Selection of SIFT Feature Points for Scene Description in Robot Vision

Yuya Utsumi, Masahiro Tsukada, Hirokazu Madokoro, and Kazuhito Sato
Faculty of Systems Science and Technology,
Akita Prefectural University
84–4 Aza Ebinokuchi Tsuchiya, Yurihonjo City, 015–0055 Japan
m11a004@akita-pu.ac.jp

***Abstract*—This paper presents an unsupervised learning-based method for selection of feature points and object category classification without previous setting of the number of categories. Our method consists of the following procedures: 1) detection of feature points and description of features using a Scale-Invariant Feature Transform (SIFT), 2) selection of target feature points using One Class-Support Vector Machines (OC-SVMs), 3) generation of visual words of all SIFT descriptors and histograms in each image of selected feature points using Self-Organizing Maps (SOMs), 4) formation of labels using Adaptive Resonance Theory-2 (ART-2), and 5) creation and classification of categories on a category map of Counter Propagation Networks (CPNs) for visualizing spatial relations between categories. Classification results of static images using a Caltech-256 object category dataset and dynamic images using time-series images obtained using a robot according to movements respectively demonstrate that our method can visualize spatial relations of categories while maintaining time-series characteristics. Moreover, we emphasize the effectiveness of our method for category classification of appearance changes of scenes.**

***Index Terms*—OC-SVMs; SIFT; SOMs; ART-2; CPNs; Unsupervised Category Classification, Robot Vision**

I. Introduction

Because of the advanced progress of computer technologies and machine learning algorithms, generic object recognition has been studied actively in the field of computer vision [1]. Generic object recognition is defined as a capability by which a computer can recognize objects or scenes to their general names in real images with no restrictions, i.e., recognition of category names from objects or scenes in images. In the study of robotics, one method to realize a robot having learning functions to adapt flexibly in various environments is to obtain brain-like memory: so-called World Images (WIs) [2]. For creating WIs, robots must classify objects and scenes in time-series images into categories and memorize them as Long-Term Memory (LTM). Additionally, in real environments for a robot, the number of categories is mostly unknown. Moreover, the categories are not known uniformly. Therefore, a robot must classify while generating additional categories.

Learning-based category classification methods are roughly divisible into supervised category classification methods and unsupervised category classification methods. Supervised category classification methods require training datasets including teaching signals extracted from ground-truth labels. However, unsupervised category classification methods require no teaching signals with which categories are automatically extracted to a problem of unknown classification categories for classifying images into respective categories. Recently, studies of unsupervised category classification methods have been active. The subject has attracted attention because it might provide technologies to classify visual information flexibly in various environments.

In recent studies of category classification, various methods have been proposed to combine the process of detecting regions or positions of an object as a target of classification and recognition. Barnard et al. proposed a word–image translation model as a method based on regions [3]. They automatically annotated segmentation images using images that assigned some keywords previously. Lampert et al. proposed an Efficient Subwindow Search (ESS) that can quickly detect a position of an object using branch and bound methods and integration images [4]. Using ESS, they realized first partial generic object detection to calculate previously output values of Support Vector Machines (SVMs) in each feature point and to localize a search range gradually. Moreover, Suzuki et al. proposed a local feature selection method used in Bag-of-Features (BoF) with SVMs [5]. This method classifies local features into background features and target features used for BoF. However, these methods require previously acquired training samples with teaching signals. Therefore, these methods are inapplicable to a real environment for which a target region and a background region can not be decided uniformly.

