PAPER

Unsupervised Feature Selection and Category Classification for a Vision-Based Mobile Robot

Masahiro TSUKADA[†a)], Yuya UTSUMI[†], *Nonmembers*, Hirokazu MADOKORO[†], *and* Kazuhito SATO[†], *Members*

SUMMARY 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 objects.

key words: *ART-2, CPN, SOM, SIFT, OC-SVMs, unsupervised category classification, robot vision*

1. 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 image maps [2]. For creating world image maps, 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.

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) [14] 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) [18] as teaching signals for Counter Propagation Networks (CPNs) [19]. 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 *de facto* standard benchmark dataset for comparing the performance of algorithms in generic object recognition, and time-series images taken by a camera on a mobile robot.

This paper presents the following. First, we describe related work in Sect. 2. Next, we explain detailed specifications of our image representation method, our category classification method, and the whole architecture of our method in Sects. 3, 4, and 5, respectively. Subsequently, we present experimental results in Sects. 6 and 7. Finally, we respectively present related discussion and salient conclusions in Sects. 8 and 9.

2. Related Work

The problem of Simultaneous Localization and Mapping (SLAM) has attracted immense attention in mobile robotics studies [3]. The objective of SLAM is to build a map and update it while simultaneously estimating locations for a robot. Cummins et al. proposed Fast Appearance Based Mapping (FAB-MAP) [4] as a probabilistic approach to recognizing places based on their appearance. The objective of FAB-MAP is similar to SLAM: to build a map of routes using appearance changes of scene images obtained using a camera on a mobile robot. Our objective is to classify images obtained using a camera on a mobile robot in categories for recognizing objects.

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

Manuscript received April 5, 2010.
Manuscript revised August 30, 2010.

[†]The authors are with the Graduate School of Systems Science and Technology, Akita Prefectural University, Yurihonjo-shi, 015–0051 Japan.
a) E-mail: m11a013@akita-pu.ac.jp

DOI: 10.1587/transinf.E94.D.127

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 [5]. 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 [6]. 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 [7]. 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.

As unsupervised category classification methods, Sivic et al. proposed an unsupervised category classification method using probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA), which are generative models from the statistical text literature [8]. They modeled an image containing instances of several categories as a mixture of topics and attempted to discover topics as object categories from numerous images. Zhu et al. introduced Probabilistic Grammar Markov Models (PGMMs) of generative models that combined Probabilistic Context-Free Grammars (PCFGs) and Markov Random Fields (MRFs) [9]. They used this method to create an object category model for object detection and unsupervised category classification. Moreover, they proposed Probabilistic Object Models (POMs) that improved their method and enabled classification, segmentation, and recognition of objects [10]. Todorovic et al. proposed an unsupervised identification method using optical, geometric, and topological characteristics of multi-scale regions consisting of two-dimensional objects [11]. They represented each image as a tree structure by division of multi-scale images. Moreover, Nakamura et al. proposed an unsupervised category classification method using multimodal information of vision, hearing, and touch [12]. They achieved category classification of objects that resemble human senses using embodied interactions of a robot.

However, these methods include the restriction of prior settings of the number of classification categories. Therefore, these methods are applied only slightly to classification problems in a real environment for which the number of categories is unknown.

3. Image Representation

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 depicted in Fig. 1, we applied OC-SVMs for selecting SIFT feature points as target regions in an image. Furthermore, we applied Self-Organizing Maps (SOMs) [16] for creating visual words and histograms in each image from selected features.

Our target is SIFT feature points on an object for recognition. Therefore, target regions and target feature points respectively mean object regions and feature points on an object. The OC-SVMs are unsupervised-learning-based binary classifiers that enable density estimation without estimating a density function. Therefore, OC-SVMs can apply to real-world images without boundary information. Detailed algorithms of SIFT, OC-SVMs, and SOMs are the following.

3.1 Description of Features Using SIFT

Generally, SIFT is used as a descriptive method of local features in generic object recognition. The SIFT [13] processing consists of two steps: detection of feature points and description of features. The procedures are the following. Difference of Gaussians (DoG) image $D(u, v, \sigma)$ as

$$\begin{aligned} D(u, v, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(u, v) \\ &= L(u, v, k\sigma) - L(u, v, \sigma). \end{aligned} \tag{1}$$

Here, $G(x, y, \sigma)$ is the convolution of a variable-scale Gaussian, $I(u, v)$ is an input image, and $L(u, v, \sigma)$ is a smoothing image. This pixel is detected as a candidate for a keypoint if an attentional pixel of DoG images is an extreme value compared with its 26-neighbor pixel.

Unnecessary keypoints are eliminated with the threshold as

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} < \frac{(\gamma_{th} + 1)^2}{\gamma_{th}}. \tag{2}$$

Here, $\mathrm{Tr}(\mathbf{H})$ is a sum of cross elements of Hessian matrix and $\mathrm{Det}(\mathbf{H})$ is a determinant. The value of DoG determines

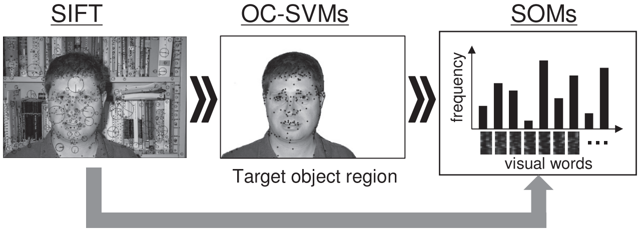


Fig. 1 Procedures of our image representation method based on BoF.

elimination of keypoints. Keypoints are eliminated if the absolute value of DoG on a position of subpixel is less than the threshold.

