Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2010, Article ID 426781, 10 pages
doi:10.1155/2010/426781

Research Article

A Multimodal Constellation Model for Object Image Classification

Yasunori Kamiya,[1] Tomokazu Takahashi,[2] Ichiro Ide,[1] and Hiroshi Murase[1]

[1] *Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan*
[2] *Faculty of Economics and Information, Gifu Shotoku Gakuen University, 1-38, Nakauzura, Gifu 500-8288, Japan*

Correspondence should be addressed to Yasunori Kamiya, kamiya@murase.m.is.nagoya-u.ac.jp

Received 8 May 2009; Revised 19 November 2009; Accepted 17 February 2010

Academic Editor: Benoit Huet

We present an efficient method for object image classification. The method is an extention of the constellation model, which is a part-based model. Generally, constellation model has two weak points. (1) It is essentially a unimodal model which is unsuitable to be applied for categories with many types of appearances. (2) The probability function that represents the constellation model requires a high calculation cost. We introduced multimodalization and speed-up technique to the constellation model to overcome these weak points. The proposed model consists of multiple subordinate constellation models so that diverse types of appearances of an object category could be described by each of them, leading to the increase of description accuracy and consequently, improvement of the classification performance. In this paper, we present how to describe each type of appearance as a subordinate constellation model without any prior knowledge regarding the types of appearances, and also the implementation of the extended model's learning in realistic time. In experiments, we confirmed the effectiveness of the proposed model by comparison to methods using BoF, and also that the model learning could be realized in realistic time.

1. Introduction

In this paper, we consider the problem of recognizing semantic categories with many types of appearances such as Car, Chair, and Dog under environment changes such as direction of objects, distance to objects, illumination, and backgrounds. This recognition task is challenging because object appearances widely vary by difference of objects in semantic categories and environment changes, which complicates feature selection, model construction, and training dataset construction. One application of this recognition task is image retrieval.

For these recognition tasks, a part-based approach, which uses many distinctive partial images as local features, is widely employed. By focusing on partial areas, this approach can handle a broad variety of object appearances. Typical well-known methods include a scheme using Bag of Features (BoF) [1] and Fergus's constellation model [2]. BoF is an analogy to the "Bag of Words" model originally proposed in the natural language processing field. Approaches using BoF have been proposed, using classifiers such as SVM (e.g., [3–5]) and document analysis methods such as probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), and Hierarchical Dirichlet Processes (HDPs) (e.g., [6–8]).

On the other hand, the constellation model represents target categories by probability functions that represent local features that describe the common regions[1] of objects in target categories and the spatial relationship between the local features. This model belongs to the "pictorial structure" approach introduced in [9]. The details will be introduced in Section 2.1.

The constellation model has the following three advantages.[2]

(a) *Adding or changing the target categories is easy.* In this research field, recognition methods are often categorized as a "generative model" or a "discriminative approach (discriminative model + discriminant function)" [10]. This advantage comes from the fact that the constellation model is a generative model. A generative model makes a model for each target

category individually. Therefore the training process for adding target categories does not affect the existing target categories. For changing the existing target categories, it is only necessary to change the models used in the tasks; no other training process is necessary.
On the other hand, discriminative approaches, which optimize a decision boundary to classify all target categories, have to relearn the decision boundary each time adding or changing the target categories. For recognition performance, the discriminative approach generally outperforms the generative model.

(b) *Description accuracy is higher than that of BoF due to continuous value expression.* Category representation by BoF is a discrete expression by a histogram formed by the numbers of local features corresponding to each codeword. On the other hand, since the constellation model is a continuous value expression by a probability function, the description accuracy is higher than BoF.

(c) *Position and scale information can be used effectively.* BoF ignores spatial information of local features to avoid complicated spatial relationship descriptions.[3] On the other hand, the constellation model uses a probability function to represent rough spatial relationships as one piece of information to describe the target categories.

In spite of the advantages, the constellation model has the following weak points.

(1) Since it is essentially a unimodal model, it has low description accuracy when objects in the target categories have various appearances.

(2) The probability function that represents the constellation model requires high computational cost.

In this paper, we propose a model that improves the weak points of the constellation model. For weak point (1), we extend the constellation model to a multimodal model. A unimodal model has to represent several types of appearances as one component. But by extension to a multimodal model, some appearances can be cooperatively described by components of the model, improving the accuracy of category description. This improvement is the same as extending a representation by Gaussian distribution to that by Gaussian Mixture Model in local feature representation. In addition, we speed-up the calculation of the probability function to solve weak point (2).

Another constellation model is proposed before Fergus's constellation model in [11]. Multimodalization of this model was done in [12], but the structure of these models considerably differs from Fergus's constellation model, and they have the following three weak points against Fergus's model.

(i) They do not have the advantage (b) of Fergus's constellation model since the way to use local features is close to BoF.

(ii) They do not use the information of common regions' scale.

(iii) They cannot learn appearance and position simultaneously since the learning of them is not independent.

However, Fergus's constellation model requires high computation cost to calculate the probability function which represents the model, so it is unrealistic to multimodalize the model since the estimation of parameters in the probability function requires high computation cost. So we realize the multimodalization of Fergus's constellation model together with the speeding-up of the calculation of the probability function. Fergus's constellation model was also improved in [13], but the improvements were made so that the model can make use of many sorts of local features and modify the positional relationship expression. For clarity, in this paper we focus on the basic Fergus's constellation model.