This paper presents unsupervised feature selection and category classification for application to a vision-based mobile robot. Our method has the following four capabilities. First, our method can localize target feature points using One Class-Support Vector Machines (OC-SVMs) without previous setting of boundary information. Second, our method can generate labels as a candidate of categories for input images while maintaining stability and plasticity together. Third, automatic labeling of category maps can be realized using labels created using Adaptive Resonance Theory-2 (ART-2) as teaching signals for Counter Propagation Networks (CPNs). Fourth, our method can present the diversity of appearance changes for visualizing spatial relations of each category on a two-dimensional map of CPNs. Through category classification experiments, we evaluate our method using the Caltech-256 object category dataset, which is the $defacto$ standard benchmark dataset for comparing the performance of algorithms in

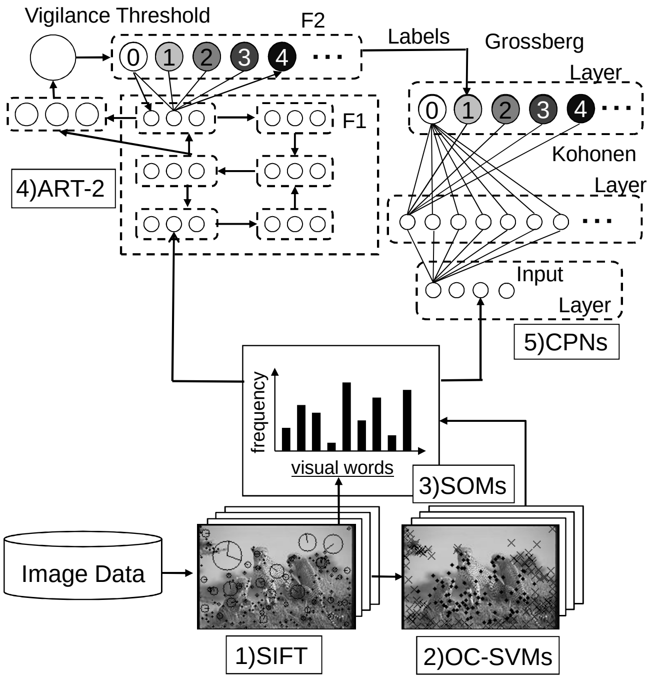


Fig. 1. Network architecture of our method.

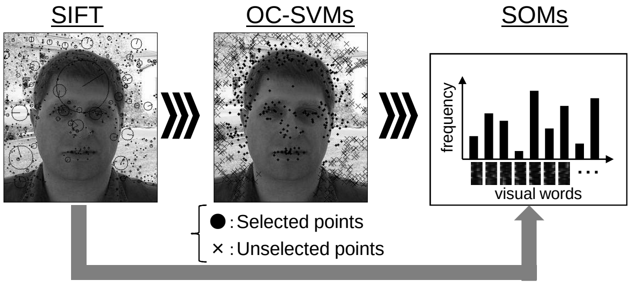


Fig. 2. Procedures of our image representation method based on BoF.

generic object recognition, and time-series images taken by a camera on a mobile robot.

II. Proposed Method

Fig. 1 depicts the network architecture of our method. The procedures are the following.

1) Extracting feature points and calculating descriptors using SIFT
2) Selecting SIFT features using OC-SVMs
3) Creating visual words of all SIFT descriptors and calculating histograms of selected SIFT descriptors matched with visual words using SOMs
4) Generating labels using ART-2
5) Creating a category map using CPNs

TABLE I
Setting values of parameters using experiments.

		Caltech-256	Robot vision
OC-SVMs	ν	0.5	0.82
ART-2	θ	0.1	0.1
	ρ	0.920	0.920
CPNs	$\alpha(t)$	0.5	0.5
	$\beta(t)$	0.5	0.5
	learning iteration	10,000	10,000

The combination of ART-2 and CPNs enables unsupervised category classification that labels a large quantity of images in each category automatically [6]. Actually, ART-2 is a theoretical model of unsupervised neural networks of incremental learning that forms categories adaptively while maintaining stability and plasticity together [7]. Features of time-series images from the mobile robot change with time. Using ART-2, our method enables an unsupervised category classification that requires no setting of the number of categories. A type of supervised neural network, CPN actualizes mapping and labeling together. Such networks comprise three layers: an input layer, a Kohonen layer, and a Grossberg layer [8]. In addition, CPNs learn topological relations of input data for mapping weights between units of the input-Kohonen layers. The resultant category classifications are represented as a category map on the Kohonen layer.

