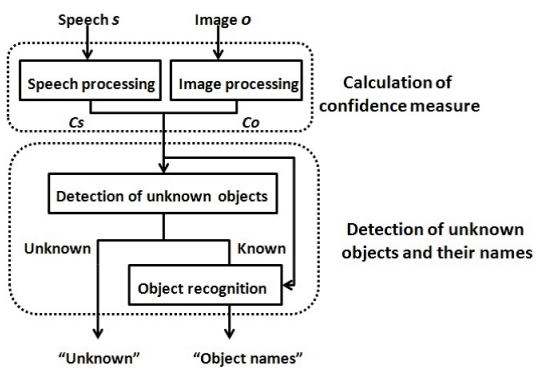


Fig. 3 Proposed system configuration diagram.

Output: "*Known*", *Object name* of i

The proposed method for unknown object discrimination uses both image and speech information in an integrated way. The confidences of the recognition results for input speeches and images, $C_s(s;\Lambda_i)$ and $C_o(o;\boldsymbol{g}_i)$, are estimated. s denotes the input speech, Λ_i denotes the speech model of i-th object, o denotes the input image, and $\boldsymbol{g}_i$ denotes the image model of i-th object. Then, the confidences are integrated via logistic regression $F(C_s, C_o)$ and the unknown objects are detected by thresholding the integrated confidence where the threshold is δ.

3.1 Confidence Measure

The proposed method integrates the confidences of speech recognition results and image recognition results, and the integrated confidence is used in detecting unknown objects and their names.

3.1.1 Speech Processing

The features used for speech recognition were Mel-frequency cepstral coefficients, which were based on short-time spectrum analysis; their delta and acceleration parameters; and the delta of short-time log power. These features were obtained by speech recognition software, Julius [14]. Speech recognition confidence is used to evaluate the reliability of the result of speech recognition and it is obtained by the following formula [17]:

$$C_s(s;\Lambda_i) = \frac{1}{n(s)} \log \frac{P(s;\Lambda_i)}{\max\limits_u P(s;\Lambda_u)} \tag{1}$$

where $P(s;\Lambda_i)$ is the likelihood of speech and Λ_i denotes the word Hidden Markov Model (HMM) for the name of the i-th object. $n(s)$ denotes the number of frames in the input speech and u denotes an arbitrary phoneme sequence. So $\max\limits_u P(s;\Lambda_u)$ means the likelihood of the result of phoneme typewriter, that is, speech recognition without a language model that allows any phoneme sequence. Since this language model does not put any restriction on the phoneme sequence, $\max\limits_u P(s;\Lambda_u)$ is considered to be the maximum of the likelihood given the input speech. So, $C_s(s;\Lambda_i)$ means how likely Λ_i is the word for the input speech[†].

[†] $C_s(s;\Lambda_i)$ equals to the ratio of the probability of Λ_i and the probability of the most likely phoneme sequence given the input speech. This is because the following holds by the Bayes' Theorem.

$$P(s;\Lambda_i) = \frac{P(\Lambda_i;s)P(\Lambda_i)}{P(s)}$$

Since we assume $P(\Lambda_i)$ is a constant in this work, $C_s(s;\Lambda_i)$ equals to

$$\frac{1}{n(s)} \log \frac{P(\Lambda_i;s)}{\max\limits_u P(\Lambda_u;s)}.$$

3.1.2 Image Processing

The features used for image recognition were L*a*b* components (three dimensions) for color, complex Fourier coefficients (eight dimensions) of contours for shape [18], and the area of an object (one dimension). Gaussian models were learned using these features by MAP adaptation. The confidence of the objects is written as follows [13]:

$$C_o(o;G_i) = \log \frac{P(o;G_i)}{P_{max}} \tag{2}$$

and G_i denotes the normal distribution of the i-th object, and $P(o;G_i)$ is the likelihood of the object, $P_{max} = ((2\pi)^{\frac{d}{2}}|\sum|^{\frac{1}{2}})^{-1}$ denotes the maximum probability density of Gaussian functions. $\sum$ denotes the covariance matrix of the Gaussian function. $C_o(o;G_i)$ is normalized by P_{max} so that it means how close the input image is to the model of the i-th object.

