Special issue: 3D image and video technology

Research Paper

Recognizing multiple objects based on co-occurrence of categories

Takahiro OKABE[1], Yuhi KONDO[2], Kris M. KITANI[3], and Yoichi SATO[4]
[1,2,4]*Institute of Industrial Science, The University of Tokyo*
[3]*Graduate School of Information Systems, The University of Electro-Communications*

ABSTRACT

Most previous methods for generic object recognition explicitly or implicitly assume that an image contains objects from a single category, although objects from multiple categories often appear together in an image. In this paper, we present a novel method for object recognition that explicitly deals with objects of multiple categories coexisting in an image. Furthermore, our proposed method aims to recognize objects by taking advantage of a scene's context represented by the co-occurrence relationship between object categories. Specifically, our method estimates the mixture ratios of multiple categories in an image via MAP regression, where the likelihood is computed based on the linear combination model of frequency distributions of local features, and the prior probability is computed from the co-occurrence relation. We conducted a number of experiments using the PASCAL dataset, and obtained the results that lend support to the effectiveness of the proposed method.

KEYWORDS

Generic object recognition, context, co-occurrence, bag of features, regression, MAP estimation

1 Introduction

With the proliferation of digital cameras, enormous numbers of digital images have been accumulated on the Internet. Since manually processing such a huge amount of data is almost impossible, automatic image classification and retrieval are research areas of increasing importance. Thus, a research topic called *generic object recognition* has recently been brought back into the spotlight in the computer vision community. In this study, we focus on the problem of object categorization among various tasks of generic object recognition.

It is generally recognized that object categorization is a very difficult task due to the following two reasons. First, objects of the same category differ in both color and shape, that is, intra-category variation. Second, the appearance of an object varies drastically depending on imaging conditions such as camera viewpoints, the object's pose, and illumination. To cope with these difficulties, previous work mainly studies feature detection [1], [2], object and category representation [3]–[5], or classifiers [2], [6] robust against changes in object appearance due to intra-category variation and variable imaging conditions.

The previous studies however share a common limitation. That is, most previous methods explicitly or implicitly assume that an image contains objects from a single category, and evaluate whether objects of each category are present or not, independent of the presence or absence of objects of the other categories. Therefore, they are not well suited for recognizing objects of various categories coexisting in an image and do not consider the fact that certain combinations of categories are more likely to appear together than others. For example, given an image of a street, it is highly probable that a "car" will coexist with a "motorbike", while it is very unlikely that a "car" and a "cow" will appear together.

Accordingly, we present a novel method for object recognition that explicitly deals with objects of multiple categories coexisting in an image. Furthermore, our proposed method aims to recognize objects by taking

Received August 28, 2009; Revised December 10, 2009; Accepted January 4, 2010.
[1)] *takahiro@iis.u-tokyo.ac.jp*, [2)]*ykondo@iis.u-tokyo.ac.jp*,
[3)]*kitani@is.uec.ac.jp*, [4)]*ysato@iis.u-tokyo.ac.jp*
DOI: 10.2201/NiiPi.2010.7.6

advantage of a *scene's context* represented by the co-occurrence relationship between object categories. The use of such contextual cues makes it possible to classify objects of different categories but with similar appearance.

In order to achieve our objective, we chose to use the bag-of-features (BoF) paradigm [3], which is now known as one of the most promising paradigms for generic object recognition. In particular, our proposed method estimates the mixture ratios of multiple categories in an image via maximum *a posteriori* (MAP) regression, where the likelihood is computed based on the linear combination model of frequency distributions (*i.e.* histograms) of local features, and the prior probability is computed from the co-occurrence relation. We conducted a number of experiments using the PASCAL dataset, and obtained the results that give support to the effectiveness of the proposed method.

The rest of this paper is organized as follows. We briefly summarize related work in Section 2. We describe our proposed method in Section 3, and report the experimental results in Section 4. Finally, in Section 5, we present concluding remarks.

2 Related work

We briefly summarize previous studies relating to the basic idea of our proposed method from two distinct points of view; *multiple categories* and *context*.

2.1 Multiple categories

In order to recognize objects of various categories coexisting in an image, a segmentation-based approach and a regression-based approach have been developed. The former approach segments an image into regions so that each segmented region contains objects of a single category, and then conducts object categorization for each region [7]. However, segmenting images of complex scenes is not necessarily an easy task, and the accuracy of classification depends on that of image segmentation.

The latter approach estimates the mixture ratios of multiple categories in an image via regression, where the mixture ratio is defined based on the number of feature points arising from each category in the BoF paradigm (see Section 3.1). For example, Sivic et al. [8] estimate the mixture ratios of various categories in an individual image by applying probabilistic Latent Semantic Analysis (pLSA) to a set of unlabeled images. Their regression-based method is similar to ours in the sense that the frequency distribution of feature points in an image is modeled by the linear combination of frequency distributions of feature points arising from various categories. However, their method finds the mixture ratios based on the framework of maximum likelihood (ML) estimation, and the prior information other than images that can be inferred from scene's context is not taken into account. Consequently, it is difficult to classify objects of different categories but with similar appearance.

