Pattern Recogn. Phys., 1, 93–103, 2013
www.pattern-recogn-phys.net/1/93/2013/
doi:10.5194/prp-1-93-2013

Pattern Recognition in Physics
Open Access

Unsupervised semantic indoor scene classification for robot vision based on context of features using Gist and HSV-SIFT

H. Madokoro, A. Yamanashi, and K. Sato

Faculty of Systems Science and Technology, Akita Prefectural University, Akita, Japan

Correspondence to: H. Madokoro (madokoro@akita-pu.acjp)

Received: 27 May 2013 – Revised: 29 July 2013 – Accepted: 1 August 2013 – Published: 6 August 2013

Abstract. This paper presents an unsupervised scene classification method for actualizing semantic recognition of indoor scenes. Background and foreground features are respectively extracted using Gist and color scale-invariant feature transform (SIFT) as feature representations based on context. We used hue, saturation, and value SIFT (HSV-SIFT) because of its simple algorithm with low calculation costs. Our method creates bags of features for voting visual words created from both feature descriptors to a two-dimensional histogram. Moreover, our method generates labels as candidates of categories for time-series images while maintaining stability and plasticity together. Automatic labeling of category maps can be realized using labels created using adaptive resonance theory (ART) as teaching signals for counter propagation networks (CPNs). We evaluated our method for semantic scene classification using KTH's image database for robot localization (KTH-IDOL), which is popularly used for robot localization and navigation. The mean classification accuracies of Gist, gray SIFT, one class support vector machines (OC-SVM), position-invariant robust features (PIRF), and our method are, respectively, 39.7, 58.0, 56.0, 63.6, and 79.4 %. The result of our method is 15.8 % higher than that of PIRF. Moreover, we applied our method for fine classification using our original mobile robot. We obtained mean classification accuracy of 83.2 % for six zones.

1 Introduction

A new lifestyle, including the coexistence of humans and robots in various environments in homes and offices, is anticipated in the near future (Kanade et al., 2004). For robots to be useful for and valuable to the existence of humans, it is necessary that they attain the ability not only to move according to programs installed previously, but also to behave autonomously in many situations and in constantly changing environments. An approach using simultaneous localization and mapping (SLAM) (Dissanayake et al., 2000) is the mainstream method used to guide automobile movements for a robot to create a map with no human assistance and to estimate its position simultaneously using various sensors to obtain range information: infrared rays, sonar, laser range finders, etc. However, because these sensors can only obtain range data from objects, walls, and obstacles in an environment, it is a challenging task for SLAM-based methods to recognize semantic categories such as kitchens, living rooms, and corridors. We regard the combination of SLAM and semantic scene category recognition as presenting the possibility of creating intelligent and autonomous behavior. Therefore, semantic scene category recognition has attracted attention as an interesting research subject in computer vision and robot vision studies (Wu et al., 2009).

In computer vision, various methods have been proposed to recognize semantic categories from numerous scene images collected through the internet (Siagian and Itti, 2007). However, classification targets are mainly static images of an outdoor environment. Therefore, recognition accuracy drops dramatically for most common indoor scenes when existing methods for outdoor scene classification are tested on indoor scene categories (Quattoni and Torralba, 2009). Human-symbiotic robots are expected to become common in our daily life in the near future. For the application of these

robots, it is desirable to improve the recognition accuracy against indoor scene categories in our living environments. Robots must have ability based on learning to adapt to an environment that is changed dynamically and momentarily according to human activities and lifestyles. In scene classification and recognition, general and adaptive methods based on machine learning have been proposed according to the progress of computers' calculation performance.

Machine learning is classifiable as supervised learning and unsupervised learning. Training datasets with teaching signals are necessary for supervised learning. The load to collect teaching signals is heavy for robot users. In contrast, unsupervised learning requires no teaching signals during the learning phase. Robots learn the environment by organizing information about the environment: information obtained from various sensors. Users provide semantic information that is assigned to learning results. In contrast to supervised learning, the load for a user is lower when using unsupervised learning. Moreover, robots can discover knowledge from organized information through unsupervised learning. Advanced communication and interaction between robots and humans can be actualized using unsupervised-learning-based methods (Tsukada et al., 2011).

This paper presents an unsupervised scene classification method that is based on the context of features. This study is intended to achieve semantic recognition of indoor scenes for an autonomous mobile robot. Our method creates visual words (VWs) of two types using Gist and scale-invariant feature transform (SIFT). Using the combination of VWs, our method creates bags of features (BoFs) to vote for a two-dimensional (2-D) histogram as context-based features. Moreover, our method generates labels as a candidate of categories while maintaining stability and plasticity together using the incremental learning function of adaptive resonance theory-2 (ART-2). Our method realizes unsupervised-learning-based scene classification using generated labels of ART-2 for teaching signals of counter propagation networks (CPNs). Spatial and topological relations among scenes are mapped on the category map of CPNs. The relations of classified scenes including categories are visualized on the category map. The experiment demonstrates the classification accuracy of semantic categories such as office rooms and corridors using an open dataset as an evaluation platform of position estimation and navigation for an autonomous mobile robot.

2 Related work

Numerous methods have been proposed for semantic scene category classification and recognition. Siagian and Itti (2007) classified these methods into three approaches: object based, region based, and context based.

In object-based scene recognition, scenes are classified based on objects used for a landmark. For this approach, it is necessary to place or describe objects in a scene in advance as landmarks (Thrun, 1998; Maeyama et al., 1997). SIFT and speeded up robust features (SURF) (Bay et al., 2008) are used widely for extraction of local features of objects from scene images. In this approach, improved robustness for environmental changes is a challenging task. Given a wide range of view, the feature points to be detected are fewer because there are few objects in a scene. Pixel information of objects is affected by sensor noise and illumination changes. For this problem, Kawewong et al. (2010) proposed position-invariant robust features (PIRFs) as local features for robustness against illumination and environmental changes caused by moving objects. In fact, PIRFs perform sequential matching to extract features of SIFT and SURF from several frames of temporally distinct images. Morioka et al. (2010) actualized vision-based steady navigation PIRFs in a situation of a crowd of people at a school restaurant.