3.2 Logistic Regression for Modality Integration

The speech and image confidences are not always reliable. For example, speech confidences can be affected by change in acoustic conditions such as noises, and image confidences can be affected by change in lighting conditions. So, it would be effective to integrate speech and image confidences to better estimating confidences. We employ logistic regression for the integration, and use the integrated confidences for unknown object discrimination.

3.2.1 Logistic Regression

The speech recognition confidence measure and object recognition confidence measure are integrated by the following logistic regression function [13]:

$$F(\mathbf{C}) = \frac{1}{1+\exp\{-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{C}\}} \tag{3}$$

Here $\mathbf{C}^{\mathrm{T}} = (1, C_s, C_o)$ and $\boldsymbol{\alpha}^{\mathrm{T}} = (\alpha_0, \alpha_1, \alpha_2)$ are logistic regression coefficients. In the training of this logistic regression function, the (i, j)-th training sample is given as the pair of input signals $(C_s(s_j;\Lambda_i), C_o(o_j;G_i))$ and teaching signal $d_{i,j}$, where i denotes the model index and j denotes the sample index. Thus, the training set T contains $N \times M$ (N models and M samples) samples.

$$\mathrm{T}^{N\times M} = \{C_s(s_j;\Lambda_i),\ C_o(o_j;G_i),\ d_{i,j} \mid i = 1,\cdots,N,\ j = 1,\cdots,M\} \tag{4}$$

The teaching signal $d_{i,j}$ is 1 when s_j is a speech of the name of the object i and o_j is a image of the object i, and 0 otherwise. When using logistic functions, we investigate only whether the input matches the model or not. Then we determine if the input is an unknown object or not using outputs

of the logistic functions each of which checks if the input matches one of the models of all known objects or not. If the input matches none of the models of the known objects, it is considered to be an unknown object. The log likelihood function of the training set using the logistic regression function is written as

$$l(\boldsymbol{\alpha}) = \sum_{j=1}^{M}\sum_{i=1}^{N}\{d_{i,j}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{C}_j^i - \log(1+\exp(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{C}_j^i))\} \tag{5}$$

Here ${\mathbf{C}_j^i}^{\mathrm{T}} = (1, C_{s_j}^i, C_{o_j}^i)$, and $C_{s_j}^i$ and $C_{o_j}^i$ are $C_s(s_j;\Lambda_i)$ and $C_o(o_j;G_i)$ respectively for abbreviation. The weight set $\boldsymbol{\alpha}$ is optimized by maximum likelihood estimation using Fisher's scoring algorithm [19].

3.2.2 Regularized Logistic Regression

Over fitting of the learning of logistic regression is a serious problem. To avoid over fitting, regularized logistic regression is used. The log likelihood function in regularized logistic regression based on ridge regression is written as follows [21]:

$$\begin{aligned} l_R(\boldsymbol{\alpha}) = &\sum_{j=1}^{M}\sum_{i=1}^{N}\{d_{i,j}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{C}_j^i - \log(1+\exp(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{C}_j^i))\} \\ &+ \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2 \end{aligned} \tag{6}$$

where λ is the coefficient of the regularized term. The weight set $\boldsymbol{\alpha}$ is optimized in the same way as that of logistic regression described in Sect. 3.2.1.

3.2.3 Kernel Logistic Regression

There are linear regression and nonlinear regression. Logistic regression is a nonlinear regression but its discrimination plane is linear. This section considers the logistic function whose discrimination plane is nonlinear. One such logistic regression is kernel logistic regression [21]. Using the basis function, kernel logistic regression is obtained. In this paper, the Gaussian basis function shown in (7) is used.

$$\phi_m(\mathbf{C}) = \exp\left(-\frac{\left\|\mathbf{C}-\boldsymbol{\mu}_m\right\|^2}{2s_m^2}\right) \tag{7}$$

where $\boldsymbol{\mu}_m$ is the center vector of the basis function, and s_m is the parameter which defines the broadening of the basis function. Kernel logistic regression is written as follows:

$$F_K(\mathbf{C}) = \frac{1}{1+\exp\{-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{C})\}} \tag{8}$$

where $\boldsymbol{\phi}(\mathbf{C})$ is the vector each element of which is the value of the Gaussian basis function (7) at the (i, j)-th training sample, namely $\boldsymbol{\mu}_m = \boldsymbol{\mu}_{i,j} = \mathbf{C}_j^i = (1, C_{s_j}^i, C_{o_j}^i,)^{\mathrm{T}}$. The log likelihood function in the kernel logistic regression is written as follows:

$$l_K(\boldsymbol{\alpha}) = \sum_{j=1}^{M}\sum_{i=1}^{N}\{d_{i,j}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{C}_j^i) - \log(1+\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{C}_j^i)))\} \tag{9}$$

The weight set $\boldsymbol{\alpha}$ is optimized in the same way as logistic regression.

