2012 International Workshop on Nonlinear Circuits,
Communications and Signal Processing
NCSP'12, Honolulu, Hawaii, USA. March 4-6, 2012

Integrated Multimodal Information for Detection of Unknown Objects and Unknown Names

Yuko Ozasa†, Naoto Iwahasi‡, Tetsuya Takiguchi†, Yasuo Ariki†, and Mikio Nakano§

†Graduate School of System Informatics, Kobe University, 1-1, Rokkodaicho, Nada-ku, Kobe, Hyogo, 657-8501, Japan
Email: {y_ozasa, takigu, ariki} @ stu.kobe-u.ac.jp

‡National Institute of Information and Communications Technology, Keihanna Research Laboratories, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
Email naoto.iwahashi@gmail.com

§Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan
Email nakano@jp.honda-ri.com

Abstract

This paper presents a new method for detecting unknown objects and their unknown names in object manipulation 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 logistic regression for the discrimination between unknown and known objects. The detection accuracy of an unknown object and its name was 97% in the case when there were about fifty known objects.

1. Introduction

The robots can recognize objects using vision with reasonable accuracy if they know those object in advance, recently. However, it is difficult to teach household robots every objects in home environments. So, robots need to learn unknown objects as well as recognize known objects. Few researchers have previously addressed such systems [1, 2, 3]. In [1], the author developed a robot that acquires multimodal information of objects, i.e. auditory, visual, and haptic information, in a fully autonomous way using its embodiment. Also an online algorithm of multimodal categorization based on the acquired multimodal information and words which are partially given by human users has been proposed. The authors summarize their ongoing project on developing an architecture for a robot that can acquire new words and their meanings while engaging in multidomain dialogues in [2, 3].

There remains a problem of how to detect unknown objects. We propose a new method that uses multimodal information, which is integrated speech and image information, for the classification of known and unknown objects. We consider a task in which a robot is told "bring me *Object Name* on the table." (Fig. 1.) From the information of the objects on the table and human speech, the robot brings the object indicated whether the objects are known or not. In this method, not only image information but also speech information is used.

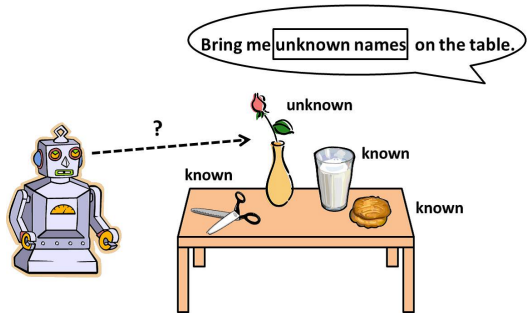

Figure 1: Autonomous Detection of Unknown Objects and Their Names by a Robot.

To use both types of information, we use logistic regression to integrate the information.

2. Proposed System

The proposed system diagram is shown in Fig. 2. It is composed of two parts, estimating the confidence and detecting unknown objects and their names. As for the confidence estimation, the confidence of the recognition results for input speeches and images is estimated. Regarding the detection of unknown objects and their names, the input object is classified into an unknown object categories and known object category using the confidence. When the input object is classified as unknown, the robot considers that an unknown object is detected, and its name is obtained. When the input object is classified as a known object, its object ID is estimated and then its name is output. The detail of the confidence estimation and the unknown object detection will be described in Sections 3 and 4, respectively.

3. Confidence Measure Integration

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

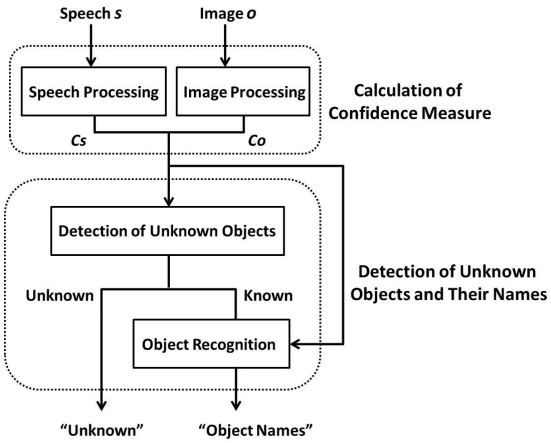

Figure 2: Proposed System Configuration Diagram

3.1. Speech Processing

The features used for speech recognition were Mel-frequency cepstral coefficients, which are based on short-time spectrum analysis; their delta and acceleration parameters; and the delta of short-time log power. These features are obtained by Julius [4]. The log likelihood of these features are calculated by HMMs and written as follows:

$$P_s(s;\Lambda_i) = \log P(s;\Lambda_i) \tag{1}$$

where $P(s;\Lambda_i)$ is the likelihood of speech. This $P(s;\Lambda_i)$ is used to estimate confidence. Speech recognition confidence is used to evaluate the reliability of the result of speech recognition and it is obtained by the following formula [5]:

$$C_s(s;\Lambda_i) = \frac{1}{n(s)} \log \frac{P(s;\Lambda_i)}{\max\limits_{u_i} P(s;\Lambda_{u_i})} \tag{2}$$

where $n(s)$ denotes the analysis frame length of the input speech, Λ_i denotes the word HMM of the i-th object, and u_i denotes a phoneme sequence of the i-th object.

