Pattern Recognition 45 (2012) 707–719

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Image representation for generic object recognition using higher-order local autocorrelation features on posterior probability images

Tetsu Matsukawa [a,*], Takio Kurita [b]

[a] The University of Tokyo, Institute of Industrial Science, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
[b] Faculty of Engineering and Undergraduate Course of Integrated Arts and Sciences, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima 739-8524, Japan

ARTICLE INFO

Article history:
Received 23 September 2009
Received in revised form
13 January 2011
Accepted 13 July 2011
Available online 28 July 2011

Keywords:
Image recognition
Higher-order local autocorrelation feature
Bag-of-features
Posterior probability image

ABSTRACT

This paper presents a novel image representation method for generic object recognition by using higher-order local autocorrelations on posterior probability images. The proposed method is an extension of the bag-of-features approach to posterior probability images. The standard bag-of-features approach is approximately thought of as a method that classifies an image to a category whose sum of posterior probabilities on a posterior probability image is maximum. However, by using local autocorrelations of posterior probability images, the proposed method extracts richer information than the standard bag-of-features. Experimental results reveal that the proposed method exhibits higher classification performances than the standard bag-of-features method.

1. Introduction

Generic object recognition technologies have many possible applications such as automatic image search. However, generic object recognition involves some very difficult problems, because one has to deal with inherent object/scene variations as well as difficulties in viewpoint, lighting, and occlusion. Thus, although many methods of generic object recognition have been developed so far, the classification performance of these conventional methods are still insufficient, and a method that can achieve high classification accuracy is strongly desired.

The bag-of-features approach is the most popular approach for generic object recognition [1] because of its simplicity and effectiveness. This approach is originally inspired from the text recognition method called "bag-of-words," and this method treats an image as an orderless collection of quantized appearance descriptors extracted from local patches. The main steps of the bag-of-features are (1) detection and description of image patches, (2) assigning patch descriptors to a set of predetermined codebooks with a vector quantization algorithm, (3) constructing a bag-of-features, which counts the number of patches assigned to each codebook, and (4) applying a classifier by treating the bag-of-features as the features vector and thus determining the category which an image can be assigned.

It is known that the bag-of-features method is robust with regard to background clutter, pose changes, and intraclass variations and offers good classification accuracy. For example, the evaluation using several local features and kernel classifiers across several object datasets showed the effectiveness of the bag-of-features method under challenging real-world conditions [2]. However, several problems still exist with regard to its application to image representation. To solve these problems, many methods have been proposed. These methods include spatial pyramid partitioning that utilizes location information [3], higher-level codebook creation based on local co-occurrence of codebooks [4–6], improvement of codebook creation [7–10], and image matching based on the region of interest [11]. All these methods are based on the histogram of local appearance, and information pertaining to semantic class labels is not used for feature representation.

In this paper, we present a novel method that improves upon the bag-of-features method. The main feature of the proposed method is that it utilizes posterior probability images for semantic feature extraction. The standard bag-of-features method is approximately thought of as a method that classifies an image to a category whose sum of posterior probabilities on a posterior probability image is maximum. This method does not utilize local co-occurrence of posterior probability images. We applied higher-order local autocorrelations [12] on posterior probability images, so as to extract richer information regarding these images. We call this image representation method as "probability higher-order local autocorrelations (PHLAC)." PHLAC has certain desirable properties for image recognition, namely, shift invariance, additivity,

[*] Corresponding author. Tel.: +81 3 5452 6461; fax: +81 3 5452 6280.
E-mail addresses: t.matsukawa@aist.go.jp, te2@iis.u-tokyo.ac.jp (T. Matsukawa).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.07.018

and synonymy [13] invariance. Furthermore, the feature dimension of PHLAC is independent of the codebook size, and it depends on the class number, which is usually much smaller than the codebook size. We confirm that the classification performance of this image representation method (PHLAC) is considerably better than that of the standard bag-of-features method and offers competitive performance to the bag-of-features using spatial information.

We also extend PHLAC to autocorrelations of posterior probability calculated from multiple image features. We call this image representation method as "multiple features probability higher-order local autocorrelations (MFPHLAC)." It is confirmed that MFPHLAC can achieve a slightly better performance than PHLAC.

This paper is an extended version of the paper cited in [14]. The extensions include an algorithm of MFPHLAC, experimental results of multiple spatial intervals, and discussions on feature dimension.

2. Related studies

We intend to improve the classification accuracy of the bag-of-features method by introducing local co-occurrence and information pertaining to semantic class labels. From these points of view, the following related studies have been reported.

Image feature extraction using local co-occurrence is recognized as an important concept [12] for image recognition. Recently, several methods have been proposed using local co-occurrence. These methods are categorized as the methods that use feature level co-occurrence and those that use codebook level co-occurrence. The examples of the methods that use feature level co-occurrence are the local self similarity method [15], gradient local autocorrelations (GLAC) [16], and color index local autocorrelation (CILAC) [17]. Low-level co-occurrence of image properties such as edge direction and color can be represented by these features, whereas the codebook level co-occurrence can capture the co-occurrence of local appearance of images. The examples of the methods that use codebook level co-occurrence are correlatons [5] and visual phrases [6]. For using codebook level co-occurrence, we need a large number of dimensions, e.g., even when the co-occurrence of only two codebooks is considered, the dimensions should be in proportion to the square of the codebook sizes. It is known that a large number of codebooks improves the classification performance [8], and a hundreds to thousands number of codebooks is generally used. Thus, the features selection method or dimension reduction method is necessary for using codebook level co-occurrence, and current researches are focused on methods to mine frequent and distinctive codebook sets [18,6,13]. The expressions of co-occurrence using a generative model such as latent Dirichlet allocation have also been proposed [4,19]. However, these methods require a complex latent model and expensive parameter estimations. A simpler method is favorable for real applications. Our proposed method can be easily implemented, and its feature dimension is relatively low (linear size of the number of categories) and effective for classifications, because it is based on autocorrelations of continuous values on posterior probability images.

From the viewpoint of the semantic feature representation using class label information, Rasiwasia et al. [20] proposed feature representation by using the bag-of-features method based on the Gaussian mixture model. In their study, each theme vector indicated the probability of each class label, and they refer to this type of scene labeling as casual annotation. Using this feature, they could achieve high classification accuracy with low feature dimensions. Methods that provide posterior probability to a codebook have also been proposed by Shotton et al. [21]. However, these methods do not employ the co-occurrence of codebooks.

3. Probability higher-order local autocorrelations

3.1. Posterior probability images

Let I be an image region, and $\mathbf{r}=(x,y)^t$ be a position vector in I. The image patches whose center is $\mathbf{r}_k$ are quantized to M codebooks $\{V_1,\ldots,V_M\}$ by local feature extraction and the vector quantization algorithm $\mathrm{VQ}(\mathbf{r}_k)\in\{1,\ldots,M\}$. These steps are the same as that of the standard bag-of-features method [3]. Posterior probability $P(c|V_m)$ of category $c\in\{1,\ldots,C\}$ is assigned to each codebook V_m using image patches on training images. Several forms of estimating the posterior probability can be used. In this study, we use two types of estimation methods.

(a) Bayes' theorem: The posterior probability is estimated by using Bayes' theorem as follows:

$$P(c|V_m)=\frac{P(V_m|c)P(c)}{P(V_m)}=\frac{P(V_m|c)P(c)}{\sum_{c=1}^{C}P(V_m|c)P(c)}, \tag{1}$$

where $P(c)=(\#$ of class c patches)/(# of all patches), $P(V_m)=(\#$ ofV_m)/(# of all patches), $P(V_m|c)=(\#$ of class $c\wedge V_m)/(\#$ of class c patches). We assume that # of class c patches are constant $(=L)$ for all class, i.e., $P(c)=(L)/(CL)=1/C$. Then, $P(c)$ becomes constant and thus we can use the following equation:

$$P(c|V_m)=\frac{P(V_m|c)}{\sum_{c=1}^{C}P(V_m|c)}. \tag{2}$$

(b) SVM weight: In our method, posterior probability is not restricted to the theoretical definition of posterior probability. Pseudo-posterior probability, which indicates the degree of support received by each category from a codebook, is also considered. The weight of each codebook, when learn by using the one-against-all linear SVM [22], is used to define pseudo-posterior probability. Assume that we use K local image patches from one image; then, the histogram of bag-of-features $\boldsymbol{H}=(H(1),\ldots,H(M))$ can be represented as follows.

$$H(m)=\sum_{k=1}^{K}\begin{cases}1 & \text{if} \quad (VQ(\boldsymbol{r}_k)=m),\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

Using this histogram, the classification function of the one-against-all linear SVM can be represented as follows:

$$\arg\max_{c\in C}\left\{f_c(\mathbf{H})=\sum_{m=1}^{M}\alpha_{c,m}H(m)+b_c\right\}, \tag{4}$$

where $\alpha_{c,m}$ is the weight of each histogram bin and b_c is the learned threshold. We transform the weight of each histogram to a non-negative value by $\alpha_{c,m}\leftarrow\alpha_{c,m}-\min\{\boldsymbol{\alpha}_c\}$ and normalize it by $\alpha_{c,m}\leftarrow\alpha_{c,m}/\sum_{m=1}^{M}\alpha_{c,m}$. Then, we can obtain the pseudo-posterior probability by using the SVM weight as follows:

$$P(c|V_m)=\frac{\alpha_{c,m}-\min\{\boldsymbol{\alpha}_c\}}{\sum_{m=1}^{M}(\alpha_{c,m}-\min\{\boldsymbol{\alpha}_c\})}. \tag{5}$$

We use the SVM weight to obtain pseudo-posterior probability, because the proposed method becomes a complete extension of the standard bag-of-features method when this pseudo-posterior probability is taken into consideration (Section 3.3).

