Object Recognition by Integrated Information Using Web Images

Hitoshi Nishimura*, Yuko Ozasa†, Yasuo Ariki‡, Mikio Nakano§
*Graduate School of System Informatics, Kobe University, Nada, Kobe 657-8501, Japan
hnishimura@me.cs.scitec.kobe-u.ac.jp
†Graduate School of System Informatics, Kobe University, Nada, Kobe 657-8501, Japan
y_ozasa@stu.kobe-u.ac.jp
‡Organization of Advanced Science and Technology, Kobe University, Nada, Kobe 657-8501, Japan
ariki@kobe-u.ac.jp
§Honda Research Institute Japan Co., Ltd., 8–1 Honcho, Wako-shi, Saitama 351–0188, Japan
nakano@jp.honda-ri.com

***Abstract*—It is an important task for a robot to bring objects requested by human via voice. In order to achieve the task, object recognition using speech and images is needed. Ozasa et al. has proposed the method for the object recognition by integrating speech and image information. Although this method requires both speech (word) and image models, the speech models are automatically constructed by combining phonemic acoustic models according to the dictionary. However, the image models have to be constructed manually in advance. In this paper, instead of the manual construction of the image models, we propose an automatic image model construction method for object recognition using Web images. The effectiveness of the proposed method is verified in the object recognition by integrating speech and image information.**

***Keywords*-object recognition; Web; integration; speech; image;**

I. Introduction and Related Work

Many countries around the world face a problem of aging population. A household robot plays an important role in an aging society, especially for the purpose of a senior assist. Yamazaki *et al.* [1] described a home-assistant robot in detail. It is essential that the home-assistant robot communicates with humans. Considering the coexistence of humans and robots, the natural communication is important. The robot needs to learn objects in a home environment in order to interact in communication with humans. Several researchers have studied leaning objects through interaction [2], [3], [4], [5], [6]. Ozasa *et al.* [6] have proposed the method for object recognition using integrated information in order to learn unknown objects through the object manipulation task that human makes a robot bring an object by human voice. In this paper, we deal with the same task and aim at learning objects in a natural way. To achieve the task, it is necessary to recognize the object name that the human speaks and the captured image of the object in front of the robot. Ozasa *et al.* performed object recognition using integrated information of speech and image. All sets of pairs of the speech and image model of the objects in the home environment are learned previously and the recognition is performed by these models.

There are two problems in their method. The first problem is about the learning of the image models. While the speech models are automatically constructed by combining phonemic acoustic models according to the dictionary, the image models have to be constructed manually in advance. To take pictures of all the objects in the home environment is unrealistic. The second problem is about the number of the models used for the recognition. When the number of the objects is increased, the object recognition becomes difficult since they use all the models prepared previously, including quite similar models in them. The number of the candidates of object recognition is increased with the number of the models used for the recognition. To solve these problems, we use Web images for the image models and perform the object recognition by the integrated information of the speech and image recognition. Since the preliminary selection of the candidates are carried out by the speech recognition results, all of the models are not used for the recognition.

Web images are different from the data sets collected manually like Caltech-101 [7] and includes many irrelevant images. We call these images *noise*. There are some works of the image recognition using Web images [8], [9], [10]. In [8], large amount of Web images are collected and the reranking of the images are performed in order to remove the noise. However, considering real-time learning, the recognition using the small amount of images is needed. In [9] and [10], the methods using k-Nearest Neighbors (kNN) are proposed. They show that the image recognition by Web images is possible when they collect enough number of Web images for the recognition even if the noise is included in the images. Based on this concept, in this paper, we use the kNN for the construction of the image models. Using this, the computational cost of the recognition is decreased. In kNN, since each training data is directly used as a prototype, the computational cost of training is lower than other methods.

978-1-4799-2190-4/13 $26.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.140

There are three contributions in our method.

