A Multimodal Constellation Model for Object Category Recognition

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
kamiya@murase.m.is.nagoya-u.ac.jp,
{ide,murase}@is.nagoya-u.ac.jp
[2] Faculty of Economics and Information, Gifu Shotoku Gakuen University
1-38, Nakauzura, Gifu, 500-8288, Japan
ttakahashi@gifu.shotoku.ac.jp

Abstract. Object category recognition in various appearances is one of the most challenging task in the object recognition research fields. The major approach to solve the task is using the Bag of Features (BoF). The constellation model is another approach that has the following advantages: (a) Adding and changing the candidate categories is easy; (b) Its description accuracy is higher than BoF; (c) Position and scale information, which are ignored by BoF, can be used effectively. On the other hand, this model has two weak points: (1) It is essentially an unimodal model that is unsuitable for categories with many types of appearances. (2) The probability function that represents the constellation model takes a long time to calculate. In this paper we propose a "Multimodal Constellation Model" to solve the two weak points of the constellation model. Experimental results showed the effectivity of the proposed model by comparison to methods using BoF.

Keywords: Constellation model, Multimodalization, Speed-up, Object category recognition, EM algorithm.

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 that include direction of objects, distance to objects, illumination, and backgrounds. This recognition task is challenging because object appearances widely varies by difference of objects in semantic categories and environment changes, which complicates feature selection, model construction, and training dataset construction. One of the 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

B. Huet et al. (Eds.): MMM 2009, LNCS 5371, pp. 310–321, 2009.

well-known methods include a scheme using Bag of Features (BoF) [3] and Fergus's constellation model [7]. 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., [9][12][16]), and document analysis methods such as probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), and Hierarchical Dirichlet Processes (HDP) (e.g., [2][6][13]).

The constellation model represents target categories by probability functions that represent local features that describe the common regions of objects in target categories and the spatial relationship between the local features. The number of regions is assumed to be five to seven. Details will be introduced in Section 2.1.

The constellation model has the following three advantages:

(a) *Adding and changing the target categories is easy.*
In this reseach field, recognition methods are often categorized as a "generative model" or a "discriminative model" [1]. This advantage is because the constellation model is a generative model. A generative model individually makes a model for each target category. Therefore the training process for adding target categories is only needed for the added target categories. For changing the already learnt 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 models, which describe 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 model generally outperforms the generative model.

(b) *Description accuracy is higher than BoF due to continuous value expression.*
Category representation by BoF is a discrete expression by 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 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. 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.

However 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 many types of appearances.
(2) The probability function that represents the constellation model takes a long time to calculate.

In this paper, we propose a model that improves the weak points of the constellation model. For weak point (1), we extend the constellation mode 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 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).

Since advantages (b) and (c) are not often described in other papers, we quantitatively show their correctness in Section 4.4.

Another constellation model is proposed before Fergus's constellation model in [15]. Multimodalization of this model was done in [14], but the structure of these models considerably differs from Fergus's constellation model, and they have three weak points: they do not have the advantage (b) of Fergus's constellation model since the way to use local features is close to BoF, they do not use the information of common regions' scale, and they can not lean appearance and position simultaneously since the learning of appearance and position has a dependence. However, Fergus's constellation model takes a long time to calculate the probability function which represents the model, so it is unrealistic to multimodalize the model since the parameter estimation needs many time of probability function calculation. In this paper, we realize the multimodalization of Fergus's constellation model with the speeding-up the calculation of the probability function. Fergus's constellation model was also improved in [8], but the improvements are to become that the model can use many sorts of local features and to 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 occupy 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 and describe our experiments in Section 4. Finally, we conclude the paper in Section 5.

2 Multimodal Constellation Model

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

2.1 Fergus's Constellation Model [7]

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 follow equation:

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

where I is an input image and Θ is the model parameters. Image I is expressed as a set of local features. Each local feature holds the feature vectors of appearance, position, and scale. A, X, and S is a set of feature vectors of appearance, position, and scale, respectively. 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 expressed as the 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 also expressed by one Gaussian distribution. For details refer to [7].