In this study, the grid sampling of local features [3] is carried out at pixel interval of p for simplicity. We denote the set of sample points as I_p and the map of (pseudo) posterior probability of the codebook of each local region as a posterior probability image. Examples of posterior probability images are shown in

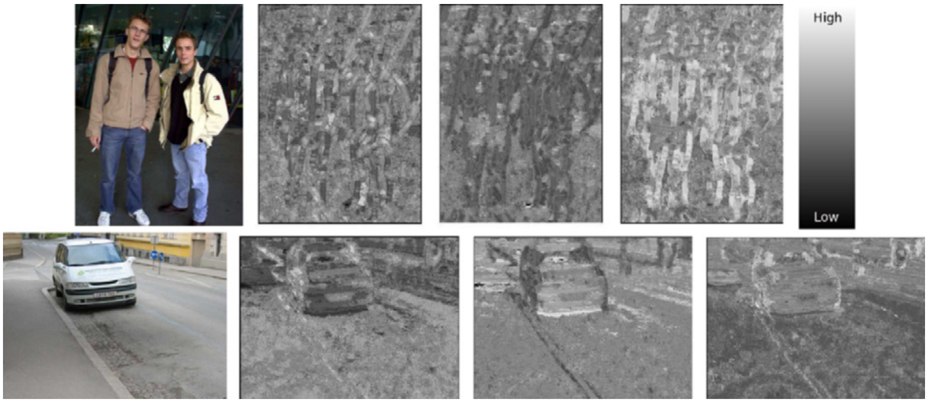

Fig. 1. Posterior probability images (Bayes' theorem): Original image, posterior probability of BIKE (left), posterior probability of CAR (middle), and posterior probability of PEOPLE (right). These posterior probability images are calculated by using a two-pixel interval ($p = 2$); for easy understanding, the original images are resized to the size of the posterior probability images. The actual size of the original images is larger than the posterior probability images by $p \times p$ pixels. Local features and the codebook are the same as those used in experiment (Section 4.1).

Fig. 1. White color represents the high probability. The data are obtained from the IG02 dataset used in the following experiment (Section 4.1). The dataset contains three categories, namely, BIKE, CAR, and PEOPLE. It is observed that the human-like contours appear in the posterior probability image of the PEOPLE category. Thus, the posterior probability images contain some spatial information about the category.

3.2. PHLAC

Autocorrelation is defined as the product of signal values from different points and represents the strong co-occurrence of these points. Higher-order local autocorrelation (HLAC) [12] has been proposed for extracting spatial autocorrelations, and its effectiveness has been demonstrated in several applications such as face and texture classification [23]. To capture the spatial autocorrelations of posterior probability, we define HLAC features of posterior probability images in terms of PHLAC. The definition of the N th order PHLAC is as follows:

$$R(c,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)=\int_{I_p} P(c|V_{VQ(\mathbf{r})})P(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_1)})\cdots P(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_N)})\,d\mathbf{r}. \tag{6}$$

In practice, many forms of Eq. (6) can be obtained by varying the parameters N and $\boldsymbol{a}_n=(a_{nx},a_{ny})^t$. In this paper, these parameters are restricted to the following subset: $N \in \{0,1,2\}$ and $a_{nx},a_{ny} \in \{\pm\Delta r \times p, 0\}$. By eliminating duplicates that arise from shifts of center positions, the mask patterns of PHLAC can be represented as shown in Fig. 2. These mask patterns are the same as the 35 HLAC mask patterns [12]. Thus, PHLAC inherits the desirable properties of HLAC for object recognition, namely, shift invariance and additivity. Note that the spatial information that HLAC uses is only local autocorrelation of 3×3 pixels and the feature value is integrated in the image. This is different from the spatial information realized by spatial partitioning of the image [3]. Although PHLAC does not exhibit scale invariance, it can be realized by using several sizes of mask patterns and local features that exhibit scale invariance.

By calculating the correlations in local regions, PHLAC becomes robust against small spatial difference and noise. These local regions can be preprocessed by calculating their values in terms of various alternatives such as their max, average, or median. We found that the optimum alternative is the average.

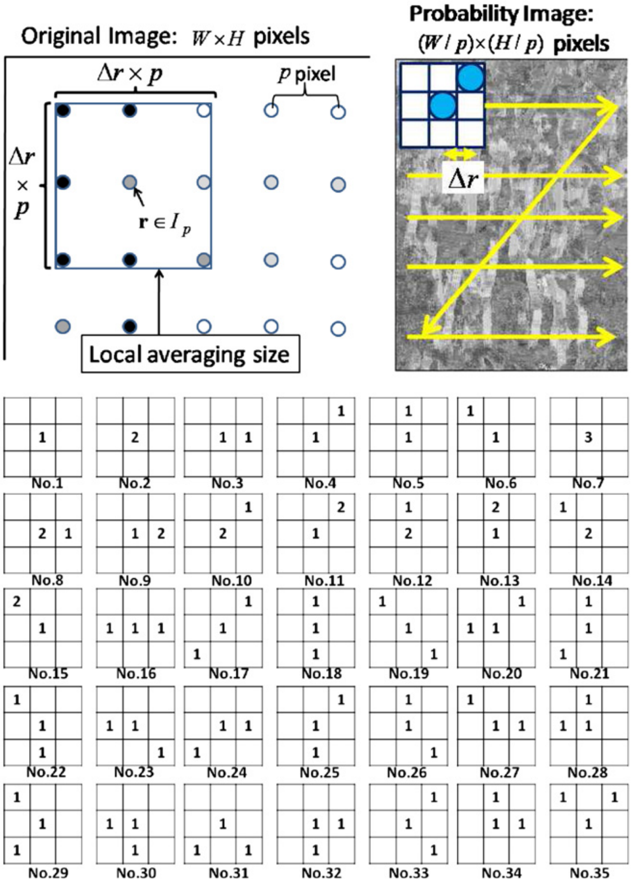

Fig. 2. PHLAC: local averaging size (left), extracting process (right) and mask patterns (bottom). The numbers {1,2,3} of the mask patterns show the frequency at which their pixel value is used for obtaining the product expressed in Eq. (6).

Thus, the practical formulation of PHLAC is given by

$$\text{0th order}: \quad R_{N=0}(c)=\sum_{r \in I_p} L_a(P(c|V_{VQ(\mathbf{r})})),$$

$$\text{1st order}: \quad R_{N=1}(c,\boldsymbol{a}_1)=\sum_{r \in I_p} L_a(P(c|V_{VQ(\mathbf{r})}))L_a(P(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_1)})),$$

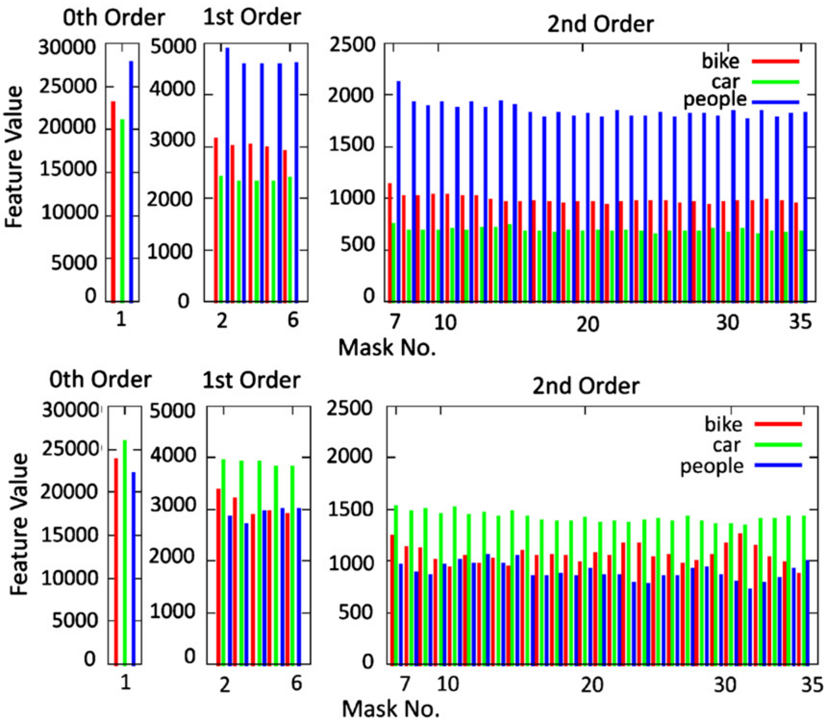

Fig. 3. Examples of PHLAC feature vector. The values $\Delta r=48$ and $p=2$ are used for the images shown in Fig. 1. Original images are those of PEOPLE (top), CAR (bottom).

$$\text{2nd order}: \; R_{N=2}(c,\boldsymbol{a}_1,\boldsymbol{a}_2)=\sum_{r\in I_p} L_a(P(c|V_{VQ(\mathbf{r})}))L_a(P(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_1)})) L_a(P(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_2)})), \tag{7}$$

where L_a represents the local averaging on a $(\Delta r\times p)\times(\Delta r\times p)$ region centered on $\boldsymbol{r}$ (Fig. 2). PHLAC is obtained by calculating the HLAC on local-averaged posterior probability images (see Algorithm 1). PHLAC is extracted from the posterior probability images of all categories; thus the total number of features of PHLAC becomes $35\times C$. Examples of PHLAC feature vector are shown in Fig. 3. It is noticed that difference in the feature values of each category is prominent, and some patterns that are different from the 0th order appear in the higher-order feature values. There are two possibilities with regard to the classification using PHLAC image representations. One is the classification using all PHLACs of all categories (PHLAC.All), and the other is using the PHLAC of one category for each one-against-all classifiers (PHLAC.Clw). We compare these classification methods in the following experiments (Section 4.1.1).