2.2 Context

Obviously the context of the scene is one of the most important clues for understanding images and has in fact been utilized in the field of generic object recognition [9], [10]. However, the co-occurrence relation of object categories has received little attention compared with other contextual information such as size and position [7].

Recently, Rabinovich et al. [7] proposed a method for object categorization based on the co-occurrence relation of object categories, and Galleguillos et al. [11] extended their method by incorporating the spatial context with respect to the relative location of objects. First, they segment an image into regions, and then tentatively estimate a category label and its confidence for each segmented region based on the BoF paradigm. Finally, they revise the label based on the confidence of the tentative label and the co-occurrence relation. As we described before, however, image segmentation itself is a potential limitation for images with complex scenes. In addition, our method differs from their segmentation-based method with respect to the manner in which we describe the co-occurrence relationship between object categories. They model the co-occurrence relation based on the presence of objects in terms of *frequencies*, that is, the number of times that certain combinations of categories appear together. In contrast, we model the co-occurrence relation in terms of *mixture ratios* based on the number of feature points arising from each category (see Sections 3.3 and 4.1 for details). The mixture ratios can capture contextual information beyond the presence or absence of categories. The ratios can be an indicator of the number of objects when using a sparse keypoint detector (*e.g.* DoG) or can represent the size of an object when using a dense set of keypoints.

From the viewpoint of co-occurrence, the method for image categorization proposed by Qi et al. [12] is related to our study. They also segment an image into regions, and represent each region by a set of low-level features such as color and size, and then classify the image based on the co-occurrence of the low-level features. Their co-occurrence describes the relationship among features arising from *a single* category, and is effective for classifying an image into *one* of given categories. On the other hand, our co-occurrence that describes the relationship between *multiple* categories is essential for estimating mixture ratios of *multiple* cate-

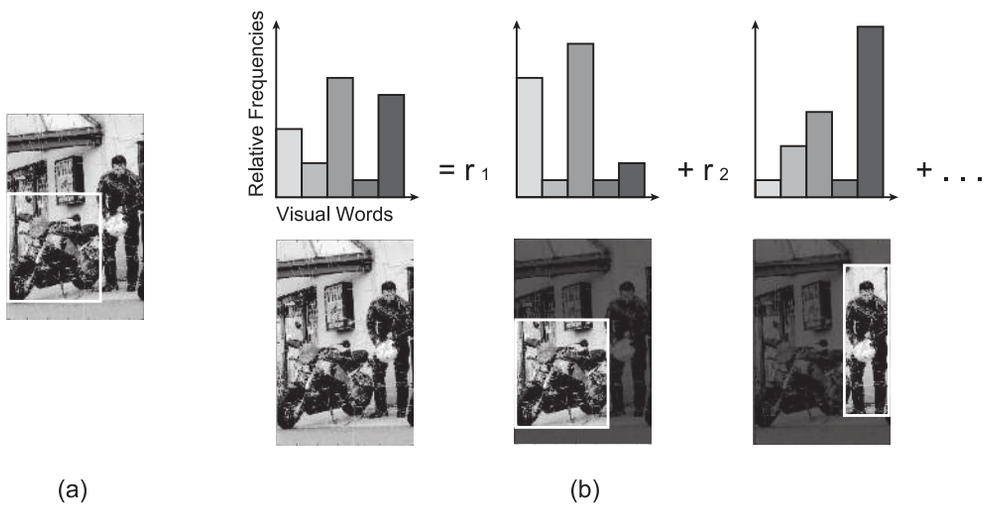

Fig. 1 (a) The mixture ratio of a "motorbike" is defined by the ratio between the number of feature points detected within the bounding box and the total number of feature points. (b) The histogram of an entire image is described by the linear combination of a motorbike's histogram, a person's histogram, *etc.*

gories in an image.

As described above, our proposed method is differentiated from related work by the following: (i) our method is a regression-based approach and avoids troublesome segmentation for images with complex scenes, and (ii) our method takes account of the co-occurrence relation of object categories in terms of mixture ratios, which captures more contextual information than that in terms of frequencies.

3 Proposed method

3.1 Overview

We represent an image as a set of local features such as SIFT [1] based on the BoF paradigm. Let us denote the label of a category by c ($c = 1, 2, 3, ..., C$), and define the mixture ratio r_c of the category in an image as the ratio between the number of feature points arising from the category c and the total number of feature points as shown in Fig. 1 (a). Here, C is the total number of categories and $\sum_{c=1}^{C} r_c = 1$ by definition. We concatenate r_c into a vector and denote the mixture ratios of all categories in the image by $\boldsymbol{r} = (r_1, r_2, r_3, \cdots, r_C)^T$.

We compress the local features via vector quantization (see Section 4.1), and call the quantized features visual words. Let us denote the label of a visual word by w ($w = 1, 2, 3, ..., W$), and the relative frequency of the visual word w arising from an image by h_w. Here, W is the total number of visual words and $\sum_{w=1}^{W} h_w = 1$ by definition. We concatenate h_w into a vector and denote the relative frequency distribution of the visual words arising from the image by $\boldsymbol{h} = (h_1, h_2, h_3, ..., h_W)^T$. Hereafter, we often call the relative frequency distribution of visual words the histogram in short.