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^{\mathrm{T}}}{\partial x}\hat{x}. \tag{3}$$

Here, D is DoG function; $\hat{x}$ is a position of a subpixel.

An orientation histogram is formed from the gradient magnitude $m(x, y)$ in local regions and gradient orientation $\theta(u, v)$ as

$$h_{\theta'} = \sum_x \sum_y w(x, y) \cdot \delta[\theta', \theta(x, y)], \tag{4}$$

$$w(x, y) = G(x, y, \sigma) \cdot m(x, y). \tag{5}$$

The 128-dimensional features are extracted from the histogram of eight detections of 4 × 4 subregions. The 128 descriptors are detected at each keypoint.

3.2 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 [14]. 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 $f(\cdot)$ is calculated to divide input feature vectors x_{i} 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.

$$f(x) = \mathrm{sgn}(\omega^{\top}\Phi(x) - \rho). \tag{6}$$

Here, ω and ρ ($\rho \in R$) represent a coefficient and a margin. Therein, z_{i} represents results of x_{i} to the high-dimension feature space.

$$\Phi : x_i \mapsto z_i \tag{7}$$

The restriction is set to the following.

$$\omega^{\top} z_i \geq \rho - \zeta_i, \quad \zeta_i \geq 0, \quad 0 < \nu \leq 1 \tag{8}$$

Here, ζ represents relaxation variable vectors. The optimization problem is solved with the following restriction

$$\frac{1}{2}\|\omega\|^2 + \frac{1}{\nu l}\sum_{i=1}^{l}\zeta_i - \rho \rightarrow \min \omega, \zeta, \text{and } \rho \tag{9}$$

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 (9) fulfills $\rho \neq 0$.

3.3 Creating Visual Words Using SOMs

For our method, we apply SOMs, not k-means, which is generally used in BoF, for creating visual words. In the learning step, SOMs update weights while maintaining topological structures of input data. Actually, SOMs create neighborhood regions around the burst unit, which demands a response of the input data. Therefore, SOMs can classify various data whose distribution resembles the training data. In addition, Terashima et al. reported that SOMs are superior to k-means as an unsupervised classification method that is useful to minimize misrecognition [15]. The SOM learning algorithm is the following.

1) $u^i_{n,m}(t)$ are weights from an input layer unit i ($i = 1, \ldots, I$) to a Kohonen layer unit (n, m) ($n = 1, \ldots, N, m = 1, \ldots, M$) at time t. The weights are initialized randomly. The training data $x_i(t)$ show input layer units i at time t. The Euclidean distance $d_{n,m}$ separating $x_i(t)$ and $u^i_{n,m}(t)$ is calculated as

$$d_{n,m} = \sqrt{\sum_{i=1}^{I}(x_i(t) - u^i_{n,m}(t))^2}. \tag{10}$$

2) The unit for which $d_{n,m}$ is smallest is defined as the winner unit c as

$$c = \mathrm{argmin}(d_{n,m}). \tag{11}$$

3) Here, $N_c(t)$ is a neighborhood region around the winner unit c. In addition, $u^i_{n,m}(t)$ of $N_c(t)$ is updated using Kohonen's learning algorithm, as

$$u^i_{n,m}(t+1) = u^i_{n,m}(t) + \alpha(t)(x_i(t) - u^i_{n,m}(t)). \tag{12}$$

In that equation, $\alpha(t)$ is the learning rate coefficients that decrease with the progress of learning. The learning of SOMs repeats up to the learning iteration that was set previously.

In this method, we used all SIFT features for creating visual words at the learning step of SOMs. We used SIFT features selected by OC-SVMs for generating histograms based on visual words. Based on our preliminary experiment, we set the learning iteration to 100,000 times. Additionally, we set the number of units of the Kohonen layer to 100 units. We created visual words to extract weights between Kohonen layer units and input layer units.

4. Unsupervised Category Classification

Figure 2 depicts the architecture of our unsupervised category classification method that combined incremental learning of ART-2 and self-mapping characteristics of CPNs. Actually, ART-2 is a theoretical model of unsupervised neural networks of incremental learning that forms categories adaptively while maintaining stability and plasticity together. 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, actual-

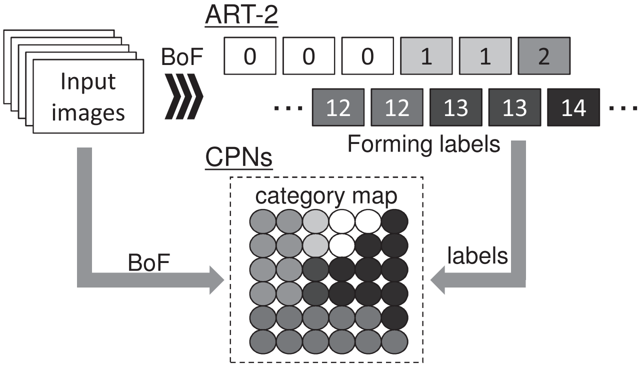


Fig. 2 Architecture of our unsupervised category classification method.

izes mapping and labeling together. Such networks comprise three layers: an input layer, a Kohonen layer, and a Grossberg layer. 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. Our method can reduce these labels using the Winner-Takes-All competition of CPNs. In addition, our method can visualize relations between categories on the category map of CPNs. Detailed algorithms of ART-2 and CPNs are the following.