Algorithm 1. PHLAC computation

```
Training Image:
(1) Create codebooks by using local features and a clustering algorithm.
(2) Configure posterior probability of each codebook.
Training and Test Image:
(3) Create $C$ posterior probability images by using $p$ pixel intervals.
(4) Preprocess posterior probability images (local averaging).
(5) Calculate HLAC features on posterior probability images by sliding HLAC mask patterns.
```

3.3. *Interpretation of PHLAC*

Bag-of-features (0th) + *local autocorrelations* (1st + 2nd): If we use SVM weights as pseudo-probabilities, then the 0th order of the PHLAC becomes the same as that obtained during the classification by the standard bag-of-features method using linear SVM. Because **H** is a histogram (see Eq. (3)), Eq. (4) is rewritten as follows.

$$\begin{aligned}&\arg\max_{c\in C}\left\{\sum_{k=1}^{K}\alpha_{c,VQ(\mathbf{r}_k)}+b_c\right\}\\&\quad=\operatorname{argmax}_{c\in C}\left\{\sum_{k=1}^{K}(\alpha_{c,VQ(\mathbf{r}_k)}-\min\{\boldsymbol{\alpha}_c\})+K\min\{\boldsymbol{\alpha}_c\}+b_c\right\}\end{aligned}\tag{8}$$

$$\arg\max_{c\in C}\left\{\sum_{k=1}^{K}\alpha_{c,VQ(\mathbf{r}_k)}+b_c\right\}=\arg\max_{c\in C}\{A_cR_{N=0}(c)+B_c\},\tag{9}$$

where $A_c=\sum_{m=1}^{M}(\alpha_{c,m}-\min\{\boldsymbol{\alpha}_c\})$ and $B_c=K\min\{\boldsymbol{\alpha}_c\}+b_c$. (To achieve the transformation from Eq. (8) to Eq. (9), the relationship $R_{N=0}(c)=\sum_{k=1}^{K}(\alpha_{c,VQ(\mathbf{r}_k)}-\min\{\boldsymbol{\alpha}_c\})/A_c$ is used.) It can be inferred from this equation that the classification by the standard bag-of-features method is possible only by using 0th order of the PHLAC and learned parameters A_c and B_c. (It was assumed that preprocessing was not carried out in the calculation of PHLAC.) In this case, the SVM weight is used as the pseudo-posterior probability; however, it is expected that other posterior probabilities may also posses a similar property of the 0th order PHLAC. Because the histogram of the standard bag-of-features is created without using local co-occurrences, the 0th order of PHLAC is almost thought of as a one-against-all bag-of-features classification. Higher-order features of PHLAC have richer information on posterior probability images (e.g., the shape of local posterior probability distributions). Thus, if any commonly existing patterns are contained in specific classes, this representation can be expected to achieve better classification performance than the standard bag-of-features method.

The relationship between the standard bag-of-features method and PHLAC classification is shown in Fig. 4. In our PHLAC classification, we train an additional classifier using the 0th order PHLAC $\{R_{N=0}(1),\ldots,R_{N=0}(C)\}$ and use the higher-order PHLAC as a feature vector. In following experiment (Section 4.1.1), the classifier is also trained when only the 0th order PHLAC is used. Thus, only the 0th order PHLAC can possibly perform better than the standard bag-of-features method.

Synonymy invariance: Synonymous codebooks are codebooks that have similar posterior probabilities [6]. PHLAC classification can be carried out directly on the posterior probability images, and the same features can be extracted even when a local appearance of an image is exchanged with other appearances whose posterior probabilities are the same as the local appearance. This synonymy invariance is important for creating compact image representations [13].

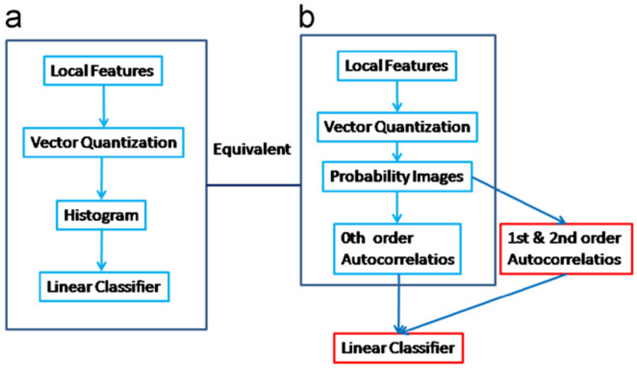

Fig. 4. Schematic comparison of the standard (a) bag-of-features classification with our proposed (b) PHLAC classification.

3.4. MFPHLAC

Recently, it has been reported that high classification performance can be achieved by implementing methods that use multiple local features in generic object recognition problems [24,25]. Although PHLAC can be calculated from posterior probability images estimated by several features independently, it is expected that richer information can be extracted by autocorrelations of posterior probability by using multiple features. We extend PHLAC to autocorrelations of posterior probability calculated from multiple image features. We call this image representation method as MFPHLAC.

Assuming that we use T ($T \geq 2$) types of local features, the definition of the N th order MFPHLAC can be expressed as follows:

$$R(c,t_0,\ldots,t_N,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N) = \int_{I_p} P_{t_0}(c|V_{VQ(\mathbf{r})})P_{t_1}(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_1)})\cdots P_{t_N}(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_N)})\,d\mathbf{r}. \qquad (10)$$

Here P_t indicates the posterior probability estimated by feature type $t \in \{1,\ldots,T\}$.

As in the case with PHLAC, the parameters N and $\boldsymbol{a}_n = (a_{nx}, a_{ny})^t$ are restricted to the following subset: $N \in \{0,1,2\}$ and a_{nx}, $a_{ny} \in \{\pm\Delta r \times p, 0\}$. Thus, the practical formulation of MFPHALC is given by

$$\text{0th order}: \; R_{N=0}(c,t_0) = \sum_{r \in I_p} L_a(P_{t_0}(c|V_{VQ(\mathbf{r})})),$$

$$\text{1st order}: \; R_{N=1}(c,t_0,t_1.\boldsymbol{a}_1) = \sum_{r \in I_p} L_a(P_{t_0}(c|V_{VQ(\mathbf{r})}))L_a(P_{t_1}(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_1)})),$$

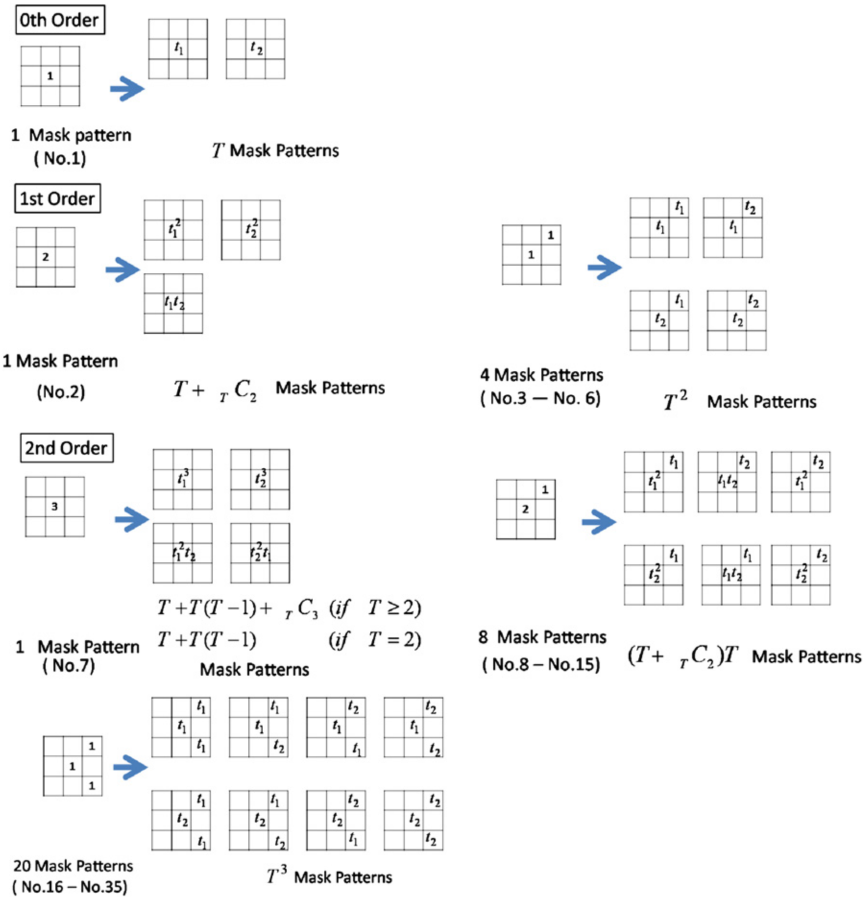

Fig. 5. Mask patterns of MFPHLACiIn the case of two features (t_1,t_2).

