IPSJ Transactions on Computer Vision and Applications Vol.7 69–73 (July 2015)

[DOI: 10.2197/ipsjtcva.7.69]

Express Paper

Mahalanobis Encodings for Visual Categorization

Tomoki Matsuzawa[1] Raissa Relator[1] Wataru Takei[1] Shinichiro Omachi[2] Tsuyoshi Kato[1,a)]

Received: March 13, 2015, Accepted: April 20, 2015, Released: July 27, 2015

Abstract: Nowadays, the design of the representation of images is one of the most crucial factors in the performance of visual categorization. A common pipeline employed in most of recent researches for obtaining an image representation consists of two steps: the encoding step and the pooling step. In this paper, we introduce the Mahalanobis metric to the two popular image patch encoding modules, Histogram Encoding and Fisher Encoding, that are used for Bag-of-Visual-Word method and Fisher Vector method, respectively. Moreover, for the proposed Fisher Vector method, a close-form approximation of Fisher Vector can be derived with the same assumption used in the original Fisher Vector, and the codebook is built without resorting to time-consuming EM (Expectation-Maximization) steps. Experimental evaluation of multi-class classification demonstrates the effectiveness of the proposed encoding methods.

Keywords: Bag-of-Visual-Word, Fisher Vector, Mahalanobis metric, visual categorization

1. Introduction

Nowadays, the design of the representation of images is one of the most crucial factors in the performance of visual categorization. A common pipeline employed in most recent researches for obtaining an image representation consists of two steps: the encoding step and the pooling step (e.g., Ref. [4]). This two-step approach stems from the Bag-of-Visual-Words (BoVW) method [4], which is the most popular representation in computer vision. As of present, the central issue in visual categorization is how to improve each step in the two-step approach.

For the last decades, many variants of the encoding modules for the two-step pipeline framework have been developed. Local image patches for the inputs of the encoding step are often extracted densely from the entire image [11], although some methods still employ interest points as the positions of local image patches [6]. Some well-known encoding modules include Histogram Encoding such as hard assignment [4], [9], [12] and soft assignment [5], [16], Fisher Encoding [11] and its variant, VLAD [7], Super Vector Encoding [18], and Sparse Encoding [1], [14], [17] such as Locality-constrained Linear Encoding [14]. Popular pooling modules are Average Pooling and Max Pooling. Max Pooling is mainly adopted together with Sparse Encoding, and Average Pooling is usually employed with Histogram Encoding and Fisher Encoding, nevertheless the pairings of the encoding module and the pooling module are interchangeable [1]. Some methods also incorporate spatial information in the pooling step [9], [17]. Chatfield et al. [2] conducted exhaustive experimental comparisons of encoding modules on large benchmarking datasets, and concluded that Fisher Vector (FV) is the most promising descriptors for visual categorization. From these facts, we focus the discussion hereinafter on the most well-known descriptor, BoVW, employing Histogram Encoding, and the most promising descriptor, FV, using Fisher Encoding.

To the best of our knowledge, almost all of the existing methods encode image patches and build a codebook in the Euclidean metric. The BoVW method minimizes the sum of the square Euclidean deviations from the nearest visual words to build a codebook. The codebook for FV method is built by maximum likelihood estimation of the model parameters for a probabilistic model defined in the Euclidean metric. However, so far no study has investigated which metric is better for image patch encoding, although there have been some works that incorporate non-Euclidean metrics for pattern classification (e.g., Ref. [8], [15]).

In this paper, we introduce the Mahalanobis metric to two popular image patch encoding modules, Histogram Encoding and Fisher Encoding, which are fundamental in the encoding step for the BoVW method and the FV method, respectively. We consider two approaches: global metric approach and local metric approach. In the first approach, a single metric is introduced to the entire local feature space, and the encoding of image patches is done in the global common metric. In the second approach, different visual words are allowed to use different local metrics to compute the deviation and the gradient of the likelihood at each image patch. In our implementation, the global metric is obtained in an unsupervised fashion, and the local metrics are computed in a class-wise manner.