In region-based scene recognition, scene features are created in each position from the hierarchical assignment of divided regions. Katsura et al. (2003) proposed an adaptable scene classification method for weather and season changes. Their method segments outdoor scene images obtained by a mobile robot into sub-regions such as sky, buildings, and trees. Matsumoto et al. (2000) proposed a scene classification method for use with vision-based navigation. They achieved navigation using a mobile robot. Compared with object-based scene classification, region-based scene classification is robust for local changes of an environment. Moreover, region-based scene classification is practical for application to navigation as an actual task for a robot because of its simple processing. However, the classification accuracy depends strongly on the results of area segmentation. In an actual environment, it is a challenging task to segment regions correctly. Shi and Malik (2000) improved the precision of segmentation using a normalized-cut method. Nevertheless, their method entails high calculation costs for segmentation related to real-time processing used for robots.

In context-based scene recognition, features of whole scenes are described after compression in a low-dimensional space based on mechanisms that humans use to recognize scenes. For this approach, the effect of the presence of objects or the precision of segmentation is low because whole-scene information can be described roughly as context. Oliva and Torralba (2006) proposed Gist as a feature to describe global features of a scene. Gist is used popularly in context-based feature description. As scene classification using Gist, Torralba et al. (2003) proposed a scene classification method that allocates the number of states on hidden Markov models (HMMs) as scene categories. In contrast, Quattoni and Torralba (2009) reported that recognition accuracy drops dramatically for most common indoor scenes when existing methods for outdoor scene classification are tested on indoor scene categories. They specially examined the classification of indoor scenes and proposed a method to improve classification accuracy for indoor scenes. Their method uses a

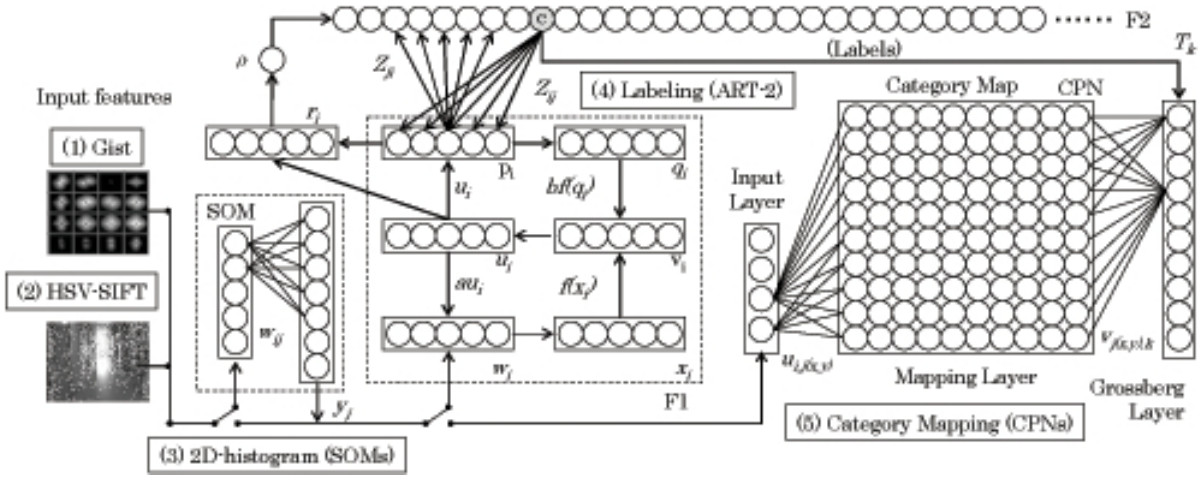


Figure 1. The whole network architecture of our method consists of the following five steps: feature description in each block using (1) Gist, feature point detection and description using (2) HSV-SIFT, creation of 2-D histograms using (3) SOMs, generation of labels using (4) ART-2, and creation of category maps using (5) CPNs.

metric function for classifying SIFT features obtained from regions of interest (ROIs) and features of Gist on the whole image. However, search results of ROI depend strongly on the classification results because their method requires manual annotation.

3 Context-based unsupervised scene classification

The aim of this study is to realize semantic recognition of indoor scenes for an autonomous mobile robot. This paper presents an unsupervised scene classification method without setting the number of categories in advance. We present the whole architecture of our method and explain each procedure as described below.

Figure 1 presents the network architecture used for our method. The procedure consists of the following five steps:

1. feature description in each block using Gist,
2. feature point detection and description using HSV-SIFT,
3. creation of 2-D histograms using SOMs,
4. generation of labels using ART-2,
5. and creation of category maps using CPNs.

Steps 1–3 correspond to creation of BoFs based on context using Gist and HSV-SIFT. Steps 4 and 5 correspond to unsupervised-learning-based scene classification. Detailed procedures in each step are described as follows.

3.1 Feature description of Gist

Gist is a general term of semantic scene categories, layout of contained objects, and attribution and knowledge related to primary objects in a scene (Torralba, 2009). The Gist of a scene is characterized as a context that exists for an object in a scene (Takeuchi, 2009). Our method uses features obtained using Gist as background features for describing context.

Gist is a feature extraction method proposed by Oliva and Torralba (2006). Primarily, Gist is used for describing structural features in outdoor scenes such as roads, mountains, and buildings. In Gist, frequencies in each block are analyzed using Fourier transformation for dividing regions to $n \times n$ blocks in an image. Moreover, filtering is conducted in each block with cutoff frequencies. Features are extracted to calculate the intensity of arbitrary directional filters for the block after filtering. In our method, we set the number of blocks n as four blocks. The cutoff frequencies are 1, 2, and 4 cycles/image. Orientation filters are 8, 8, and 4 directions in each cutoff frequency. Features are calculated in each color space. Therefore, the feature dimensions per block are 60 dimensions in our method.

3.2 Feature description of HSV-SIFT

In generic object recognition, SIFT is widely used for describing local features (Yanai, 2006). In object-based scene classification, features of objects in a scene are described using SIFT (Siagian and Itti, 2007). Our method uses features obtained using HSV-SIFT (Bosch et al., 2008) as foreground features for describing the context.