$$\text{2nd order}: \; R_{N=2}(c,t_0,t_1,t_2,\boldsymbol{a}_1,\boldsymbol{a}_2) = \sum_{r \in I_p} L_a(P_{t_0}(c|V_{VQ(\mathbf{r})}))L_a(P_{t_1}(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_1)}))L_a(P_{t_2}(c|V_{VQ(\mathbf{r}+\boldsymbol{a}_2)})). \tag{11}$$

Here, MFPHLAC is calculated by sliding extended mask patterns from PHLAC (Algorithm 2). By eliminating duplicates that arise from the second and third power of a certain pixel, the mask patterns of MFPHLAC can be represented as shown in Fig. 5. In Fig. 5, the mask pattern with two features is shown. The independent number of feature values that arise from the second power of a certain pixel is $T+{}_TC_2$, because there exist T combinations of the second power of the same features and ${}_TC_2$ combinations obtained by the multiplication of different feature values. For example, the number of mask patterns become 233 when $T=2$ and 739 when $T=3$. Since MFPHLAC involves the calculation of autocorrelation from multiple features, these features contain richer information than PHLAC features calculated from multiple features independently. Thus, it is expected that better classification performance can be achieved by using MFPHLAC.

Algorithm 2. MFPHLAC computation

Training Image:
(1) Create T types of codebooks by using local features and a clustering algorithm.
(2) Configure T posterior probabilities of each codebook type.

Training and Test Image:
(3) Create $C \times T$ posterior probability images by using p pixel intervals.
(4) Preprocess posterior probability images (local averaging).
(5) Calculate MFPHLAC on posterior probability images by sliding MFPHLAC mask patterns.

4. Experiment

We compared the classification performances of the standard bag-of-features method and PHLAC using three commonly used image datasets: IG02 [26], a dataset having 15 natural scene categories (Scene-15) [3], and Caltech101 dataset [32].

To obtain reliable results, we repeated the experiment 10 times except for Caltech101 dataset. Ten random subsets were selected from the data to create 10 pairs of training and test data. For each of these pairs, a codebook was created by using k-means clustering on the training set. For classification, a linear one-against-all SVM was used. For the implementation of SVM, we used LIBSVM. Five-fold cross validation was carried out on the training set to tune the parameters of SVM. The classification rate reported by us is the average of the per-class recognition rates, which in turn are averaged over 10 random test sets. With regard to Caltech101 dataset, we repeated the experiment five times.

As local features, we used a SIFT descriptor [27] sampled on a regular grid. The modification by the dominant orientation was

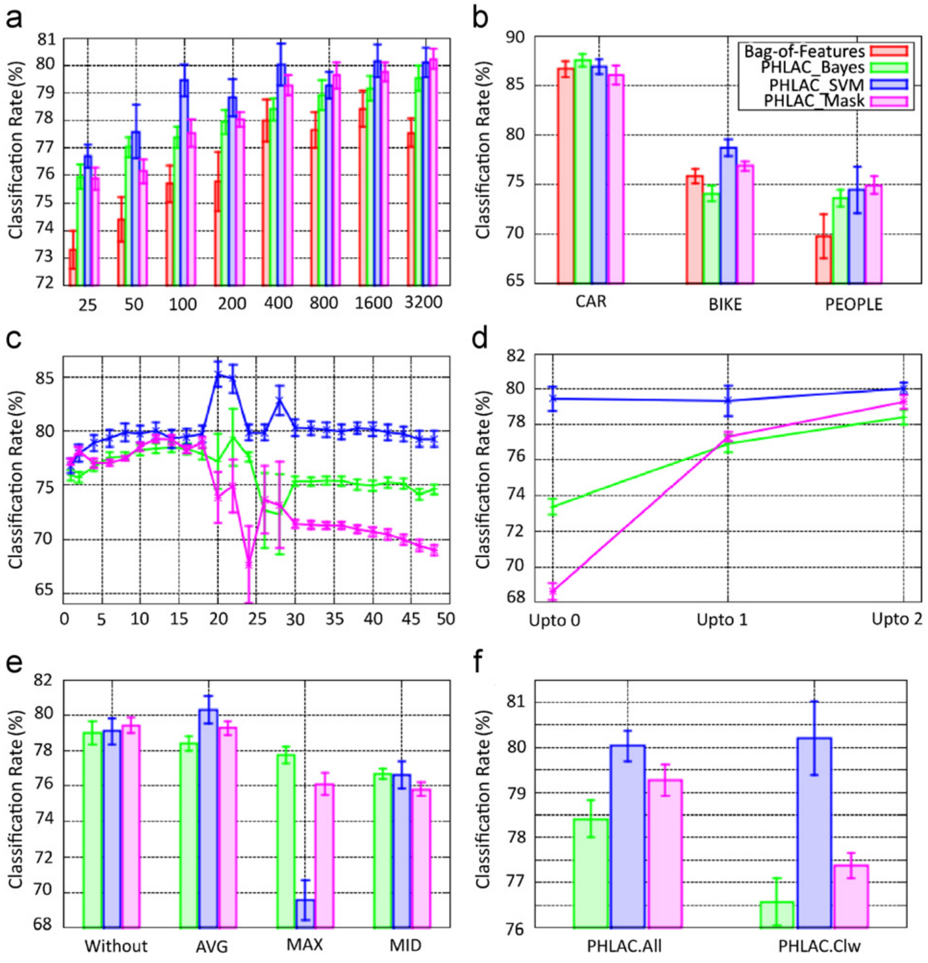

Fig. 6. Recognition rates of IG02. The basic settings are codebook size = 400 ((b)–(f)), spatial interval $\Delta r=12$ ((a), (b), (d)–(f)), and PHLAC. All (a) number of codebooks, (b) category, (c) spatial interval, (d) autocorrelation order, (e) preprocessing and (f) classification type.

not used and the descriptor was computed on a 16×16 pixel patch sampled every 8 pixels ($p = 8$). In the codebook creation process, all the features sampled every 16 pixels on all training images were used for k-means clustering. We used the L2-norm normalization method for both the standard bag-of-features method and PHLAC. In PHLAC, the features were L2 normalized by each order of autocorrelations. We denote the classification of PHLAC using posterior probability by Bayes' theorem as PHLAC_{Bayes} and PHLAC using pseudo-probability by SVM weight as PHLAC_{SVM}. It should be noted that although the SVM of the standard bag-of-features method is used in Eq. (4) of PHLAC_{SVM}, the result of the 0th order PHLAC_{SVM} is different from the result of the standard bag-of-features method because we train an additional linear SVM as mentioned in Section 3.3.

4.1. Results of IG02 dataset

4.1.1. Basic property

First, we used the IG02 [26] (INRIA Annotations for Granz-02) dataset, which contains large variations of the target size. The classification task is to classify the test images into three categories, i.e., CAR, BIKE, and PEOPLE. The number of training images in each category is 162 for CAR, 177 for BIKE, and 140 for PEOPLE. The number of test images is the same as that of the training images. We resampled 10 sets of training and test sets from all images. The image size was 640×480 pixels or 480×640 pixels. Maraszalek et al. prepared mask images that indicated the locations of the target objects. We also attempted to estimate the posterior probability of Eq. (1) by using only the local features of the target object region. We denote these PHLAC features as PHLAC_{MASK}. The experimental results are shown in Fig. 6.

Overall performance: The basic settings used were a spatial interval $\Delta r = 12$ and the classification using PHLACs of all categories (PHLAC.All). In all the codebook sizes, all types of PHLACs achieve higher classification performances than the standard bag-of-features method (Fig. 6(a)). PHLAC_{SVM} achieves higher classification rates than PHLAC_{Bayes}. By using mask images for estimating the posterior probability, the performance of PHLAC_{MASK} improves when the codebook size is larger than 400.

Recognition rates per category: The classification rates of PHLAC are higher than those of the standard bag-of-features method in almost all cases (Fig. 6(b)). Especially, the classification rates of the PEOPLE category using PHLAC are higher than those using the standard bag-of-features method for any settings of PHLAC. This is because human-like contours (shown in Fig. 1) appear in the posterior probability images obtained from images of PEOPLE; these contours were less visible in the posterior probability images obtained from images of other categories.