1) By collecting images from Web, it is possible to construct image models automatically.
2) The proposed method can reduce the computational cost because it carries out the image recognition using only P image models selected by speech recognition result among Q total image models, where $2 \leq P < Q$. The method proposed by Ozasa *et al.* must search all models in the dictionary. Since their method set up 50 object recognition, it was computationally feasible. However, the object recognition becomes computationally infeasible if the number of candidates increases.
3) The ambiguity of speech recognition can be reduced by image recognition since the method selects the top P candidates of the speech recognition results and confirms the most likely candidate by the image recognition using the corresponding P image models. This indicates that even when the true one is not the top candidate, it can be pushed up to the top candidate owing to the image recognition by the corresponding model at the integration stage.

II. Proposed Method

We assume that there are several objects in front of a robot, and a human tells the robot "bring me <object name>". The number of the image models needed for the recognition is narrowed down by speech recognition, and object images corresponding to speech recognition results are collected from Web. Then, the object image models are constructed by the collected images for each object, and the object recognition is performed by integrating the speech and image recognition results. The proposed system configuration diagram is shown in Fig. 1, and the procedure is described as follows.

1) When the speech of the object name is given by a human, the robot performs speech recognition by HMM, then the top P object candidates and their confidences are obtained from the recognition results.
2) The robot checks whether it has all image models of the top P candidates in the image model database or not. If not, the robot gets the object images from Web for each object class, and constructs the image models using the images.
3) The robot calculates the image confidences of the input image using image models.
4) The robot integrates the confidences of speech and image recognition by a logistic regression, and performs the object recognition by integrated confidences.

A. Image Processing

In this paper, we use Scale Invariant Feature Transform (SIFT) [11] for local features which are invariant to image translation, scaling and rotation. Bag of Features (BoF) [12],

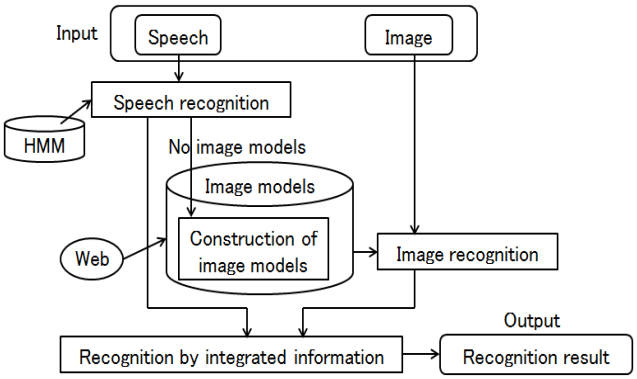

Figure 1: Proposed system configuration diagram

Sparse Coding SPM (ScSPM) [13], Locality-constrained Linear Coding (LLC) [14] have been proposed and their methods represent the image using local features according to a codebook. Except for BoF, these methods use Spatial Pyramid Matching (SPM) [15]. SPM method partitions the image into increasingly fine sub-regions and computes histograms of local features found inside each sub-region. The spatial pyramid framework also offers insights into the success of several recently proposed image descriptions. Among them, we use LLC which is the most efficient coding method, and image recognition is performed by kNN. The histogram obtained by LLC is used as the prototypes of kNN. These prototypes are stored as the image models. In the image recognition, k of the prototypes closest in distance to the input histogram are chosen. Let k_i be the number of histograms among the k nearest neighbors (to input histogram v), that belong to the image model o_i, where i denotes the index of the object. Then the image confidence $C_v(v; o_i)$ is given by the following formula [16]:

$$C_v(v; o_i) = \frac{k_i}{k} \tag{1}$$

The result of the image recognition is obtained as follows:

$$\hat{i} = \arg\max_i C_v(v; o_i). \tag{2}$$

B. Object Recognition by Integrated Information

The proposed system integrates the confidences of speech recognition results and image recognition results, and the integrated confidence is used in the object recognition. Speech features are Mel-frequency cepstral coefficients (MFCC), their delta and log power. MFCC is based on short-time spectrum analysis. Speech recognition confidence is used to evaluate the reliability of the result of speech recognition and it is obtained by the following formula [17]:

$$C_s(s; \Lambda_i) = \frac{1}{n(s)} \log \frac{P(s; \Lambda_i)}{\max_{u_i \neq i} P(s; \Lambda_{u_i})}. \tag{3}$$

where $P(s; \Lambda_i)$ is the likelihood of a speech s and Λ_i denotes the word HMM for the name of the i-th object. This

$P(s;\Lambda_i)$ is used to estimate the confidence. $n(s)$ denotes the number of frames in the input speech, and u_i denotes the best phoneme sequence.