The SIFT algorithm consist of two steps: feature point extraction and feature description. Actually, difference of Gaussian (DoG) is used for feature point extraction. A pixel is detected as a candidate of a feature point if the attentional pixel that is compared with pixels of 26 neighboring pixels using DoG is selected for the extreme value. Detected feature points as candidates are refined because numerous feature points are included on linear edges. Subsequently, weighted

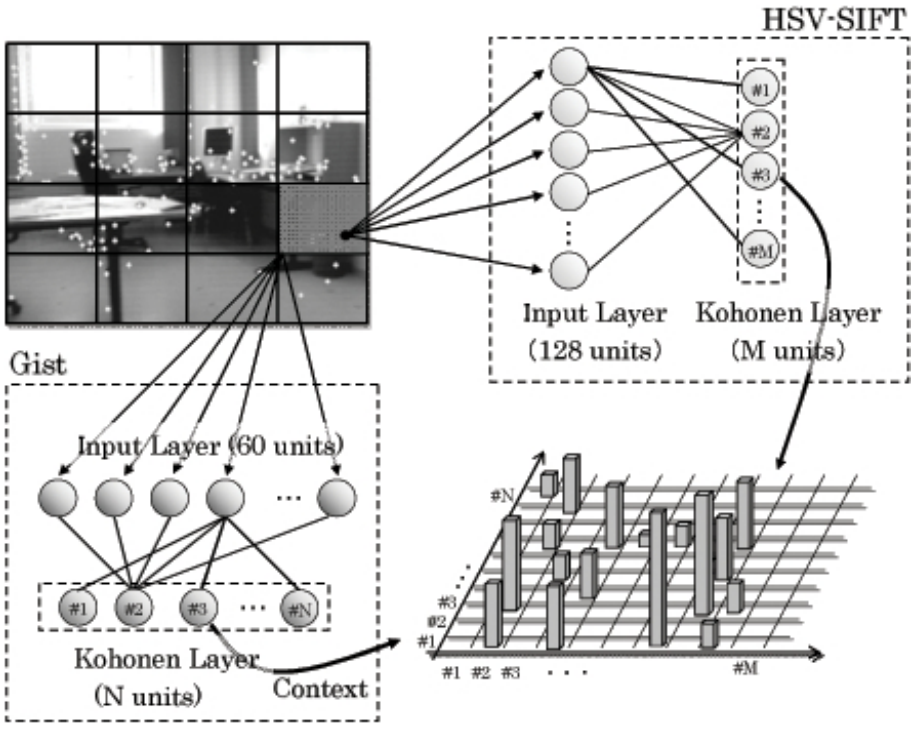


Figure 2. Procedure for creating 2-D histograms. The vertical and horizontal axes show VWs of Gist and VWs of HSV-SIFT, respectively.

orientation histograms are calculated from the gradient intensity and orientation on surrounding regions of a feature point. In the step of feature description, histograms of eight directions are created in the region of 4×4 blocks. Therefore, 128×3 dimensional features are calculated. Features of all points are calculated using this procedure.

3.3 Creation of 2-D histograms using SOMs

We create 2-D histograms as BoFs. Figure 2 portrays the procedure for creating 2-D histograms. The vertical and horizontal axes respectively portray VWs of Gist and VWs of HSV-SIFT. Herein, x_i is an x coordinate position of a VW on the i-th HSV-SIFT feature point. The block of Gist on which the i-th point is located is specialized. Subsequently, y_i is a y coordinate position of a VW on Gist features. The 2-D histogram is created with voting the position of (x_i, y_i) for all HSV-SIFT feature points in each image. Our method can describe local and global features as contexts using part-based description of objects as a foreground region and its global feature description as a background region.

Nagahashi et al. (2009) proposed a context-based feature description method using 2-D histograms. Using their method, using SIFT, foreground features are extracted from annotated object regions and background features are extracted from the region on the scale of six times from each foreground feature. However, their method requires boundaries between foreground and background regions in advance. In contrast, our method can apply images without boundaries between foreground and background regions for mapping HSV-SIFT as foreground features and Gist as background features.

For creating VWs, we use self-organizing maps (SOMs) proposed by Kohonen (1995), although k means (McQueen, 1967) is widely used. Terashima et al. (1996) reported that the false recognition rate is minimized using SOMs rather than k means for unsupervised classification. In the preliminary experiment associated with this study, we confirmed that SOMs is superior to k means for creating VWs. In the learning step, SOMs update weights with maintained topological structures of input data. A unit that contains similar weights to those of input data is designated as a burst. The burst unit forms neighborhood regions for updating weights. SOMs can classify input data of various patterns that are similar to training data. The training algorithm of SOMs is identical to the algorithm between the input layer and the Kohonen layer of CPNs.

3.4 Creation labels using ART-2

Actually, ART-2 proposed by Carpenter and Grossberg (1987) is a theoretical model of unsupervised neural networks used to form categories for time-series datasets while maintaining stability and plasticity together. Additionally, ART-2 creates labels as a candidate of categories. View images obtained from a mobile robot are changed dynamically

according to its movements. We considered that application of ART-2, which can additionally learn time-series datasets, is useful for scene classification by a mobile robot.

The network of ART-2 consists of two fields: field 1 (F1) for feature representation and field 2 (F2) for category representation. The F1 consists of sublayers. The sublayers actualize short-term memory (STM), which enhances features of input data and removes noise for a filter. The F2 actualizes long-term memory (LTM) based on finer or coarser recognition categories. In this study, we use these categories as labels.

The learning algorithm of ART-2 is the following. Points F1 and F2 are connected via the sub-layer p_i. Input data I_i are presented to F1. After propagating F1, the maximum active unit T_J is searched as

$$T_J(t) = \max\left(\sum_j p_i(t) Z_{ij}(t)\right). \tag{1}$$

Then top-down weights Z_{ji} and bottom-up weights Z_{ij} are updated as shown below.

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_{ji}(t) = d[p_i(t) - Z_{ji}(t)] \tag{2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_{ij}(t) = d[p_i(t) - Z_{ij}(t)] \tag{3}$$

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

$$r_i(t) = \frac{u_i(t) + cp_i(t)}{e + \|u\| + \|cp\|} \qquad \frac{\rho}{e + \|r\|} > 1. \tag{4}$$

The active unit is reset and goes back to the searching step again if Eq. (4) is true. Repeat propagation in F1 until the change of F1 is sufficiently small if Eq. (4) is not true.