4.1 Generating of Labels Using ART-2

In ART of various types [17], we use ART-2, into which it is possible to input continuous values [18]. The learning algorithm of ART-2 is the following.

1) Top-down weights Z_{ji}, bottom-up weights Z_{ij}, and outputs p_i, q_i, and u_i on the F1 of sublayers are initialized as

$$Z_{ji}(0) = 0, \qquad Z_{ij}(0) = \frac{1}{(1-d)\sqrt{M}}, \tag{13}$$

$$p_i(0) = q_i(0) = u_i(0) = v_i(0) = w_i(0) = x_i(0) = 0.0. \tag{14}$$

2) Input data I_i are presented to the F1; the sublayers are propagated as

$$w_i(t) = I_i(t) + au_i(t-1), \tag{15}$$

$$x_i(t) = \frac{w_i(t)}{e + \|w\|}, \tag{16}$$

$$v_i(t) = f(x_i(t)) + bf(q_i(t-1)), \tag{17}$$

$$u_i(t) = \frac{v_i(t)}{e + \|v\|}, \tag{18}$$

$$p_i(t) = \begin{cases} u_i(t) & \text{(inactive)} \\ u_i(t) + dZ_{Ji}(t) & \text{(active)}, \end{cases} \tag{19}$$

$$q_i(t) = \frac{p_i(t)}{e + \|p\|}, \tag{20}$$

$$f(x) = \begin{cases} 0 & \text{if} \quad 0 \le x < \theta \\ x & \text{if} \quad x \ge \theta. \end{cases} \tag{21}$$

3) Search for the maximum active unit T_j as

$$T_J(t) = max\left(\sum_j p_i(t) Z_{ij}(t)\right). \tag{22}$$

4) Top-down weights Z_{ji} and bottom-up weights Z_{ij} are updated as

$$\frac{d}{dt} Z_{Ji}(t) = d[p_i(t) - Z_{Ji}(t)], \tag{23}$$

$$\frac{d}{dt} Z_{iJ}(t) = d[p_i(t) - Z_{iJ}(t)]. \tag{24}$$

5) The vigilance threshold ρ is used to judge whether input data correctly belong to a category.

$$\frac{\rho}{e + \|r\|} > 1, \quad r_i(t) = \frac{u_i(t) + cp_i(t)}{e + \|u\| + \|cp\|}. \tag{25}$$

When (25) is true, the active units reset and return (15) to search again. Repeat (14) and (16) until the rate of change of F1 is sufficiently small if (25) is not true. In addition, a and b are coefficients of feedback loops from u to w and from q to v. Here, c is a propagation coefficient from p to r, and d is a learning rate coefficient. Furthermore, $cd/(1-d) \le 1$ is the constraint between them, and θ is a parameter to control a noise detection level in v.

4.2 Creating Category Maps Using CPNs

The CPNs [19] perform pattern mapping, i.e. CPNs map one pattern into another pattern in all sets of patterns. When a pattern is presented, learned networks classify patterns into specific categories using weights. Our method can automate labeling with generation of labels as teaching signals to the units of the Grossberg layer on CPNs. The CPN learning algorithm is the following.

1) $u^i_{n,m}(t)$ are weights from an input layer unit i ($i = 1, \dots, I$) to a Kohonen layer unit (n, m) ($n = 1, \dots, N, m = 1, \dots, M$) at time t. Therein, $v^j_{n,m}(t)$ are weights from a Grossberg layer unit j to a Kohonen layer unit (n, m) at time t. These weights are initialized randomly. The training data $x_i(t)$ show input layer units i at time t. The Euclidean distance $d_{n,m}$ separating $x_i(t)$ and $u^i_{n,m}(t)$ is calculated as

$$d_{n,m} = \sqrt{\sum_{i=1}^{I} (x_i(t) - u^i_{n,m}(t))^2}. \tag{26}$$

2) The unit for which $d_{n,m}$ is smallest is defined as the winner unit c as

$$c = \text{argmin}(d_{n,m}). \tag{27}$$

3) Here, $N_c(t)$ is a neighborhood region around the winner unit c. In addition, $u^i_{n,m}(t)$ of $N_c(t)$ is updated using Kohonen's learning algorithm, as

$$u^i_{n,m}(t+1) = u^i_{n,m}(t) + \alpha(t)(x_i(t) - u^i_{n,m}(t)). \tag{28}$$

4) In addition, $v^{j}_{n,m}(t)$ of $N_c(t)$ is updated using Grossberg's outstar learning algorithm, as

$$v^{j}_{n,m}(t+1) = v^{j}_{n,m}(t) + \beta(t)(t_j(t) - v^{j}_{n,m}(t)). \quad (29)$$

In that equation, $t_j(t)$ is the teaching signal to be supplied to the Grossberg layer. Furthermore, $\alpha(t)$ and $\beta(t)$ are the learning rate coefficients that decrease with the progress of learning. The learning of CPNs repeats up to the learning iteration that was set previously.