Procedures 1. through 3., which correspond to preprocessing, are based on the representation of BoF. In fact, BoF, which represents features for histograms of visual words with local features as typical patterns extracted from numerous images, is widely used to emphasize the effectiveness in image representation methods of generic object recognition. In BoF of our method, we applied OC-SVMs for selecting SIFT feature points as target regions in an image. Furthermore, we applied SOMs for creating visual words and histograms in each image from selected features. The OC-SVMs are unsupervised-learning-based binary classifiers that enable density estimation without estimating a density function. Therefore, OC-SVMs can apply to a real environment without boundary information. Table I shows parameters of OC-SVMs, ART-2, and CPNs with each experiment. Detailed algorithms of OC-SVMs is the following.

A. Selected feature points using OC-SVMs

As described earlier, the OC-SVMs are unsupervised learning classifiers that estimate the dense region without estimation of the density function. The OC-SVMs set a hyperplane that separates data points near the original point and the other data points using the characteristic by which the outlier data points are mapped near the original point on a feature space with a kernel function. The discriminant function is calculated to divide input feature spaces into two parts. The position of the hyperplane is changed according to parameter ν, which controls outliers of input data with change, and which has range of 0–1.

$$f(x) = sgn(\omega^{\top}\Phi(x) - \rho) \qquad (1)$$

The restriction is set to the following.

$$\begin{gathered} \omega^{\top} z_i \geq \rho - \zeta_i, i = 1, ..., l \\ \zeta_i \geq 0, i = 1, ..., l, 0 < \nu \leq 1 \end{gathered} \qquad (2)$$

The optimization problem is solved with the following restriction

$$\frac{1}{2}\|\omega\|^2+\frac{1}{\nu l}\sum_{i=1}^{l}\quad\zeta_i\quad-\rho \\ \rightarrow\quad \min \omega,\zeta,\text{and } \rho \tag{3}$$

Therein, z_{i} represents results of the mapping input vector x_{i} to the high-dimension feature space.

$$\varPhi : x_i \mapsto z_i \tag{4}$$

In those expressions, ω and ρ are results of the optimization problem. The Lagrangian function of the optimization problem is calculated to solve the optimization problem.

$$L(\omega,\zeta,\rho,\alpha,\beta)=\frac{1}{2}\|\omega\|^2+\frac{1}{\nu l}\sum_{i=1}^{l}\zeta_i-\rho \\ -\sum_{i=1}^{l}\alpha_i((\omega^\top z_i)-\rho+\zeta_i)-\sum_{i=1}^{l}\beta_i\zeta_i \tag{5}$$

In those expressions, α and β of the Lagrangian function are maximized. Ω, ρ and ζ of the Lagrangian function are minimized. Lagrangian functions that are partially differentiated by ω, b, ρ and ζ are 0 for an optimized solution.

$$\frac{\partial}{\partial\omega}L=0\rightarrow\omega=\sum_{i=1}^{l}\alpha_i z_i \tag{6}$$

$$\frac{\partial}{\partial\zeta_i}L=0\rightarrow\alpha_i=\frac{1}{\nu l}-\beta_i \tag{7}$$

$$\frac{\partial}{\partial\rho}L=0\rightarrow\sum_{i=1}^{l}\alpha_i=1 \tag{8}$$

$$\left.\begin{array}{l} \alpha_i\cdot[\rho-\zeta_i-\omega^\top z_i]=0,\quad i=1,...,l \\ \rho-\zeta_i-\omega^\top z_i\leq 0,\quad i=1,...,l \\ 0\leq\alpha_i\leq\frac{1}{\nu l},\quad i=1....,l \\ \beta_i\cdot\zeta_i=0,\quad -\zeta_i\leq 0,\quad \beta_i\geq 0,\quad i=1,...,l \end{array}\right\} \tag{9}$$

