Unknown Object Detection Using Multimodal Information Integrated by Kernel Logistic Regression

Yuko OZASA†, Yasuo ARIKI††, Naoto IWAHASHI†††, and Mikio NAKANO††††

† Graduate School of System Informatics, Kobe University, 1–1, Rokkodaicho, Nada-ku, Kobe, Hyogo, 657–8501, Japan
†† National Institute of Information and Communications Technology, Keihanna Research Laboratories, 3–5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619–0289, Japan
††† Honda Research Institute Japan Co., Ltd., 8–1 Honcho, Wako-shi, Saitama 351–0188, Japan
E-mail: †yozasa@stu.kobe-u.ac.jp, ††ariki@kobe-u.ac.jp, †††naoto.iwahashi@gmail.com, ††††nakano@jp.honda-ri.com

Abstract This paper presents a new method to detect unknown objects and their unknown names in object manipulation through man-robot dialog. In the method, the detection is carried out by using the information of object images and user’s speech in an integrated way. Originality of the method is to use kernel logistic regression and multiclass logistic regression for the discrimination between unknown and known objects. From the experimental results, the method using the kernel logistic regression is most efficient in the experiments.

Key words unknown detection, multimodal, logistic regression

1. Introduction

The image features and names of the objects are associated when humans memorize the objects. The object name is represented as speech features spoken in the respective language. Therefore, image features and speech are memorized at the same time.

When the object is recognized by computers, there is an ambiguity that the object is recognized as the different object with the similar image or speech features when recognized using the image or speech features separately. Since the image and speech features exist associatively, if they are used simultaneously, the ambiguity of object recognition will be dissolved and the recognition accuracy can be improved.

In a task where a robot brings a requested object to a human, the speech feature spoken by the human and image feature of the objects the robot watches can be utilized simultaneously. The most powerful method to integrate the features is the logistic regression method using feature confidences. If the integrated confidence measures are higher a certain threshold, the robot can bring the object correctly. Otherwise, the robot does not know the object, namely, it is unknown object to the robot.

We have already proposed an object recognition method using integrated confidence measure of image and speech, and the unknown object detection method by the extension of the object recognition method in [1]. The task which we set up has several objects on a table, and the user tells the robot "bring me ⟨object name⟩." as shown in Fig.1, whether the robot knows the objects or not. 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. Then, the robot brings the object indicated, no matter whether the objects are known or not.

In this paper, we extend the method in [1] by employing the kernel logistic regression and multiclass logistic regression instead of the simple logistic regression, and compare them. Two types of the two class logistic regression are used for multiple unknown object detection.
in [1]. The proposed methods use two types of the two class kernel logistic regression and three class kernel logistic regression in this paper.

2. Proposed System

The proposed system diagram is shown in Fig. 2. It is composed of two parts, estimating confidence and detecting unknown objects and their names. The proposed method for unknown object detection uses both image and speech information in an integrated way. The confidence of the recognition results for input speeches and images are estimated. Then, the confidences are integrated via logistic regression and are detected by thresholding the integrated confidence.

3. 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 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 [2]. The log likelihood of HMMs were calculated by these features and written as follows:

$$P_s(s; \Lambda_i) = \log P(s; \Lambda_i)$$

where \(P(s; \Lambda_i)\) is the likelihood of speech and \(\Lambda_i\) denotes the word HMM for the name of the \(i\)-th object. This \(P(s; \Lambda_i)\) is used to estimate the confidence. Speech recognition confidence is used to evaluate the reliability of the result of speech recognition and it is obtained by the following formula [3]:

$$C_s(s; \Lambda_i) = \frac{1}{n(s)} \log \frac{P(s; \Lambda_i)}{\max_{u_i} P(s; \Lambda_{u_i})}$$

where \(n(s)\) denotes the number of frames in the input speech, \(\Lambda_i\) denotes the word HMM for the name of the \(i\)-th object, and \(u_i\) denotes the best phoneme sequence.