Image classification tasks can be classified into the following two types.

(1) Classify images with target objects occupying most area of an image, and the object scales are similar (e.g., Caltech101/256).

(2) Classify images with target objects occupying partial area of an image, and the object scales may differ (e.g., Graz, PASCAL).

The method proposed in this paper targets type (1) images. It can, however, also handle type (2) images using methods such as the sliding window method and then handle them as type (1) images.

The remainder of this paper is structured as follows. In Section 2, we describe the Multimodal Constellation Model, the speeding-up techniques, and the training algorithm. In Section 3, we explain the classification method and describe the experiments in Section 4. Finally, we conclude the paper in Section 5.

Note that this paper is an extended version of our work [14], which includes additional experiments and discussions, about number of effective components (part of Section 4.3), object appearances described in each component (Section 4.5), and comparison with Fergus's model (Section 4.6).

2. Multimodal Constellation Model

In this section, we describe Fergus's constellation model, then explain its multimodalization, and finally describe the speeding-up technique for the calculation.

2.1. Fergus's Constellation Model [2]. The constellation model describes categories by focusing on the common object regions in each category. The regions and the positional relationships are expressed by Gaussian distributions.

The model is described by the following equation:

$$\begin{aligned} p(I|\Theta) &= \sum_{\mathbf{h}\in H} p(A, X, S, \mathbf{h}|\Theta) \\ &= \sum_{\mathbf{h}\in H} p(A\mathbf{h}, \theta_A)\, p(X|\mathbf{h}, \theta_X) \\ &\quad \cdot\, p(S|\mathbf{h}, \theta_S)\, p(\mathbf{h}|\theta_{\text{other}}), \end{aligned} \tag{1}$$

where I is an input image and Θ is the model parameters. Image I is expressed as a set of local features which are extracted from image I by a local feature detector (e.g., [15]). Each local feature holds the feature vectors of appearance, position, and scale. The feature vectors of each local feature are brought together according to appearance, position, and scale, and shown as A, X, and S. In addition, as a hyperparameter, the model has the number of regions for description: R. $\mathbf{h}$ is a vector that expresses the combination of correspondences between local features extracted from image I and each region of the model. H is a set of all the combinations of correspondences. By $\sum_{\mathbf{h}\in H}$ all combinations are covered. $p(A|\mathbf{h},\theta_A)$ is a probabilistic distribution which expresses appearances of regions of the model by multiplication of R Gaussian distributions. $p(X|\mathbf{h},\theta_X)$ expresses a pair of x, y coordinates of each region as a $2R$ dimensional Gaussian distribution. $p(S|\mathbf{h},\theta_S)$ is a probabilistic distribution which expresses scale of regions as one Gaussian distribution. For details refer to [2].

The part of the equation, which cyclopedically exhaustively calculates all combinations between all local features and each region of the model ($\sum_{\mathbf{h}\in H}$), is in the form of a summation. However, the part of the equation that describes a target category, $p(A,X,S,\mathbf{h}|\Theta)$, is substantively represented by a multiplication of the Gaussian distributions. Therefore, Fergus's constellation model can be considered as a unimodal model.

2.2. Multimodalization. For improving the description accuracy, we extend the constellation model from a unimodal model to a multimodal model. We formulate the proposed "Multimodal Constellation Model" as follows:

$$\begin{aligned} p_m(I|\Theta) &= \sum_k^K \left\{ \prod_l^L G\left(\mathbf{x}_l|\theta_{k,\hat{r}_{k,l}}\right) \right\} \cdot \pi_k \\ &= \sum_k^K \left\{ \prod_l^L G\left(\mathbf{A}_l|\theta^{(A)}_{k,\hat{r}_{k,l}}\right) G\left(\mathbf{X}_l|\theta^{(X)}_{k,\hat{r}_{k,l}}\right) G\left(\mathbf{S}_l|\theta^{(S)}_{k,\hat{r}_{k,l}}\right) \right\} \\ &\quad \cdot \pi_k, \\ \hat{r}_{k,l} &= \arg\max_r G(\mathbf{x}_l|\theta_{k,r}), \end{aligned} \tag{2}$$

where K is the number of components. If $K \geq 2$, then the model becomes multimodal. Each type of appearance in a target object category is described by each component, so the description accuracy is expected to be improved. L is the number of local features extracted from image I, and $G(\cdot)$ is the Gaussian distribution. Also, $\Theta = \{\theta_{k,r}, \pi_k\}$, $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, $I = \{\mathbf{x}_l\}$, and $\mathbf{x} = (\mathbf{A}, \mathbf{X}, \mathbf{S})$. $\theta_{k,r}$ is a set of parameters of the Gaussian distribution of region r in component k. $\mathbf{x}_l$ is the feature vector of the lth local feature. $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{S}$, which are the feature vectors of appearance, position, and scale, respectively, are subvectors of $\mathbf{x}$. π_k is the existence probability of component k, which assumes $0 \leq \pi_k \leq 1$ and $\sum_k^K \pi_k = 1$. $\hat{r}_{k,l}$ is the index of the most similar region to the local feature l of image I, in component k. Moreover, R (number of regions) exists as a hyperparameter, though it does not appear explicitly in the equation.

