IPSJ Transactions on Computer Vision and Applications Vol.6 93–97 (July 2014)

[DOI: 10.2197/ipsjtcva.6.93]

Express Paper

Image Classification Using a Mixture of Subspace Models

Takashi Takahashi[1,a] Takio Kurita[2]

Received: January 21, 2014, Accepted: April 24, 2014, Released: July 29, 2014

Abstract: This paper introduces a novel method for image classification using local feature descriptors. The method utilizes linear subspaces of local descriptors for characterizing their distribution and extracting image features. The extracted features are transformed into more discriminative features by the linear discriminant analysis and employed for recognizing their categories. Experimental results demonstrate that this method is competitive with the Fisher kernel method in terms of classification accuracy.

Keywords: image classification, object recognition, subspace, principal component analysis, linear discriminant analysis

1. Introduction

This paper addresses the problem of image classification using local feature descriptors. The typical methods encode local descriptors (e.g., SIFT descriptors [1]) of an image into a global image feature and then classify the image feature by a classifier such as support vector machine (SVM). In the well-known bag-of-features (BoF) or bag-of-keypoints approach, for instance, the distribution of local descriptors of an image is summarized into a histogram that counts the occurrence of visual words [2], [3]. This histogram is employed as an image feature for recognizing categories of the image. After the successful application of such BoF approach to image classification, many studies have been devoted to developing image feature encoding methods that achieve higher classification accuracy. The Fisher kernel method [4], [5] and the super-vector coding [6] are representative of such methods.

In this study, we also investigate a image classification method based on the above-mentioned feature encoding approach. We propose a novel method to extract expressive image features, which utilizes a certain similarities between local descriptors and some prototypes of their distribution. For modeling the distribution of local descriptors, we employ probabilistic principal component analysis (probabilistic PCA or PPCA) [7], [8]. Since the PPCA prototype models can be approximated by linear subspace models, we can encode the population of the local descriptors by using the linear subspaces. The extracted features are then transformed by linear discriminant analysis (LDA) in order to obtain discriminative global image features. Classification performance of this method is evaluated through experiments on two standard datasets: PASCAL-VOC2007 [9] and Caltech-256 [10]. In the recent empirical study [11], it is reported that the Fisher kernel method attains the highest classification accuracy. Therefore, we compare the accuracy of our method with that of the Fisher kernel method.

2. Method

2.1 Overview of the Proposed Method

Figure 1 outlines our image classification method. In the same way as the BoF approach, we extract local feature descriptors from an image and assign them to one of K visual words. We employ the standard k-means clustering for generating the codebook of visual words. Then local descriptors of an image are divided into K bags, which are denoted as $X_1, X_2, \ldots, X_K$. For each X_k ($k = 1, 2, \ldots, K$), we compute the feature vector $\boldsymbol{y}_k$ which characterizes the population. Parameters for this feature extraction process are estimated by unsupervised learning. Next, the obtained features are transformed into more discriminative features z_k via LDA. The resulting K features, $z_1, z_2, \ldots, z_K$, are then fused into a single feature vector $\bar{z}$. Finally, this vector is employed as an input for a classifier such as SVM.

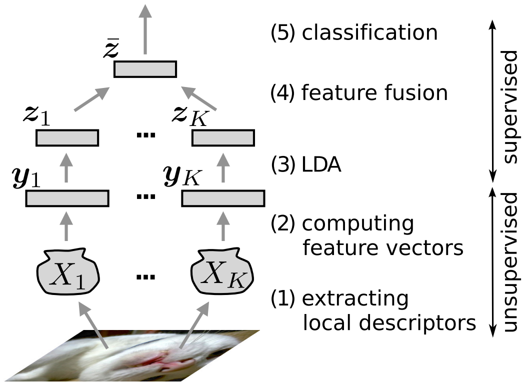

Fig. 1 Overview of the proposed image classification method.