3.2 Image Processing

The features used in image recognition were L*a*b* components (three dimensions) for the color, complex Fourier coefficients (eight dimensions) of contours for the shape [4], and the area of an object (one dimension). Gaussian Models were learned using these features with MAP adaptation. The log likelihood of object \(P_o(o; g_i)\) is obtained by the following formula [5]:

$$P_o(o; g_i) = \log P(o; g_i)$$

where \(P(o; g_i)\) is the likelihood of the object. The confidence of the objects is written as follows:

$$C_o(o; g_i) = \log \frac{P(o; g_i)}{P_{max}}$$

where \(g_i\) denotes the normal distribution of the \(i\)-th object, and \(P_{max} = ((2\pi)^{d/2} | \sum |^{1/2})^{-1}\) denotes the maximum probability densities of Gaussian functions.

4. Logistic Regression for Modality Integration

The speech and image confidence measures are integrated by the logistic regression. The integrated confidence measures are used for unknown object detection.
4.1 Logistic Regression

The speech recognition confidence measure and object recognition confidence measure are integrated by the following logistic regression function [5]:

\[
F = \frac{1}{1 + \exp(-\alpha^T \mathbf{C})}
\]  

(5)

Here \(\mathbf{C}^T = (1, C_s, C_o)\) are the confidence measures and \(\alpha^T = (\alpha_0, \alpha_1, \alpha_2)\) 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 signal \((C_s(s_j; \Lambda_i), C_o(o_j; g_i))\) and teaching signal \(d_{i,j}\), where \(i\) denotes the model index and \(j\) denotes the sample index for the model \(i\). Thus, the training set \(\mathbf{T}\) contains \((N\) models and \(M\) samples\) data.

\[
\mathbf{T}^{N \times M} = \{C_s(s_j; \Lambda_i), C_o(o_j; g_i), d_{i,j}; i = 1, \ldots, N, j = 1, \ldots, M\}
\]

(6)

where \(d_{i,j}\) is 0 or 1, depending on whether the object is unknown or known. The log likelihood function of the training set by the logistic regression function is written as

\[
l(\alpha) = \sum_{j=1}^{M} \sum_{i=1}^{N} \{d_{i,j} \alpha^T \mathbf{C}_j^i - \log(1 + \exp(\alpha^T \mathbf{C}_j^i))\}
\]

Here \(\mathbf{C}_j^T = (1, C_{s_j}, C_{o_j})\), and \(C_{s_j}\) and \(C_{o_j}\) are \(C_s(s_j; \Lambda_i)\) and \(C_o(o_j; g_i)\) respectively for abbreviation. The weight set \(\alpha\) is optimized by maximum likelihood estimation using Fisher’s scoring algorithm [6].

4.2 Regularized Logistic Regression

Over fitting of the learning of logistic regression is serious problem. To avoid the over fitting, the regularized logistic regression is used. The log likelihood function in regularized logistic regression based on the ridge regression is written as follows [7]:

\[
l_R(\alpha) = \sum_{j=1}^{M} \sum_{i=1}^{N} \{d_{i,j} \alpha^T \mathbf{C}_j^i - \log(1 + \exp(\alpha^T \mathbf{C}_j^i))\} + \frac{\lambda}{2} \| \alpha \|^2
\]

(8)

where \(\lambda\) is the coefficient of the regularized term. The weight set \(\alpha\) is optimized by the same way to that of the logistic regression described in Section 4.1.

4.3 Kernel Logistic Regression

There are linear regression and nonlinear regression. The logistic regression is nonlinear function but the discrimination plane is linear. In this section, let us consider about the logistic function which discrimination plane is nonlinear. One of such logistic regression is kernel logistic regression [7]. Using the basis function, the kernel logistic regression is obtained. In this paper, the Gaussian basis function shown in Eq. (9) is used.

\[
\phi_j(\mathbf{C}) = \exp(-\frac{\| \mathbf{C} - \mu_j \|^2}{2\sigma_j^2})
\]

(9)

where \(\mu_j\) is the center vector of the basis function, and \(s_j\) is the parameter which defines the broadening of basis function. The kernel logistic regression is written as follows:

\[
F_K = \frac{1}{1 + \exp(-\alpha^T \phi(\mathbf{C}))}
\]

(10)

where \(\phi(\mathbf{C})\) is the vector where element is the value of the Gaussian basis function at the \((i, j)\)-th training sample. The log likelihood function in the kernel logistic regression is written as follows:

\[
l_K(\alpha) = \sum_{j=1}^{M} \sum_{i=1}^{N} \{d_{i,j} \alpha^T \phi(\mathbf{C}_j^i) - \log(1 + \exp(\alpha^T \phi(\mathbf{C}_j^i)))\}
\]

(11)

The weight set \(\alpha\) is optimized by the same way to that of the logistic regression.