2.3. Speeding-Up Techniques. Since the probability function that represents Fergus's constellation model requires high computation cost, estimating the model parameter is also time consuming. In addition, this complicates multimodalization because multimodalization increases the number of parameters and thus completing the training in realistic time becomes impossible. Here we describe two speeding-up techniques.

Simplifying Matrix Calculation. For simplification, we approximated all covariance matrices to be diagonal. This is equivalent of assuming independence. This modification considerably decreases the calculation cost of $(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ and $|\boldsymbol{\Sigma}|$ needed for calculating the Gaussian distributions. The total calculation cost is reduced from $O(D^3)$ to $O(D)$ for $D \times D$ matrices. Although the approximation decreases the individual description accuracy of each component, we expect that the multimodalization increases the overall description accuracy. In particular, when assuming that $\boldsymbol{\Sigma}$ is a diagonal matrix whose diagonal components are σ_d^2,

$$\begin{aligned} (\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \sum_d^D \frac{1}{\sigma_d^2}(x_d-\mu_d)^2, \\ |\boldsymbol{\Sigma}| &= \prod_d^D \sigma_d^2. \end{aligned} \tag{3}$$

Modifying $\sum_{h\in H}$ to $\prod_l^L$ and $\arg\max_r$. The order of $\sum_{h\in H}$ in equation (1) is $O(L^R)$, where L is the number of local features and R is the number of regions. In actuality, even though A* search method is used for speeding-up in [2], the total calculation cost is still large. In the proposed method we changed $\sum_{h\in H}$ to $\prod_l^L$ and $\arg\max_r$. As a result, the cost is reduced to $O(LR)$. This approach is same with the calculation cost reduction in [16] which targeted the classification of identical view angle car images captured by a fixed camera and modified the constellation model for this task.

Here we compare the expression of each model. Fergus's model exhaustively calculates probabilities of all combinations of correspondences between regions and local features. The final probability is calculated as a sum of these probabilities ($\sum_{h\in H}$). On the other hand, our model calculates the final probability using all the local features at once. This is expressed as $\prod_l^L$. After the region which is most similar to each local feature is selected ($\arg\max_r$), the probability to the region is calculated for each local feature. The final probability is calculated as a multiplication of these probabilities. For the detail of the modification refer to [16].

2.4. Parameter Estimation. Model parameter estimation is carried out using the EM algorithm [17]. Algorithm 1 shows the model parameter estimation algorithm for the Multimodal Constellation Model. N denotes the number of training images, and n denotes the index of the training image. $\mathbf{x}_{n,l}$ denotes a feature vector of local feature l in training image n. $\hat{r}_{k,n,l}$ denotes $\hat{r}_{k,l}$ in training image n.

(1) Initialize model parameter $\theta_{k,r}(=\{\boldsymbol{\mu}_{k,r}, \boldsymbol{\Sigma}_{k,r}\}), \pi_k$.
(2) **E step**:

$$q_{k,n} = \frac{\pi_k p(I_n|\theta_k)}{\sum_k^K \pi_k p(I_n|\theta_k)}, \quad \text{where } p(I_n|\theta_k) = \prod_l^L G(\mathbf{x}_{n,l}|\theta_{k,\hat{r}_{k,n,l}}).$$

(3) **M step**:

$$\begin{aligned}
\boldsymbol{\mu}_{k,r}^{\text{new}} &= \frac{1}{Q_{k,r}} \sum_n^N \sum_{l:(\hat{r}_{k,n,l}=r)} q_{k,n}\mathbf{x}_{n,l}, \\
\boldsymbol{\Sigma}_{k,r}^{\text{new}} &= \frac{1}{Q_{k,r}} \sum_n^N \sum_{l:(\hat{r}_{k,n,l}=r)} q_{k,n}(\mathbf{x}_{n,l} - \boldsymbol{\mu}_{k,r}^{\text{new}})(\mathbf{x}_{n,l} - \boldsymbol{\mu}_{k,r}^{\text{new}})^t, \\
\pi_k^{\text{new}} &= \frac{N_k}{N}, \\
&\text{where } Q_{k,r} = \sum_n^N \sum_{l:(\hat{r}_{k,n,l}=r)} q_{k,n}, \quad N_k = \sum_n^N q_{k,n}.
\end{aligned}$$

(4) If parameter updating converges, the estimation process is finished, and $p(k) = \pi_k$, otherwise return to (2).

Algorithm 1: Model parameter estimation algorithm for the multimodal constellation model.

We explain the initial values in initialization (1). The initial values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (diagonal matrix with only diagonal σ^2) are initialized based on the range of feature values. $\boldsymbol{\mu}$ is initialized as random values considering the range of feature values. $\boldsymbol{\Sigma}$ is initialized as static values also considering the range of feature values. π is initialized as $1/K$.

One difference with the general EM algorithm for the Gaussian Mixture Model is that the data that update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are not per image but per local feature extracted from the images. Degree of belonging $q_{k,n}$ of training image n to component k is calculated in the E step, and then all local features extracted from training image n participate in the updating of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ based on the value of $q_{k,n}$. In addition, local feature l participates in the updating of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ of only region $\hat{r}_{k,n,l}$ to which local feature l corresponds.