Related Work. Preceding the encoding step, dimension reduction using principal component analysis (PCA) is often performed [7], [11]. The pre-processing may yield a similar effect of introducing the global Mahalanobis metric. PCA eliminates the minor elements and whitens the major elements. Indeed, the classification performance of the global Mahalanobis Fisher encoding did not differ largely from the approach of the PCA fol-

[1] Graduate School of Engineering, Gunma University, Kiryu, Gunma 376–8515, Japan
[2] Graduate School of Engineering, Tohoku University, Sendai 980–8579, Japan
[a)] katotsu@cs.gunma-u.ac.jp

lowed by Fisher encoding (called *PCA FV*) in our experiments reported in Section 5. On the other hand, the use of local metrics in FV yielded a significant improvement in the categorization performance.

Tanaka et al.'s study [13] is relevant to the proposed local Mahalanobis Fisher encoding. Fisher vectors are given by the gradient of the log likelihood of the codebook expressed as a probabilistic model. Tanaka et al. [13] employed the full covariance Gaussian mixture as the probabilistic model, which may give rise to an expectation that their method already has a potential for the local Mahalanobis metrics. One noteworthy difference between their method and ours lies in the construction of the codebook. Tanaka et al.'s method for codebook construction is an expectation-maximization (EM) algorithm that contains a computationally expensive process of matrix inversion at every iteration for every visual word. In this study, an alternative faster technique that still allows us to reuse the fast EM implementation for classical diagonal covariance Gaussian mixture is developed. Tanaka et al. also presented a solution to make the EM algorithm faster, which is by replacing the covariance matrices with block-diagonal matrices. We experimentally show that the proposed formulation significantly outperforms both the full covariance FV (*FCFV*) and the block diagonal covariance FV (*BDFV*).

Contributions. The main contributions of this study can be summarized as follows.

- We explore how to introduce both the global Mahalanobis metric and the local Mahalanobis metrics to BoVW method and FV method.
- We show that a close-form approximation of FV can be derived with the same assumption used in the original FV.
- We develop a method to build a codebook of FV method for both the global Mahalanobis metric and the local Mahalanobis metric, without resorting to very time-consuming steps.
- Experimental evaluation of multi-class classification demonstrates the significantly superior performance of the proposed local Mahalanobis FV, not only in classical image representations, but also in state-of-the-art encoding methods including PCA FV, FCFV, and BDFV.

2. Fisher Vector Revisited

In this section, we review the FV method as a preliminary to our formulations.

FV method is based on the Fisher kernel that exploits a probabilistic model $p(\boldsymbol{X}\,|\,\boldsymbol{\theta})$ with model parameter $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_F]^\top$, to define an inner product of data $\boldsymbol{X}$. FV of a set of local features $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \in \mathbb{R}^{d\times T}$ extracted from an image is defined as a normalized gradient vector

$$\boldsymbol{f}(\boldsymbol{X}) := \left[\frac{\nabla_{\theta_i} \log p(\boldsymbol{X}|\boldsymbol{\theta})}{\sqrt{\mathcal{E}_X\left[(\nabla_{\theta_i} \log p(\boldsymbol{X}|\boldsymbol{\theta}))^2\right]}}\right]_{i\in\{1,\ldots,F\}}$$

where $\mathcal{E}_X(\cdot)$ is the expectation operator with respect to $\boldsymbol{X}$. The probabilistic model $p(\cdot\,|\,\boldsymbol{\theta})$ is an alternative to a codebook of the BoVW method. FV allows us to choose an arbitrary probabilistic model as $p(\cdot\,|\,\boldsymbol{\theta})$ [3], [13], although Gaussian mixture with diagonal covariance has been adopted in most researches [11]. The probabilistic densities with K mixture components are expressed as

$$p_{\mathrm{euc}}(\boldsymbol{X}\,|\,\boldsymbol{\theta}) = \prod_{t=1}^{T}\sum_{k=1}^{K}\pi_k \mathcal{N}(\boldsymbol{x}_t\,;\,\boldsymbol{\mu}_k, \boldsymbol{D}^2_{\boldsymbol{\sigma}_k})$$