Our proposed method finds the mixture ratios $\boldsymbol{r}$ from the histogram $\boldsymbol{h}$ of a given image based on the framework of MAP estimation. The posterior probability $p(\boldsymbol{r}|\boldsymbol{h})$ is given by the Bayes' rule as

$$p(\boldsymbol{r}|\boldsymbol{h}) \propto p(\boldsymbol{h}|\boldsymbol{r})p(\boldsymbol{r}). \tag{1}$$

Here, as described in Sections 3.2 and 3.3, the likelihood $p(\boldsymbol{h}|\boldsymbol{r})$ is derived from the relative frequency distribution of visual words, and the prior probability $p(\boldsymbol{r})$ is derived from the co-occurrence relation of object categories.

3.2 Likelihood

As shown in Fig. 1 (b), the histogram of an image which includes a motorbike and a person is represented by the linear combination of a motorbike's histogram, a person's histogram, *etc.* Therefore, it is clear that the relative frequency distribution $\boldsymbol{h}$ arising from the entire image is described by the linear combination of relative frequency distributions $\boldsymbol{h}_c$ arising from various categories in the image:

$$\boldsymbol{h} = \sum_{c=1}^{C} r_c \boldsymbol{h}_c, \tag{2}$$

where the mixture ratios are the coefficients of the linear combination.

Assuming that the relative frequency of each visual word is independent of those of the other visual words,

the likelihood $p(\boldsymbol{h}|\boldsymbol{r})$ is represented by the product of individual likelihoods $p(h_w|\boldsymbol{r})$ as

$$p(\boldsymbol{h}|\boldsymbol{r}) = \prod_{w=1}^{W} p(h_w|\boldsymbol{r}). \tag{3}$$

In addition, let us assume that each component h_{cw} of $\boldsymbol{h}_c$ obeys a normal distribution $\mathcal{N}(\mu_{cw}, \sigma_{cw}^2)$ with the mean μ_{cw} and the variance σ_{cw}^2. Then, the linear combination of relative frequency, that is, $h_w = \sum_{c=1}^{C} r_c h_{cw}$ also obeys the normal distribution $\mathcal{N}(\sum_{c=1}^{C} r_c \mu_{cw}, \sum_{c=1}^{C} r_c^2 \sigma_{cw}^2)$ due to the reproductive property of the normal distribution. Hence, the likelihood is given by

$$\begin{aligned} p(\boldsymbol{h}|\boldsymbol{r}) &= \prod_{w=1}^{W} \frac{1}{\sqrt{2\pi \sum_{c=1}^{C} r_c^2 \sigma_{cw}^2}} \\ &\times \exp\left[-\frac{(h_w - \sum_{c=1}^{C} r_c \mu_{cw})^2}{2 \sum_{c=1}^{C} r_c^2 \sigma_{cw}^2} \right]. \end{aligned} \tag{4}$$

For the sake of simplicity in the following discussion, we define $\mathcal{E}_{\text{like}}$ as

$$\begin{aligned} \mathcal{E}_{\text{like}} &= -\ln p(\boldsymbol{h}|\boldsymbol{r}) \\ &\simeq \sum_{w=1}^{W} \left[\frac{(h_w - \sum_{c=1}^{C} r_c \mu_{cw})^2}{\sum_{c=1}^{C} r_c^2 \sigma_{cw}^2} \right. \\ &+ \left. \ln\left(\sum_{c=1}^{C} r_c^2 \sigma_{cw}^2 \right) \right]. \end{aligned} \tag{5}$$

Here, we omit constants for estimation.

3.3 Prior probability

We address the co-occurrence relationship between two object categories. Specifically, we assume that the mixture ratios obey a C-dimensional normal distribution $\mathcal{N}_C(\boldsymbol{\nu}, \Sigma)$ with the mean vector $\boldsymbol{\nu}$ and the covariance matrix Σ. In the similar way to the above, we define $\mathcal{E}_{\text{pri}}$ as

$$\mathcal{E}_{\text{pri}} = -\ln p(\boldsymbol{r}) \simeq (\boldsymbol{r} - \boldsymbol{\nu})^T \Sigma^{-1} (\boldsymbol{r} - \boldsymbol{\nu}). \tag{6}$$

3.4 Cost function

Substituting (5) and (6) into the negative logarithm of (1) and introducing a parameter λ, we define the empirical cost function $\mathcal{E}_{\text{pos}}$ as

$$\mathcal{E}_{\text{pos}} = \mathcal{E}_{\text{like}} + \lambda \mathcal{E}_{\text{pri}}. \tag{7}$$

Our proposed method estimates the mixture ratios of multiple categories in an image by minimizing this empirical cost function. Because the mixture ratios are non-negative and their summation is equal to 1, our method results in a nonlinear minimization problem with the following constraints:

$$\begin{aligned} &\text{minimize} && \mathcal{E}_{\text{pos}} \\ &\text{subject to} && r_c \geq 0 \ (c = 1, 2, 3, ..., C) \\ & && \sum_{c=1}^{C} r_c = 1. \end{aligned} \tag{8}$$

The parameter λ is a relative weight between $\mathcal{E}_{\text{like}}$, which represents the degree by which the linear combination of histograms fits the data, and $\mathcal{E}_{\text{pri}}$, which represents the statistical constraints enforced by the co-occurrence relationship between object categories. The ML estimation (*i.e.* without the prior probability) corresponds to the case when $\lambda = 0$.