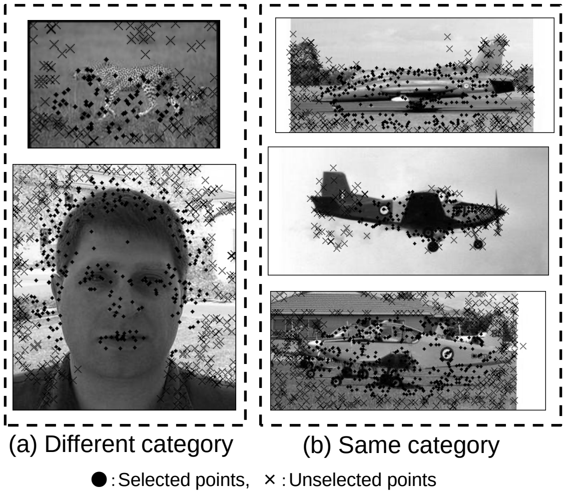
Equations (6)–(9) are substituted to Lagrangian function. A binary optimization problem is developed if the inner product is transposed to the kernel.

$$\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j k(z_i^\top z_j), \\ 0\leq\alpha_i\leq\frac{1}{\nu l}, i=1,...,l,\sum_{i=1}^{l}\alpha_i=1 \tag{10}$$

Support vectors are learning data z_{i} fulfilling assumptions of (9) , $\alpha_{\mathrm{i}} > 0$ and ζ_{i}=0. The equation (6) is expanded. An equality is true if α_{i} and β_{i} are not 0 for an optimized solution and ρ is calculated as

$$f(z)=\sum_{i=1}^{l}\alpha_i k(x_i,z)-\rho, \tag{11}$$

where $\zeta_{\mathrm{i}}=0$. Points of $\varPhi(x)$ are not apparent in the discriminant function that is a binary problem using a kernel trick. Therefore, huge calculation costs of the inner product can be avoided and the number of calculations can be reduced. Parameter ν of OC-SVMs is a high limit of unselected data and lower limit of support vectors if the solution of the optimization problem (3) fulfills $\rho \neq$0.


(a) Different category

(b) Same category

Fig. 3. Results of selected SIFT feature on two sample images in different category and three sample images in the same category of Caltech-256.

III. Experimental results

This section presents experiment results obtained using two datasets: the Caltech-256 object dataset, which is the $defacto$ standard dataset for object recognition, and our original time-series image dataset taken using a mobile robot. We show that selecting feature points with OC-SVMs is efficient for category classification using these two datasets.

A. Classification results of caltech-256

This section presents experimental results of image classification using Caltech-256, which is the $defacto$ standard benchmark dataset, to compare the performance of algorithms in generic object recognition. The target of this experiment is category classification of static images because Caltech-256 has no temporal factors in each category. We use the highest 20 categories with the number of images in 256 categories.

Fig. 3 depicts results of selected feature points using OC-SVMs on five sample images of Caltech-256. Fig. 3(a) shows that our method can select feature points of target objects in images of the Leopards and Face categories. In addition, Fig. 3(b) shows that our method can select feature points around the wings that characterize airplanes for various images of the Airplane category. Fig. 4 depicts labels generated by ART-2. The vertical and horizontal axes respectively represent labels and images. The independent labels in each category without confusion are generated among different categories. Moreover, for the Airplane, Motorbike, and Face categories one label is generated; for the Car-side and Leopards categories

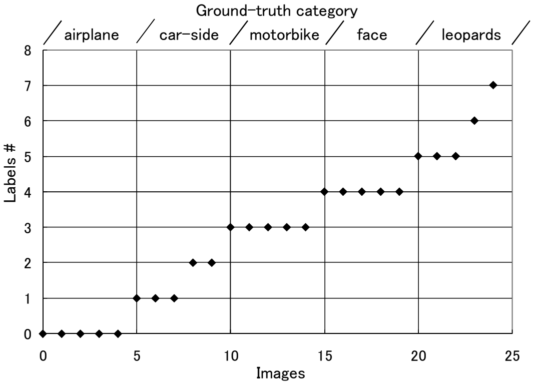


Fig. 4. Results of formed labels using ART-2 at five categories.