3. Classification

The classification is performed by the following equation:

$$\hat{c} = \arg\max_c \; p_m(I|\Theta_c)p(c), \tag{4}$$

where $\hat{c}$ is the resultant category, c is a candidate category for classification, and $p(c)$ is the prior probability of category c, which is calculated as the ratio of training image of category c to all candidate categories.

Since the constellation model is a generative model, it is easy to add categories or change candidate categories, and thus the training process is only independently needed for the first time a category is added. For changing already learnt candidate categories, it is only necessary to change the models used in the tasks. On the other hand, discriminative approaches make one classifier using all of the data for all candidate categories. Therefore it has the following two weak points: a training process is needed every time candidate categories are added or changed, and for relearning, all of the training data need to be kept.

4. Experiments

4.1. Conditions. We evaluate the effectivity of multimodalization for constellation models by comparing two models Multimodal Constellation Model ("Multi-CM") and Unimodal Constellation Model ("Uni-CM"). Uni-CM is equivalent to the proposed model when $K = 1$ (unimodal).

We also compare the proposed model's performance to two methods using BoF. "LDA + BoF" is a method using LDA. Each category c is described by LDA probabilistic model individually ($p(I|\Theta_c)$, like a model for bag of words), and an image I is classified by (4). "SVM + BoF" is a method using SVM. In the feature space of BoF (codebook size dimension), SVM classifies an image I described by a BoF feature vector. Multi-CM, Uni-CM, and LDA + BoF are generative models, SVM + BoF is a discriminative approach, and LDA is a multimodal model.

Next, we discuss the influence of hyperparameters K and R on the classification rate, compare the proposed model's performance to Fergus's model with limitation due to the difficulty of Fergus's model calculation time, and quantitatively validate the two previously mentioned advantages (b) and (c) of the constellation model.

Two image datasets were used for the experiments. The first is the Caltech Database [2] ("Caltech"), and the other is the dataset used in the PASCAL Visual Object Classes Challenge 2006 [18] ("Pascal"). As a preparation for the experiments, object areas were clipped from the images as target images using the object area information available in the dataset, because these datasets do not assume the task targeted in this paper (classifying images with target objects occupying most area of an image to correct categories). We defined the task as classifying target images into correct categories (i.e., for ten categories dataset, it is ten-class classification). The classifying process was carried out for each dataset.[4] Half of the target images were used for training and the rest for testing.

Caltech consists of four categories. Figure 1 shows examples of the target images. The directions of the objects in these images are roughly aligned but their appearances widely vary. Table 1 shows number of object area in each category. Pascal has ten categories. Figure 2 shows examples of the target images. The direction and the appearance of objects in Pascal vary widely. Furthermore, the posees of objects in some categories (e.g., Cat, Dog, and Person) vary considerably.

Figure 1: Target images in Caltech [2].

Figure 2: Target images in Pascal [18].

Therefore classification of Pascal images is considered more difficult than that of Caltech images. And Table 2 shows number of object area in each category.

The identical data of local features are used for all methods compared here to exclude the influence of difference of local features on the classification rate. In addition, we experimented ten times by changing training and test images randomly and used the average classification rate of ten times for comparison.

In this paper, we empirically determined K (number of components) as 5 and R (number of regions) as 21. For the local features, we used the KB detector [15] for detection and the Discrete Cosine Transform (DCT) for description. The KB detector outputs positions and scales of local features. Patch images are extracted using these information and are described by the first 20 coefficients calculated by DCT excluding the DC. Therefore, the dimension of a feature vector $\mathbf{x}$ is 23 ($\mathbf{A}:20$, $\mathbf{X}:2$, $\mathbf{S}:1$).

4.2. Effectivity of Multimodalization and Comparison to BoF. For validating the effectivity of multimodalization, we compared the classification rates of Multi-CM and Uni-CM and applied Student's t-test to verify the effectivity. We also compared the proposed method to LDA + BoF and SVM + BoF, which are related methods. These related methods have hyperparameters to represent the codebook size (k of k-means) for BoF. The number of assumed topics for LDA corresponds to the number of components K of Multi-CM. We show the best classification rates obtained by changing these hyperparameters in the following results.

Table 1: Number of object area in Caltech [2].

Category name	Number of object area
Airplanes	1074
Cars Rear	1155
Faces	450
Motorbikes	826

Table 2: Number of object area in Pascal [18].

Category name	Number of object area
Bicycle	649
Bus	469
Car	1708
Cat	858
Cow	628
Dog	845
Horse	650
Motorbike	549
Person	2309
Sheep	843

Table 3 shows classification rates of Multi-CM and Uni-CM together with the standard deviations over ten trials. In addition, we verified the significance of Multi-CM and Uni-CM for both datasets by Student's t-test ($P < 0.01$). The reason for this is considered that multimodalization to a constellation model is effective to such datasets as Caltech and Pascal which contain various appearances in a category (e.g., Caltech-Faces: different persons, Pascal-Bicycle: direction of bicycles).