4.4 Multiclass Logistic Regression

The logistic regression described above is two class logistic regression and that can discriminate the objects into two classes. Multiclass logistic regression discriminates the objects into multiclass [7].

Let us consider about the \(K\) class logistic regression. The \(k\)-th class logistic function is written as follows:

\[
F_{M_{i,j,k}} = \frac{\exp(\alpha_k^T \mathbf{C}_j^i)}{\sum_{k=1}^{K} \exp(\alpha_k^T \mathbf{C}_j^i)}
\]

(12)

Then, the log likelihood function is written as follows:

\[
l_M(\alpha) = \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{k=1}^{K} d_{i,j,k} \log F_{M_{i,j,k}}
\]

(13)

where \(d_{i,j,k}\) is teaching signal, and 0 or 1.
5. Detection of Unknown Objects and Their Names

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

5.1 Detection of Unknown Objects

Fig. 3 shows the joint distribution of speech recognition confidence and image recognition confidence. It indicates that discriminating unknown and known objects would be possible by using both confidences simultaneously. Given a threshold $\delta$, the object is classified as unknown or known.

The logistic regression function $F(C_s, C_o)$ and its family are used for the classification of unknown and known objects. If the following condition is satisfied,

$$\max_i F(C_s(s; \Lambda_i), C_o(o; g_i)) < \delta, \quad (14)$$

the input object is classified as an unknown object, else as a known object.

5.2 Object Recognition

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

$$\hat{i} = \arg \max_i F(C_s(s; \Lambda_i), C_o(o; g_i)) \quad (15)$$

Then, the object name is output.

5.3 Detection of Multiple Unknown Objects and Their Names

5.3.1 The Cases of The Multiple Unknown Object Detection

The method for detecting an unknown object was proposed in Section 5.1. This method can be extended to methods which detect multiple unknown objects and their names.

The proposed method described above uses only one detector trained using a set of pairs of a known image and a known speech with $d_{i,j} = 1$, and a set of pairs of an unknown image and an unknown speech with $d_{i,j} = 0$. Since the input set of pairs are a known image and speech or an unknown image and speech, this case is suitable for the case where only one object on a table. This training method is not robust enough to detect multiple unknown objects.

Let us consider about the cases of the multiple unknown object detection shown in Fig.4 and 5.

(1) In the 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 other objects do not. When the objects do not correspond to the input speech even if the objects are known, the objects are treated as unknown objects for the robot because only one image confidence which corresponds to the input speech becomes the known image confidence and other image confidences become the unknown image confidence. If the sets of pairs of a known image and a known speech and that of an unknown image and a known speech can be discriminated, the robot can bring the target known object.

(2) In the 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 a known image and a known speech and that of an unknown image and a known speech can be discriminated, the robot can bring the target known object.

(3) In the case 4, the object corresponding to the input speech is an unknown object, and the known objects are not corresponding to the input speech. If the sets of pairs of an unknown image and an unknown speech and that of a known image and an unknown speech can be discriminated, the robot can bring the target unknown object.

(4) In the case 5, the object corresponding to the input speech is an unknown object, and other objects are not corresponding to the input speech. If the sets of pairs of an unknown image and an unknown speech and that of a known image and an unknown speech can be discriminated, the robot can narrow down the selections of the target object.

(5) In the case 6, the objects on the table and input speech are unknown. All input sets of pairs are an unknown image and an unknown speech. If the unknown objects are detected, the robot can know all objects on the table are unknown objects.

Considering the all cases, three sets of pairs; 1) unknown image and speech ($C_1$) and 2) an unknown image and a known speech and an unknown speech ($C_2$) and, 3) known image and speech ($C_3$) are needed to be discriminated.

To detect multiple objects, three sets of pairs $C_1$, $C_2$ and $C_3$ are needed to be discriminated, and these sets are described as in Fig. 6. We propose two methods to
discriminate the three sets of pairs. The first method uses two of the two class logistic regression, and the second method uses three class logistic regression for multiple unknown object detection. Using these methods, the robot can execute the task no matter whether the objects are known or not as shown in Fig. 4 and 5.