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

2.2 Multimodalization

We define the proposed "Multimodal Constellation Model" as follows:

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

where K is the number of components. If $K \geq 2$, then the model becomes multimodal. L is the number of local features extracted from image I, and $G(\ )$ is the Gaussian distribution. Also, $\Theta = \{\theta_{k,r}\ , p(k)\}$, $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, $I = \{\mathbf{x}_l\}$, 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 l-th local feature. $\mathbf{A}, \mathbf{X}$, and $\mathbf{S}$, which are the feature vectors of appearance, position, and scale, respectively, are subvectors of $\mathbf{x}$. $p(k)$ is the existence probability of component k. $\hat{r}_{k,l}$ is the index of the most similar region to the local feature l of the 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 takes a long time to calculate, 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 assumed all covariance matrices to be diagonal as an approximation. 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. In particular, when assuming that $\boldsymbol{\Sigma}$ is a diagonal matrix whose diagonal components are σ_d^2,

$$(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \sum_d^D \frac{1}{\sigma_d^2}(x_d-\mu_d)^2 \tag{3}$$

$$|\boldsymbol{\Sigma}| = \prod_d^D \sigma_d^2. \tag{4}$$

[Modifying $\Sigma_{h\in H}$ to $\prod_l^L$ and $\arg\max_r$]. The order of $\Sigma_{h\in H}$ in (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 [7], the total calculation cost is still large. In the proposed method we changed $\Sigma_{h\in H}$ to $\prod_l^L$ and $\arg\max_r$. As a result, the cost is reduced to $O(LR)$.

This approach was inspired from [11] who targeted the classification of identical view angle car images captured by a static single camera, and modified the constellation model for this task. We referred to the part of calculation cost reduction.

Here we compare the expression of each model, and describe that Fergus's model and our model approximately have an equivalent description ability. First we describe each model with its calculation procedure. Fergus's model cyclopedically calculates probabilities of all combinations of correspondences between regions and local features. The final probability is calculated as a sum of these probabilities. The cyclopedic search of corresponding local features is done by $\Sigma_{h\in H}$. 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.

Next, we describe each model with its handling of occlusions (in particular, lack of necessary local features). Fergus's constellation model explicitly handles occlusion. When calculating probabilities for combinations of correspondences, some regions do not correspond to any local feature. This expresses the existence

of occlusions. The probability of the occluded regions' combination is also modeled. For Fergus's constellation model, by such explicit handling, the probability considering occlusion is calculated. On the other hand, our model calculates the final probability using all the local features in an image at once. Therefore the final probability is calculated as a probability without the occluded local features. In addition, frequent occlusion patterns are learnt as one appearance of an object by multimodalization. This corresponds to the modeling the probability of occluded region's combination in Fergus's model. For our model, by these implicit handling, the probability considering occlusion is calculated.

At last, we consider images with unnecessary local features. For Fergus's constellation model, at the cyclopedic searches of corresponding local features, the probability for the combination of correspondences with unnecessary local features becomes small, therefore the final probability is almost not affected by unnecessary local features since it is calculated as a sum of probabilities of the combinations. On the other hand, for our model, since the final probability is calculated as a multiplication of each local feature's probability, unnecessary local features decrease the final probability. However, the probability decreases simultaneously for all candidate categories. Therefore the classification results is not affected by unnecessary local features.

According to [7], the actual computation time of Fergus's constellation model to estimate model parameters is 24–36 hours per model for R=6–7, L=20–30 per image, and using 400 training images. However, our model that applies the above two techniques takes around ten seconds to estimate the parameters in the same condition and K=1 (unimodal). In addition, even when $K \geq 2$ (multimodal), it only takes a few dozen seconds to estimate the parameters.

2.4 Parameter Estimation

Model parameter estimation is carried out using the EM algorithm [4]. Fig. 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.

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:

$$\hat{c} = \arg\max_{c} p_m(I|\Theta_c)p(c), \tag{5}$$

(1) Initialize model parameter $\theta_{k,r}$, and π_k.
(2) **E step** :

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

(3) **M step** :

$$\boldsymbol{\mu}_{k,r}^{new} = \frac{1}{Q_{k,r}} \sum_n^N \sum_{l:(\hat{r}_{k,n,l}=r)} q_{k,n} \mathbf{x}_{n,l},$$

$$\boldsymbol{\Sigma}_{k,r}^{new} = \frac{1}{Q_{k,r}} \sum_n^N \sum_{l:(\hat{r}_{k,n,l}=r)} q_{k,n} \left(\mathbf{x}_{n,l} - \boldsymbol{\mu}_{k,r}^{new}\right) \left(\mathbf{x}_{n,l} - \boldsymbol{\mu}_{k,r}^{new}\right)^t,$$

$$\pi_k^{new} = \frac{N_k}{N},$$

$$\text{where} \quad Q_{k,r} = \sum_n^N \sum_{l:(\hat{r}_{k,n,l}=r)} q_{k,n} \quad , \quad N_k = \sum_n^N q_{k,n}.$$

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

Fig. 1. Model parameter estimation algorithm for the Multimodal Constellation Model

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 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 models makes one classifier (decision boundary) 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 and changed, and for relearning, all of the training data, needs to be keept.