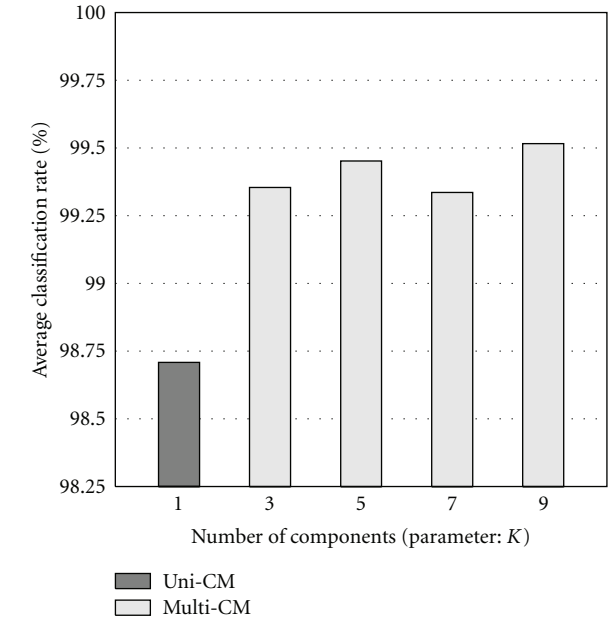
Since the proposed model shows better classification rate than that of LDA + BoF (generative model) or SVM + BoF (discriminative approach), it indicates that the constellation model has better classification ability than the methods based on BoF, for either generative or discriminative approaches.

4.3. Influence of the Number of Components K. Here we discuss the influence of K, one of the hyperparameters of the proposed method, on the classification rate. K is changed in the range of 1 to 9 in increments of 2, to compare the classification rates at each K. When $K = 1$, it is Uni-CM, and when $K \geq 2$, they are Multi-CM. The number of regions R is fixed to 21.

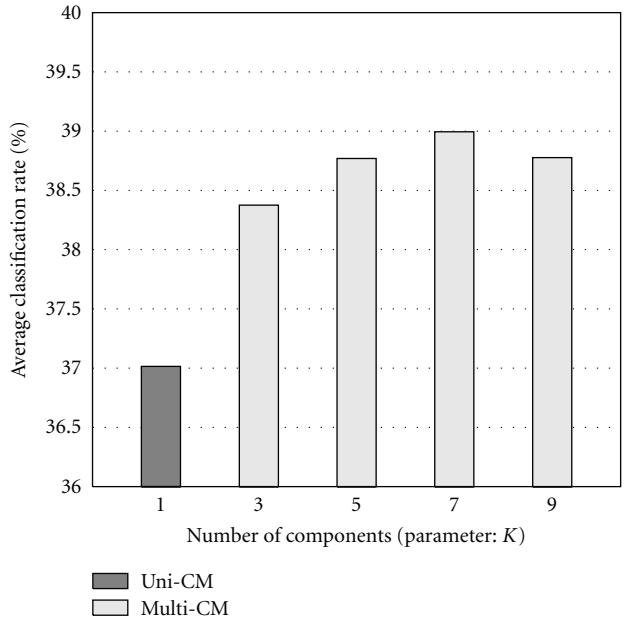
Figure 3 shows the results. Note that the scale of the vertical axis for each graph differs because the difficulty of each dataset differs greatly. By comparing the graphs, we can see that the classification rates roughly saturate at $K = 5$ (Caltech) and 7 (Pascal). We can understand this because the appearance variation of objects for Pascal is larger than that for Caltech. However, we can choose $K = 5$ as a constant setting because these classification rates only differ slightly when $K \geq 2$.

Table 3: Effectivity of multimodalization and comparison to BoF, by average classification rates and standard deviations over ten trials (%).

Dataset	LDA + BoF	SVM + BoF	Uni-CM	Multi-CM
Caltech	94.7 ± 0.66	96.4 ± 0.47	98.7 ± 0.32	99.5 ± 0.10
Pascal	29.6 ± 0.78	27.9 ± 0.44	37.0 ± 0.39	38.8 ± 1.00

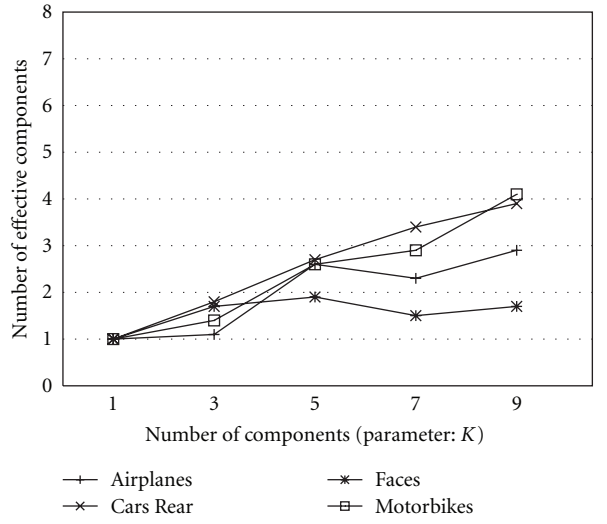


(a) Caltech

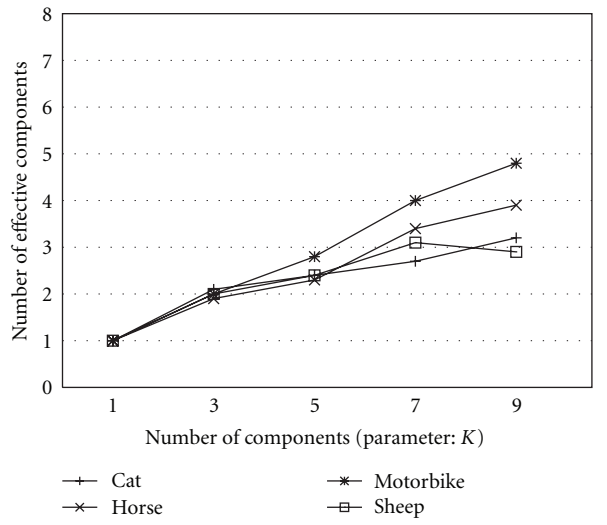


(b) Pascal

Figure 3: Influence of K (number of components) on average classification rate. (Note that the scale of the vertical axis for each graph differs because the difficulty of each dataset differs greatly.)