3.5 Category map formation using CPNs

The CPNs proposed by Nilsen (1987) are supervised and self-organizing neural networks that combine Kohonen's competitive learning algorithm and Grossberg's outstar learning algorithm. The network comprises three layers: an input layer, a Kohonen layer, and a Grossberg layer. Our method uses CPNs for unsupervised learning to provide labels created by ART-2 for teaching signals to the Grossberg layer. The CPNs perform automatic labeling with this mechanism. Our method can create labels as a candidate of a category without setting the number of categories in advance. Moreover, our method can visualize spatial relations among categories based on their similarities.

The CPN learning algorithms are the following. $u^i_{n,m}(t)$ are weights from an input layer unit $i (i = 1, ..., I)$ to a Kohonen layer unit $(n, m)(n = 1, ..., N, m = 1, ..., M)$ 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

$$d_{n,m} = \sqrt{\sum_{i=1}^{I} (x_i(t) - u^i_{n,m}(t))^2}. \tag{5}$$

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

$$c = \mathrm{argmin}(d_{n,m}). \tag{6}$$

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

$$u^i_{n,m}(t+1) = u^i_{n,m}(t) + \alpha(t)(x_i(t) - u^i_{n,m}(t)). \tag{7}$$

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

$$v^j_{n,m}(t+1) = v^j_{n,m}(t) + \beta(t)(t_j(t) - v^j_{n,m}(t)). \tag{8}$$

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 concomitantly with the learning progress. The learning of CPNs repeats up to the learning iteration that was set previously.

4 Experimental results obtained using KTH-IDOL

The KTH-IDOL (image database for robot localization) (Luo et al., 2006) is an open image database used for navigation, localization, and position estimation for a mobile robot in an indoor environment. This database is used as a benchmark in indoor scene recognition (Pronobis et al., 2010). In this experiment, we evaluated the classification accuracy of semantic scene categories using this database.

4.1 KTH-IDOL

KTH-IDOL consists of time-series images obtained using two robots: Dumbo and Mannie. The resolution of images is 320×240 pixel captured at 5 fps. Target scenes were of five categories: printer area (PA), one-person office (EO), two-person office (BO), kitchen (KT), and corridor (CR). The images at CR are numerous because the robot accesses each room via the corridor according to the route assigned preliminarily.

For environmental variation, this database has three weather and illumination conditions: cloudy, night, and sunny. Moreover, the robot turned 360° to the center of each room to obtain various view data. We used images obtained using Mannie under the sunny weather condition.

Table 1. Setting values of major parameters.

θ	A	B	$N \times M$ [unit]	O [epoch]
0.10	0.50	0.50	30×30	10 000

Table 2. Setting values of ρ in each feature representation method.

Gist	Gray SIFT	OC-SVM	PIRF	Our method
0.80	0.80	0.93	0.95	0.95

4.2 Experimental setup

We compared our method to four existing feature representation methods: (1) Gist, (2) gray SIFT by Lowe (1999), (3) SIFT features selected using one-class support vector machine (OC-SVM) proposed by Tsukada et al. (2011), and (4) PIRF proposed by Kawewong et al. (2010). Our earlier study (Madokoro et al., 2011, 2012; Utsumi et al., 2010) used comparison with (1) and (2). For this experiment, we add (3) and (4) to verify the superiority of our method.

Herein, for quantitative evaluation of the classification performance, we defined classification accuracy CA [%] as

$$\mathrm{CA} = \frac{S_{\mathrm{correct}}}{N_{\mathrm{total}}} \times 100, \tag{9}$$

where S_{correct} and N_{total} respectively denote the number of correct and total images.

Table 1 portrays setting values of parameters on ART-2 and CPN used for this experiment. We set respective values of ρ shown in Table 2 because classification accuracy is strongly affected by this parameter, which controls classification granularity of ART. We decided these setting values based on our preliminary experiment.

4.3 Category formation results

Figure 3 presents sample images in each scene and their feature distributions of BoF. The 2-D histograms show the intensity of BoF as frequencies of features according to grayscale values. The positions of features differ in each scene as sparse distributions of BoF. Moreover, similar distributions are shown in the same scene depicted in Fig. 3e and f.

Figure 4 portrays the category maps created using each method. The category maps of the existing four methods generated numerous labels with discrete distributions. Moreover, categories are mapped as separate areas. We consider that these results mean low similarity of features in the same category. In contrast, our method created a category map including global clusters in each scene category, although local clusters are shown partially in PA or CR. This result means that our method represents similar features as the same categories.

Table 3. Classification accuracy of each method and semantic category [%].

Methods	PA	EO	BO	KT	CR	Avg.
Gist	44.9	48.4	51.0	50.4	70.5	39.7
Gray SIFT	62.9	59.2	40.8	51.5	76.1	58.0
OC-SVM	68.1	60.8	40.8	42.0	68.2	56.0
PIRF	68.1	62.4	73.5	50.4	70.5	63.6
Our Method	80.9	89.6	70.6	69.8	86.2	79.4

4.4 Classification results

Table 3 portrays comparison results of the classification accuracy in each method and category. The classification accuracy of Gist is 39.7 %, which is the lowest among the existing methods. This result shows a similar tendency of dramatic performance decrease in indoor environments as that reported by Quattoni and Torralba (2009), although Gist has a remarkable global performance in representing features used for scene classification in outdoor environments. The classification accuracy of gray SIFT, which can extract features from objects, remains at 58.0 %, although the performance of gray SIFT is superior to that of Gist. The classification accuracy of PIRF is 63.6 %. This result is higher than the result of gray SIFT because PIRF considers continuity among images. However, this result is 15.8 % lower than our method because PIRF is used only for foreground features of SIFT and SURF. The classification accuracy of our method is 79.4 %. This result is the highest of any obtained using these five methods. For respective scene categories, our method is superior to the existing methods. Particularly, the classification accuracies of EO and PA are 89.6 and 80.9 %, which are higher than 27.2 and 12.8 % of PIRF, respectively.