[1] Department of Applied Mathematics and Informatics, Ryukoku University, Otsu, Shiga 520–2194, Japan

[2] Department of Information Engineering, Hiroshima University, Higashihiroshima, Hiroshima 739–8527, Japan

[a] takataka@math.ryukoku.ac.jp

2.2 Feature Extraction from the Local Descriptor Population

Let $\boldsymbol{\xi}_{k,n}$ denote the n-th local descriptor of an image assigned to the k-th visual word and X_k denote the bag consisting of descriptors centered at the cluster centroid $\boldsymbol{c}_k$, that is, $\boldsymbol{x}_{k,n} = \boldsymbol{\xi}_{k,n} - \boldsymbol{c}_k$, where $k = 1, 2, \ldots, K$, $n = 1, 2, \ldots, N_k$ and N_k is the number of descriptors in X_k. The descriptors are assumed to be D-dimensional. Our objective is to extract a feature vector $\boldsymbol{y}_k$ with a fixed dimension from the bag X_k regardless of the number N_k. To this end, let us suppose that each $\boldsymbol{x}_{k,n}$ is generated from some probability distribution. We also assume that we have L "prototypes" of such distribution, P_ℓ ($\ell = 1, 2, \ldots, L$). Then we define a "similarity" $S(X_k, P_\ell)$ between the population of X_k and each of P_ℓ. If the L prototypes are representative enough, the L values $S(X_k, P_1), \ldots, S(X_k, P_L)$ could be used for positioning the distribution of X_k against them. Hence, we make use of these values to characterize the bag X_k.

Provided that each prototype P_ℓ is a Gaussian with mean $\boldsymbol{\mu}_\ell$ and covariance $\boldsymbol{\Sigma}_\ell$, it is possible to define $S(X_k, P_\ell)$ by using the probability density function $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$ as follows:

$$S(X_k, P_\ell) = \frac{1}{N_k} \sum_{n=1}^{N_k} \log \mathcal{N}(\boldsymbol{x}_{k,n}|\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell). \tag{1}$$

The variables $\boldsymbol{x}_{k,n}$ are assumed to be independent each other. Although there are no constraints on the choice of covariance for Eq. (1), it may be unfavorable to adopt the full-covariance model, as it requires considerable computational costs. In this study, therefore, we employ PPCA [7], [8] for modeling each P_ℓ. In this case, under some assumptions including $\boldsymbol{\mu}_\ell = \boldsymbol{0}$ for all ℓ that assumes the mean of each P_ℓ equals to the k-th cluster centroid, $S(X_k, P_\ell)$ can be simplified to the following form (see A.1):

$$\widetilde{S}(X_k, P_\ell) = \frac{1}{N_k} \sum_{n=1}^{N_k} \|\boldsymbol{U}_\ell^\top \boldsymbol{x}_{k,n}\|^2 = \mathrm{tr}(\boldsymbol{U}_\ell^\top \boldsymbol{R}_k \boldsymbol{U}_\ell). \tag{2}$$

Here $\boldsymbol{U}_\ell$ is the $D \times H$ column orthonormal matrix which defines the $H < D$ dimensional subspace for P_ℓ, and $\boldsymbol{R}_k$ is the auto-correlation matrix: $\boldsymbol{R}_k = \sum_{n=1}^{N_k} \boldsymbol{x}_{k,n}\boldsymbol{x}_{k,n}^\top / N_k$. We refer to these subspaces as prototype subspaces. We have adopted such a linear subspace model for the sake of computational efficiency, though more complex models (e.g., an affine subspace model having different means) may show some improvement in terms of classification accuracy. As seen in the next section, this linear subspace model is sufficient for obtaining discriminative image features competitive to Fisher vectors.

The value $\widetilde{S}(X_k, P_\ell)$ measures the similarity between the population of X_k and P_ℓ by the average distance of the descriptors to the prototype subspace defined by $\boldsymbol{U}_\ell$. In this paper, we examine to use the L-dimensional vector consisting of

$$\hat{y}_\ell = \sqrt{\widetilde{S}(X_k, P_\ell)} = \sqrt{\mathrm{tr}(\boldsymbol{U}_\ell^\top \boldsymbol{R}_k \boldsymbol{U}_\ell)} \tag{3}$$

as a feature for image classification. Furthermore, we also investigate another feature composed of the following H-dimensional vectors:

$$\hat{\boldsymbol{y}}_\ell = \left(\sqrt{\boldsymbol{u}_{\ell,1}^\top \boldsymbol{R}_k \boldsymbol{u}_{\ell,1}}, \ldots, \sqrt{\boldsymbol{u}_{\ell,H}^\top \boldsymbol{R}_k \boldsymbol{u}_{\ell,H}} \right)^\top, \tag{4}$$

where $\boldsymbol{u}_{\ell,h}$ is the h-th column of $\boldsymbol{U}_\ell$. It is to be noted that $\|\hat{\boldsymbol{y}}_\ell\|^2 = \widetilde{S}(X_k, P_\ell)$. We obtain the LH-dimensional feature by concatenating $\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \ldots, \hat{\boldsymbol{y}}_L$. In both cases, the resulting L or LH dimensional vector $\boldsymbol{y}$ is normalized so that $\|\boldsymbol{y}\|^2 = 1$.