5. Whole Architecture of Our Method

In generic object recognition, it is a challenging task to develop a unified model to address all steps from feature representation to creation of classifiers. The aim of our study is the realization of category classification for generic object recognition to apply theories with different characteristics for each step. Figure 3 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 SOM
4. Generating labels using ART-2
5. Creating a category map using CPNs

Procedures 1. through 3., which correspond to preprocessing, are based on the representation of BoF. We apply OC-SVMs to select SIFT feature points for localizing target regions in an image. For producing visual words, we use SOMs, which can learn neighborhood regions while updating the cluster structure, although k-means must decide data of the center of a cluster. Actually, SOMs can represent visual words that minimize misclassification [15]. Furthermore, the combination of ART-2 and CPNs enables unsupervised category classification that labels a large quantity of images in each category automatically. Table 1 shows parameters of OC-SVMs, ART-2, and CPNs with each experiment.

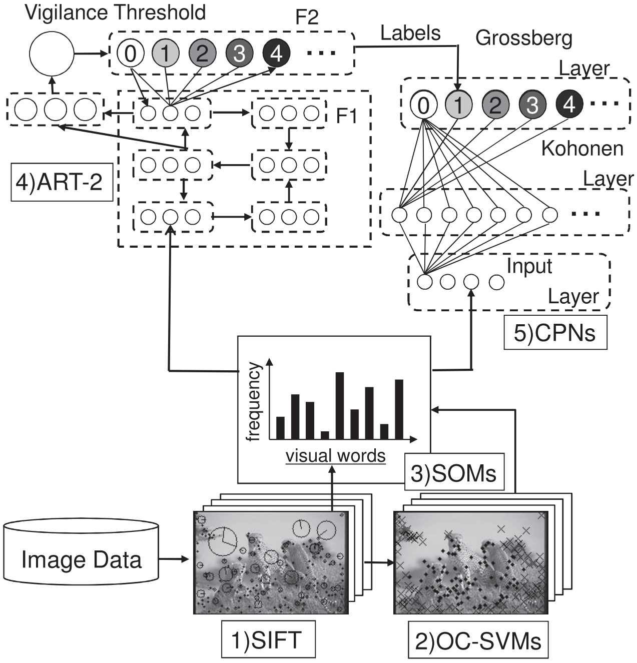


Fig. 3 Whole architecture of our method.

6. Experimental Results Obtained Using the Caltech-256 Dataset

This section presents experimental results of image classification using Caltech-256 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. The results of selection of SIFT features and recognition rates for classification of 5, 10, and 20 categories are the following.

6.1 Selection of Feature Points and Generation of Labels

Figure 4 depicts results of selected feature points using OC-SVMs on five sample images of Caltech-256. Figure 4 (a) shows that our method can select feature points of target objects in images of the Leopards and Face categories. In

Table 1 Setting values of parameters used in experiments.

Parameters		Setting values
OC-SVMs	ν	0.5
ART-2	θ	0.1
	ρ	0.920
CPNs	$\alpha(t)$	0.5
	$\beta(t)$	0.5
	learning iteration	10,000

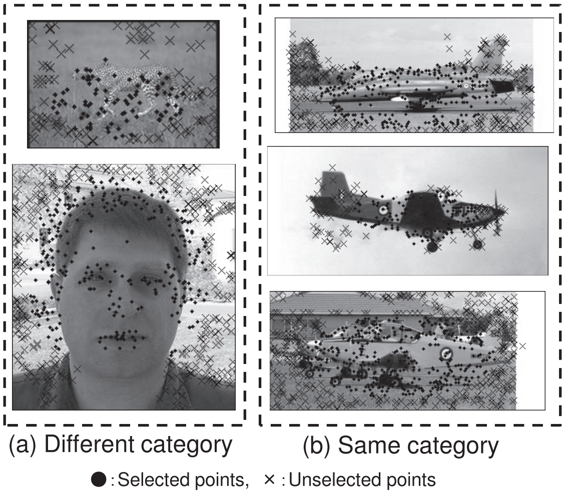


Fig. 4 Results of selected SIFT feature on two sample images in different category and three sample images in the same category of Caltech-256.

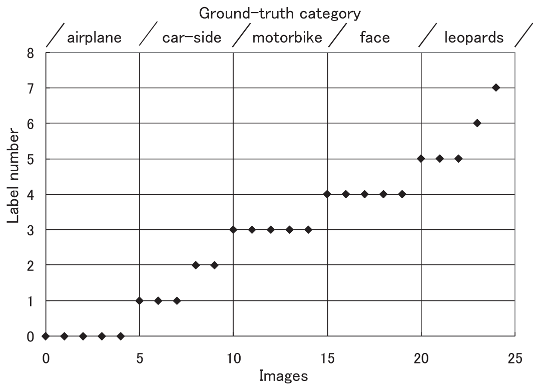


Fig. 5 Results of formed labels using ART-2 at five categories.