4.5 Discussion

Table 4 portrays the confusion matrix in each method and scene category for analyzing details of classification results. For this matrix, the number of correct images is shown on the diagonal cells that are marked as boiled numbers. Other cells refer to the number of incorrect images and its category name shown in the column. The maximum numbers of incorrect categories are marked as underlined.

As an overall tendency, the incorrect images of CR are numerous. The classification accuracy of BO is the lowest in our method. For this matrix, 32 images of BO are misclassified as CR. The CR images are numerous because Minnie went through CR every time to access each room. We consider that the misclassification caused by the number of units labeled CR is increased on the category map shown in Fig. 4. In contrast, no misclassification exists between EO and BO, although both features are similar in original images. We consider that definitive distinguishable categories

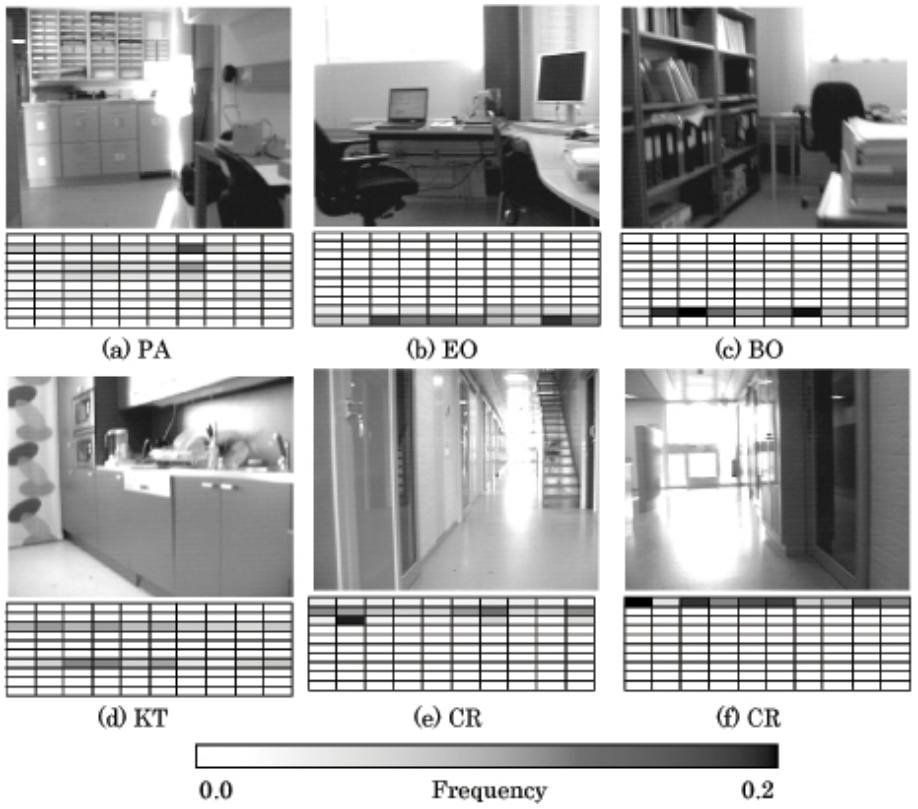


Figure 3. Sample images in each scene and their feature distributions of BoF. The 2-D histograms show the intensity of BoF as frequencies of features according to grayscale values.

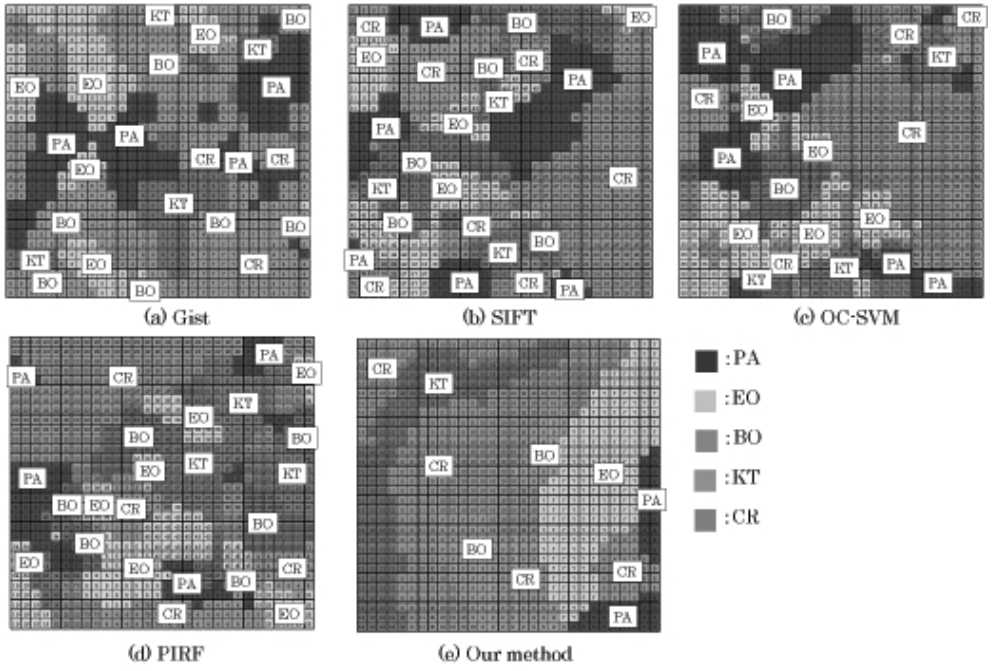


Figure 4. Category maps created using Gist, gray SIFT, OC-SVM, PIRF, and our method.

Table 4. Confusion matrix in each method [images]. Bold values show the number of correct images. Underlines show the maximum numbers of incorrect images.

Methods	Zone	PA	EO	BO	KT	CR
Gist	PA	**53**	10	25	10	18
	EO	25	**59**	4	5	29
	BO	10	14	**50**	2	22
	KT	23	24	38	**19**	25
	CR	54	17	123	39	**152**
Gray SIFT	PA	**73**	9	4	4	26
	EO	14	**74**	7	13	17
	BO	12	13	**40**	16	17
	KT	14	11	18	**67**	21
	CR	38	20	10	23	**289**
OC-SVM	PA	**79**	6	9	7	15
	EO	17	**76**	1	7	24
	BO	18	6	**40**	10	24
	KT	8	22	16	**55**	30
	CR	43	13	18	15	**159**
PIRF	PA	**71**	11	6	11	15
	EO	4	**78**	14	15	14
	BO	6	0	**72**	10	10
	KT	6	16	21	**66**	22
	CR	19	17	24	20	**268**
Our Method	PA	**114**	9	3	3	12
	EO	0	**129**	3	12	0
	BO	0	9	**87**	21	111
	KT	0	27	12	**111**	9
	CR	15	6	15	27	**393**

are created, although these features are mapped onto neighboring areas on the category map.