We note here that the solution of the optimization problem is influenced by the initializing value. Our current implementation finds the initial value by minimizing $\sum_{w=1}^{W} (h_w - \sum_{c=1}^{C} r_c \mu_{cw})^2$ under the constraints $r_c \geq 0$ $(c = 1, 2, 3, ..., C)$ and $\sum_{c=1}^{C} r_c = 1$. Then, we optimize the exact cost function by using *fmincon* in the MATLAB toolbox. The initial value is the solution of (2) when the histogram of the c-th category $\boldsymbol{h}_c$ is replaced by its mean $\boldsymbol{\mu}_c$.

4 Experiments

4.1 Procedures

4.1.1 Dataset

We used the PASCAL2006 dataset [13] for evaluating the performance of our proposed method. This dataset contains objects of ten categories; "bicycle", "bus", "car", "cat", "cow", "dog", "horse", "motorbike", "person", and "sheep". The dataset consists of a set of data for training and another set for test. In addition, the annotations describing the labels and bounding boxes of those objects are given for all images.

4.1.2 Bag of features

We used DoG[1)] and SIFT [1] for detecting and describing local features in images, and k-means clustering algorithm for vector quantization. Although other feature detectors, descriptors [2], and quantization algorithms [14] could be used as well, we implemented the above standard BoF since the main purpose of our experiments is to confirm the advantage of incorporating the co-occurrence relation into generic object recognition.

First, we prepared 50 images for each category from the training data by cropping regions inside the bounding boxes. Then, local features were detected and vector-quantized via k-means algorithm. The number

[1)] A dense set of keypoints would be better suited for recognizing uniform areas such as sky.

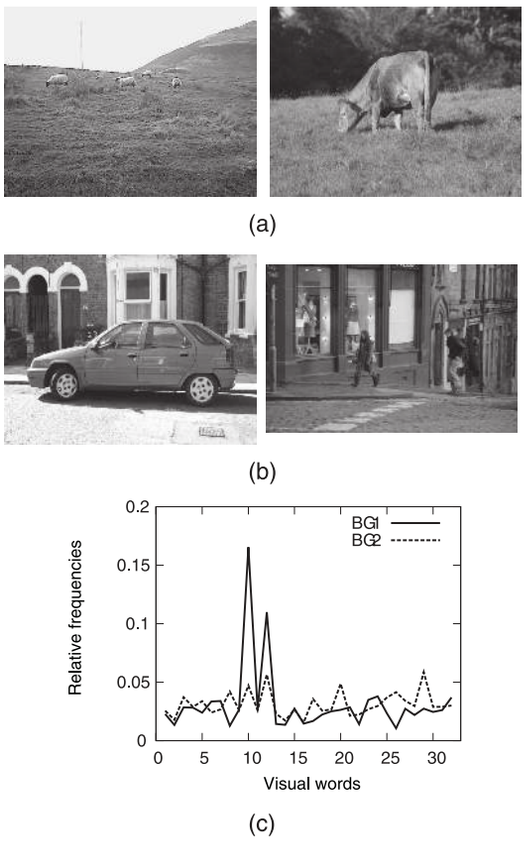

Fig. 2 Example images from (a) BG1, (b) BG2, and (c) the histograms of these categories when $W = 32$.

of visual words W are 32, 64, 128, 256, 512, and 1024. We computed the histograms of those 500 images and finally obtained the means μ_{cw} and variances σ^2_{cw} of relative frequencies for describing the likelihood in (5).

So far, we implicitly assume that images contain objects of only given categories. However, objects of other categories generally appear in images. Accordingly, we consider those objects as backgrounds, and investigate the effects of adding background categories to the ten object categories. We manually classified backgrounds into two categories: one contains artificial materials such as buildings (BG1) and the other contains natural objects such as grass (BG2). Then, we selected 50 images containing each background category and detected local features from the outside of the bounding boxes. After that, we calculated the statistics of the background histograms. In Fig.2, we show example images from (a) BG1, (b) BG2, and (c) the histograms of these categories when $W = 32$. We can find that the background histograms significantly differ from each other.

4.1.3 Co-occurrence of categories

We acquired the following two co-occurrence relations of object categories from 2618 images in the training data. The first type of co-occurrence relation is described in Section 3.3. Because the labels and bounding boxes are given, calculating the mixture ratio of each category is straightforward. We denote the mean vector and the covariance matrix of the mixture ratios by $\boldsymbol{\nu}_r$ and Σ_r.