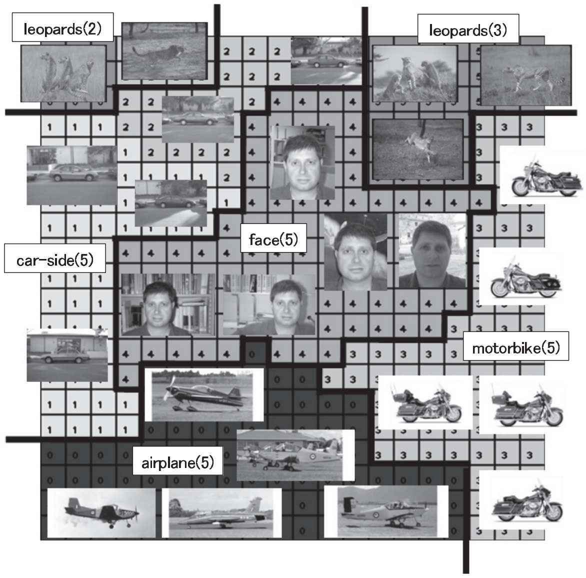


Fig. 6 Result of category mapping using CPNs of five categories.

addition, Fig. 4 (b) shows that our method can select feature points around the wings that characterize airplanes for various images of the Airplane category.

Figure 5 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 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.

6.2 Category Classification

Figure 6 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. Figure 6 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. The Car-side category is mapped to neighborhood units. On the other hand, the Leopards category is divided into two regions.

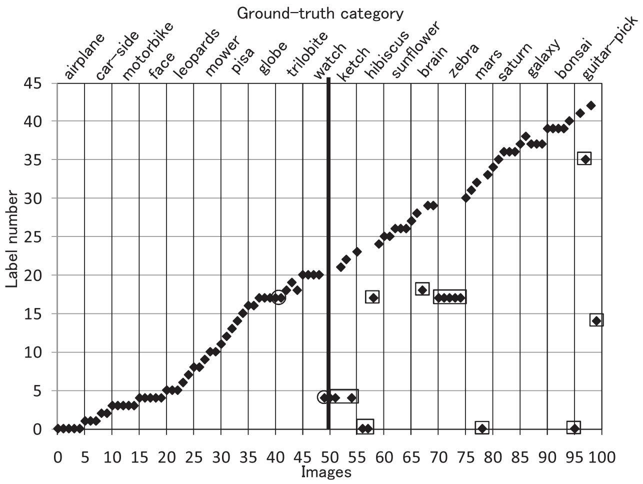


Fig. 7 Results of formed labels using ART-2 at 10 and 20 categories.

Figure 7 depicts labels by ART-2 on 20-category classification. The bold line shows the number of images in 10 categories. The circles and squares portray images for which ART-2 confused labels on 10 and 20 categories, respectively. In the 10-category classification, ART-2 generated independent labels in all categories, although three images of two labels are confused. In the 20-category classification, independent labels of 19 categories are generated, except for the Zebra category that is confused of all images, although 16 images of five labels are confused. Confusion of labels occurs often in images of Ketch, Hibiscus, and Guitar-pick categories. Although confused labels are restrained until 10-category classification, numerous confused labels are apparent in the 20-category classification.

Figure 8 depicts a category map generated by CPNs on 20-category classification. The names of categories and the number of images are shown on the category map. For all images in each category, 11 categories are mapped to neighborhood units. The CPNs created categories for mapping neighborhood units on the category map in images of each category by which ART-2 generated several labels. In addition, categories without their names are mapped images of different categories. Here, for quantitative evaluation of the classification performance of our method, we use the following recognition rate.

$$(RecognitionRate) = \frac{(CorrectData)}{(AllData)} \times 100. \qquad (30)$$

Figure 9 portrays recognition rates in 5, 10, and 20 categories without OC-SVMs and with OC-SVMs for training and testing datasets. The recognition rates without OC-SVMs were, respectively, 84%, 70%, and 64% for training

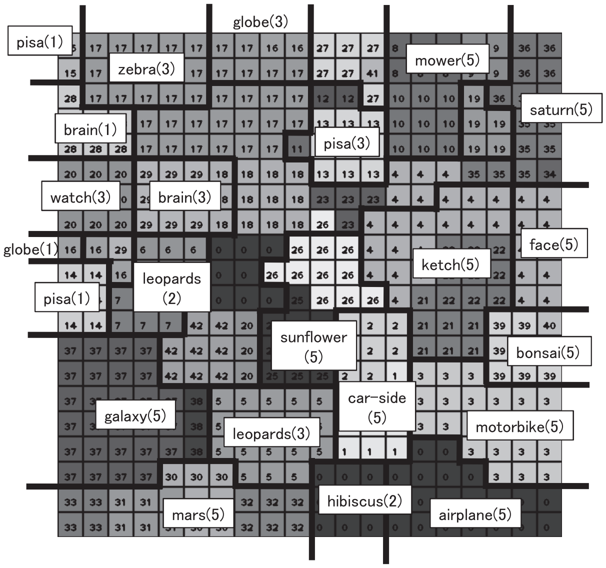


Fig. 8 Result of a category map of 20 categories.