The speech recognition confidences $C_s(s;\Lambda_i)$ and image recognition confidences $C_v(v;o_i)$ are integrated by the following logistic regression [18]:

$$F(C_s, C_v) = \frac{1}{1+e^{-(\alpha_0+\alpha_1 C_s+\alpha_2 C_v)}}. \quad (4)$$

where s and v indicate the input speech and image respectively. Λ_i and o_i denote the i-th speech and image models. i denotes an object index. In the training of this logistic regression function, the (i,j)-th training sample is given as the pair of input signal $(C_s(s_j;\Lambda_i), C_v(v_j;o_i))$ and teaching signal $d_{i,j}$, where i denotes the object model index, and j denotes the sample index for the object model i. Thus, the training set T contains (N models and M samples) data.

$$\mathrm{T}^{N\times M} = \{C_s(s_j;\Lambda_i), C_v(v_j;o_i), d_{i,j} \mid i=1,\cdots,N\ , j=1,\cdots,M\} \quad (5)$$

If the j-th speech s_j and j-th image v_j are related to the object "i" so that their speech and image models are Λ_i and o_i, and in this case, the supervising signal is set to 1. Otherwise it is set to 0. The log likelihood function is described as

$$P(\mathbf{d}|\alpha_0,\alpha_1,\alpha_2) = \prod_{j=1}^{M}\prod_{i=1}^{N}\left(F(C_s(s_j;\Lambda_i), C_v(v_j;o_i))\right)^{d_{i,j}} \left(1-F(C_s(s_j;\Lambda_i), C_v(v_j;o_i))\right)^{1-d_{i,j}} \quad (6)$$

where $\mathbf{d} = (d_{1,1},\cdots,d_{N,M})$. α_0, α_1, α_2 are optimized by maximum likelihood estimation using Fisher's scoring algorithm [19].

The result of the object recognition is obtained as follows:

$$\hat{i} = \arg\max_i F(C_s(s;\Lambda_i), C_v(v;o_i)). \quad (7)$$

III. Experiment

We conducted experiments with images collected from Web in order to ensure that our system works in practice. The experimental procedure and condition are described as follows.

First, speaker-independent isolated word recognition was performed using Julius software [20]. As an acoustic model, we used the speaker-independent PTM triphone HMM (Hidden Malkov Model). The acoustic model is trained using the JNAS [21] speech data. The feature vector is composed of 12-dimensional MFCC, their delta and log energy. Speakers spoke 100 words randomly selected from the speech dictionary (1000 words). The number of speakers was set to 8. One set of them was used for training the logistic regression parameters and the remaining was used for test data.

Table I: An accuracy of the object recognition using Web images and Caltech-101 (%)

	Preliminary Selection (Speech)	Image	Integration
(a) Web	94.00	41.71	94.33
(b) Caltech-101	85.00	71.00	94.00

Next, image recognition was performed using the image models selected by the result of spoken word recognition. We used two software applications, ImageSpider [22] and ImageGeter [23] to collect Web images, and used Scikit-learn software [24] for kNN. 210 images were collected using these software application for each object. 10 images were used for test data and the remaining was used for training data for each object. We resized the images to be less than 150×150 pixels with fixed aspect ratio. The SIFT features were extracted from patches densely located at every 10 pixels on the image, under three scales, 1×1, 2×2, 4×4 respectively. We trained a codebook with 100 bases by k-means [25]. The other parameters in LLC were set the same as the J. Wang's implementation [14]. For image model parameters, the value of k was decided as 3 based on the preliminary experiments, where k denotes the value for kNN. By preliminary selection, only top 10 candidates are used for the object recognition, so the image recognition result is assumed as a failure if correct object is ranking below top 10 in speech recognition.