(a) Caltech

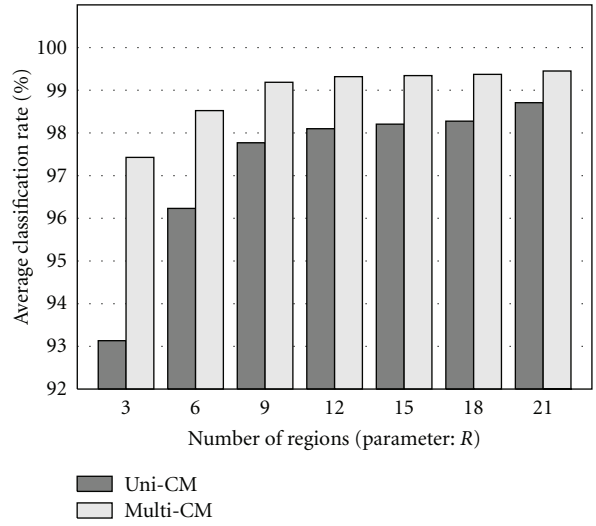


(b) Pascal

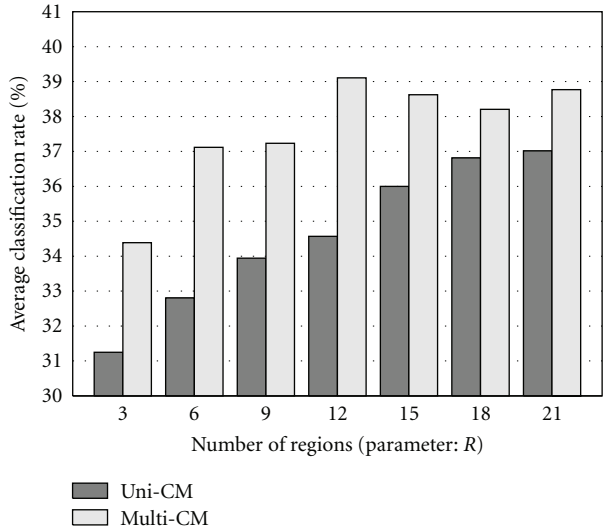
Figure 4: Number of effective components.

In addition, the fact that the classification rates when $K \geq 2$ are better than $K = 1$ shows the effect of multimodalization.

Next, we discuss the number of effective components for each category. We decided that the effective component is a component that satisfies $\pi_k > (1/K) \cdot 0.9$. $1/K$ is the value of π_k when all components are effective and the effect levels are equal. We decided this value as the minimum value,



(a) Caltech



(b) Pascal

Figure 5: Influence of R (number of regions) on average classification rate. (Note that the scale of the vertical axis for each graph differs because the difficulty of each dataset differs greatly.)

and applied 0.9 for allowing some variation. Figure 4 shows graphs with horizontal axes of the number of components K and vertical axes of the number of effective components. The graphs show all categories for Caltech, and some categories for Pascal. From the graphs, we can see that the number of effective components saturates at a certain point, and also the number of effective components for each category varies. We consider that this value roughly indicates the number of object appearances for each category. From the result, we can see that if K increases beyond necessity, the number of components which are learnt as effective components does not change.

Moreover, from this result, we can see that the variation of appearance in Pascal is generally larger than that in Caltech. Actually, when $K = 9$, the average numbers of effective components for all categories are 3.2 for Caltech and 4.0 for Pascal.

4.4. Influence of the Number of Regions R. To discuss the influence of R, another hyperparameter of the proposed method on the classification rate, we evaluated the classification rates by increasing R in the range of 3 to 21 in increments of 3. The classification rate at each R is shown in Figure 5. The number of components K is fixed to 5. The results show the classification rates of both Uni-CM and Multi-CM.