3.2.4 Multiclass Logistic Regression

The logistic regression described above is two-class logistic regression and that can discriminate multimodal inputs into two classes. Here we mention multiclass logistic regression [21] which will be used in Sect. 3.4.

Let us consider K class logistic regression. The k-th class logistic function is written as follows:

$$F_{M,k}(\mathbf{C}_j^i) = \frac{\exp(\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{C}_j^i)}{\sum_{p=1}^{K}\exp(\boldsymbol{\alpha}_p^{\mathrm{T}}\mathbf{C}_j^i)} \tag{10}$$

Then, the log likelihood function is writen as follows:

$$\begin{aligned} l_M(\boldsymbol{\alpha}) = &\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=1}^{K} d_{i,j,k}\log F_{M,k}(\mathbf{C}_j^i) \\ &+ (1-d_{i,j,k})\log(1-F_{M,k}(\mathbf{C}_j^i)) \end{aligned} \tag{11}$$

where $d_{i,j,k}$ is a teaching signal, and 0 or 1.

3.3 Discrimination of Unknown Objects and Their Names

In the discrimination phase, the multimodal input is classified as an unknown object or a known object using integrated confidence. When the multimodal input is classified as unknown, it is considered that an unknown object is detected and its name is obtained by the input speech. When multimodal input is classified as known, then the object with its name is output.

3.3.1 Discrimination of Unknown Objects

Figure 4 shows the joint distribution of speech recognition confidence and image recognition confidence of the experiment data described in Sect. 4. This graph plots data of

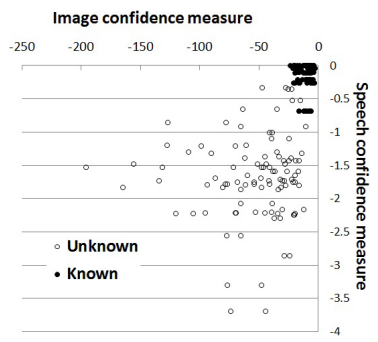


Fig. 4 Joint distribution of values of speech and object confidence.

10 objects. For each object, 11 images and one speech are used to form 11 pairs of an image and a speech. Their confidences for 10 object models are obtained, so in total 110 data are plotted. The sets of pairs of speech and image confidence measure when the input is unknown or known are plotted. It indicates that discriminating unknown and known objects would be possible by using both confidences simultaneously. Given a threshold δ, the object is classified as unknown or known.

The logistic regression function $F(C_s, C_o)$ is used for the classification of unknown and known objects. If the following condition is satisfied, the input object is classified as an unknown object, otherwise as a known object.

$$\max_i F(C_s(s;\Lambda_i), C_o(o;G_i)) < \delta, \tag{12}$$

δ denotes the boundary of the classification based on the logistic regression. There are two kinds of thresholds that are often used; one is the confidence boundary which is 0.9 and the other is decision boundary which is 0.5 [20]. Although the decision boundary is a standard, the confidence boundary is known to work better for the classification of real data [20]. So, the confidence boundary is used in the experiments in this paper.

3.3.2 Object Recognition

When the input is classified as a known object, it is recognized and its ID is obtained as follows:

$$\hat{i} = \arg\max_i F(C_s(s;\Lambda_i), C_o(o;G_i)) \tag{13}$$

Then, the object name is output.

3.4 Discrimination of Multiple Unknown Objects and Their Names

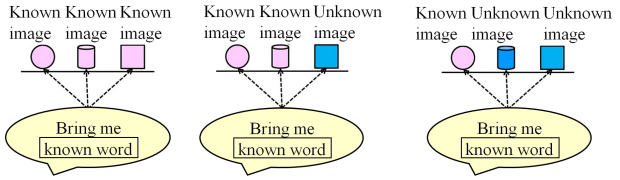
3.4.1 Cases of Multiple Unknown Object Discrimination

The method for detecting an unknown object proposed in Sect. 3.3 can be extended to methods which detect multiple unknown objects and their names.