Finally, object recognition was performed by logistic regression, integrating the scores obtained from speech recognition and image recognition. A parameter in logistic regression shown in Eq.4 was estimated commonly for all the objects using one speech data and 10 images taken from each object. 10 images were used in order to improve the robustness to variety of input images.

A. Comparison between Image Dataset Collected from Web and Caltech-101

The experiment was carried out to compare the image databases. One is (a) Web images we collected and the other is (b) Caltech-101. In (a), 1000 objects supposed to be in houses were chosen and their names were stored in the speech dictionary, and their images were collected from Web. In (b), the 100 object images included in Caltech-101 were selected and their names were stored in the speech dictionary. In (a), the object images corresponding to top 10 candidates of speech recognition were collected from Web. In (b), instead of collecting images from Web, the Caltech-101 images were directly used. Table I shows the result. In (a), an overall average accuracy of the object recognition using integrated information is 52.62 points higher than that by image information. In (b), it is 23.00 points. It can be seen from this result that the integrating information is effective even when image models are constructed by Web images. In

Table II: A Computational cost of each process with the preliminary selection (sec)

(1) Extracting SIFT descriptors	180.0503
(2) Coding descriptors in LLC	26.9253
(3) Training kNN	0.6023
(4) Recognition by kNN	0.0004
Total	207.57

terms of the image recognition, even if the object recognition fails by the image recognition, it can succeed owing to the help of speech recognition confidences.

B. Reduction of the Object Recognition Cost by Speech Processing

In the experiment, the efficiency of the object recognition with the preliminary selection by speech recognition was verified. We used a 3.5GHz Intel(R) Core(TM) i7 CPU machine with 32GB of RAM. TABLE II shows the overall average computational cost for one image recognition process with the preliminary selection of the candidates. The time for training kNN denotes the time for making a ball-tree data structure [26]. Ball-tree is tree-based data structures for efficient neighbors searches and is efficient in higher dimensions. The total computational cost from (1) to (4) in 1000 object recognition becomes 100 times as high as that in 10 object recognition. The total computational cost observed actually was 207.57 seconds with the preliminary selection (10 object recognition). Thus, the total computational cost is inferred more than 20757 seconds without the preliminary selection (1000 object recognition). The computational cost of the object recognition without the preliminary selection is infeasible in view of the real-time interaction. Since the cost of extracting SIFT descriptors is overrepresented in whole, we should reduce it in the future work.

C. Disambiguation of Speech Recognition by Integrating with Image Recognition Result

The purpose of the experiment is to evaluate the proposed method in terms of the disambiguation of speech recognition. 7000 tests were carried out totally. In 140 of the 7000 tests, correct objects are ranking below top 10 in speech recognition, so other 6860 tests are analyzed. Details of the object recognition by integrated information is shown in TABLE III. The number of the object recognition failures is 31 when the speech recognition is true, and the number of the object recognition successes is 64 even when the speech recognition is false. This result confirms that the proposed method is effective for the disambiguation of speech recognition.

D. Comparison with Other Methods

We compared the kNN with Support Vector Machine (SVM). SVM is one of the most efficient discriminative

Table III: Details on object recognition by integrated information (frequency)

		Object recognition	
		True	False
Speech recognition	True	6479	31
	False	64	286

Table IV: Comparison of the recognition accuracy and computational cost between SVM and kNN

	Image (%)	Integration (%)	Computational cost (sec)
SVM	40.58	94.86	1.058
kNN	41.71	94.33	0.575

model. It is easy to compute and gives superior image classification performance than many existing approaches. In the experiment, one-vs-rest SVMs for each object were trained by the histograms. C was decided as 1 based on the preliminary experiments, where C is the penalty parameter of the error term in SVM. We used Scikit-learn software [24] for SVM. An overall average accuracy and computational cost are shown in TABLE IV. The computational cost is the total time for the training kNN (SVM) and recognition by kNN (SVM). kNN achieved as good object recognition accuracy as SVM, and achieved lower computational cost than SVM.

IV. Discussion

The experimental results suggested that Web images can be used for image models for the object recognition by integrated information. In order to improve the recognition accuracy by integrated information, improvements to speech and image recognition accuracy are needed.