5 Application experiment using a mobile robot

For expanding the application range of our method, we obtained the original dataset using a mobile robot. KTH-IDOL includes ground truth (GT) labels for each room and a corridor. For this experiment, our classification targets are local areas in a corridor to actualize fine classification.

5.1 Mobile robot

As a platform for scene classification and vision-based autonomous locomotion, we developed the mobile robot prototype MEGURI (Madokoro and Takahashi, 2012) shown in Fig. 5. The objective for this robot design is to recognize an environment with similar height to that of the human eyeline. Therefore, we set the height of the robot as 170 cm, which is the average height of an adult Japanese male. For this robot, stereo camera Bumblebee2 by Point Gray Research Inc. and omni-directional sensor VS-C14U by Vstone Co., Ltd. are

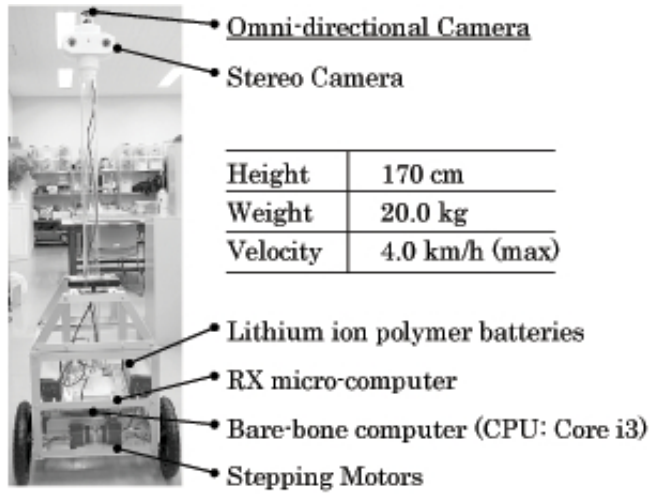


Figure 5. Prototype of our developed mobile robot and the assignment of sensors, motors, and computer systems.

used. To ensure a wide range of view, we used the omni-directional sensor to obtain time-series images. Moreover, this sensor can reduce variation related to a positional shift or turning.

The major specifications of the camera are the following: 30 mm mirror length, 15° upper view angle, 55° lower view angle, 1/3 inch interlaced CCD camera sensor, 640×480 pixel imaging resolution, and 30 fps frame capture rate. The drive unit consists of two stepping motors controlled separately. As an independent calculation environment for moving and vision processing, we equipped a RX microcomputer by Renesas Technology Corp. for motor control and a barebones computer equipped Core i5 by Intel Corp. for real-time image processing. Regarding power consumption and vibration resistance, we used a solid state drive for data storage. The power source is two lithium polymer batteries of 20 Ah. The continuous running time is about two hours. The maximum moving velocity is 4.0 km h^{-1}. The total weight is 20.0 kg including batteries.

5.2 Experimental conditions

Figure 6a portrays the experimental environment and the route for the robot. In a corridor at the fourth floor of a building in our campus, the robot ran two rounds to produce time-series images. The width, longitudinal, and lateral lengths are respectively 2, 74, and 13 m. The inside of the corridor is a void space from the second floor to the top floor.

We divided six zones from zone A to zone F shown in Fig. 6b. This is the GT for this dataset. In zone A, laboratories and staff offices are lined along the route. Zones B and F include a resting space with benches and vending machines. Zones C and E are the lateral side. Herein, both sides of zone C and the outside of zone E are, respectively, a void space

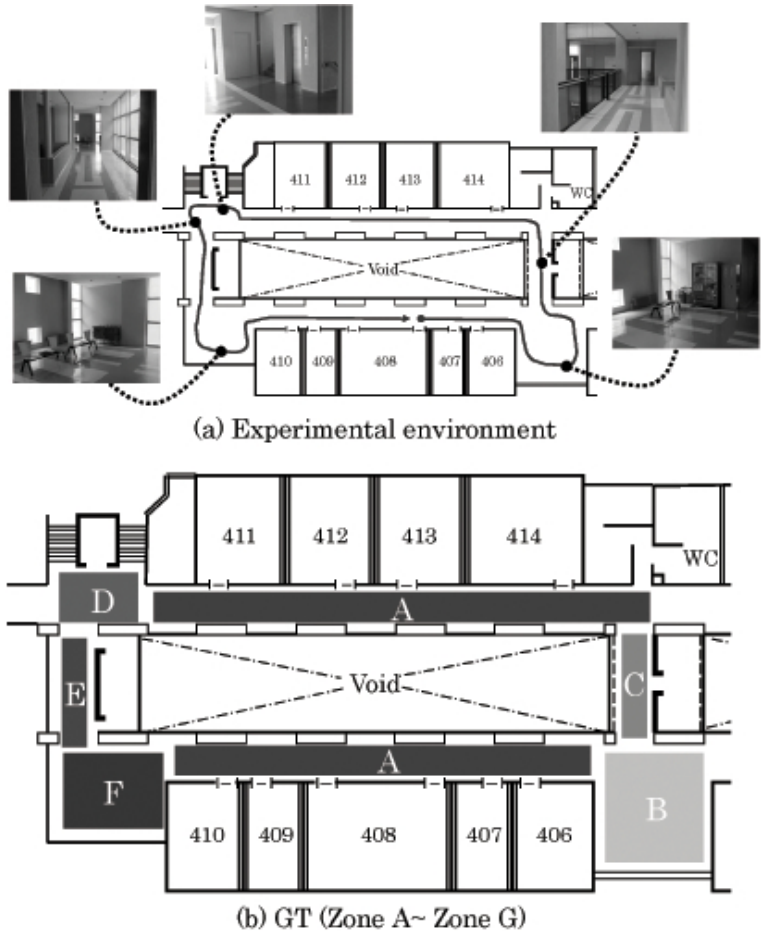


Figure 6. Experimental environment (zones A–F).