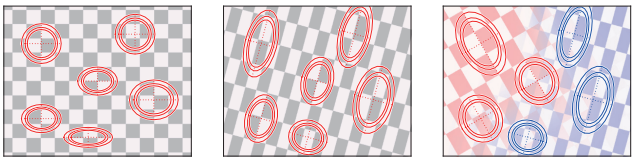
where $\boldsymbol{\theta}$ is the $(2d+1)K$-dimensional parameter vector

$$\boldsymbol{\theta} = \left[\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_K^\top, \boldsymbol{\sigma}_1^\top, \ldots, \boldsymbol{\sigma}_K^\top\right]^\top,$$

π_k are the model coefficients such that $\sum_k \pi_k = 1$, and $\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k \in \mathbb{R}^d$ are the mean and the variance parameters, respectively, and the notation $\boldsymbol{D}_{\boldsymbol{x}}$ is used to denote a diagonal matrix whose diagonal entries are $\boldsymbol{x}$.

The sub-vector of FV in the first K dimensions, associated with $\pi_1, \ldots, \pi_K$, becomes the soft assignment of Histogram Encoding in some setting, but these K dimensions are often discarded and only the remaining $2dK$-dimensional sub-vector is used [11].

From the nature of diagonal Gaussian, one might think that the use of FV with the diagonal Gaussian mixture adjusts the metric by changing the variance parameters, although the directions of rescaling the metric are still limited to the original axes (**Fig. 1** (a)).



(a) EucFV (b) GMahaFV (c) LMahaFV

Fig. 1 Visual words on different metric spaces. In this example, six visual words are in the codebook with different metrics. (a) EucFV uses visual words on the Euclidean metric. (b) GMahaFV method replaces the Euclidean metric to a single Mahalanobis metric. (c) LMahaFV method divides the image patch space softly into two subsets each of which is on a different metric, and three visual words are given to each of the two subsets.

3. Mahalanobis Encodings

This study introduces the Mahalanobis metric [8], [15] to BoVW and FV frameworks to enhance classification performance. The Mahalanobis metric is a distance metric among vectors. The distance metric is parameterized with a positive definite matrix $\boldsymbol{A}$ referred to as a *Mahalanobis matrix*. The distance between $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ is defined as

$$D(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{A}) := \sqrt{(\boldsymbol{x}-\boldsymbol{x}')^\top \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}')}.$$

Mahalanobis BoVW. We propose adopting the Mahalanobis metric in the encoding step of BoVW method. Namely, when each image patch is assigned to a visual word, the Mahalanobis distance from the image patch to all the visual words are computed to find the nearest visual words. We refer to the encoding module based on the Mahalanobis metric as *Mahalanobis encoding*, whereas a new term, *Euclidean encoding*, is also used in this paper to distinguish the classical encoding from the proposed encoding.

The codebook used for Mahalanobis encoding consists of not only a set of visual words, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K$, but also of several Mahalanobis matrices, $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_M$, where K and M denote the number of visual words and number of Mahalanobis matrices, respectively. Moreover, $M \le K$ is assumed. Each visual word

$\boldsymbol{v}_k \in \mathbb{R}^d$ is associated with one of M Mahalanobis matrices, $\boldsymbol{A}_{m_k}$. An input image patch $\boldsymbol{x} \in \mathbb{R}^d$ is assigned into a single visual word $\boldsymbol{v}_{k_*}$ such that

$$k_* \in \operatorname*{argmin}_{k\in\{1,\dots,K\}} D(\boldsymbol{x}, \boldsymbol{v}_k; \boldsymbol{A}_{m_k}). \tag{1}$$

Letting $\boldsymbol{A}_m = \boldsymbol{U}_m \boldsymbol{\Lambda}_m \boldsymbol{U}_m^\top$ be the spectral decomposition of $\boldsymbol{A}_m$, the whitening matrix $\boldsymbol{W}_m := \boldsymbol{\Lambda}_m^{1/2} \boldsymbol{U}_m^\top$ offers another equivalent form to Eq. (1) for finding the nearest visual word: $k_\star \in \operatorname*{argmin}_{k\in\{1,\dots,K\}} \|\boldsymbol{W}_{m_k}\boldsymbol{x} - \boldsymbol{W}_{m_k}\boldsymbol{v}_k\|$.