Spatial interval: The spatial interval appears to be better near $\Delta r = 12$ ($12 \times 8 = 96$ pixels) for all settings except for PHLAC_{SVM} (Fig. 6(c)). The classification rates of PHLAC_{Bayes} and PHLAC_{MASK} decrease as the spatial interval is increased from $\Delta r = 20$. In the case of PHLAC_{SVM}, classification rates are high even when the spatial interval increases, and the peak of the classification rates appears near $\Delta r = 20$. However, at $\Delta r = 20$, the classification rates for PHLAC_{Bayes} and PHLAC_{MASK} reduce; therefore, as a basic settings, we set the spatial interval to $\Delta r = 12$. In practice, a multiscale spatial interval is more useful than a single spatial interval, because there are several optimal spatial intervals (Section 4.1.2).

Order of autocorrelation: In the cases of PHLAC_{Bayes} and PHLAC_{MASK}, the classification rates increase with the order of autocorrelation (Fig. 6(d)). PHLAC_{SVM} exhibit a higher classification performance than other PHLACs using only 0th order autocorrelations. Thus, the PHLAC_{SVM} did not decrease the classification performance compared to other PHLACs in the non-optimal spatial intervals ($\Delta r > 22$). For experiments using up to 2nd order autocorrelations, PHLAC_{SVM} can achieve the best classification performance. Especially in the optimal spatial interval of PHLAC_{SVM} ($\Delta r = 20$), the classification using the 2nd order autocorrelation was 5.01% better than 0th order autocorrelation (Fig. 6(c)).

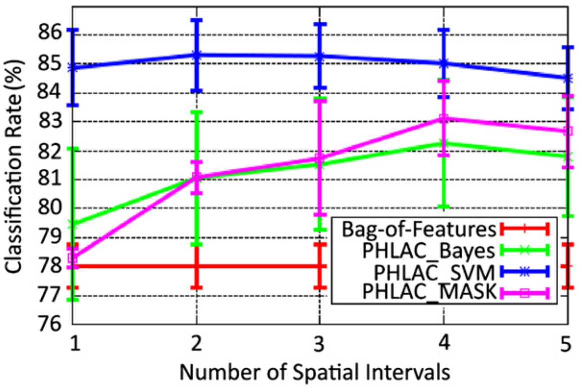

Fig. 7. Recognition rates of multiscale spatial interval (IG02).

Preprocessing: As can be observed from Fig. 6(e), the graphs of the local averaging and no preprocessing cases appear to be comparable. However, when the codebook size and spatial intervals are changed, the local averaging often outperformed the no preprocessing case. Thus, we recommend the use of local averaging for preprocessing.

Classification type: Of the different classification types, PHLAC.All exhibits better performance than PHLAC.Clw (Fig. 6(f)) in PHLAC_{Bayes} and PHLAC_{MASK}. On the other hand, when the PHLAC_{SVM} is used, the PHLAC.Clw classification performs better than the PHLAC.All. This indicates that the number of dimensions for the training of each SVM can be reduced to 35 when PHLAC_{SVM} is used.

4.1.2. Multiscale spatial interval

A multiscale spatial interval can capture several spatial co-occurrences. Thus, such an interval is expected to exhibits a higher classification performance than a single spatial interval, described in the paper cited in [23]. We concatenated the feature vector calculated from different sizes of mask patterns by varying the spatial interval Δr. The number of spatial intervals shows how many Δr is used. We experimented with all combinations of Δr by using the values {2,4,8,16,22} for each number of spatial intervals. The classification result reported in this paper is the best classification rate selected from the results obtained for these combinations.

The classification rates of PHLAC using a multiple spatial interval are shown in Fig. 7. In Fig. 7, PHLAC.All was used. It is confirmed that the performance of PHLAC_{Bayes} and PHLAC_{MASK} improved when the number of spatial intervals was increased to four. The use of PHLAC_{SVM} does not increase the accuracy because only $\Delta r = 22$ is higher than other spatial intervals. However, the performance did not decrease when a multiple spatial interval of four was used. These results indicate that the use of a multiscale spatial interval is desirable for both reducing the setting cost of Δr and improving the classification accuracy.

4.2. Results of Scene-15 dataset

4.2.1. Results of PHLAC

Next, we performed experiments on the Scene-15 dataset [3]. The Scene-15 dataset consists of 4485 images spread over 15 categories. The 15 categories contain 200–400 images each and range from natural scenes like mountains and forests to man-made environments like kitchens and offices. We selected 100 random images from each category as a training set and the

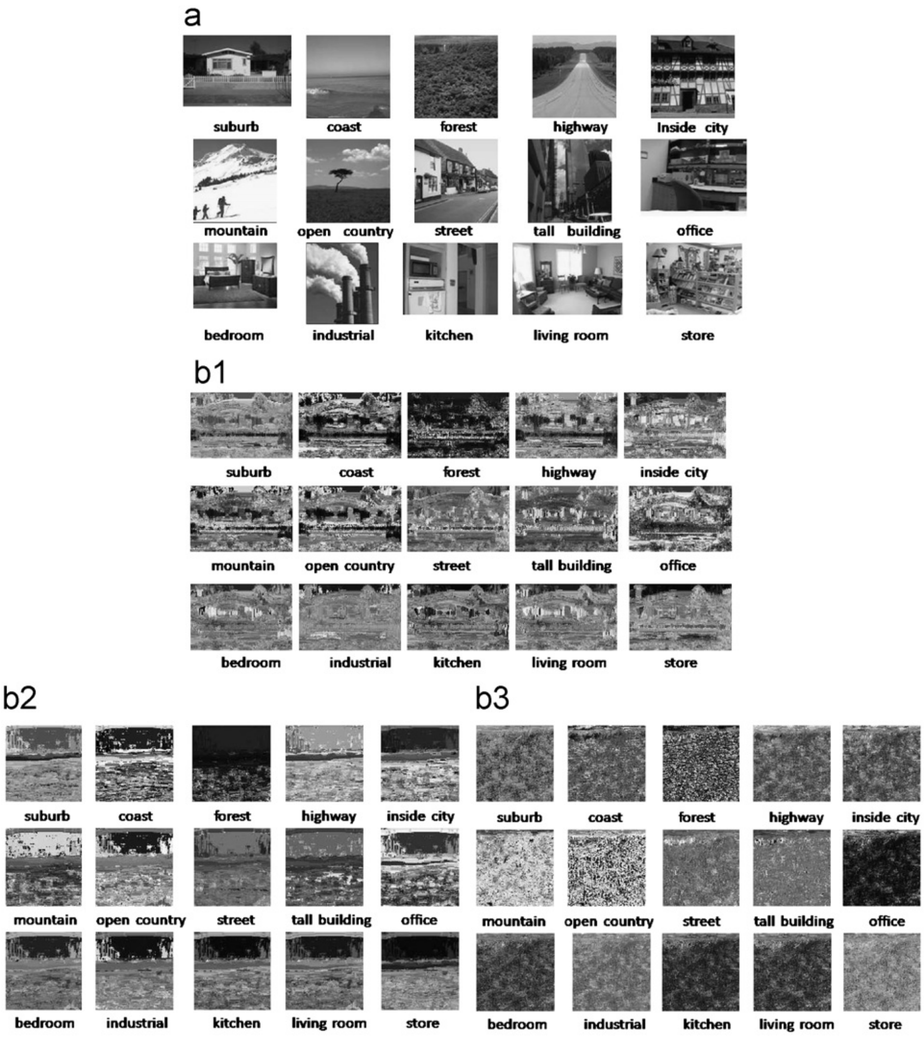

Fig. 8. Examples of Scene-15 dataset. Examples of the original images (a) and probability images (b). The original images of (b) are suburb (b-1), coast (b-2), and forest (b-3).

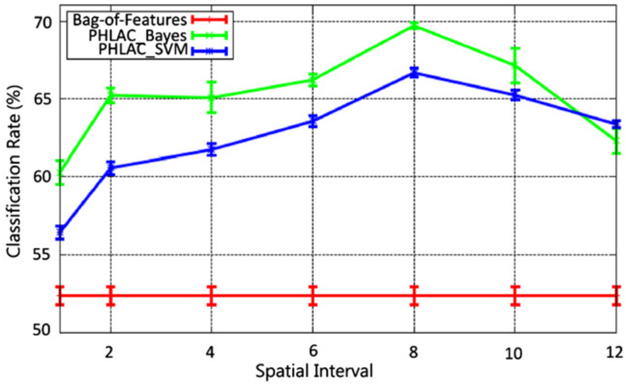

Fig. 9. Recognition rates of Scene 15 per spatial interval (codebook size is 200).

remaining images as the test set. Fig. 8 shows some examples of dataset (Fig. 8(a)) and posterior probability images (Fig. 8(b)). It is observed that each posterior probability image contains some spatial patterns.

We used PHLAC.All, and experimentally set the spatial interval to $\Delta r = 8$. This was determined by comparing the result of $\Delta r = \{1,2,4,6,8,12\}$ in codebook size 200. The classification rates in each Δr are shown in Fig. 9. The classification rates are improved as to increase the spatial interval upto $\Delta r = 8$. The actual size of each mask pattern corresponding to $\Delta r = \{1,4,8\}$ is shown in Fig. 10. Figs. 9 and 10 show the larger regions correlation produce better performance, e.g., the width of the mask pattern corresponding to $\Delta r = 8$ is about half of the original image. However, the minimum size of mask pattern ($\Delta r = 1$) already outperformed the standard bag-of-features. This shows the local autocorrelation is effective even the size of mask pattern is small.