The second type of co-occurrence relation is used for (partially) comparing our proposed method with the method proposed by Rabinovich [7]. Specifically, we confirm the advantage of the co-occurrence relation in terms of mixture ratios over that in terms of frequencies. We calculate the mean vector $\boldsymbol{\nu}_f$ and covariance matrix Σ_f based on the presence of objects: $r_c = 1$ if objects of the category c are present and $r_c = 0$ otherwise.

Fig. 3 shows the two covariance matrices Σ_r and Σ_f (we show only the lower left values due to symmetry). The combinations of categories with positive covariance tend to appear together, whereas those with negative covariance have a tendency not to appear at the same time. For example, a "person" often appears with a "motorbike" and a "horse", but a "cat" rarely appears with a "dog". Interestingly, we observe that the sign of covariance differ between Σ_r and Σ_f for a few combinations of categories.

4.1.4 Measure for quantitative evaluation

We used all of 2686 images from the test data. For quantitative evaluation, we use a measure known as the Area Under Curve (AUC), *i.e.* the area under the Receiver Operating Characteristic (ROC) curve, which is commonly used in the field of generic object recognition. Specifically, we consider the estimated *mixture ratio* of a given category as the *probability* that objects of that category are present in an image. Namely, we consider objects of the category c to be present if r_c is greater than a certain threshold, and draw the ROC curve by varying the threshold.

In general, performance is considered to be better as the AUC grows closer to one. However, the way of evaluation that regards the ratio as the probability has some limitations. For example, an object with a small mixture ratio will be considered to be a false negative even though its mixture ratio is accurately estimated by our method, and as a result would degrade the AUC. We note that because our method characterizes the mixture ratios of multiple categories (*i.e.* not the presence and absence of objects), the AUC may not provide a holistic measure.

bicycle	+3.4									
bus	-0.2	+1.9								
car	-0.3	-0.2	+2.8							
cat	-0.5	-0.3	-0.6	+5.9						
cow	-0.2	-0.1	-0.2	-0.3	+2.0					
dog	-0.4	-0.2	-0.5	-0.6	-0.3	+4.5				
horse	-0.2	-0.1	-0.2	-0.3	-0.1	-0.3	+1.9			
motorbike	-0.2	-0.1	-0.2	-0.4	-0.2	-0.3	-0.2	+2.7		
person	-0.2	-0.1	-0.4	-0.5	-0.2	-0.3	**+0.0**	**+0.1**	+2.3	
sheep	-0.2	-0.1	-0.3	-0.3	-0.1	-0.3	-0.1	-0.2	-0.2	+2.0

bicycle	+9.2									
bus	-0.6	+6.2								
car	-1.0	**+1.4**	+16.7							
cat	-1.5	-1.0	-3.1	+12.6						
cow	-0.8	-0.5	-1.6	-1.1	+7.2					
dog	-1.4	-0.9	-2.7	-1.8	-1.1	+12.0				
horse	-0.9	-0.6	-1.7	-1.4	-0.7	-1.2	+8.5			
motorbike	-0.7	-0.4	-0.3	-1.3	-0.7	-1.2	-0.8	+8.2		
person	**+0.5**	**+1.1**	-1.2	-3.3	-1.4	-1.4	**+2.6**	**+3.0**	+19.0	
sheep	-1.0	-0.6	-1.9	-1.4	-0.7	-1.2	-0.9	-0.9	-1.8	+8.7
	bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep

Fig. 3 The covariance matrices in terms of mixture ratios (top) and frequencies (bottom). The numerical values are multiplied by 100 for display purpose.

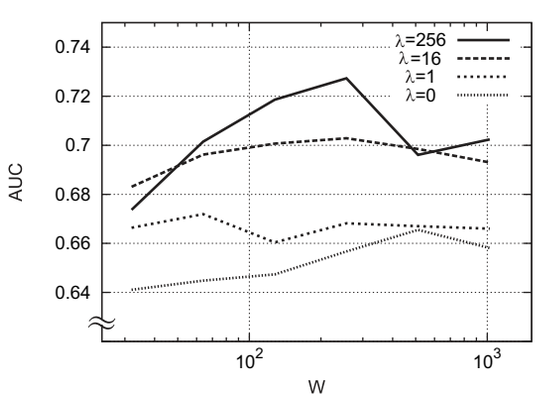

Fig. 4 AUC: incorporating the co-occurrence relation in terms of mixture ratios.

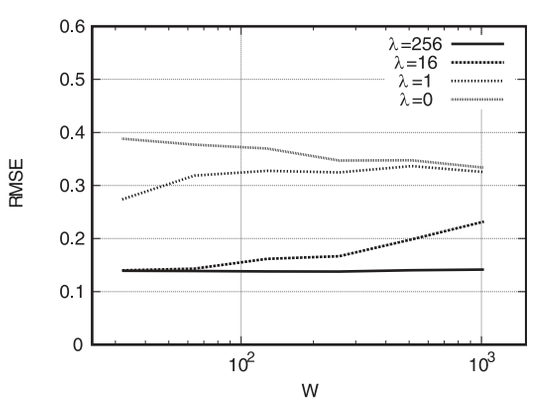

Fig. 5 RMSE: incorporating the co-occurrence relation in terms of mixture ratios.