When $M = 1$, the deviations from visual words are computed with a common single metric everywhere in the entire d-dimensional space of local descriptors. Some studies (e.g., Refs. [8], [15]) employ multiple local metrics to adjust the metrics in each local region, which motivated us to also employ multiple metrics. We refer to the BoVW method with a single Mahalanobis metric as the *global Mahalanobis BoVW* (*GMahaBoVW*), whereas the BoVW method with multiple Mahalanobis metrics is called the *local Mahalanobis BoVW* (*LMahaBoVW*).

Mahalanobis FV. We now present another proposed method, *Mahalanobis FV*, that incorporates the effects of the Mahalanobis metrics in the FV framework. The Mahalanobis FV employs a probabilistic model of an image patch set representing the densities of whitened local features as

$$\begin{aligned} &p_{\mathrm{maha}}(\boldsymbol{X}|\boldsymbol{\theta}) \\ &:= \prod_{t=1}^{T} \sum_{k=1}^{K} \pi_k |\det(\boldsymbol{W}_{m_k})| \mathcal{N}(\boldsymbol{W}_{m_k}\boldsymbol{x}_t\,;\, \boldsymbol{\mu}_k, \boldsymbol{D}^2_{\sigma_k}). \end{aligned} \tag{2}$$

This becomes the standard diagonal covariance Gaussian mixture provided that $\forall m, \boldsymbol{A}_m = \boldsymbol{I}$. This formulation gives rise to a soft division of the image patch space into M regions, and the encoding to each visual word is performed on the metric specific to its region (Fig. 1 (c)).

The classical diagonal covariance FV method assumes almost hard assignments to visual words and that the number of image patches on an image is fixed, to express the normalized gradient vector in a simple form. A simple form of our formulation (2) can be derived under the same assumption as the classical FV. Let us partition the $2dK$-dimensional FV into $2K$ sub-vectors as

$$\begin{aligned} &\boldsymbol{f}(\boldsymbol{X}) \\ &:= [\boldsymbol{f}^{\mu_1}(\boldsymbol{X})^\top, \dots, \boldsymbol{f}^{\mu_K}(\boldsymbol{X})^\top, \boldsymbol{f}^{\sigma_1}(\boldsymbol{X})^\top, \dots, \boldsymbol{f}^{\sigma_K}(\boldsymbol{X})^\top]^\top \end{aligned}$$

to give the approximations

$$\begin{aligned} \boldsymbol{f}^{\mu_k}(\boldsymbol{X}) &\approx \frac{1}{\sqrt{T\pi_k}} \boldsymbol{D}^{-1}_{\sigma_k} \boldsymbol{Y}_k \boldsymbol{\gamma}_k, \\ \boldsymbol{f}^{\sigma_k}(\boldsymbol{X}) &\approx \frac{1}{\sqrt{2T\pi_k}} \boldsymbol{D}^{-2}_{\sigma_k} \left(\boldsymbol{Y}_k \odot \boldsymbol{Y}_k - \boldsymbol{1}_d \boldsymbol{1}_T^\top\right) \boldsymbol{\gamma}_k, \end{aligned} \tag{3}$$

for $k = 1, \dots, K$, where $\odot$ denotes the entrywise product, and $\boldsymbol{Y}_k := \boldsymbol{W}_{m_k}\boldsymbol{X} - \boldsymbol{\mu}_k \boldsymbol{1}_T^\top$ has been defined.