4 Experiments

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 the two methods using BoF. "LDA+BoF" is a method using LDA, one document analysis method. "SVM+BoF" is a method using SVM. Multi-CM, Uni-CM, and LDA+BoF are generative models, SVM+BoF is a discriminative model, and LDA is a multimodal model.

Fig. 2. Target images in Caltech [7]

Fig. 3. Target images in Pascal [5]

Next, we discuss the influence of hyperparameters K and R on the classification rate and quantitatively show the two previously mentioned advantages of the constellation model.

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. We defined the task as classifying target images into correct categories. The classifying process was carried out for each dataset. Half of the target images were used for training and the rest for testing.

Two image datasets were used for the experiments. The first is the Caltech Database [7] ("Caltech"), and the other is the dataset used in the PASCAL Visual Object Classes Challenge 2006 [5] ("Pascal"). Caltech consists of four categories. Fig. 2 shows examples of the target images. The directions of the objects in these images are roughly aligned but their appearances widely varies. Pascal has ten categories. Fig. 3 shows examples of the target images. The direction and the appearance of objects in Pascal vary widely. Furthermore, the pose of objects in some categories (e.g., Cat, Dog, and Person) vary considerably. Therefore Pascal is considered more difficult than Caltech.

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 varying training and test images and used the average classification rate of ten times for comparison.

In this paper we empirically determined K (number of components) as five and R (number of regions) as 21. For the local features we used the KB detector [10] for detecting and the Discrete Cosine Transform (DCT) for describing. 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 feature vector $\mathbf{x}$ is 23 ($\mathbf{A}$:20, $\mathbf{X}$:2, $\mathbf{S}$:1).

4.1 Effectivity of Multimodalization and Comparison to BoF

For validating the effectivity of multimodalization, we compared the classification rates of Multi-CM and Uni-CM. We also compared the proposed method to LDA+BoF and SVM+BoF, which are related methods. These related methods

Table 1. Effectivity of multimodalization and comparison to BoF, by average classification rates (%)

Dataset	LDA+BoF	SVM+BoF	Uni-CM	Multi-CM
Caltech	94.7	96.4	98.7	99.5
Pascal	29.6	27.9	37.0	38.8

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 while changing these hyperparameters in the following results.

Table 1 shows that the classification rate of Multi-CM outperforms Uni-CM. This shows that multimodalization to a constellation model is effective to such datasets as Caltech and Pascal which contain many types of appearances in a category (e.g., Caltech-Face: differences of persons, Pascal-Bicycle: direction of bicycle).

Since the results also show that the proposed model obtains better classification rate than LDA+BoF (generative model) and SVM+BoF (discriminative model), we can obtain better classification performance with the constellation model than using methods based on BoF, for either generative or discriminative models.

4.2 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 4 shows the results. Note that the scale of the vertical axis for each graph differs because the difficulty of each dataset differs greatly. The classification rates