4.2 Results

4.2.1 Effects of the co-occurrence relation in terms of mixture ratios

First, we examined the effects of incorporating the co-occurrence relationship in terms of mixture ratios (ν_r, Σ_r) into generic object recognition. Fig. 4 shows the average of AUCs with respect to the ten object categories for various combinations of the weight λ and the number of visual words W. We can find that the results using the prior probability are better than those of

bicycle	0	0	0	0	0	0	0	0	0	0	0	0
bus	0	0	0	0	0.09	0.03	0.01	0	0	0	0.01	0
car	0	0.14	0	0	0	0	0	0	0	0	0	0
cat	0	0	0	0	0	0	0	0	**0.51**	**0.50**	**0.52**	**0.82**
cow	0	0	0	0	0	0	0	0	0	0	0.03	0
dog	0	0	0	0	0	0	0	0	0.40	0.36	0.28	0
horse	0	0	0	0	**0.70**	**0.60**	**0.41**	**0.38**	0	0.02	0.05	0
motorbike	**1.00**	**0.86**	**0.63**	**0.61**	0	0	0.19	0	0	0	0.02	0
person	**0**	**0**	**0.37**	**0.28**	**0**	**0.11**	**0.35**	**0.31**	0	0	0.04	0
sheep	0	0	0	0	0.20	0.27	0.05	0	0.09	0.12	0.05	0
	ML	MAP: freq	MAP: ratio	ground truth	ML	MAP: freq	MAP: ratio	ground truth	ML	MAP: freq	MAP: ratio	ground truth

Fig. 6 Mixture ratios found via ML estimation (ML), via MAP estimation using the co-occurrence relation in terms of frequencies (MAP: freq)/mixture ratios (MAP: ratio), and ground truth.

ML estimation ($\lambda = 0$). Our proposed method and ML estimation achieve maximum AUCs of 0.73 and 0.66 respectively. Thus, we can say that the co-occurrence relation in terms of mixture ratios works well for recognizing multiple objects.

In Fig. 6, we show the estimated mixture ratios and the ground truth for some images. For example, the ML estimation (ML) yields the result of "motorbike" for the left image. On the other hand, our method based on MAP estimation (MAP: ratio) yields the result of "motorbike" and "person", which is consistent with the ground truth. Fig. 5 shows the root-mean-square errors (RMSEs) of the estimated mixture ratios. One can see that the use of the co-occurrence relation decreases RMSEs. These results also support the effectiveness of the proposed method. The nonlinear optimization takes about 0.3 seconds per image on a typical Core2 PC when $W = 256$ and $\lambda = 256$.

4.2.2 Effects of the co-occurrence relation in terms of frequencies

Second, we examined the effects of incorporating the co-occurrence relationship in terms of frequencies $(\boldsymbol{\nu}_f, \Sigma_f)$. In the similar manner to the above, we show the average of AUCs in Fig. 7. Also in this case, the

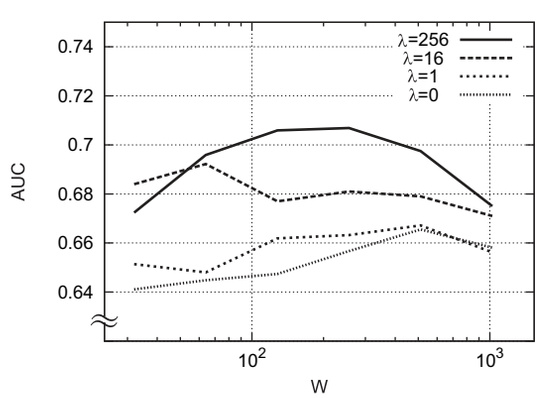

Fig. 7 AUC: incorporating the co-occurrence relation in terms of frequencies.

results that make use of the prior probability are better than those of ML estimation in most combinations. However, the performance of the method using the co-occurrence relation in terms of frequencies is worse than that using the relation in terms of ratios. Therefore, one can conclude that the co-occurrence relation in terms of frequencies (*i.e.* based only on the presence of categories) is also effective for object recognition,

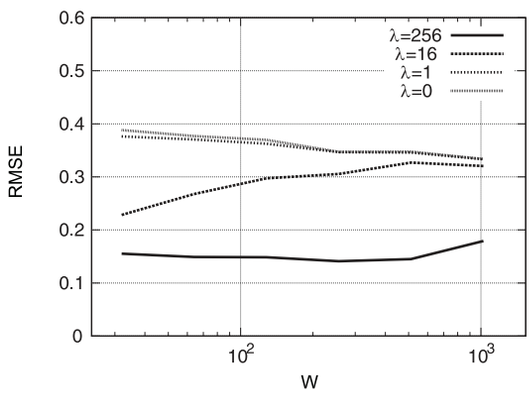

Fig. 8 RMSE: incorporating the co-occurrence relation in terms of frequencies.