The vector $\boldsymbol{\gamma}_k \in \mathbb{R}^T$ stores the responsibilities to k-th mixture component, where the t-th entry in $\boldsymbol{\gamma}_k$ is expressed as

$$\frac{\pi_k |\det(\boldsymbol{W}_{m_k})| \mathcal{N}(\boldsymbol{W}_{m_k}\boldsymbol{x}_t\,;\, \boldsymbol{\mu}_k, \boldsymbol{D}^2_{\sigma_k})}{\sum_{k'} \pi_{k'} |\det(\boldsymbol{W}_{m_{k'}})| \mathcal{N}(\boldsymbol{W}_{m_{k'}}\boldsymbol{x}_t\,;\, \boldsymbol{\mu}_{k'}, \boldsymbol{D}^2_{\sigma_{k'}})}.$$

The FV method with a single metric (i.e., $M = 1$) and the FV method with multiple metrics (i.e., $M > 1$) are referred to as the *global Mahalanobis FV* (*GMahaFV*) and the *local Mahalanobis FV* (*LMahaFV*), respectively.

4. Construction of Codebooks

To apply the aforementioned methods based on the Mahalanobis encoding to visual categorization, the Mahalanobis matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_M$, need to be given in the codebook in advance. Additionally, for Mahalanobis FV methods, the values of the parameter $\boldsymbol{\theta}$ of the probabilistic model (2) need to be determined. In this section, we discuss how to determine the Mahalanobis matrices and the parameters $\boldsymbol{\theta}$.

Codebook for GMahaBoVW. GMahaBoVW method requires a single Mahalanobis matrix $\boldsymbol{A}_1$. In this study, the covariance matrix of image patches in a training dataset is computed and the inverse of the covariance matrix is chosen as the Mahalanobis matrix. To ensure the existence of the inverse of the covariance matrix, a small positive number is added to the diagonal entries of the covariance matrix. In our experiments, we set this small constant as 0.05 times the largest eigenvalue of the covariance matrix. K visual words are determined using a K-means-like algorithm so that the distortion function based on the Mahalanobis metric $D(\cdot, \cdot; \boldsymbol{A}_1)$ is minimized.

Codebook for LMahaBoVW. This study aims at multi-class classification scenario as the application of Mahalanobis encodings. Letting n_c be the number of classes, LMahaBoVW method use $(n_c + 1)$ Mahalanobis matrices (i.e., $M = n_c + 1$). For the first n_c Mahalanobis matrices, $\boldsymbol{A}_c$ is associated with class c, and is determined as the inverse of the covariance matrix of the image patches from class c, while the inverse of the covariance matrix of the entire set of image patches is set to the last Mahalanobis matrix $\boldsymbol{A}_{n_c+1}$. The K visual words are divided in two halves. The first half is evenly divided into n_c groups, each of which has a one-to-one correspondence to one of n_c classes. Visual words associated with each class are determined using the image patches from the corresponding class. The last half of K visual words is determined in an unsupervised fashion.

Codebook for GMahaFV. The Mahalanobis matrix $\boldsymbol{A}_1$ used in the GMahaFV method is obtained in the same manner as in the GMahaBoVW method. The model parameter $\boldsymbol{\theta}$ is determined by maximum likelihood estimation from the entire training data. Advantageously, our probabilistic model supports the reuse of the publicly available very fast implementation of the EM algorithm for diagonal Gaussian mixture provided, for example, in `VLFeat 0.9.20`.

Codebook for LMahaFV. Similar to the LMahaBoVW method, half of the K mixture components is evenly divided to n_c classes so that each class has a smaller Gaussian mixture model, and the mixture components in the class-specific Gaussian mixture has a single whitening matrix computed from the class-specific covariance matrix. The model parameters of the class-specific Gaussian mixture is determined by the EM algorithm. The n_c Gaussian mixtures are fused again and, furthermore, a Gaussian mixture fitted to the entire training data is added to obtain the final Gaussian mixture (2). Notice that this approach

Table 1 Classification accuracies on FMD and LSP15.

	EucBoVW	GMahaBoVW	LMahaBoVW	EucFV	GMahaFV	LMahaFV	PCA FV	FCFV	BDFV
FMD	0.423	0.428	0.428	0.485	0.507	**0.525**	0.508	0.492	0.488
LSP15	0.768	0.757	0.736	0.768	0.781	**0.796**	0.785	0.767	0.773

again allows reuse of the resource for the fast EM implementation for diagonal covariance Gaussian mixture in order to determine the model parameters of our model.