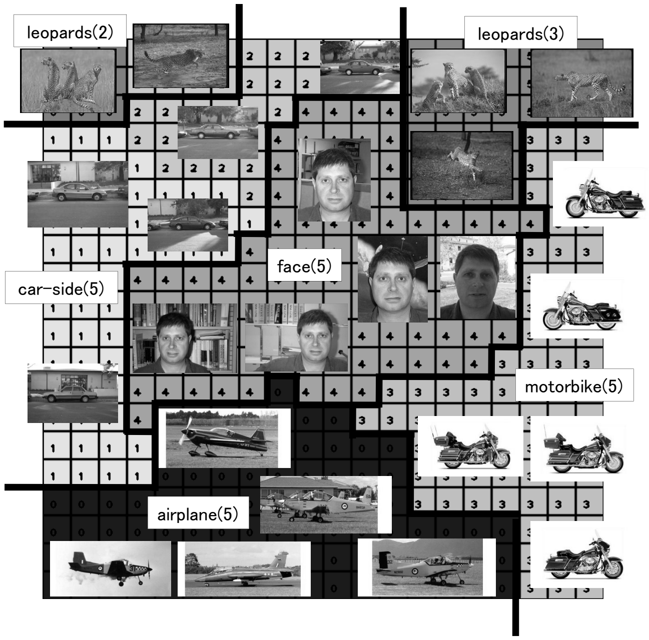


Fig. 5. Result of category mapping using CPNs of five categories.

several labels are generated. These results demonstrate that OC-SVMs can select SIFT features of target objects and show that ART-2 can generate independent labels to images for which backgrounds and appearances of objects differ in each category.

Fig. 5 depicts a category map generated by CPNs for classifications of five categories: Airplane, Car-side, Motorbike, Face, and Leopards. We show images that mapped each unit and mapping regions in each category on the category map. Fig. 5 depicts that CPNs created categories for mapping to neighborhood units on the category map in each image with labels generated by ART-2. The Car-side and Leopards categories contain several labels by ART-2. The Car-side category is mapped to neighborhood units. On the other hand, the Leopards category is divided into two regions.

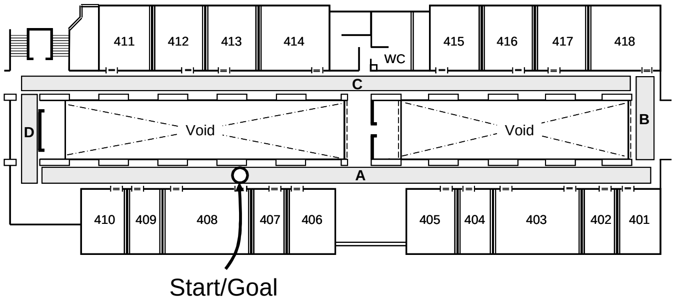


Fig. 6. Experimental environment

Here, for quantitative evaluation of the classification performance of our method, we use the following recognition rate.

$$(Recognition Rate) = \frac{(CorrectData)}{(AllData)} \times 100. \tag{12}$$

Recognition rates for training datasets are, respectively, more than 90% for 10 categories and more than 80% for 20 categories. Although recognition rates for testing datasets reached 76% for five categories, it decreased less than 50% for 10 and 20 categories.

B. Classification results of time-series images

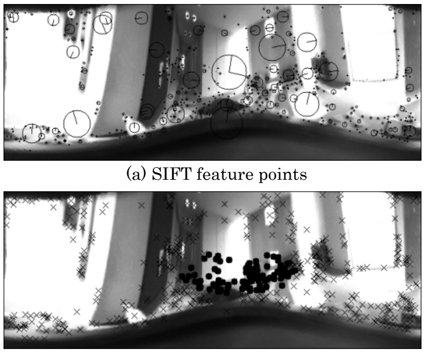
We used an omnidirectional camera to take time-series images with running a corridor shown in Fig. 6. Specifications of the camera are: an imaging device, 1/3” interline CCD; resolution, 640 × 480 pixel; and frame rate, 30 fps. The camera height on the robot is 1,500 mm from the floor. The mean velocity of movement is 30 m/min. The corridor width is 1,830 mm. The robot runs once around counterclockwise.

Fig. 7 depicts the result of selected feature points using our method. Feature points apply to the direction of movement; its surrounding areas are selected. This tendency is the same as those of other scene images. As an indication for category classification, we annotated Zones A through D that resemble appearances.