Table 1 AUC for each category: ML estimation (ML), MAP estimation using the co-occurrence relation in terms of frequencies (MAP: freq)/mixture ratios (MAP: ratio), and the winner of PASCAL2006.

Category	ML	MAP: freq	MAP: ratio	PASCAL 2006
bicycle	0.778	0.823	0.784	0.948
bus	0.834	0.859	0.851	0.981
car	0.648	0.704	0.828	0.975
cat	0.577	0.670	0.679	0.937
cow	0.551	0.787	0.778	0.938
dog	0.584	0.630	0.607	0.876
horse	0.624	0.589	0.604	0.926
motorbike	0.715	0.724	0.765	0.969
person	0.542	0.476	0.619	0.855
sheep	0.804	0.808	0.757	0.956

but the relation considering mixture ratios works better. We show the estimated mixture ratios (MAP: freq) in Fig. 6 and their RMSEs in Fig. 8.

In Table 1, we show the AUC for each category: ML estimation (ML) with $W = 512$, MAP estimation using the co-occurrence in terms of frequencies (MAP: freq) with $W = 256$ and $\lambda = 256$, and that in terms of ratios (MAP: ratio) with $W = 256$ and $\lambda = 256$. Similar to the average AUCs show in Fig. 4 and Fig. 7, one can find the effectiveness of incorporating the co-occurrence relation into object recognition. Especially, by using the co-occurrence in terms of ratios, the AUCs for "car" and "person" are significantly improved. We also show the AUCs of the PASCAL2006 winner in Table 1. As described in Section 4.1, it is not appropriate to directly compare these AUCs with our AUCs computed by interpreting mixture ratios as probabilities. Nevertheless,

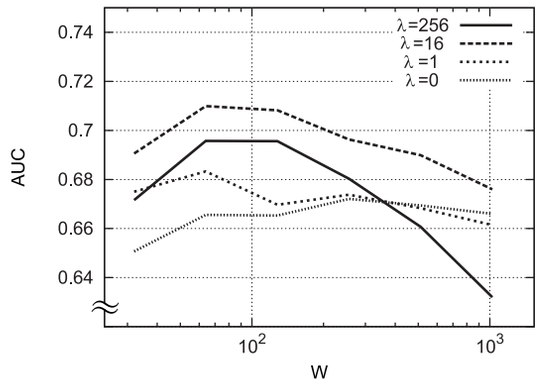

Fig. 9 AUC: adding background categories.

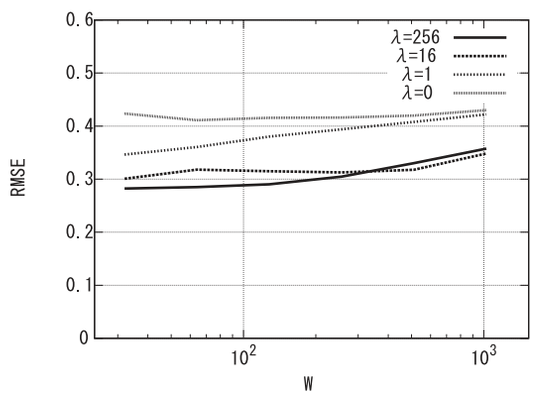

Fig. 10 RMSE: adding background categories.

some common tendencies are observed: the AUCs for "dog", "horse", and "person" are relatively smaller than others.

4.2.3 Effects of background categories

Finally, we examined the effects of adding background categories to the ten object categories. Fig. 9 shows the results obtained by using the co-occurrence relation (ν_r, Σ_r). Although the results are similar to the previous experiments in the sense that the co-occurrence relation works well, the performance becomes slightly worse than the case without the background categories. As described in Section 4.1, this is because the background categories lower the mixture ratios of the object categories, and therefore increase the number of false negatives.

We show the estimated mixture ratios and the ground truth in Fig.11. Here, "bg" stands for the summation of the mixture ratios of two background categories. When we ignore the background categories (C=10), the estimated ratios are significantly different from the ground

	C=10	C=12	ground truth	C=10	C=12	ground truth	C=10	C=12	ground truth
bicycle	0	0	0	0	0	0	0	0	0
bus	0.35	0	0	0	0	0	0	0	0
car	0	0	0	**0**	**0.22**	**0.20**	0.43	0.08	0
cat	0	0	0	0	0	0	0	0	0
cow	0	0	0	0	0	0	0	0.03	0
dog	0	0	0	0	0	0	0	0	0
horse	0	0	0	0	0	0	0	0	0
motorbike	0	0	0	0.37	0	0	0.57	0	0
person	0	0	0	0.63	0	0	**0**	**0.07**	**0.11**
sheep	**0.65**	**0.06**	**0.05**	0	0	0	0	0	0
bg	**---**	**0.94**	**0.95**	**---**	**0.78**	**0.80**	**---**	**0.82**	**0.89**

Fig. 11 Mixture ratios found via MAP estimation without (C=10)/with (C=12) background categories, and ground truth.

truth, because the histogram of visual words arising from backgrounds is forced to be described by those arising from the object categories. On the other hand, when the background categories are combined (C=12), the mixture ratios of the backgrounds have larger values, and those of the object categories come closer to the ground truth. These results imply the effectiveness of the background categories for recognizing images with large background area.