The recognition rates for the Scene-15 dataset are shown in Fig. 11. For the Scene-15 dataset, PHLAC achieves higher recognition performances than the standard bag-of-features classification for all categories and codebook sizes. For this dataset, PHLAC_{Bayes} exhibits higher accuracy than PHLAC_{SVM}. When the codebook size is 200, the recognition rate of PHLAC_{Bayes} is 15% higher than that of the standard bag-of-features classification.

In our experiment, the classification rates of PHLAC_{Bayes} are around 69.48 ($\pm$0.27)% by using linear SVM for a codebook size of 200, and that the classification rates of the standard bag-of-features classification using a histogram intersection kernel [3] are around 66.31 ($\pm$0.15)%. Lazebnik reported differences in the 72.2 ($\pm$0.6)%; this difference can be attributed to the differences in the implementations such as feature extraction and codebook creation. The proposed method and the standard bag-of-features method use the same codebook and features used in our experiments.

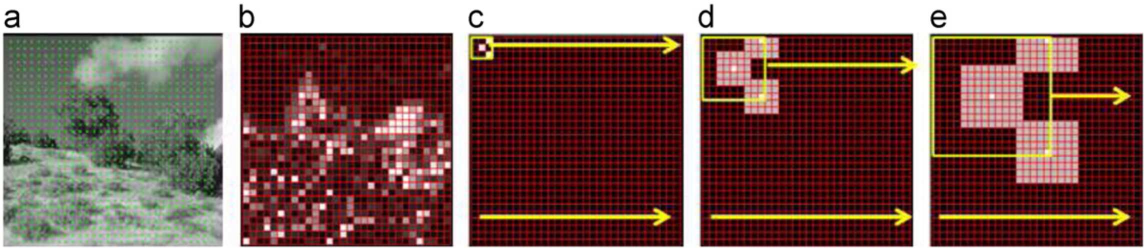

Fig. 10. Actual size of mask patterns: (a) original image, (b) probability image, (c) mask pattern of $\Delta r=1$, (d) mask pattern of $\Delta r=4$, (e) mask pattern of $\Delta r=8$, where green points of (a) is the sampling points of local features and gray areas of (c)–(e) show the local-averaged areas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

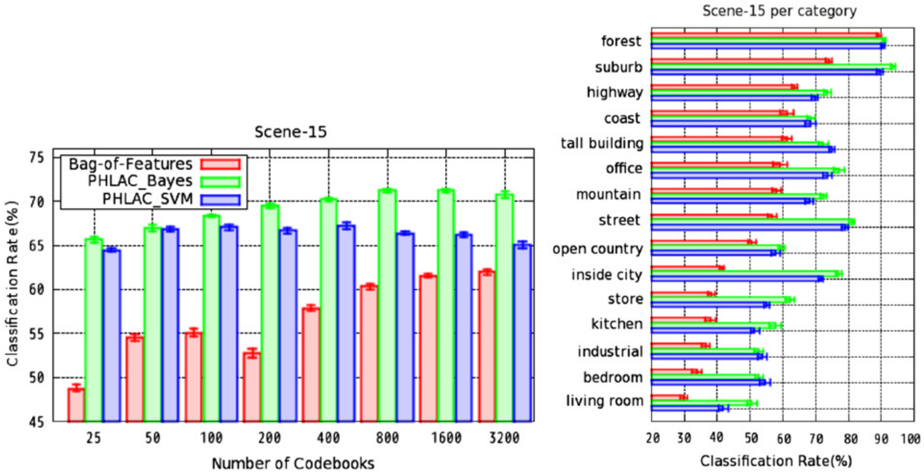

Fig. 11. Recognition rates of Scene 15 per codebook size (left) and per category (right) when codebook size is 200.

The examples of PHLAC$_{Bayes}$ features are shown in Fig. 12. These examples are of those samples that are classified correctly by PHLAC$_{Bayes}$; the bag-of-features method failed to classify these samples. It is noticed that the posterior probabilities of correct category are not maximum in 0th order; the 1st order feature values of the correct category increase for some samples (inside city and street). However, it is not necessary that the posterior probabilities of correct category are high. We can also use the other categories evidences such as mountain likely contains forest and open country like regions in both 0th and higher-order feature values for final classifiers. On the basis of all these evidences, the PHLAC classification outperformed the classification carried out using the standard bag-of-features method.

4.2.2. Results of MFPHLAC

Next, we compared MFPHLAC and PHLAC using a multiscale spatial interval. The number of features used simultaneously is restricted to 2 (T=2). We used five features as local features. These were Intensity, GLAC [16], CS-LBP [28], Texton in addition to the SIFT-like features described in the beginning of Section 4.

Intensity: A 128-dimensional intensity histogram in a 4×4 cell obtained from a 16×16 pixel patch was used. The intensity level of a pixel was divided to eight level from the original 0–255 intensity value. L1 normalization was used in each cell.

GLAC: A 256-dimensional co-occurrence histogram of gradient direction that contains four types of local autocorrelation patterns was used. We calculated the feature values from a 16×16 pixel patch, and histogram of each autocorrelation pattern was L2-Hys normalized.

CS-LBP: A 256-dimensional histogram of 64 types of intensity patterns per 4×4 cells obtained from 16×16 pixel patch was used. We applied L2-Hys normalization to each cell.

Texton: The histogram of filter responses in a 16×16 pixel patch was used. We used 13 types of Schmid filters [29] and eight directions and three sizes of the multi-resolution Gabor filter [30]. We considered the positive and negative responses of the Schmid filter; thus, the number of dimensions of the filter was 26. We considered the amplitude of the responses of Gabor filter; thus, the dimension of the filter was 24. In total, the number of dimensions of Texton was 50. We applied L2 normalization to each filter type.

For all features, we created 200 codebooks by k-means clustering. In PHLAC and the bag-of-features method using multiple features, the results were obtained by using a concatenated feature vector having multiple feature type. Posterior probability images were created by using Bayes' theorem. PHLAC.All was used for the classification method.

We concatenated the feature vector calculated from different sizes of mask patterns, as described in Section 4.1.2. The number of spatial intervals shows how many Δr is used. We experimented with all combinations of Δr by using the values {1,2,4,8,12} for each number of spatial intervals. The classification result reported in this paper is the best classification rate selected from the results obtained for these combinations. Since MFPHLAC requires a large number of dimensions, we restricted the number of the spatial intervals for MFPHLAC to 2. The features of MFPHLAC were L2 normalized by each order of autocorrelations.

It is known that the use of spatial information is very effective [3] in achieving the high accuracy for Scene-15 dataset. We also compared the proposed methods with the bag-of-features using spatial information. Spatial information is realized by spatial partitioning of an image, and then, a bag-of-features histogram is created in each spatial partition. The settings for the spatial partitioning are SI1(2×2), SI2(4×4), and PSI(1×1, 2×2, 4×4). The features of the bag-of-features method with spatial information are L2 normalized by each partitioning setting. These setting