Table 5. Classification accuracy in each zone [%].

A	B	C	D	E	F	Avg.
96.8	58.1	69.7	78.0	100	96.6	83.2

and a glass wall. Therefore, the view differs in each zone. Zone D has an elevator outside of the corridor.

We set ρ to 0.90 after optimizing against this dataset. Other parameters were set to the same values as the former experiment. Regarding the amount of storage and processing time, we set the sampling rate for capturing time-series images to 3 fps.

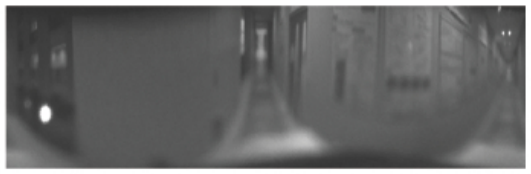
5.3 Classification results

Figure 8 depicts the category map created using our method. For this dataset, the number of images in zone A is numerous because of the longitudinal side. Therefore, numerous units on the category map were allocated to zone A. Zones B and F and zones, C, D, and E are allocated respectively to the upper left and the left side on the category map. Zones B and F consist of a wide space as a resting space with vending machines and benches. In contrast, zones C and E are the lateral side of the corridor. The similarity of scenes is represented on the category map.

Table 5 presents the classification accuracy in each zone. The mean classification accuracy is 83.2 % for six zones. The classification accuracy in zone E is 100 %. All images were classified correctly. Classification accuracy performance values in zones B and C respectively remained at 58.1 and 69.7 %. In contrast, classification accuracy performance values in zones A and F were, respectively, 96.8 and 96.6 %.

5.4 Discussion

Table 6 portrays the confusion matrix for the detail of the classification accuracy in Table 5. Zone B shows the lowest classification accuracy. From the matrix, 22 and 21 images in zone B were misclassified respectively to zones A and D. Considering the numerous images in zone A, which is the longitudinal side, zone B shows a tendency to be misclassified to zone D. Several images in zone B are mapped into



(a) Zone A (Along with labs and professor's rooms)

(b) Zone E (The right-side is glass walls)

(c) Zone F (Wide space for resting)

Figure 7. Scene images obtained using our prototype mobile robot.

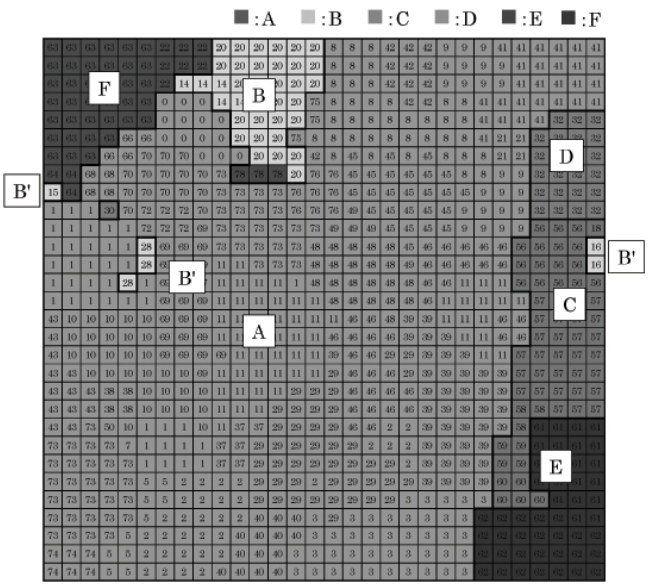


Figure 8. Formation result of category map.

discrete units inside of other categories. This tendency is apparent on the category map shown in Fig. 8. In this figure, we marked B' for these discrete distribution units. We consider that misclassification occurred from these units. The mapping of zone B to the category map can create clusters globally. However, several images are similar to other categories to some degree. Therefore, these discrete distribution units are apparent.

A challenging task is the creation of a rule for unsupervised learning-based methods to use these data as specific features or to delete them as a noise. For this classification result, these distributed data connect misclassification. Therefore, the classification accuracy can be improved to introduce a mechanism to delete them. These images are classified correctly if we change the label from B' to A or D. Classification accuracy is improved to optimize category maps. We consider that category maps can be optimized using the **U** matrix (Ultsuh, 2005) to calculate the boundary distance among categories (Honma et al., 2012).

Table 6. Confusion matrix in each zone [images]. Bold values show the numbers of correct images. Underlines show the maximum number of incorrect images.

Zone	A	B	C	D	E	F
A	**390**	4	0	0	0	0
B	22	**36**	0	21	0	3
C	7	3	**23**	0	0	0
D	6	0	3	**46**	4	0
E	0	0	0	0	**61**	0
F	0	0	0	0	2	**57**

6 Conclusions

This paper presented an unsupervised scene classification method using Gist and HSV-SIFT as a context for semantic recognition of indoor scenes used for a mobile robot. Our method represents spatial relationships among categories for mapping neighborhood units on category maps of CPNs while maintaining sequential information using labels generated from ART-2. We evaluated the classification accuracy of semantic categories using KTH-IDOL. The mean classification accuracies of Gist, gray SIFT, OC-SVM, PIRF, and our method respectively reached 39.7, 58.0, 56.0, 63.6, and 79.4 %. The results obtained using our method were 15.8 % higher than those of PIRF. For the application experiment using time-series images obtained using our developed mobile robot, the mean classification accuracy is 83.2 % for six zones. We consider that our proposed feature representation method based on the context and category formation method, combined with ART-2 and CPNs, is useful for indoor scene classification for robot vision.

We will determine a suitable number of categories from extracting category boundaries from the category map on CPNs. Moreover, we must extend the application of our method to a dynamic environment and environments in which numerous pedestrians are present.

Edited by: L. Aliouane
Reviewed by: two anonymous referees

References

Bay, H., Ess, A., Tuytelaars, T., and Gool, L.: SURF: Speeded Up Robust Features, Comput. Vis. Image Und., 110, 346–359, 2008.