Fig. 8 portrays a comparison of labeling results obtained using our method and our former method without OC-SVMs. The results show that our method decreases not only mixed labels, but also the total number of created labels. Fig. 9(a) portrays classification results obtained using our method. Numbers in each unit on the category map correspond to the labels portrayed in Fig. 8. Using topological mapping of CPNs, some labels are integrated on the category map. Comparison with the result obtained using our former method shown in Fig. 9(b) shows that the total number of categories is decreased. We consider that feature points selected using OC-SVMs are effective for category classification in robot vision.

IV. Discussion

Experimental results of Caltech-256 and time-series images of the robot show that OC-SVMs select feature points not only of the whole object, but also of the background and surrounding regions, and of partial objects. These results signify that OC-SVMs can select a region to concentrate specific information in an image, i.e. features that characterize an image, not



(a) SIFT feature points

(b) Selected SIFT feature points

Fig. 7. Selected feature points.

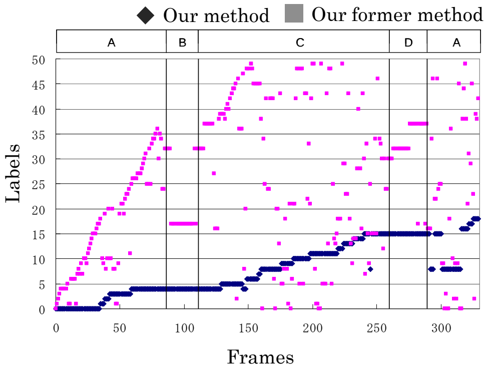
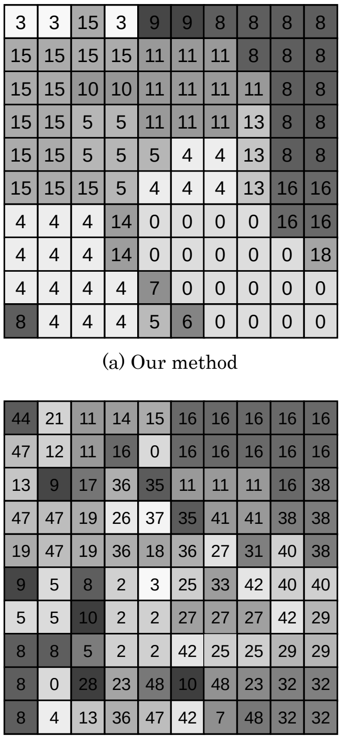


Fig. 8. Comparison of labeling results obtained using the presented method and our former method.



3	3	15	3	9	9	8	8	8	8
15	15	15	15	11	11	11	8	8	8
15	15	10	10	11	11	11	11	8	8
15	15	5	5	11	11	11	13	8	8
15	15	5	5	5	4	4	13	8	8
15	15	15	5	4	4	4	13	16	16
4	4	4	14	0	0	0	0	16	16
4	4	4	14	0	0	0	0	0	18
4	4	4	4	7	0	0	0	0	0
8	4	4	4	5	6	0	0	0	0

(a) Our method

44	21	11	14	15	16	16	16	16	16
47	12	11	16	0	16	16	16	16	16
13	9	17	36	35	11	11	11	16	38
47	47	19	26	37	35	41	41	38	38
19	47	19	36	18	36	27	31	40	38
9	5	8	2	3	25	33	42	40	40
5	5	10	2	2	27	27	27	42	29
8	8	5	2	2	42	25	25	29	29
8	0	28	23	48	10	48	23	32	32
8	4	13	36	47	42	7	48	32	32

(b) Our former method

Fig. 9. Category maps with CPN.

feature points to be classified into the object and background. Humans, when classifying objects, devote attention to a region that gathers information for characterizing an object, not the whole object. We consider that selection of SIFT features using OC-SVMs can describe features effectively for category classification to represent features and can thereby improve classification accuracy.