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 [6], 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 [7]:

$$P_o(o;g_i) = \log P(o;g_i) \tag{3}$$

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

$$C_o(o;g_i) = \log \frac{P(o;g_i)}{P_{max}} \tag{4}$$

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

3.3. Logistic Regression for Modality Integration

The speech recognition confidence measure and object recognition confidence measure are integrated by the following logistic regression function [7]:

$$F_c(C_s, C_o) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 C_s + \alpha_2 C_o)}}. \tag{5}$$

Here α_0, α_1 and α_2 are logistic regression coefficients. In the training of this logistic regression function, the i-th training sample is given as the pair of input signal $(C_s(s;\Lambda_i), C_i(o;g_i))$ and teaching signal d_i. Thus, the training set T contains N samples:

$$\mathrm{T}^N = \{C_s(s_j;\Lambda_i), C_o(o_j;g_i), d_i | i = 1, \cdots, N\} \tag{6}$$

where d_i is 0 or 1, which respectively represents the object is unknown or known. The likelihood function is written as

$$P(\mathbf{d}|\alpha_0,\alpha_1,\alpha_2) = \prod_{j=1}^{M}\prod_{i=1}^{N} (F_c(C^i_{s_j}, C^i_{o_j}))^{d_{i,j}} (1 - F_c(C^i_{s_j}, C^i_{o_j}))^{1-d_{i,j}} \tag{7}$$

where $\mathbf{d} = (d_{i,j}, \cdots, d_{N,j})$. The weights $(\alpha_0, \alpha_1, \alpha_2)$ are optimized by maximum likelihood estimation using Fisher's scoring algorithm [8].

4. 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 obtained from Section 4.3. When the input object is classified as unknown, it is considered an unknown object is detected and its name is obtained. When the input object is classified as known, then the object recognized to get its name is output.

4.1. 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 with these confidences using simultaneous use of both confidences. Given a threshold δ, the object is classified as unknown or known.

$F_c(C_s, C_o)$ is used for the classification of unknown and known objects. If

$$\max_i (F_c(C_s(s;\Lambda_i), C_o(o;g_i))) < \delta, \tag{8}$$

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

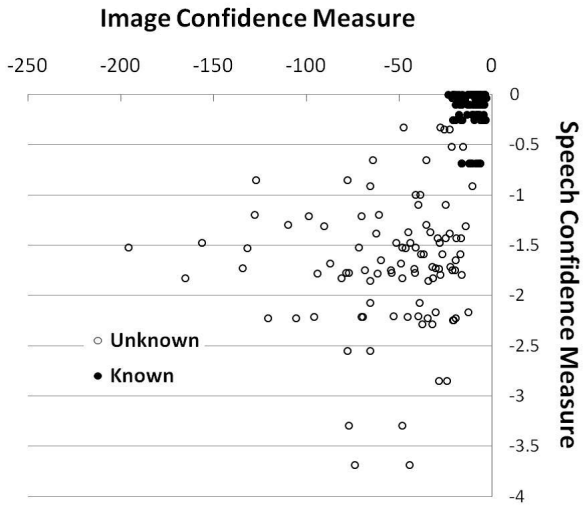

Figure 3: Joint Distribution of Values of the Speech and Object Confidence

Figure 4: Examples of Object Image Used in the Experiment

4.2. Object Recognition

When the input object is classified as a known object, the object is recognized and its ID is obtained. The ID of an object is obtained as follows:

$$\hat{i} = \arg\max_{i}(F_C(C_s(s;\Lambda_i), C_o(o;g_i))) \quad (9)$$

Then, the object name is output.

5. Experimental Evaluation

We first evaluated unknown object detection, and then we evaluated object recognition. The coefficients α_0, α_1, and α_2, and threshold δ were also optimized in the experiment.

We prepared 50 objects. For each object, we collected one utterance including its name and 10 images. Some of the images are shown in Fig. 4. All utterances are made by one speaker.

5.1. Evaluation of Unknown Object Detection

The evaluation is performed by leave-one-out cross validation. We investigated (1) if known objects are classified as known objects and then (2) if unknown objects are classified as unknown objects, and averaged their accuracies. For (1), we chose one image for each of the 50 objects as a test data, and other images are treated as training data. We carried out the experiment for all 500 images. For (2) we chose one object for testing, and other objects were treated as training data. We also carried out the experiments for each of the 500 images.