5. Experimental Results

We experimentally study the categorization performance of Mahalanobis encoding by comparing the following six methods: EucBoVW, GMahaBoVW, LMahaBoVW, EucFV, GMahaFV, and LMahaFV, where EucBoVW and EucFV are the classical Euclidean BoVW and FV methods, respectively.

Experimental Settings. In all experiments, 128-dimensional SIFT features are extracted densely from a grid with 3×3 spacing. The size of the codebook is fixed to $K = 1{,}024$ for BoVW method and $K = 256$ for FV method. After the encoding step and the pooling step, the obtained feature vectors are transformed by the combination of signed square rooting [10] and ℓ_2-normalization. One-vs-rest SVM is employed for multi-class classification. The six encoding methods are first examined on two datasets, Flickr Material Database (FMD) and LSP15. From each class, 30 training images and 10 testing images are picked randomly. This procedure is repeated 20 times and the average accuracies are reported. One-sample t-test is performed to detect the difference between two accuracies.

Comparisons of Different Methods. **Table 1** summarizes the average accuracies of the six methods. Although the Mahalanobis encodings do not necessarily improve the classification performance of the BoVW method, FV method achieves some performance improvement by Mahalanobis encodings. In particular, the local Mahalanobis encoding in FV outperforms the other methods. Table 1 also shows the comparisons with the other three state-of-the-art methods: PCA FV, FCFV and BDFV. The results suggest that the local Mahalanobis FV has an advantage over all the other methods on the two datasets. The difference of LMahaFV to the best method among the three state-of-the-art methods is statistically tested, and the P-values are 3.52×10^{-5} and 2.22×10^{-4}, respectively, on FMD and LSP15.

Effects on Number of Classes. We conducted visual categorization experiments on Caltech-101 as well as on FMD and LSP15. To investigate the performances on different numbers of classes, five subsets, Set A, Set B, Set C, Set D, and Set E, are made from the 101-class problem in Caltech-101. From the 101 ordered classes, the first 10 classes are picked to construct Set A. Similarly, Set B, Set C, Set D, and Set E are the subsets of Caltech-101 containing the first 20, 30, 40, and 50 classes, respectively. Therefore, the number of classes in the five sets are $n_c = 10, 20, 30, 40, 50$, respectively. Top-N accuracies are used to evaluate the categorization performance, where N is set to $n_c/10$. The other settings are same as those for FMD and LSP15.

Figure 2 plots the average of top-N accuracies over 20 different training/testing sets. It can be observed that the local Mahalanobis metric performs better especially when the number

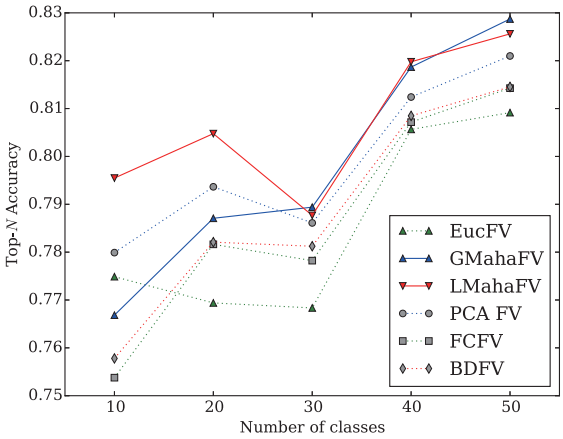


Fig. 2 Accuracies with different numbers of classes.

of classes is small. When $n_c = 10$ and $n_c = 20$, the difference of the top-N accuracies between LMahaFV and the best of the other methods are statistically detected with $P = 3.79 \times 10^{-3}$ and $P = 3.24 \times 10^{-4}$, respectively. On the other hand, the performances of GMahaFV and LMahaFV are comparable in cases where $n_c = 30, 40, 50$.

6. Conclusions

In this paper, we focused on two popular encoding modules, Histogram Encoding and Fisher Encoding, and introduced the Mahalanobis metric to the two encoding modules. Aside from these two modules, many extensions of the encoding-pooling approach for image classification are discussed, as in Section 1. Combinations of these extensions with the Mahalanobis metric open possibilities of further improvement of each method, and thereby the concept of Mahalanobis metric has a potential of being a new axis of the computer vision community.