Regarding results of the static category classification, the accuracy of our method reached 81% for training and 50% for testing of 20-category classification. The unsupervised category classification method proposed by Chen et al. [9] showed respective performances of 76.9% for training and 67.4% for testing of 26-category classification for the Caltech dataset. The accuracy of our method is apparently inferior to that of the existing method. Nevertheless, our method can classify objects without previous setting of the number of categories. Therefore, our method is effective for application to problems that are known as challenging tasks of classification of categories whose ranges and types are unclear. In this experiment, we observed 10 categories for which multiple labels are generated on ART-2. The images of Caltech-256 have no time-series factors, although ART-2 learns time-series changes of input data positively. Therefore, we inferred that ART-2 maintains no continuity of labels. For the relation of labels generated by ART-2 and a category map on CPNs, categories that maintained continued and non-continued labels are mapped respectively to neighborhood and separated units on the category map of CPNs.

For a mobile robot, category classification of scene images is necessary to acquire WIs. In this situation, the number of categories is mostly unknown in a real environment. Therefore, extracting the number of categories is necessary for category classification. In this section, we analyze extraction of boundaries for which topological structures of categories change widely using classification results of ART-2 and CPNs.

The labeling results of ART-2 shown in Fig. 8 have two characteristic parts: rapidly changing parts and gradually changing parts. Labels of ART-2 change rapidly while changing appearances in a scene rapidly. We set labels that change rapidly to candidates of boundaries. Fig. 10 depicts labeling results of ART-2 and CPNs. The candidates of boundaries are two parts: integrated parts and nonintegrated parts. We examine positional relations of Labels 7, 8, 16, and 18 that are selected as candidates of boundaries on the category map. Labels 7 and 8 are mapped onto distant units on the category map. Therefore, we consider that boundaries exist between them, although labels of ART-2 are integrated by CPNs. In contrast, we consider that no category boundary exists to map Labels 16 and 18 into neighborhood units, although both labels are integrated by ART-2. Fig. 11 depicts category classification results to consider extracted boundaries with changing topological structures of categories.

Fig. 11 portrays extracted boundaries and categories. Four categories are extracted from temporal and spatial relations

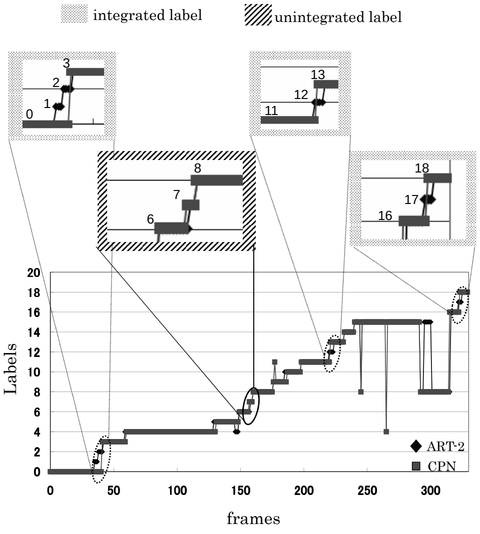


Fig. 10. Extracting boundaries of categories from labeling results of ART-2 and integrated labels with CPNs.

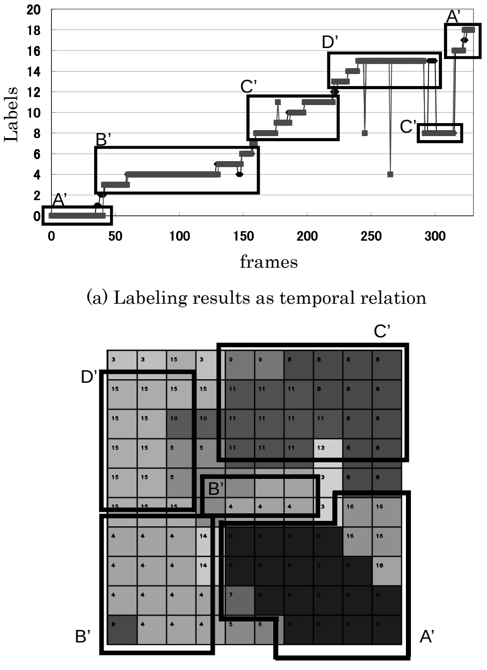
of labels. Fig. 12 portrays classification results of zones in the experimental environment. Comprehensive categories with extracted boundaries are mapped to neighborhood units on the category map using labels generated by ART-2 and CPNs in Fig. 11. The scene images of the environment are categorized into four categories. Labels A', B' and D' correspond to one zone. However, Label C' is divided into two zones. Using temporal relations of labels by ART-2 and spatial relations of categories on CPNs, we ascertained the possibility of extracting global boundaries among categories. However, we extracted it manually. Automatic extraction of categories is a subject to be addressed in our future work.