In respect of the speech recognition, the correct object did not rank in the top 10 in some tests at the speech recognition stage. In order to solve this problem, it is necessary for a robot to ask a human to utter the object name again, and correct object rank in the more higher.

In respect of the learning image models, each time object recognition is performed, a robot should be taught whether the recognition successes or not and incrementally reconstructs the image models using the images. Moreover, Web includes variety of images and noise images. For example, “cheek” means cosmetics cheek, teak-wood, cheek brush, and products made from teak-wood, etc. Our database does not take account of a conceptual structure [27]. The conceptual structure represents the semantic relations among words. It is indispensable that a robot learns a conceptual structure since this knowledge can help the object recognition.

V. Conclusion

In this paper, we proposed the object recognition method by collecting the Web images based on the speech recog-

nition, constructing the image models using them and integrating the scores of speech and image recognition. We used kNN for construction of image models, and kNN achieved as a good accuracy as SVM and achieved lower computational cost than SVM. In order to improve the recognition accuracy by integrating the information, the image recognition accuracy needs to be improved. In the future work, to achieve that, we will make a system that can learn object in more natural way and use other new information.

REFERENCES

[1] K. Yamazaki, R. Ueda, S. Nozawa, M. Kojima, K. Okada, K. Matsumoto, M. Ishikawa, I. Shimoyama, and M. Inaba, "Home-Assistant Robot for an Aging Society," at Proceedings of the IEEE, pp. 2429-2441, 2012.

[2] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, 26(1):pp. 113-146, 2002.

[3] 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.

[4] 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.

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

[6] Y. Ozasa, Y. Ariki, M. Nakano, and N. Iwahashi, "Disambiguation in Unknown Object Detection by Integrating Image and Speech Recognition Confidences," in *Proc. ACCV*, 2012.

[7] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," in *IEEE CVPR Workshop on Generative-Model Based Vision*, 2004.

[8] K. Yanai, "Generic Image Classification Using Visual Knowledge on the Web," in *Proc. ACM International Conference Multimedia*, pp. 67-76, 2003.

[9] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *PAMI*, 30(11):pp. 1958-1970, November 2008.

[10] X. Wang, L. Zhang, M. Liu, Y. Li, and W. Ma, "Arista - image search to annotation on billions of web photos," in *Proc. CVPR*, 2010.

[11] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, pp. 1150-1157, 1999.

[12] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Workshop on Statistical Learning in Computer Vision, *ECCV*, pp. 1-22, 2004.

[13] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. CVPR*, 2009.

[14] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained Linear Coding for Image Classification," in *Proc. CVPR*, 2010.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006.

[16] A. F. Atiya, "Estimating the Posterior Probabilities Using the K-Nearest Neighbor Rule," at *Neural Computation*, pp. 731-740, 2005.

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

[18] 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.

[19] T. Kurita, "Interactive Weighted Least Squares Algorithms for Neural Networks Classifiers," in *Proc. Workshop on Algorithmic Learning Theory*, pp. 77-86, 1992.

[20] A. Lee, T. Kawahara and K.Shikano, "Julius - an Open Source Real-Time Large Vocabulary Recognition Engine," in *Proc. EUROSPEECH*, pp. 1691-1694, 2001. http://julius.sourceforge.jp/

[21] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi,, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *JOURNAL-ACOUSTICAL SOCIETY OF JAPAN-ENGLISH EDITION-*, vol. 20, pp. 199-206, 1999.

[22] ImageSpider, http://kurima.sakura.ne.jp/sb/log/eid123.html

[23] ImageGeter, 'http://uwa.potetihouse.com/soft/imagegeter.html

[24] F. Pedregosa, "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, 12, pp. 2825-2830, 2011.

[25] J. HARTIGAN, "A K-means clustering algorithm," *Applied Satistics*, pp. 100-108, 1979.

[26] S. M. Omohundro, "Five balltree construction algorithms," *International Computer Science Institute Technical Report*, TR-89-063, 1989.

[27] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. CVPR*, 2009.