References

[1] Boureau, Y.-L., Bach, F., LeCun, Y. and Ponce, J.: Learning mid-level features for recognition, *IEEE Conf. Comput. Vis. Pattern Recogn.* (*CVPR*), pp.2559–2566 (2010).
[2] Chatfield, K., Lempitsky, V., Vedaldi, A. and Zisserman, A.: The devil is in the details: An evaluation of recent feature encoding methods, *Proc. British Machine Vision Conference*, pp.76.1–76.12, BMVA Press (Sep. 2011).
[3] Cinbis, R.G., Verbeek, J. and Schmid, C.: Image categorization using Fisher kernels of non-iid image models, *IEEE Conf. Comput. Vis. Pattern Recogn.* (*CVPR*), pp.2184–2191 (2012).
[4] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Workshop on Statistical Learning in Computer Vision* (*ECCV*), Vol.1, p.22 (2004).
[5] Farquhar, J., Szedmak, S., Meng, H. and Shawe-Taylor, J.: Improving bag-of-keypoints image categorisation: Generative models and pdf-kernels, Technical report, Univ. of Southampton (2005).
[6] Fraz, M., Edirisinghe, E.A. and Sarfraz, M.S.: Mid-level-Representation Based Lexicon for Vehicle Make and Model Recognition, *2014 22nd International Conference on Pattern Recognition* (*ICPR*), pp.393–398 (2014).
[7] Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P. and Schmid, C.: Aggregating local image descriptors into compact codes, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.34, No.9, pp.1704–1716 (2012).

[8] Kato, T., Takei, W. and Omachi, S.: A Discriminative Metric Learning Algorithm for Face Recognition, *IPSJ Trans. Computer Vision and Applications*, Vol.5, pp.85–89 (2013). Presented at MIRU2013 as Oral Presentation.
[9] Lazebnik, S., Schmid, C. and Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *IEEE Conf. Comput. Vis. Pattern Recogn.* (*CVPR*), Vol.2, pp.2169–2178 (2006).
[10] Perronnin, F., Sanchez, J. and Mensink, T.: Improving the Fisher kernel for large-scale image classification, *Computer Vision–ECCV 2010*, pp.143–156, Springer (2010).
[11] Sánchez, J., Perronnin, F., Mensink, T. and Verbeek, J.: Image classification with the Fisher vector: Theory and practice, *International Journal of Computer Vision*, Vol.105, No.3, pp.222–245 (2013).
[12] Sivic, J. and Zisserman, A.: Efficient Visual Search of Videos Cast as Text Retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.31, No.4, pp.591–606 (2009).
[13] Tanaka, M., Torii, A. and Okutomi, M.: Fisher Vector based on Full-covariance Gaussian Mixture Model, *IPSJ Trans. Computer Vision and Applications*, Vol.5, pp.50–54 (2013).
[14] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T. and Gong, Y.: Locality-constrained linear coding for image classification, *IEEE Conf. Comput. Vis. Pattern Recogn.* (*CVPR*), pp.3360–3367 (2010).
[15] Weinberger, K.Q. and Saul, L.K.: Distance Metric Learning for Large Margin Nearest Neighbor Classification, *J. Mach. Learn. Res.*, Vol.10, pp.207–244 (2009).
[16] Winn, J., Criminisi, A. and Minka, T.: Object categorization by learned universal visual dictionary, *IEEE Int'l. Conf. Comp. Vis.* (*ICCV*), Vol.2, pp.1800–1807 (2005).
[17] Yang, J., Yu, K., Gong, Y. and Huang, T.: Linear spatial pyramid matching using sparse coding for image classification, *IEEE Conf. Comput. Vis. Pattern Recogn.* (*CVPR*), pp.1794–1801 (2009).
[18] Zhou, X., Yu, K., Zhang, T. and Huang, T.S.: Image classification using super-vector coding of local image descriptors, *Computer Vision–ECCV 2010*, pp.141–154, Springer (2010).

(Communicated by *Tatsuya Harada*)