V. Conclusion

This paper presented an unsupervised method of SIFT feature points selection using OC-SVMs and category classification combined with incremental learning of ART-2 and self-mapping characteristic of CPNs. Our method enables feature representation that contributes to improved accuracy of classification for selecting feature points to concentrate characterized information of an image. Moreover, our method can visualize spatial relations of labels and integrate redundant and similar labels generated by ART-2 as a category map using self-mapping characteristics and neighborhood learning of CPNs. Therefore, our method can represent diverse categories. Future studies must be conducted to develop methods to extract boundaries among clusters automatically and to determine a suitable number of categories from category maps of CPNs. Additionally, we will examine approaches that include generation of robot behavior for classification and recognition of objects.

References

[1] K. Yanai, ”The Current State and Future Directions on Generic Object Recognition,” Journal of Information Processing: The Computer Vision and Image Media, vol. 48 no. SIG16 (CVIM 19), Nov. 2007.



(a) Labeling results as temporal relation

(b) Category map as spatial relation

Fig. 11. Extracted boundaries and categories from temporal and spatial relations.

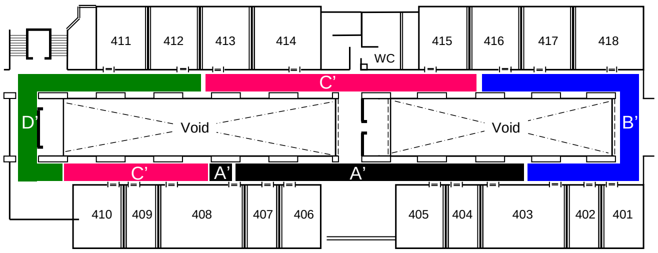


Fig. 12. Classification results of zones in the experimental environment.

[2] K. Nakano, ”Making of a Brain – Thinking about Biotechnology from a Making of a Robot –,” Kyoritsu Shuppan Co., Aug. 1995.

[3] K. Barnard, P. Duygulu, N. D. Freitas, D. Forsyth, D. Blei, and M. Jordan, Matching Words and Pictures, Journal of Machine Learning Research, vol. 3, pp. 1107-1135, 2003.

[4] Lempert, C. H., Blaschko, M. B. and Hofmann, T. Beyond Sliding Windows: Object Localization by Efficient Subwindow Search, Proc. of IEEE Computer Vision and Pattern Recognition, 2008.

[5] K. Suzuki, T. Matsukawa, and T. Kurita, "Bag-of-features car detection based on selected local features using Support Vector Machine," The Institute of Electronics, Information and Communication Engineers Technical Report PRMU, 2009.

[6] M. Tsukada, H. Madokoro, and K. Sato, ”Unsupervised Category Classification Based on Appearance Changes Using Mobile Robot,” IEICE Technical Report PRMU2009-124, pp.213-218, Nov. 2009.

[7] G. A. Carpenter and S. Grossberg, ”ART 2: Stable Self-Organization of Pattern Recognition Codes for Analog Input Patterns,” Applied Optics, vol. 26, pp. 4919-4930, 1987.

[8] R. Hetch-Nielsen, ”Counterpropagation networks,” Proc. of IEEE First Int'l. Conference on Neural Networks, 1987.

[9] Y. Chen, L. Zhu, A. Yuille, and H. Zhang, ”Unsupervised Learning of Probabilistic Object Models(POMs) for Object Classification, Segmentation, and Recognition Using Knowledge Propagation,” IEEE Trans. PAMI vol. 31, no. 10, Oct. 2009.