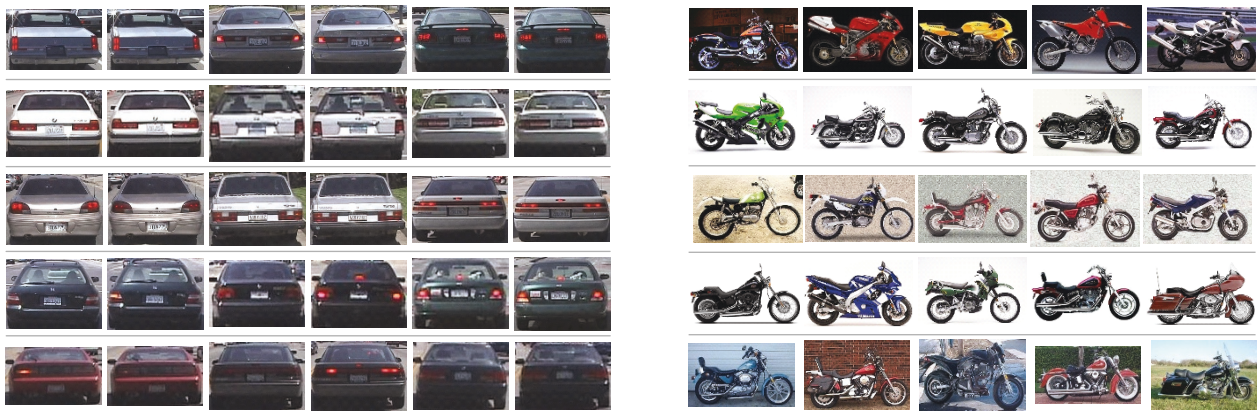
The improvement of classification rates saturates at around $R = 9$ for Caltech and at $R = 21$ for Pascal. In addition, at all R, the classification rates of Multi-CM are higher than those for Uni-CM, so the effectivity of multimodalization is also confirmed here.

For Fergus's constellation model, $R = 6$–7 is the extent that the training process can be finished in realistic time. Thanks to the proposed method with the speed-up techniques, we increased R (number of regions) until the improvement of the classification rate saturated and at the same time in realistic time. Therefore the proposed speeding-up techniques not only contributed to the realization of multimodalization but also to the improvement of the classification performance.

4.5. Object Appearances Described in Each Component. We discuss object appearances described as model components. The model was learnt as $K = 10$ to make it easy to understand what appearances are learnt as component. We apply the learnt multimodal constellation model to test images of the same category that was learnt and calculate the contribution rate $\{\prod_l^L G(\mathbf{x}_l|\theta_{k,\hat{r}_{k,l}})\} \cdot \pi_k$ for each test image for each component. A component with the largest contribution rate is decided as the component that the test image belongs.

Figures 6 and 7 show example images belonging to each component; five dominant components out of ten components are shown. In Caltech-Cars Rear, the groups seem to be constructed mainly by difference of car types. In contrast, Caltech-Motorbikes seem to be constructed by the difference of background appearances because the differences of the backgrounds are larger than those of the bike appearances. In Pascal-Car, the direction of objects and the luminance of the object bodies seem to have affected the group construction. The reason that the luminance affects the grouping is that DCT of luminance is used for local feature description. In Pascal-Motorbike, direction of objects affects the grouping. Pascal-Cow and Pascal-Cat have a wide appearance variation and are difficult to make groups. But the direction of bodies and the texture roughly form groups.

4.6. Comparison with Fergus's Model. Because Fergus's model requires high computation cost and does not run in realistic time under the same experimental condition as ours, we separately discuss this comparison. For this comparison, we limit L (the number of local features in an image) to 20 and set R (the number of parts) to 3 (original setting for our model is 21)[5] for both models. In addition, we set our model to the unimodal condition ($K = 1$) to examine the



(a) Cars Rear (b) Motorbike

Figure 6: Example of groupings for each component of the model (Caltech). Each row shows each component. (In Cars Rear, it seems as if images are shown twice, but this is because Caltech database consists of a lot of images which include same object at same angle but shot timings differ.)

(a) Car (b) Motorbike

(c) Cow (d) Cat

Figure 7: Example of groupings for each component of the model (Pascal). Each row shows each component.

Table 4: Comparison with Fergus's constellation model, by average classification rate and standard deviations over ten trials (%), under limited condition ($L = 20$, $R = 3$) to compare with Fergus's model.

	Our model (unimodal)	Fergus's model
Caltech	93.0 ± 0.68	71.1 ± 0.60
Pascal	31.3 ± 0.34	19.5 ± 0.68

Table 5: Validation of the effectivity of continuous value expression and position-scale information, by average classification rate and standard deviations over ten trials (%).

Dataset	LDA + BoF	Multi-CM no-X,S	Multi-CM
Caltech	94.7 ± 0.66	96.5 ± 0.51	99.5 ± 0.10
Pascal	29.6 ± 0.78	33.5 ± 0.50	38.8 ± 1.00

effects of the simplifications (Section 2.3). The differences between these models are only the simplifications. Same as the experiment in Section 3, we experimented ten times by varying training and test images and used the average classification rate of ten times for comparison.

Table 4 shows the experimental result. For both Caltech and Pascal, the classification rates of the proposed method are higher than those of Fergus's model. First, this result shows that our model outperforms Fergus's model in spite of the limited condition which is favorable for Fergus's model. Second, this result also shows that the effects of the simplifications (Section 2.3) do not affect the recognition performance. Note that Fergus's model implemented by Fergus et al. would give better performance than our implementation, thus a better performance than this result would be given.

4.7. Discussion of Computation Time. First we compare the computation time required for the experiments in Section 4.6. The computation time of Fergus's constellation model to estimate model parameters is five minutes per model for $R = 3$ and $L = 20$ per image. However, our model that applies the above two techniques takes only a second per model to estimate the parameters in the same condition and $K = 1$ (unimodal).

For reference we also compare with the computation time reported in [2]. Note that this is not an accurate comparison because each experimental condition probably does not match (performance of computers used and implementations). According to [2], Fergus's model takes 24–36 hours per model for $R = 6$–7, $L = 20$–30 per image, using 400 training images. However, our model for $K = 1$ (unimodal) takes around ten seconds per model in the same condition. In addition, even when $K \geq 2$ (multimodal), it only takes a few scores of seconds.