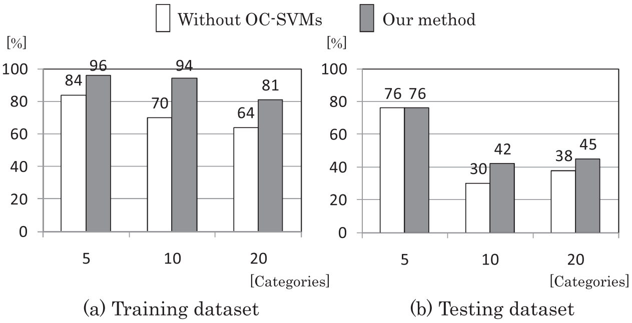


Fig. 9 Recognition rates of training and testing datasets used in Caltech-256.

datasets and 76%, 30%, and 38% for testing datasets in 5, 10, and 20 categories. In our method, the recognition rates were, respectively, 96%, 94%, and 81% for training datasets and 76%, 42%, and 45% for testing datasets in 5, 10, and 20 categories. These results address the effectiveness to select SIFT feature points using OC-SVMs.

The unsupervised category classification method proposed by Chen et al. [10] 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.

7. Experimental Results Obtained Using a Mobile Robot

In this section, we applied our method to category classification experiments using time-series images taken by a camera with movements of a robot. In this experiment, we evaluated our method for category classification of dynamic images because the target is time-series images according to the change of appearances. We built an original experimental environment to take images of datasets. This section presents the experimental environment and results of our method as the following.

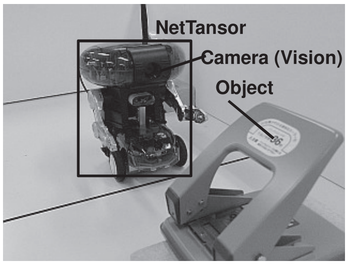


Fig. 10 Robot used for experiments (NetTansor; Bandai Co. Ltd.).

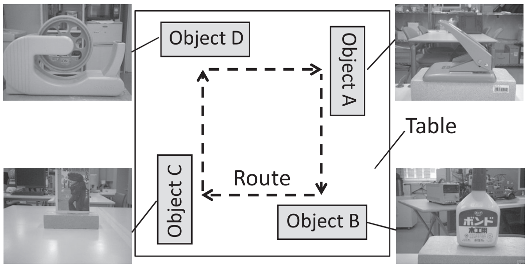


Fig. 11 Four objects and the robot route used for our experiment.

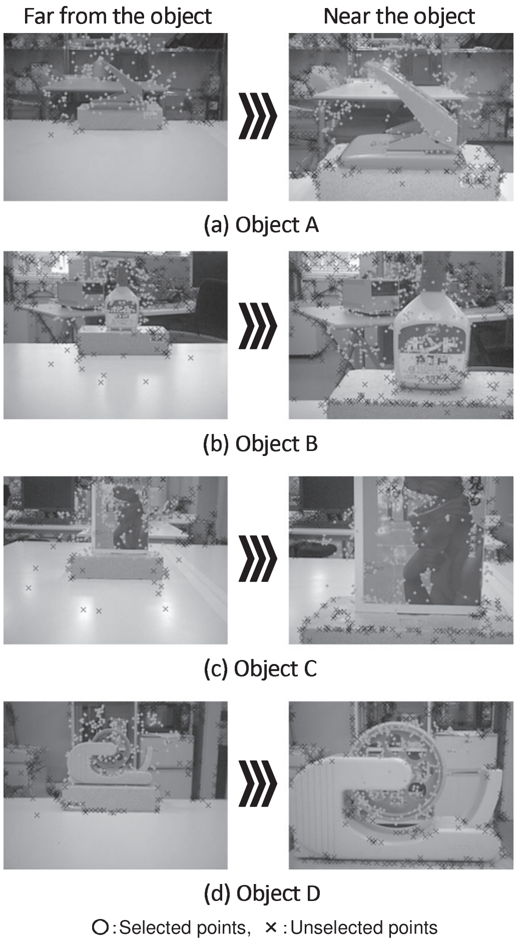
7.1 Experimental Environment

Figure 10 portrays a home robot (NetTansor; Bandai Co. Ltd.) used in this experiment. The robot is 190 mm high, 160 mm long, and 160 mm wide. The camera specifications are the following: imaging device, 1/4 inch CMOS; image format, JPEG; resolution, 320 × 240 pixels; and frame rate, 15 fps. The moving environment is 1,150 × 1,150 mm.

Figure 11 shows the assignment of objects in the environment and the roughly determined goals of routes for the robot. We assumed the environment for moving of this robot as a desk. In consideration of the robot height, we used office supplies with characteristic shapes. Target objects were a hole punch (Object A), a plastic bottle of glue (Object B), a book (Object C), and a cellophane tape holder (Object D) shown in Fig. 11. For this experiment, we created datasets consisting of time-series images as shown in the behavior of Fig. 11. Datasets comprise RUN1 and RUN2 for which the robot runs twice around in the environment.

7.2 Selection of Feature Points and Generation of Labels

Figure 12 depicts results of selected feature points using OC-SVMs on four samples of time-series images taken by the robot. Our method can select feature points near objects against various appearance changes. In images of Object D,



(a) Object A

(b) Object B

(c) Object C

(d) Object D

Fig. 12 Results of selected SIFT feature points of time-series images.