5.3.2 Proposed Method By Two Class Logistic Regression

In the first proposed method, we use two types of the two class logistic regression. The first logistic regression is trained by the set of pairs of a known image and a known speech confidence measure with $d_{i,j} = 1$, and the set of others with $d_{i,j} = 0$, namely the set of pairs of a known image and a known speech, a known image and an unknown speech, an unknown image and a known speech confidence measure. The second logistic regression is trained by the set of pairs of an unknown image and an unknown speech confidence measure with $d_{i,j} = 1$, and the sets of others with $d_{i,j} = 0$, namely the set of pairs of a known image and a known speech, a known image and an unknown speech, an unknown image and a known speech confidence measure. Using these two logistic regressions, the input sets of pairs of speech and image are classified into three classes, 1) unknown image and speech and 2) an unknown image and a known speech and a known image and unknown speech, and 3) known image and speech.

5.3.3 Proposed Method By Three Class Logistic Regression

Three class logistic regression is used in the second proposed method. The three class logistic regression classifies the input sets of pairs of speech and image into three classes, 1) unknown image and speech and 2) an unknown image and a known speech and a known image and unknown speech, and 3) a known image and unknown speech. Three class logistic regression is trained by using each class sets of pairs of speech and image with $d_{i,j} = 1$ and other class sets of pairs with $d_{i,j} = 0$.

6. Experimental Evaluation

We first evaluated the unknown object detection method, and then evaluated object recognition. The coefficients $\alpha_0$, $\alpha_1$, and $\alpha_2$, and threshold $\delta$ were also 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.7. Fig. 8 shows the samples of 11 images of bear taken from 11 angles. Two types of image datasets, data set 1 and data set 2 were prepared. The data set 1 consists of the images of the objects taken from 5 angles. The data set 2 consists of the images of the objects taken from 11 angles. All utterances were spoken by one speaker.

6.1 Evaluation of Object Recognition

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

The features used in image recognition were $L^*a^*b^*$ components for the color, complex Fourier coefficients of contours for the shape, and the area of an object. The accuracy of object recognition of each feature is shown in Table 1. In Table 1, the accuracies of the object recognition by the image confidence measure, speech confidence measure, and integrated confidence measure by logistic regression are shown. Among the accuracies of the image confidence measure of each feature, $L^*a^*b^*$ components were most efficient in both data sets. The accuracy of the integrated confidence measure is most efficient in Table 1.

6.2 Evaluation of Multiple Unknown Objects Detection Method

The evaluation was also performed by leave-one-out cross validation in this section, too. The data set 2 is used in this experiment. The input sets of the experiment were all combinations namely pairs of known image and speech, a known image and an unknown speech, and an unknown image and a known speech, and unknown image and speech. The accuracy of the method using two types of two class logistic regression and three class logistic regression was compared in this section. The four types of the logistic regression are used in this experiment, logistic regression, regularized logistic regression, kernel logistic regression, and regularized kernel logistic regression.

The experimental result is shown in Table 2. In Table 2, L, RL, KL, and RKL denote the 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.

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

7. Conclusion

Acquiring new knowledge through interactive learning mechanisms is a key ability for robots in a real environment. To acquire new knowledge, the detection and learning of the unknown objects and their names are
Tbl. 1 Accuracy of object recognition for respective feature (%)

<table>
<thead>
<tr>
<th>Features</th>
<th>Image</th>
<th>Speech</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lab</td>
<td>Area</td>
<td>Fourier</td>
</tr>
<tr>
<td>Data set 1</td>
<td>69.6</td>
<td>11.4</td>
<td>29.2</td>
</tr>
<tr>
<td>Data set 2</td>
<td>73.6</td>
<td>14.6</td>
<td>38.0</td>
</tr>
</tbody>
</table>

Tbl. 2 Accuracy of multiple unknown object detection (%)

<table>
<thead>
<tr>
<th>Two types of two class logistic regression</th>
<th>Three class logistic regression</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>L RL KL RKL</td>
<td>L RL KL RKL</td>
<td>L RL KL RKL</td>
</tr>
<tr>
<td>82.3 85.8 92.4</td>
<td>97.6 88.2 90.1 91.5</td>
<td>98.0 95.6</td>
</tr>
</tbody>
</table>

needed. The proposed method makes it possible for the robot to detect unknown objects and their names online using multimodal information. From the experimental result, it was shown that the regularized kernel logistic regression was most efficient. We will pursue a method for learning unknown objects in a real environment.

References