Bosch, A., Zisserman, A., and Munoz, X.: Scene Classification Using a Hybrid Generative Discriminative Approach, IEEE T. Pattern Anal., 30, 712–727, 2008.

Carpenter, G. A. and Grossberg, S.: ART 2: Stable Self-Organization of Pattern Recognition Codes for Analog Input Patterns, Appl. Optics, 26, 4919–4930, 1987.

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

Honma, K., Madokokoro, H., and Sato, K.: Estimation of Interests and Classification of Behavior Patterns with Trajectory Analysis Used for Event Sites, The IEICE transactions on information and systems, J95-D 10, 1848–1858, 2012.

Kanada, T., Hirano, T., Eaton, D., and Ishiguro, H.: Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial, Hum.-Comput. Interact., 19, 61–84, 2004.

Katsura, H., Miura, J., Hild, M., and Shirai, Y.: A View-Based Outdoor Navigation Using Object Recognition Robust to Changes of Weather and Seasons, Proc. IEEE/RSJ Int'l Conf. Intelligent Robot and Systems, 2974–2979, 2003.

Kawewong, A., Tangruamsub, S., and Hasegawa, O.: Position-invariant Robust Features for Long-term Recognition of Dynamic Outdoor Scenes, IEICE Trans. Information and Systems, E93-D, 9, 2587–2601, 2010.

Kohonen, T.: Self-Organizing Maps, Springer Series in Information Sciences, 1995.

Lowe, D. G.: Object Recognition from Local Scale-Invariant Features, Proc. IEEE I. Conf. Com. Vis., 2, 1150–1157, 1999.

Luo, J., Pronobis, A., Caputo, B., and Jensfelt, P.: The KTH-IDOL2 database, Technical Report CVAP304, Kungliga Tekniska Hoegskolan, CVAP/CAS, 2006.

Madokoro, H., Utsumi, Y., and Sato, K.: Unsupervised Scene Classification Based on Context of Features for a Mobile Robot, Proc. 15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Part I, 446–455, 2011.

Madokoro, H. and Takahashi, J.: Generic Object Recognition Based on Unsupervised Learning Using Mobile Robot, Proc. Tateishi Science and Technology Fundation, 21, 87–90, 2012.

Madokoro, H., Utsumi, Y., and Sato, K.: Scene Classification Using Unsupervised Neural Networks for Mobile Robot Vision, Proc. Society of Instrument and Control Engineers Annual Conference, 1568–1573, 2012.

Maeyama, S., Ohya, A., and Yuta, S.: Long Distance Outdoor Navigation of an Autonomous Mobile Robot by Playback of Perceived Route Map, Proc. Fifth Int'l Symp. Experimental Robotics, 185–194, 1997.

Matsumoto, Y., Inaba, M., and Inoue, H.: View-Based Approach to Robot Navigation, Proc. Int'l Conf. Intelligent Robots and Systems, 1702–1708, 2000.

McQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations, Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297, 1967.

Morioka, H., Yi, S., Tongprasit, N., and Hasegawa, O.: Visual SLAM in Crowded Environments and Mobile Robot Navigation, Proc. the 28th Annual Conference of the Robotics Society of Japan, 2010.

Nagahashi, T., Ihara, A., and Fujiyoshi, H.: Tendency of Image Local Features that are Effective for Discrimination by using Bag-of-Features in Object Category Recognition, IPSJ SIG Notes Computer Vision and Image Media, 3, 13–20, 2009.

Nielsen, R. H.: Counterpropagation networks, Appl. Optics, 26, 4979–4983, 1987.

Oliva, A. and Torralba, A.: Building the gist of a scene: the role of global image features in recognition, Visual Perception, Prog. Brain Res., 155, 23–26, 2006.

Pronobis, A., Xing, L., and Caputo, B.: Overview of the CLEF 2009 Robot Vision Track, Proc. 10th international conference on Cross-language evaluation forum: multimedia experiments, 2010.

Quattoni, A. and Torralba, A.: Recognizing Indoor Scenes, Proc. Computer Vision and Pattern Recognition, 413–420, 2009.

Shi, J. and Malik, J.: Normalized Cut and image Segmentation, IEEE T. Pattern Anal., 22, 8881–8905, 2000.

Siagian, C. and Itti, L.: Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attentinon, IEEE T. Pattern Anal., 29, 300–312, 2007.

Takeuchi, T.: Underlying Mechanisms of Scene Recognition and Visual Search, ITE Technical Report, 33, 24, 7–14, 2009.

Terashima, M., Shiratani, F., and Yamamoto, K.: Unsupervised Cluster Segmentation Method Using Data Density Histogram on Self-Organizing Feature Map, The transactions of the Institute of Electronics, Information and Communication Engineers, J79-D-II, 7, 1280–1290, 1996.

Thrun, S.: Finding Landmarks for Mobile Robot Navigation, Proc. IEEE Int'l Conf. Robotics and Automation, 958–963, 1998.

Torralba, A., Murphy, K. P., Freeman, W. T., and Rubin, M. A.: Context-Based Vision System for Place and Object Recognition, Proc. IEEE Int'l Conf. Computer Vision, 1023–1029, 2003.

Torralba, A.: How many pixels make an image?, Visual Neurosci., 26, 123–131, 2009.

Tsukada, M., Utsumi, Y., Madokoro, H., and Sato, K.: Unsupervised Feature Selection and Category Classification for a Vision-Based Mobile Robot, IEICE Trans. Inf. & Sys., E94-D, 1, 127–136, 2011.

Ultsch, A.: Clustering with SOM U×C, Proc. Workshop on Self-Organizing Maps, 75–82, 2005.

Utsumi, Y., Tsukada, M., Madokoro, H., and Sato, K.: Selection of SIFT Feature Points for Scene Description in Robot Vision, Proc. IEEE Sys. Man Cybern., 2276–2281, 2010.

Wu, J., Christensen, H. I., and Rehg, J. M.: Visual Place Categorization: Problem, Dataset, and Algorithm, Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems, 4763–4770, 2009.

Yanai, K.: The Current State and Future Directions on Generic Object Recognition, IPSJ SIG Notes Computer Vision and Image Media, 121–134, 2006.