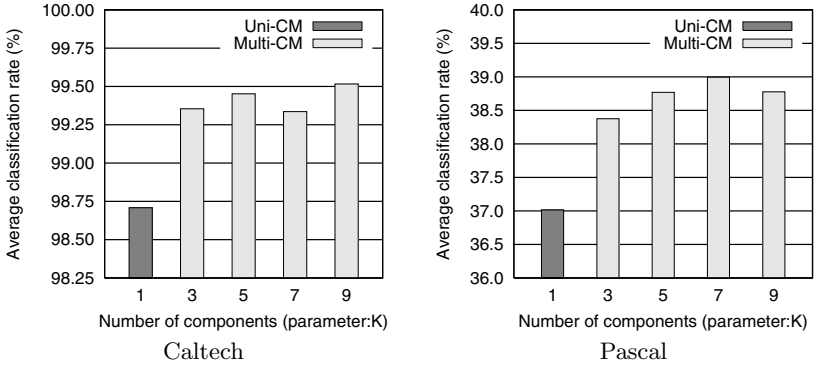

Fig. 4. Influence of K (number of components) on average classification rate

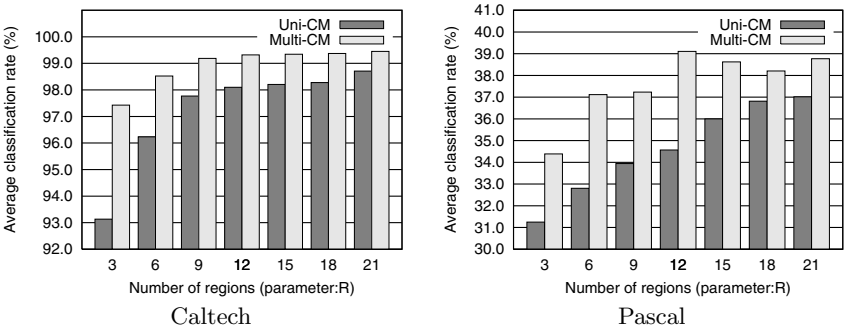

Fig. 5. Influence of R (number of regions) on average classification rate

saturate at K=5 for Caltech and at K=7 for Pascal because the appearance variation of objects for Pascal is bigger than Caltech. However, we can choose K=5 as a constant setting because these classification rates only differ slightly when $K \geq 2$.

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

4.3 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, and the classification rate at each R is shown in Fig. 5. The number of components K is fixed to 5. The results contain the classification rates of Uni-CM and Multi-CM.

The improvement of classification rates saturate at R=9 for Caltech and at R=21 for Pascal. In addition, at all R, the classification rates of Multi-CM are better than 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. For the proposed method with the speed-up techniques, we increased R (number of regions) until the improvement of the classification rate saturated, in realistic time. Therefore the speeding-up techniques not only contributed to the realization of multimodalization but also to the improvement of the classification performance.

4.4 Continuous Value Expression and Position-Scale Information

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, (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.

Table 2. Validation of effectivity of continuous value expression and position-scale information, by average classification rate (%)

Dataset	LDA+BoF	Multi-CM no-X,S	Multi-CM
Caltech	94.7	96.5	99.5
Pascal	29.6	33.5	38.8

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 (c) we compared Multi-CM no-X,S and the normal Multimodal Constellation Model.

Table 2 shows the classification rates of these three methods. The classification rate of Multi-CM no-X,S is better than 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 many types of appearances.

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

- Performance improvement by multimodalization
- 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. Furthermore we quantitatively showed 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 models, we also showed that the advantage (a) of the constellation model is that candidate categories can be easily added and changed.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
2. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)

3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. ECCV International Workshop on Statistical Learning in Computer Vision, pp. 1–22 (2004)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statistical Society, Series B 39(1), 1–38 (1977)
5. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L.: The PASCAL Visual Object Classes Challenge 2006 Results (VOC 2006) (2006), http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf
6. Fei-Fei, L., Perona, A.P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 524–531 (2005)
7. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 264–271 (2003)
8. Fergus, R., Perona, P., Zisserman, A.: A sparse object category model for efficient learning and exhaustive recognition. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 380–387 (2005)
9. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: Proc. IEEE Int. Conf. on Computer Vision, vol. 2, pp. 1458–1465 (2005)
10. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. of Computer Vision 45(2), 83–105 (2001)
11. Ma, X., Grimson, W.E.L.: Edge-based rich representation for vehicle classification. In: Proc. IEEE Int. Conf. on Computer Vision, vol. 2, pp. 1185–1192 (2005)
12. Varma, M., Ray, D.: Learning the discriminative power-invariance trade-off. In: Proc. IEEE Int. Conf. on Computer Vision (2007)
13. Wang, G., Zhang, Y., Fei-Fei, L.: Using dependent regions for object categorization in a generative framework. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 1597–1604 (2006)
14. Weber, M., Welling, M., Perona, P.: Towards automatic discovery of object categories. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 101–108 (2000)
15. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, pp. 18–32. Springer, Heidelberg (2000)
16. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. Int. J. of Computer Vision (2), 213–238 (2007)