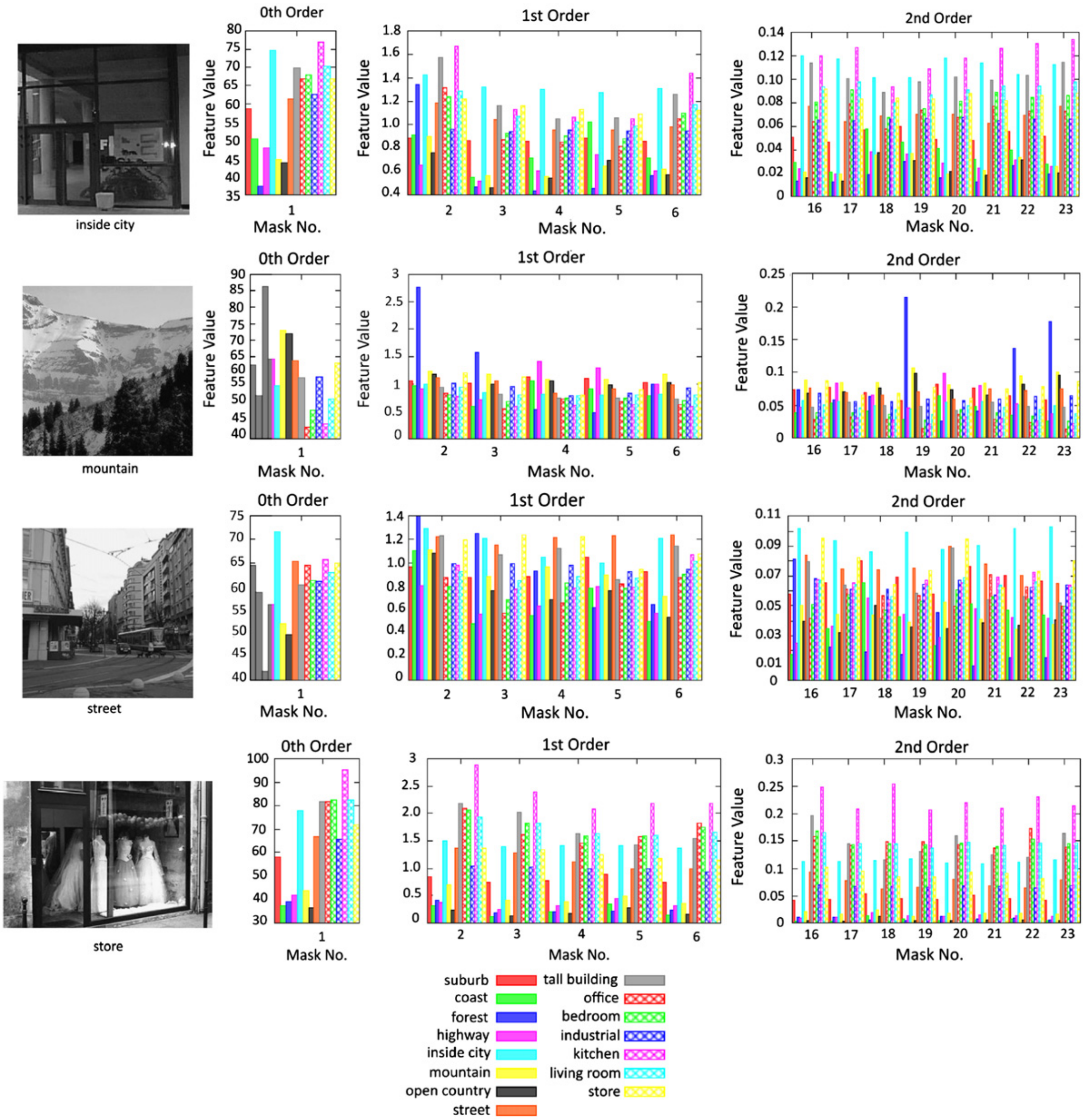

Fig. 12. Examples of PHLAC features (PHLAC$_{Bayes}$); All examples are those of the samples that were recognized correctly by PHLAC and not recognized by the bag-of-features method.

of the spatial partitioning is the same as the setting cited in [3]; however, to compare only the goodness of feature representation, linear SVM is used for all the methods. Note that spatial partitioning of an image was used only for the bag-of-features and this was not used for PHLACs.

The results are shown in Fig. 13. In all features, PHLAC achieved a considerably higher classification performance than the standard bag-of-features method. The classification performance improves better as the number of multiple spatial intervals increases. MFPHLAC achieved better performance than PHLAC for the same number of multiple spatial intervals. PHLAC performs slightly better than the spatial pyramid bag-of-features method with a single feature. The performance of MFPHLAC and PHLAC is competitive compared to that of the spatial pyramid bag-of-features method with two features.

4.3. *Results of Caltech101 dataset*

Finally, we compared PHLAC and the bag-of-features method using Caltech101 dataset [32]. The Caltech101 dataset contains 8677 images spread over 101 object categories, where the number of images in each category varies from 31 to 800 images. We used 30 images for training per category, and 50 images per category were used for testing. We repeated the random selection five times and report the average classification accuracy. Because the image size differs per image in this dataset, we resized the original images so that the all images have almost the same pixels ($z \times z$ pixels). To extract three sizes of local feature, we use three image size z and we changed the sampling interval p so that $(z,p) \in \{(100,2),(200,4),(400,8)\}$. In this set up, we used PHLAC.All and PHLAC$_{Bayes}$, and experimentally set the spatial interval to $\Delta r = 8$ for all image sizes. The concatenated feature of the features calculated by each size was used for both bag-of-features and PHLAC. As local features, we used SIFT-like feature and following OpponentSIFT feature [31].

OpponentSIFT: The rgb color space is converted to the opponent color space. Then calculate SIFT-like feature over the all opponent color spaces, independently. This gives 3×128 dimensional feature. We applied L2-Hys normalization to each color space.

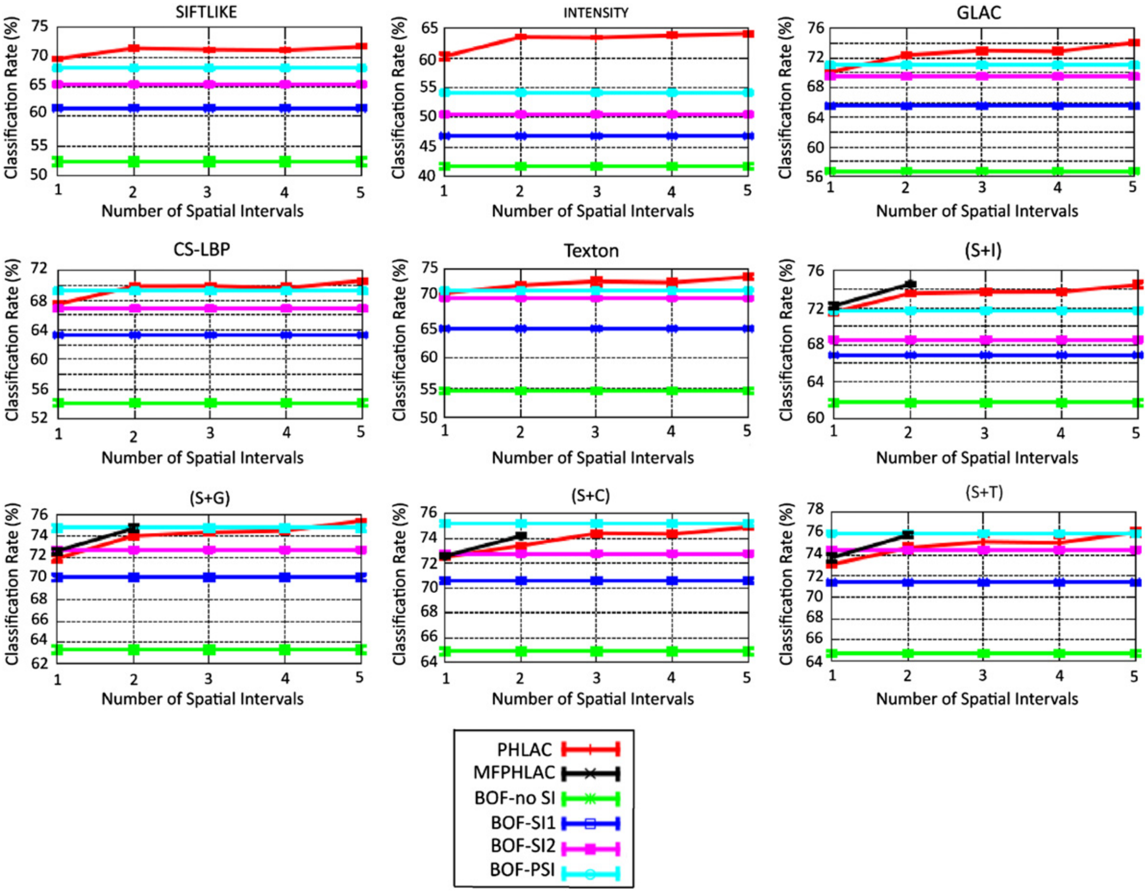

Fig. 13. Recognition rates of MFPHLAC and comparison with those of bag-of-features method with spatial information (Scene-15). SI1 (Spatial Information 2 × 2), SI2 (Spatial Information 4 × 4), PSI (Pyramid Spatial Information (1 × 1, 2 × 2, 4 × 4)).

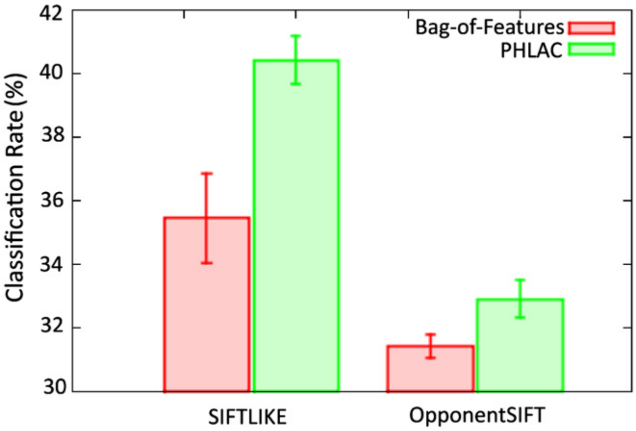

Fig. 14. Recognition rates of Caltech101 dataset.

We used 400 codebooks created by k-means clustering. The results are shown in Fig. 14. In this dataset, the PHLAC achieved also better performances when both SIFT-like and OpponentSIFT features were used for local features. SIFT-like feature exhibited better performance than OpponentSIFT. When SIFT-like feature was used for local feature, PHLAC achieved 40.42(± 0.76)% average recognition rate while that of the standard bag-of-features was 35.46 (± 1.41)%.

The comparison to the other recent proposed methods in the same setting is shown in Table 1. Our recognition rate is lower than that of the other methods because the classification rule is so simple. Despite the linear classification, the method achieved comparable results to that of Grauman et al. [33].

Table 1
Comparison to the other methods on the Caltech101 dataset.