The proposed method described in Sect. 3.3 assumed that the input speech refers to the input image. However, when there are multiple objects, this assumption does not hold. For the image of each object on the table, we need to check if the pair of the input speech and the input image matches one of the known objects or not. Even if the speech is the name of known object, the input image may not a known object. So the method described in Sect. 3.3 is not applicable when there are multiple objects on the table.

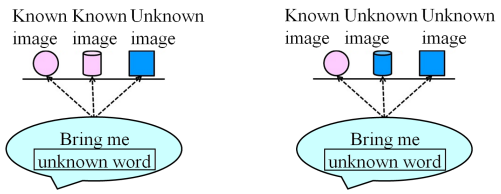
Let us consider the cases of multiple unknown object discrimination shown in Figs. 5 and 6. In this setting, we assume that the spoken name is always the name of one of the objects on the table.

Case 1: There are three known objects on the table and a known speech is input. One of the objects corresponds to the input speech, and the other objects do not. If the robot discriminates the corresponding pair of speech and image, it can get the target known object.

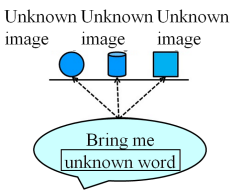


(a) Case 1 (b) Case 2 (c) Case 3

Fig. 5 Cases where the input word is known.



(a) Case 4 (b) Case 5



(c) Case 6

Fig. 6 Cases where the input word is unknown.

Case 2 and 3: The object corresponding to the input speech is one of the known objects. The objects not corresponding to the input speech are treated as unknown objects for the robot. If the sets of pairs of an image of a known object and a speech of a known name and that of an image of an unknown object and a speech of a known name can be discriminated, the robot can get the targeted known object.

Case 4: The object corresponding to the input speech is an unknown object, and the known objects do not correspond to the input speech. If the sets of pairs of an image of an unknown object and a speech of an unknown name and that of an image of a known object and a speech of an unknown name can be discriminated, the robot can get the targeted unknown object.

Case 5: The object corresponding to the input speech is an unknown object, and other objects do not correspond to the input speech. If the sets of pairs of an image of an unknown object and a speech of an unknown name and that of an image of a known object and a speech of an unknown name can be discriminated, the robot can narrow down the selections of the targeted object.

Case 6: The objects on the table and the input speech are

unknown. All input sets of pairs are an image of an unknown object and a speech of an unknown name. If the unknown objects are detected, the robot learn all the objects on the table are unknown objects.

We propose a method for dealing with these cases. It consists of the following three parts.

The first part checks if each input pair of speech and image matches the model of an object. This process classifies each input into one of the three classes C_1, C_2, and C_3. C_1 means neither the speech or image matches the model, C_2 means either of the speech or the image match the model, and C_3 means both of the speech and image match the model (Fig. 7). We employed two methods for this discrimination. One method is to use two two-class logistic regression functions; one for discriminating C_3 from C_1 and C_2 and the other is for discriminating C_1 from C_2 and C_3. The other method uses three-class logistic regression mentioned in Sect. 3.2.4 to classify C_1, C_2 and C_3.

The second part checks if the input is an unknown object or not based on the results of the first part by the following procedure.

(U) If the results of the first part for the models of all known objects are C_1, the input is considered to be an unknown object.
(K) Else if the results of the first part for at least one of the models of all known objects is C_3, the input is considered to be a known object. If the results of the first part were C_3 for more than one model, the model that matches with the highest confidence is selected in the object identification.
(O) Otherwise, the input speech is an unknown name and the input image is as a known object, or the input speech is a known name and the input image is as an unknown object.

The third part identifies the requested object based on the results of second part for the images of all objects on the table. For example, in Case 1 of Fig. 5, the results of the second part should be K for one object on the table and O for the remaining two objects, so the robot can select the object that are assigned K.