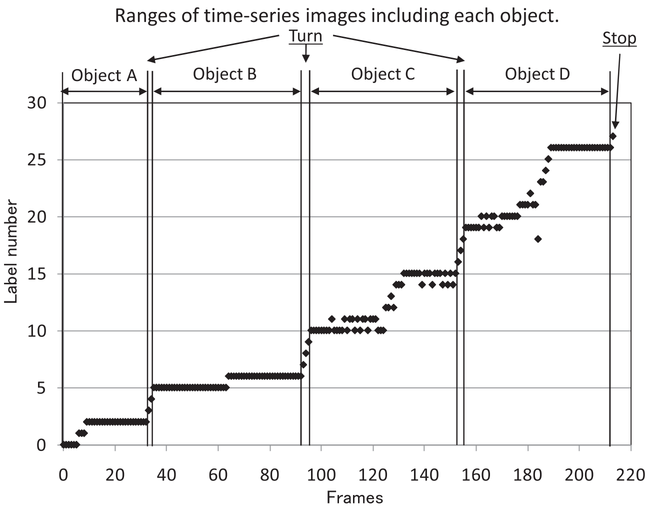


Fig. 13 Results of labels created using ART-2 from time-series images.

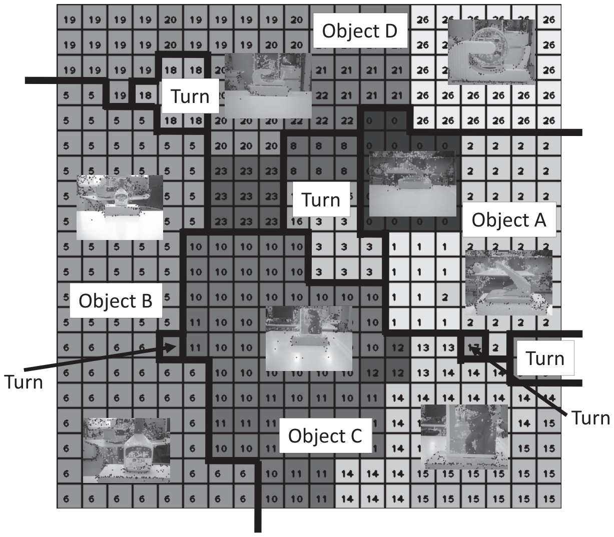


Fig. 14 Mapping result of images on the category map of CPNs used in labels generated by ART-2.

Table 2 Recognition rates of learning and testing datasets of time-series images.

		Testing Datasets		Mean
		RUN1	RUN2	
Training Datasets	RUN1	98.1%	96.2%	96.7%
	RUN2	97.2%	98.8%	

feature points of whole and a part of Object D are, respectively, selected distant from the object and near the object. In addition, feature points are selected not only of the object, but also around the object.

7.3 Category Classification

Figure 13 depicts labels generated by ART-2 on the experiment using time-series images of RUN1. The vertical and horizontal axes respectively represent labels of ART-2 and frames in images. The top parts portray ranges including objects and parts of the robot turned 90 deg as time-series images. In this result, 27 labels are generated from time-series images of 220 frames. In addition, the labels are more numerous than the target objects because labels are assigned to each image taken by the robot turned 90 deg from the four corners in the environment. Objects A, B, C, and D respectively generated 3, 2, 6, and 8 labels.

Figure 14 depicts a category map generated by CPNs. On the category map, we show mapping regions of images in each object. Each object classified with different labels with ART-2 is mapped to neighborhood units on the category map of CPNs shown in Fig. 14. In addition, images of turning of labels 3 and 4 are mapped around border units between categories.

Table 2 portrays recognition rates for training and testing calculated using Eq. (30). This experiment evaluated recognition rates for all combinations of datasets of RUN1 and RUN2 for learning and testing. Underlined values are the recognition rates for training.

In [20], the recall rate of SIFT is less than 50% when objects are occluded more than 30%. We annotated images including defective objects of more than 30% as being of the category of backgrounds and other objects. Table 2 shows that recognition rates for training and testing datasets are more than 90%. Moreover, the mean recognition for testing datasets is 96.7%. In contrast, images of turning include misrecognitions and confused labels in each object.

7.4 Computational Costs

The robot we used for this experiment has a wireless LAN system that enables it to communicate with a PC as an external computation environment. Therefore, we conducted calculations for learning and testing on a PC. Computational costs of our method are as follows.

- SOMs: 7 min per 1,000 frames
- SIFT and OC-SVMs: 11 min per 1,000 frames
- Training for ART-2 and CPNs: 45 s per 1,000 frames
- Testing for CPNs: 0.15 s per frame

Some important parameters of our computational environment are Core 2 Duo 2.2 GHz CPU (Intel Corp.); 1.7 G bytes memory,; Vine Linux 4.2 OS,; and the Eclipse 3.4 development tool with OpenCV 1.0. The mean calculation cost for SIFT and OC-SVMs is 0.66 s per frame, although it depends on the number of feature points. The mean calculation cost for CPN testing is 0.15 s per frame, which enables calculation in real-time for the 30 fps input image.

8. 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 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.

In the static category classification using Caltech-256, the accuracy of our method reached 81% for training and 50% for testing of 20-category classification. 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.