Method	Ours	Lazebnik et al. [3]	Zhang et al. [2]	Grauman et al. [33]
Classification	Linear SVM	Kernel SVM	Kernel SVM	Kernel SVM
Avg.	40.4%	64.6%	53.9%	43%

5. Discussion on feature dimension

One of the advantages of PHLAC is its feature dimension. The comparison of the dimension of different feature representation is listed in Table 2. The dimension of the bag-of-features method depends on the codebook size M. Thus, to achieve high accuracy, the training time of a classifier should be increased and a large memory size is required. Furthermore, it is necessary for larger dimensions to utilize spatial grid information. On the other hand, the dimension of PHLAC depends on the number of categories C, and it is independent of the codebook size M. At least, the 0th order of PHLAC can reflect the reliable estimation of large codebook size; thus, the accuracy of PHLAC can be increased by not increasing the feature dimension. PHLAC$_{SVM}$ must train SVM using bag-of-features for estimations posterior probability of codebook;

Table 2
Dimensions of feature representations.

Feature	General	IG02 ($M=400$, $C=3$)	Scene-15 ($M=200$, $C=15$)	Caltech-101 ($M=400$, $C=101$)
PHLAC	35C	105	525	3535
MFPHLAC	233C	–	3495	–
BOF	M	400	200	400
BOF (with SI1)	4M	–	800	–
BOF (with SI2)	16M	–	3200	–
BOF (with PSI)	21M	–	4200	–

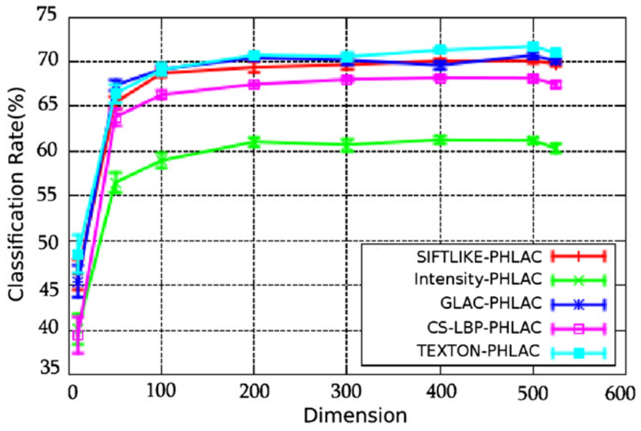

Fig. 15. Recognition rates of compressed PHLAC by PCA (Scene-15 dataset): the points of the extreme right indicate original PHLAC without PCA.

However, PHLAC$_{SVM}$ is not effective to Scene-15 dataset which contains large number of category compared to PHLAC$_{Bayes}$. Thus, we highly recommend the use of PHLAC using Bayes' theorem when the codebook size and number of categories are large. Although it is obvious that the dimension of PHLAC for all categories becomes large for a problem which involving a very large number of categories, the number of the category that is classified once undergoes reduction by hierarchal category recognition.

Furthermore, the PHLAC feature can be compressed effectively by principle component analysis (PCA). The recognition rates per compressed dimension by PCA are shown in Fig. 15. In this experiment, PHLAC$_{Bayes}$ and PHLAC.All were used. Because redundancy exists owing to similar properties of mask patterns and similar posterior probability images of different categories, the performances do not decrease even when the dimension is less than 40% of the original PHLAC dimension. Thus, the feature dimension of PHLAC can be further reduced from linear size of the categories with maintaining the classification accuracy.

6. Conclusion

In this paper, we proposed an image description method using higher-order local autocorrelations on posterior probability images called "probability higher-order local autocorrelations (PHLAC)." This method is regarded as an extension of the standard bag-of-features method. Our method overcomes the limitation of spatial information by utilizing the co-occurrence of local spatial patterns in posterior probabilities. This method possesses the properties of shift invariance and additivity as does HLAC [12]. Experimental results revealed that the proposed method achieved a higher classification performance than the standard bag-of-features method by an average of 2% and 15% in the case of the IG02 and Scene-15 datasets, respectively, using 200 codebooks. In Caltech-101, the proposed method improved 5% using 400 codebooks. We also extended PHLAC to autocorrelations of posterior probability calculated from multiple image features, which is called "multiple features probability higher-order local autocorrelations (MFPHLAC)." MFPHLAC was able to achieve a slightly better performance than PHLAC.

We also compared the proposed methods with the bag-of-features method using spatial information. PHLAC was able to achieve a competitive result compared to the bag-of-features method using spatial information.

References

[1] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bag of keypoints, in: European Conference on Computer Vision, Workshop on Statistical Learning in Computer Vision, 2004, pp. 59–74.
[2] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, International Journal of Computer Vision 73 (2) (2007) 213–238.
[3] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.
[4] A. Agarwal, B. Triggs, Multilevel Image coding with hyperfeatures, International Journal of Computer Vision 78 (2008) 15–27.
[5] S. Savarse, J. Winn, A. Criminisi, Discriminative object class models of appearance and shape by correlatons, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2033–2040.
[6] J. Yuan, Y. Wu, M. Yang, Discovery of collocation patterns: from visual words to visual phrases, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[7] F. Jurie, B. Triggs, Creating efficient codebooks for visual recognition, IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 604–610.
[8] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: European Conference on Computer Vision, 2006, pp. 490–503.
[9] J.C.V. Gemert, J.-M. Geusebroek, C.J. Veenman, A.W.M. Smeulders, Kernel codebooks for scene classification, in: European Conference on Computer Vision, Part III, Lecture Notes on Computer Science, vol. 5304, 2008, pp. 696–709.
[10] F. Perronnin, Universal and adapted vocabularies for generic visual categorization, IEEE Transaction on Pattern Analysis and Machine Intelligence 30 (7) (2008) 1243–1256.
[11] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[12] N. Otsu, T. Kurita, A new scheme for practical flexible and intelligent vision systems, in: IAPR Workshop on Computer Vision, 1988, pp. 431–435.
[13] Y.-T Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, Q. Tian, Visual synset: towards a higher-level visual representation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[14] T. Matsukawa, T. Kurita, Image classification using probability higher-order local auto-correlations, in: Asian Conference on Computer Vision, Part III, Lecture Notes on Computer Science, vol. 5996, 2009, pp.384–394.
[15] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 511–518.
[16] T. Kobayashi, N. Otsu, Image feature extraction using gradient local auto-correlations, in: European Conference on Computer Vision, Part I, Lecture Notes on Computer Science, vol. 5302, 2008, pp. 346–358.
[17] T. Kobayashi, N. Otsu, Color Image feature extraction using color index local auto-correlations, in: International Conference on Acoustics, Speech, and Signal Processing, 2009, pp. 1057–1060.
[18] T. Quack, V. Ferrari, B. Leibe, L. Van-Gool, Efficient mining of frequent and distinctive feature configurations, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[19] X. Wang, E. Grimson, Spatial latent dirichlet allocation, Advances in Neural Information Processing Systems, vol. 20, 2007.

[20] N. Rasiwasia, N. Vasconcelos, Scene classification with low-dimensional semantic spaces and weak supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[21] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[22] V. Vapnik, Statistical Learning Theory, John Wiley & Sones, New York, USA, 1998.

[23] T. Toyoda, O. Hasegawa, Extension of higher order local autocorrelation features, Pattern Recognition 40 (2007) 1466–1473.

[24] M. Varma, D. Ray, Learning the discriminative power-invariance trade-off, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[25] A. Bosch, A. Zisserman, X. Munouz, Representing shape with a spatial pyramid kernel, in: International Conference on Image and Video Retrieval, 2007, pp. 401–408.

[26] M. Marszalek, C. Schmid, Spatial weighting for bag-of-features, IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2118–2125.

[27] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[28] M. Heikkila, M. Pietikainen, C. Schmid, Description of interest regions with local binary patterns, Pattern Recognition 42 (2009) 425–436.

[29] C. Schmid, Constructing models for content-based image retrieval, IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2001, pp. 39–45.

[30] K. Hotta, Scene classification based on multi-resolution orientation histogram of gabor features, in: International Conference on Computer Vision Systems, Lecture Notes on Computer Science, vol. 5008, 2008, pp. 291–301.

[31] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluation of color descriptors for object and scene recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[32] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, in: IEEE CVPR Workshop on Generative-Model Based Vision, 2004.

[33] K. Grauman, T. Darrell, The Pyramid match kernels: discriminative classification with sets of image features, in: International Conference on Computer Vision, vol. 2, 2005, pp. 1458–1465.

Tetsu Matsukawa received the B.Eng., the M.Eng. and Ph.D. degrees from University of Tsukuba, Japan, in 2006, 2008, and 2011, respectively. He joined the Institute of Industrial Science in The University of Tokyo from 2011 as a project research associate. His research interest includes pattern recognition, machine learning, and generic object recognition.

Takio Kurita received the B.Eng. degree from Nagoya Institute of Technology and the Dr. Eng. degree from the University of Tsukuba, in 1981 and in 1993, respectively. He joined the Electrotechnical Laboratory, AIST, MITI in 1981. From 1990 to 1991 he was a visiting research scientist at Institute for Information Technology, National Research Council Canada. From 2001 to 2009, he was a deputy director of Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology (AIST). Also he was a Professor at Graduate School of Systems and Information Engineering, University of Tsukuba from 2002 to 2009. He is currently a Professor at Hiroshima University. His current research interests include statistical pattern recognition and its applications to image recognition. He is a member of the IEEE, the IPSJ, the IEICE of Japan, Japanese Neural Network Society, The Japanese Society of Artificial Intelligence.