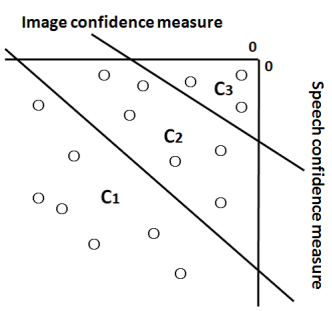


Fig. 7 Matching input pairs of a speech and an image with object models.

4. Experimental Evaluation

We first evaluated the unknown object discrimination method, and then evaluated object recognition. The coefficients α_0, α_1, and α_2 were optimized in the experiment.

50 objects were prepared and for each object, one utterance including its name and 11 images were collected. Some of the images are shown in Fig. 8. Two types of image datasets, data set 1 and data set 2 were prepared. Data set 1 consists of images of objects taken from 11 angles. Data set 2 consists of images of objects taken from 5 angles. Figure 9 shows samples of 11 images of a bear taken from 11 angles. The size of the image is 640×480 pixels. The RGB image and depth map are taken by Kinect [22], and the object region is automatically extracted by both the RGB image and depth map. Examples of the RGB image, depth map, and extracted object region are shown in Fig. 10. The extracted object regions are used in the experiment. All utterances were spoken by one speaker.

4.1 Evaluation of Method to Detect Unknown Objects

Evaluation was also performed using leave-one-out cross validation in this section. The features used in image recognition were L*a*b* components for color, complex Fourier coefficients of contours for shape, and the area of an object as described in Sect. 3.2.1. Data set 1 is used in this experiment. We evaluated the unknown object detection with the

Fig. 8 Examples of objects used in the experiment.

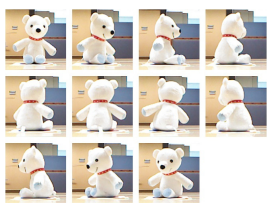
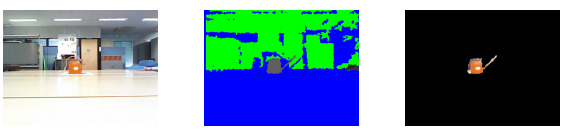


Fig. 9 11 images of a bear taken from 11 different angles.



(a) RGB image (b) Depth map (c) Extracted object region

Fig. 10 Example of information obtained by Kinect.

Table 1 Accuracy of multiple unknown object discrimination (%).

Two types of two-class logistic regression				Three class logistic regression				SVM
L	RL	KL	RKL	L	RL	KL	RKL	
82.3	85.8	92.4	97.6	88.2	90.1	91.5	98.0	95.6

Table 2 Accuracy of object recognition for respective feature (%).

Features	Image					Speech	Logistic
	L*a*b	Area	Fourier	L*a*b+Fourier	All		
Data set 1	73.6	14.6	38.0	89.6	93.0		100.0
Data set 2	69.6	11.4	29.2	84.8	88.2	96.0	100.0

four types of pairs of all combinations of a speech and an image. The four types of the pairs are the pairs of a speech of a known name and an image of a known object, a speech of an unknown name and an image of a known object, a speech of a known name and an image of an unknown object, and a speech of an unknown name and an image of an unknown object. The data excluding a test data (a pair of a speech and an image) were used for the training data. When a test data was a pair of known speech and image, 9 images of each 50 object (450 images) were used for the training data of the image models. When a test data was not a pair of known speech and image, the unknown object of the test data was excluded, and 9 images of each 49 object (441 images) were used for the training data of the image models. The pairs of a speech and an image excluding the test data and training data of the image models were used for the training data of the logistic regression. The pairs of 1 speech and 1 image of each 49 object (49 pairs) were used for the training data of the logistic regression. The accuracy of the method using two-class logistic regression and three class logistic regression was compared in this section. Four types of logistic regression are used, logistic regression, regularized logistic regression, kernel logistic regression, and regularized kernel logistic regression. The training of RKL is time-consuming, so the least-squares probabilistic classifier (LSPC) [23] is used for the determination of the kernel parameter and the regularization parameter. The parameters are optimized in each experiment one by one.

The experimental result is shown in Table 1. In Table 1, L, RL, KL, and RKL denote logistic regression, regularized logistic regression, kernel logistic regression, and regularized kernel logistic regression, respectively. The accuracy of the method using polynomial kernel SVM (Support Vector Machine) was also compared, too.