In the dynamic category classification using time-series images of the robot, the accuracy shows high performance of better than 90% for training and testing datasets. This result means that our method can classify time-series images into categories used for characteristics of ART-2. Category classification for generic object recognition is necessary to classify categories for assigning one label to one category. However, category classification for robot vision is necessary to classify categories for assigning labels positively to changes in appearance with sensing in an environment. We consider that ART-2 can learn changes in appearance positively for generation of labels. Nevertheless, the number of labels of ART-2 is greater because the appearance changes in the environment increase along with the behavior of turning 90 deg.

The CPNs created categories in each object whose appearance differs from that of neighboring units. In addition, with the topological mapping characteristic based on the neighborhood learning of CPNs, images that characterized each object and images for which the robot is turning are mapped respectively near the center in each category and near borders between categories. This result means that our method can represent the diversity of categories on category classification.

9. 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), pp.1–24, Nov. 2007.

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

[3] M.W.M.G. Dissanayake, P. Newman, S. Clark, H.F. Durrant-Whyte, and M. Csorba, “An experimental and theoretical investigation into simultaneous localisation and map building (SLAM),” in Lecture Notes in Control and Information Sciences: Experimental Robotics VI, Springer, 2000.

[4] M. Cummins and P. Newman, “FAB-MAP: Probabilistic localization and mapping in the space of appearance,” Int. J. Robot. Res.,

vol.27, no.6, pp.647–665, June 2008.

[5] 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.

[6] C.H. Lempert, M.B. Blaschko, and T.B. Hofmann, "Sliding windows: Object localization by efficient subwindow search," Proc. IEEE Computer Vision and Pattern Recognition, pp.1–8, 2008.

[7] K. Suzuki, T. Matsukawa, and T. Kurita, "Bag-of-features car detection based on selected local features using support vector machine," IEICE Technical Report, PRMU, vol.108, no.484, pp.7–12, 2009.

[8] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering objects and their localization in images," IEEE International Conference on Computer Vision, pp.370–377, 2005.

[9] L. Zhu, Y. Chen, and A. Yuille, "Unsupervised learning of probabilistic grammar – Markov models for object categories," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.1, pp.114–128, Jan. 2009.

[10] 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. Pattern Anal. Mach. Intell., vol.31, no.10, pp.1747–1761, Oct. 2009.

[11] S. Todorovic and N. Ahuja, "Unsupervised category, modeling, recognition, and segmentation in images," IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.12, pp.2158–2174, Dec. 2008.

[12] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal object categorization by a robot," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J91-D, no.10, pp.2507–2518, Oct. 2008.

[13] D.G. Lowe, "Object recognition from local scale-invariant features," Proc. IEEE International Conference on Computer Vision (ICCV), pp.1150–1157, 1999.

[14] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol.20, no.3, pp.273–297, 1995.

[15] M. Terashima, F. Shiratani, and K. Yamamoto, "Unsupervised cluster segmentation method using data density histogram on self-organizing feature map," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J79-D-II, no.7, pp.1280–1290, July 1996.

[16] T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, Springer, 1995.

[17] G.A. Carpenter and S. Grossberg, Pattern Recognition by Self-Organizing Neural Networks, The MIT Press, 1991.

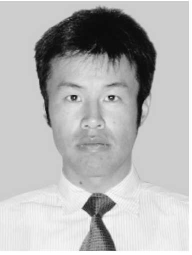
[18] G.A. Carpenter and S. Grossberg, "ART 2: Stable self-organization of pattern recognition codes for analog input patterns," Appl. Opt., vol.26, pp.4919–4930, 1987.

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

[20] K. Mikolajczyk and C. Schmid, "A Performance evaluation of local descriptors," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.10, pp.1615–1630, Oct. 2005.

Masahiro Tsukada received his B.S. degree in mechanical engineering from Akita Prefectural University in 2009. Currently, he has been a master student in the Graduate School of Machine Intelligence and Systems Engineering, Akita Prefectural University. His research interests include in neural networks and robot vision systems.

Yuya Utsumi received his B.S. degree in mechanical engineering from Akita Prefectural University in 2010. Currently, he has been a master student in the Graduate School of Machine Intelligence and Systems Engineering, Akita Prefectural University. His research interests include in computer vision systems for human sensing and robot vision systems.



Hirokazu Madokoro received his M.E. degree in information engineering from Akita University in 2000 and joined Matsushita Systems Engineering Corporation. He moved to the Akita Prefectural Industrial Technology Center in 2002 and Akita Research Institute of Advanced Technology in 2006. Since 2008, he has been an assistant in the Department of Machine Intelligence and Systems Engineering, Akita Prefectural University. His research interests include in neural networks and robot vision systems. He is a member of the Japan Society for Welfare Engineering and the IEEE.

Kazuhito Sato received his M.E. degree in electrical engineering from Akita University in 1975 and joined Hitachi Engineering Corporation. He moved to the Akita Prefectural Industrial Technology Center in 1979 and Akita Research Institute of Advanced Technology in 2005. He received his D. Eng. degree in information system engineering from Akita University in 1997. Since 2008, he has been an associate professor in the Department of Machine Intelligence and Systems Engineering, Akita Prefectural University. He is engaged in the development of equipment for noninvasive inspection of electronic pats, various kinds of expert systems, and MRI brain image diagnostic algorithms. His current research interests include in biometrics, medical image processing, facial expression analysis, computer vision. He is a member of the Medical Information Society, the Medical Imaging Technology Society, the Japan Society for Welfare Engineering and the IEEE.