To evaluate the confidences, we compared the accuracies of the proposed method using the confidences and the method using the log likelihood. The latter uses a measuere

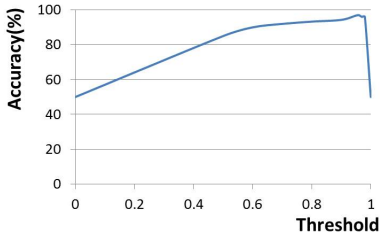

Figure 5: The Variation in Accuracy by Threshold ($F_c(C_s, C_o)$)

that integrates log likelihoods of image and speech recognition by logistic regression. The coefficient set $\{\alpha_0, \alpha_1, \alpha_2\}$ are $\{7.64, 5.22, 5.16e-03\}$ in the proposed method and $\{9.17, 0.02, 0.15\}$ in the log likelihood method. For every cross validation, we evaluate the accuracy with one threshold. The variation in accuracy by the threshold are shown in Fig. 5 and 6. The optimized threshold δ of the proposed method is 0.96, and the threshold of the log likelihood based method is 0.98. The experimental result using the optimized weight set is shown in Table 1. The accuracy of the proposed method is 7.6% higher than that of the method which uses the log likelihood integrated by logistic regression $F_p(P_s, P_o)$ and the most efficient as shown in Table 1.

5.2. Evaluation of Object Recognition

The evaluation was also performed by leave-one-out cross validation. As the condition that unknown object is input, we chose one image for testing for each of the 50 objects, and other images are treated as training data. We carried out the experiment for all 500 images. To evaluate the confidence, we compared the accuracy of the proposed method using the confidence measure and the method using the log likelihood. The

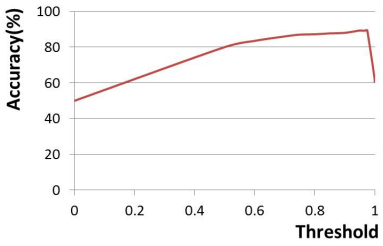

Figure 6: The Variation in Accuracy by Threshold ($F_p(P_s, P_o)$)

Likelihood		$P_s + P_o$	$F_p(P_s, P_o)$
Object P_o	93.20%	78.80%	89.40%
Speech P_s	66.00%		

Confidence		$C_s + C_o$	$F_c(C_s, C_o)$
Object C_o	93.20%	94.60%	97.00%
Speech C_s	95.00%		

Table 1: Accuracy of Unknown Object Detection

same weight sets in Section 6.1 are used in this experiment. The experimental result is shown in Table 2. The accuracy of the proposed method and the method using the log likelihood is the same and it is 100%.

6. Discussion

We detect an unknown objects and its names as preliminary experiment. This method can be extended to the method which detects multiple unknown objects and their names. From the experimental result, we can see the possibility for the extension of the proposed method for multiple objects. In future work, we extend the proposed method to that for the multiple objects and their names.

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 needed. The proposed method makes it possible for the robot to detect unknown objects and their names online using the multimodal information. We will extend the proposed method for the detection of multiple unknown objects and also pursue a method for learning unknown objects in a real environment.

Likelihood		$P_s + P_o$	$F_p(P_s, P_o)$
Object P_o	98.80%	99.40%	100.00%
Speech P_s	96.00%		

Confidence		$C_s + C_o$	$F_c(C_s, C_o)$
Object C_o	98.80%	99.40%	100.00%
Speech C_s	96.00%		

Table 2: Accuracy of Object Recognition

References

[1] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, N. Iwahashi, "Autonomous Acquisition of Multimodal Information for Online Object Concept Formation by Robots," *IEEE International Conference on Intelligent Robots and Systems*, 2011.

[2] H. Holzapfel, D. Neubig, and A. Waibel, "A Dialogue Approach to Learning Object Descriptions and Semantic Categories," *Robotics and Autonomous Systems*, vol.56, Issue.11, pp.1004-1013, 2008..

[3] M. Nakano, N. Iwahashi, T. Nagai, T. Sumii, X. Zuo, R. Taguchi, T. Nose, A. Mizutani, T. Nakamura, M. Attamimi, H. Narimatsu, K. Funakoshi, and Y. Hasegawa, "Grounding New Words on The Physical World in Multi-Domain Human-Robot Dialogues," *Dialog with Robots: Papers from the AAAI Fall Symposium*, 2010.

[4] Julius, http://julius.sourceforge.jp/.

[5] H. Jiang, "Confidence Measures for Speech Recognition: A survey," *Speech Communication*, vol. 45, pp. 455-470, 2005.

[6] E. Persoon and K.S. Fu, "Shape Discrimination Using Fourier Descriptors," *IEEE Trans. Accoust. Speech Signal Processing*, vol. 28, no. 4, pp. 170-179, 1977.

[7] 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," *Journal of Artificial Intelligence*, vol. 25, no. 25, pp. 670-682, 2010.

[8] T. Kurita, "Interactive Weighted Least Squares Algorithms for Neural Networks Classifiers," in *Proc. Workshop on Algorithmic Learning Theory*, 1992.