When comparing regularized logistic regression with logistic regression, the former is more effective. Kernel logistic regression is more effective than regularized logistic regression. This result shows that pairs of confidence measures of data sets varies widely, and in such data sets, the method using kernel logistic regression is effective. Regularized kernel logistic regression is the most effective, compared to logistic regression, regularized logistic regression, kernel logistic regression, and SVM. SVM is more effective than logistic regression, regularized logistic regression, and kernel logistic regression but less effective than regularized kernel logistic regression. The method using three class regularized kernel logistic regression is the most effective in this experiment.

4.2 Evaluation of Object Recognition

Evaluation was performed using leave-one-out cross validation. Under the condition that a known object was input, we chose one image as test data from 50 objects, and the remaining images were used as training data. When data set 1 was used, the number of training data was 549 and when data set 2 was used, the number of training data was 249. The experiment was carried out for all images. The parameters $\boldsymbol{\alpha}, \lambda$ are optimized in the experiments in this paper.

The features used in image recognition were L*a*b* components for color, complex Fourier coefficients of contours for shape, and the area of an object as described in Sect. 3.2.1. The accuracy of object recognition of each feature is shown in Table 2. In Table 2, the accuracies of object recognition by image confidence measure, speech confidence measure, and integrated confidence measure using logistic regression are shown. Among the accuracies of image confidence measure of each feature, L*a*b* components were the most efficient in both data sets. The accuracy of the integrated confidence measure is the most efficient in Table 2.

5. Conclusion

Acquiring new knowledge through interactive learning mechanisms is a key ability for robots in a real environment. To acquire new knowledge, discrimination and learning of unknown objects and their names are needed. The proposed method makes it possible for a robot to detect unknown objects and their names online using multimodal information. Though the method is based on well-known logistic regression techniques, how to apply them to detecting unknown objects and identifying known objects was not trivial. Experimental results show that regularized kernel logistic regression was the most efficient. We will pursue a method for learning unknown objects in a real environment.

Our method employs a simple way of integrating different modalities to investigate if the input matches one of the models that the system has. Thanks to this simplicity, this method is expected to be applied to other task domains such as person identification from his/her face image and voice.