4.8. Validation of the Advantage of the Constellation Model. Here, we quantitatively validate the advantages of the constellation model described in Section 1; (b) Description accuracy is higher than BoF due to continuous value expression, and (c) position and scale information ignored by BoF can be used effectively.

First, advantage (b) is validated. The comparison of BoF and the constellation model should be performed on the condition only with the difference that a continuous value expression by a probability function and a discrete expression by a histogram, formed by the numbers of local features, correspond to each codeword. Therefore we compared LDA + BoF, which is a generative multimodal model identical to a constellation model, and Multi-CM without position and scale information that are not used in LDA + BoF ("Multi-CM no-X,S"). Next, to validate advantage (c) we compared Multi-CM no-X,S and the normal Multimodal Constellation Model.

Table 5 shows the classification rates of these three methods. The classification rate of Multi-CM no-X,S is better than that of LDA + BoF, demonstrating the superiority of continuous value expression. The Multi-CM classification rate outperforms Multi-CM no-X,S. This shows that the constellation model can adequately use position and scale information.

5. Conclusion

We proposed a multimodal constellation model for object category recognition. Our proposed method can train and classify faster than Fergus's constellation model and describe categories with a high degree of accuracy even when the objects in the target categories have various appearances.

The experimental results show the following effectivities of the proposed method:

(i) performance improvement by multimodalization

(ii) performance improvement by speeding-up techniques, enabling use with more regions in realistic time.

We also compared Multi-CM to the methods using BoF, LDA + BoF, and SVM + BoF. Multi-CM showed higher performance than these methods. We also compared Multi-CM in the unimodal condition with Fergus's model and confirmed that the simplification of the model structure for the speeding-up in the proposed model does not affect the classification performance. Furthermore, we quantitatively verified the advantages of the constellation model; (b) Description accuracy is higher than BoF due to continuous value expression, and (c) position and scale information ignored by BoF can be used effectively. In Sections 1 and 3, by comparing generative and discriminative approaches, we also showed that the advantage (a) of the constellation model is that candidate categories can be easily added and changed.

In future works, we try to apply our method to object detection, and to investigate deeply the relationship between the appearance variations which seem to differ for each category and the hyperparameters.

Endnotes

1. The number of regions is assumed to be five to seven.

2. Since advantages (b) and (c) are not often described in other papers, we validate them quantitatively in Section 4.8
3. There are some extended BoF methods that consider spatial information (e.g., [19, 20]).
4. Caltech101, 256 exist as datasets considering the task targeted in this paper, but these are not suitable for experiments of this paper because the number of image in each category is small.
5. Fergus's original paper [2] set R to $R = 6$–7. But our paper set R to 3 because of computational cost. For evaluation, the paper in [2] calculated one classification rate only, but our paper used average rate of ten time classifications, thus $R = 6$–7 was not a realistic setting for our paper.

References

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of the International Workshop on Statistical Learning in Computer Vision (ECCV '04)*, pp. 1–22, Prague 1, Czech Republic, 2004.

[2] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, pp. 264–271, Madison, Wis, USA, 2003.

[3] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proceedings of the 10th IEEE International Conference on Computer Vision*, vol. 2, pp. 1458–1465, Beijing, China, October 2005.

[4] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proceedings of the 11th IEEE International Conference on Computer Vision*, pp. 1–8, October 2007.

[5] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[6] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *Proceedings of the European Conference on Computer Vision*, vol. 3954 of *Lecture Notes in Computer Science*, pp. 517–530, 2006.

[7] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 524–531, San Diego, Calif, USA, 2005.

[8] G. Wang, Y. Zhang, and L. Fei-Fei, "Using dependent regions for object categorization in a generative framework," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 1597–1604, New York, NY, USA, June 2006.

[9] M. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 22, no. 1, pp. 67–92, 1973.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, London, UK, 2006.

[11] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proceedings of the 6th European Conference on Computer Vision*, vol. 1, pp. 18–32, Dublin, Ireland, June 2000.

[12] M. Weber, M. Welling, and P. Perona, "Towards automatic discovery of object categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 2, pp. 101–108, Hilton Head Island, SC, USA, 2000.

[13] R. Fergus, P. Perona, and A. Zisserman, "A sparse object category model for efficient learning and exhaustive recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 380–387, San Diego, Calif, USA, June 2005.

[14] Y. Kamiya, T. Takahashi, I. Ide, and H. Murase, "A multi-modal constellation model for object category recognition," in *Proceedings of the 15th International Multimedia Modeling Conference (MMM '09)*, vol. 5371 of *Lecture Notes in Computer Science*, pp. 310–321, Sophia-Antipolis, France, January 2009.

[15] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[16] X. Ma and W. E. L. Grimson, "Edge-based rich representation for vehicle classification," in *Proceedings of the 10th IEEE International Conference on Computer Vision*, vol. 2, pp. 1185–1192, Beijing, China, October 2005.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[18] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool, "The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results," http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006/results.pdf.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 2169–2178, New York, NY, USA, June 2006.

[20] T. Li, T. Mei, I. Kweon, and X. S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*. In press.