Fig. 10 shows the RMSEs of the estimated mixture ratios. One can see that the co-occurrence relation works well in this case. However, similar to the AUCs, the performance becomes worse than the case without the background categories. The performance could be improved by incorporating the co-occurrence relationship between object categories and background categories.

5 Conclusions and future work

In this paper, we proposed a novel method for recognizing objects of multiple categories coexisting in an image. In particular, our proposed method estimates the mixture ratios of multiple categories in a single image via regression by incorporating the co-occurrence relationship between object categories. We conducted a number of experiments by using the PASCAL dataset and confirmed the effectiveness of the proposed method.

Future directions of this study include incorporating the co-occurrence relationship among more than three categories and modeling background categories via unsupervised learning. In addition, individual elements of BoF such as feature detection, description, and vector quantization could be improved.

Acknowledgement

A part of this work was supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (No. 20700153).

References

[1] D. Lowe, “Distinctive image features from scale-invariant keypoints”, *Int’l Journal of Compute Vision*, vol.60, no.2, pp.91–110, 2004.

[2] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: a comprehensive study”, *Int’l Journal of Compute Vision*, vol.73, no.2, pp.213–238, 2007.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints”, In *Proc. ECCV04 Workshop on Statistical Learning in Computer Vision*, pp.1–22, 2004.

[4] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning”,

In *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition* (*CVPR03*), pp.II-264–II-271, 2003.

[5] S. Savarese and L. Fei-Fei, "3D generic object categorization, localization and pose estimation", In *Proc. IEEE Int'l Conf. Computer Vision* (*ICCV07*), pp.1–8, 2007.

[6] A. Frome, Y. Singer, F. Sha, and J. Malik, "Learning globally-consistent local distance functions for shape-based image retrieval and classification", In *Proc. IEEE Int'l Conf. Computer Vision* (*ICCV07*), pp.1–8, 2007.

[7] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context", In *Proc. IEEE Int'l Conf. Computer Vision* (*ICCV07*), pp.1–8, 2007.

[8] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images", In *Proc. IEEE Int'l Conf. Computer Vision* (*ICCV05*), pp.370–377, 2005.

[9] I. Biederman, R. Mezzanotte, and J. Rabinowitz, "Scene perception: detecting and judging objects undergoing relational violations", *Cognitive Psychology*, vol.14, no.2, pp.143–177, 1982.

[10] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective", In *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition* (*CVPR06*), pp.2137–2144, 2006.

[11] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance", In *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition* (*CVPR08*), pp.1–8, 2008.

[12] G.-J. Qi, X.-S. Hua, Y. Rui, T. Mei, J. Tang, and H.-J. Zhang, "Concurrent multiple instance learning for image categorization", In *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition* (*CVPR07*), pp.1–8, 2007.

[13] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, "The 2006 PASCAL Visual Object Classes Challenge (VOC2006) Results", *http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf*.

[14] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying discriminative visual codebook generation with classifier training for object category recognition", In *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition* (*CVPR08*), pp.1–8, 2008.

Takahiro OKABE

Takahiro OKABE received the BS degree in physics from the School of Science, the University of Tokyo, Japan in 1997, and the MS degree in physics from the Graduate School of Science, the University of Tokyo in 1999. In 2001, he joined the Institute of Industrial Science at the University of Tokyo, where he is currently a research associate. His primary research interests are in the fields of computer vision, pattern recognition, and computer graphics, especially in their physical and mathematical aspects.

Yuhi KONDO

Yuhi KONDO received the BS degree in information and communication engineering from the School of Engineering, the University of Tokyo, Japan in 2006. In 2008, he received the MS degree in information and communication engineering from the Graduate School of Information Science and Technology, the University of Tokyo, where he was engaged in research on generic object recognition. He is currently working at Sony Corporation, Japan.

Kris M. KITANI

Kris M. KITANI graduated with a BS in electrical engineering from the University of Southern California, Los Angeles, USA in 1999. He worked as an engineer at KLA-Tencor from 2000 to 2003 and later earned his MS and PhD in information and communications engineering from the University of Tokyo in 2005 and 2008, respectively. Dr. Kitani is currently an assistant professor at the University of Electro-communications in Tokyo, Japan.

Yoichi SATO

Yoichi SATO is an associate professor jointly affiliated with the Graduate School of Interdisciplinary Information Studies, and the Institute of Industrial Science, at the University of Tokyo, Japan. He received the BSE degree from the University of Tokyo in 1990, and the MS and PhD degrees in robotics from the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, in 1993 and 1997 respectively. His research interests include physics-based vision, reflectance analysis, image-based modeling and rendering, tracking and gesture analysis, and computer vision for HCI.