References

[1] D.K. Roy and A.P. Pentland, "Learning words from sights and sounds: A computational model," Cognitive Science, vol.26, no.1, pp.113–146, 2002.
[2] L. Steels and M. Loetzsch, The grounded naming game. Experiments in Cultural Language Evolution, John Benjamins, 2012.
[3] L. Steels and F. Kaplan, "Aibos first words: The social learning of language and meaning," Evolution of Communication, vol.4, no.1, pp.3–32, 2002.
[4] T. Araki, T. Nakamura, T. Nakai, K. Funakoshi, M. Nakano, and N. Iwahashi, "Autonomous acquisition of multimodal information for online object concept formation by robots," 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.1540–1547, 2011.
[5] Y. Chen and D.H. Ballard, "On the integration of grounding language and learning objects," AAAI, pp.488–493, 2004.
[6] N. Iwahashi, "Interactive learning of spoken words and their meanings through an audio-visual interface," IEICE Trans. Inf. & Syst., vol.E91-D, no.2, pp.312–321, Feb. 2008.
[7] D.O. Johnson and A. Agah, "Human robot interaction through semantic integration of multiple modalities, dialog management, and contexts," Int. J. Social Robotics, vol.1, no.4, pp.283–305, 2009.
[8] D.K. Roy, New horizons in the study of child language acquisition, International Speech Communication Association, 2009.
[9] M. Nakano, N. Iwahashi, T. Nagai, T. Sumii, X. Zuo, R. Taguchi, T. Nose, A. Mizutani, T. Nakamura, M. Attamim, H. Narimatsu, K. Funakoshi, and Y. Hasegawa, "Grounding new words on the physical world in multi-domain human-robot dialogues," 2010 AAAI Fall Symposium Series, pp.74–79, 2010.
[10] H. Holzapfel, D. Neubig, and A. Waibel, "A dialogue approach to learning object descriptions and semantic categories," Robotics and Autonomous Systems, vol.56, no.11, pp.1004–1013, 2008.
[11] D. Skocaj, M. Janicek, M. Kristan, G.M. Kruijff, A. Leonardis, P. Lison, A. Vrecko, and M. Zillich, "A basic cognitive system for interactive continuous learning of visual concepts," ICRA 2010 Workshop ICAIR-Interactive Communication for Autonomous Intelligent Robots, pp.30–36, 2010.
[12] F. Lomker and G. Sagerer, "A multimodal system for object learning," Pattern Recognition, pp.490–497, 2002.
[13] X. Zuo, N. Iwahashi, K. Funakoshi, M. Nakano, R. Taguchi, S. Matsuda, K. Sugiura, and N. Oka, "Detecting robot-directed speech by situated understanding in physical interaction," Information and Media Technologies, vol.5, no.4, pp.1314–1326, 2010.
[14] Julius, http://julius.sourceforge.jp/
[15] T. Nakamura, K. Sugiura, T. Nagai, N. Iwahashi, T. Toda, H. Okada, and T. Omori, "Learning novel objects for extended mobile manipulation," J. Intelligent & Robotic Systems, vol.66, no.1-2, pp.187–204, 2012.
[16] K. Okada, S. Kagami, M. Inaba, and H. Inoue, "Plane segment finder: Algorithm, implementation and applications," Proc. 2001 ICRA, IEEE International Conference on Robotics and Automation, pp.2120–2125, 2001.
[17] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Commun., vol.45, no.4, pp.455–470, 2005.
[18] E. Persoon and K. Fu, "Shape discrimination using Fourier descriptors," IEEE Trans. Syst. Man Cybern., vol.7, no.3, pp.170–179, 1977.
[19] T. Kurita, "Iterative weighted least squares algorithms for neural networks classifiers," New Generation Computing, vol.12, no.4, pp.375–394, 1994.
[20] D. Barber, Bayesian reasoning and machine learning, Cambridge University Press, 2012.
[21] C.M. Bishop and N.M. Nasrabadi, Pattern recognition and machine learning, Springer, New York, 2006.
[22] Kinect, http://www.microsoft.com/en-us/kinectforwindows/
[23] M. Sugiyama, H. Hachiya, M. Yamada, J. Simm, and H. Nam, "Leastsquares probabilistic classifier: A computationally efficient alternative to kernel logistic regression," Proc. International Workshop on Statistical Machine Learning for Speech Processing, pp.1–10, 2012.

Yuko Ozasa received her B.E. and M.E. degrees in computer science from Kobe University in 2007 and 2009, respectively, after which she continued her research as a doctoral student at 2011. She was interested in computer security and information theory until 2011. In 2011, she changed her major to media recognition and understanding. Her research interests include speech and image signal processing, pattern recognition, multimodality, and symbol grounding problems. She is currently a 3rd-year doctoral student at Kobe University.

Mikio Nakano is a Principal Researcher at Honda Research Institute Japan Co., Ltd. (HRI-JP). He received his M.S. degree in Coordinated Sciences and Sc.D. degree in Information Science from the University of Tokyo, respectively in 1990 and 1998. From 1990 to 2004, he worked for Nippon Telegraph and Telephone Corporation. In 2004, he joined HRI-JP. His research interests include speech understanding, spoken dialogue systems, and conversational robots. He is a member of ACM, IEEE, AAAI, ISCA and other academic societies.

Yasuo Ariki received his B.E., M.E. and Ph.D. in information science from Kyoto University in 1974, 1976 and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990, and stayed at Edinburgh University as visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. Since 2003 he has been a professor at Kobe University. He is mainly engaged in speech and image recognition and interested in information retrieval and database. He is a member of IEEE, IPSJ, JSAI, ITE and IIEEJ.

Naoto Iwahashi received the B.E. degree in engineering fromreceived the B.E. degree in engineering from Keio University in 1985, Yokohama, Japan. He received Ph.D. degree in engineering from Tokyo Institute of Technology in 2001. In 1985, he joined Sony Corporation, Tokyo, Japan. From 1990 to 1993, he stayed at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. From 1998 to 2003, he was with Sony Computer Science Laboratories Inc., Tokyo, Japan. In 2003, he joined Advanced Telecommunications Research Laboratories International. In 2006, he joined National Institute of Information and Communications Technology. His research areas include interactive speech system, language acquisition, human-robot interaction.