2.3 Learning of the Prototype Subspaces

We adopt an unsupervised learning approach for constructing representative prototype subspaces from learning images. One candidate method is the EM algorithm for mixture of PPCA [12]; however, its computational costs might be too expensive to apply to large-scale data. Thus, we propose to use a variant of subspace clustering algorithms [13], [14], [15]. The procedure of the proposed algorithm is as follows. Here, the matrix $\boldsymbol{R}_i$ denotes the auto-correlation matrix of the local feature descriptors (assigned to a cluster) of the i-th learning image (the subscript k is omitted).

(1) Initialize the L prototypes $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_L$ so that each of which becomes a $D \times H$ column orthonormal matrix.

(2) Find the cluster index $\ell_i^* \in \{1, 2, \ldots, L\}$ for each $\boldsymbol{R}_i$ as follows:

$$\ell_i^* = \underset{\ell}{\mathrm{argmax}}\, \mathrm{tr}\left(\boldsymbol{U}_\ell^\top \boldsymbol{R}_i \boldsymbol{U}_\ell\right) \tag{5}$$

(3) Update the prototypes so that each matrix $\boldsymbol{U}_\ell$ is composed of the eigenvectors corresponding to the H largest eigenvalues of the following matrix ($\ell = 1, 2, \ldots, L$):

$$\bar{\boldsymbol{R}}_\ell = \sum_{i:\ell_i^*=\ell} \boldsymbol{R}_i \tag{6}$$

(4) Repeat (2) and (3) until the termination condition is met.

Note that this algorithm is distinct from the conventional subspace clustering algorithms, as it clusters not vectors but matrices.

2.4 Feature Fusion via Linear Discriminant Analysis

By applying the feature extraction method described in Section 2.2, an image is represented by K vectors, $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K$. We can simply concatenate these vectors and input the resulting single high dimensional vector to a classifier such as linear SVM. Our preliminary experiments revealed, however, that such concatenated features did not achieve competitive classification performance. The main reason seems to be that the elements of $\boldsymbol{y}_k$ are highly correlated each other because the set $\{\boldsymbol{u}_{\ell,h}\}$ is overcomplete, that is, the number of the basis vectors LH is larger than their dimensionality D in our parameter settings.

In this study, we attempt to solve this problem by applying LDA to the feature vectors in one-vs-rest manner as suggested in Ref. [16]. It is expected that this approach improves the discriminative power of the features. When we have C categories, the direction for discriminating the c-th category ($c = 1, 2, \ldots, C$) from the others is given as

$$\boldsymbol{w}_c = \boldsymbol{\Sigma}^{-1}(\boldsymbol{m}_c - \boldsymbol{m}_{\bar{c}}). \tag{7}$$

The matrix $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\boldsymbol{y}$ (the subscript k is omitted), and the vectors $\boldsymbol{m}_c$ and $\boldsymbol{m}_{\bar{c}}$ denote the means of $\boldsymbol{y}$ belonging to the c-th category and not belonging to the c-th category, respectively. In this approach, it is necessary for estimating $\boldsymbol{\Sigma}$; however, its computation might be numerically unstable due to high data dimensionality. For this reason, we employ the shrinkage covariance estimation method proposed by Ledoit and Wolf [17].

Using the obtained discriminant vectors, we can transform $\boldsymbol{y}_k$ into a C-dimensional vector $\hat{z}_k$ for each $k = 1, \ldots, K$:

$$\hat{z}_k = \boldsymbol{W}_k^\top \boldsymbol{y}_k \qquad (8)$$

where $\boldsymbol{W}_k$ is the matrix composed of the discriminant vectors for $\boldsymbol{y}_k$. Then each element of $\hat{z}_k$ is standardized to have mean 0 and unit variance; furthermore, the standardized vector is L2-normalized. The resulting C-dimensional vector z_k defines the alternative feature of $\boldsymbol{X}_k$. In our experimental conditions, the number of local features N_k is distributed between a wide range including 0. The feature z_k may not be informative if they are extracted from $\boldsymbol{X}_k$ consisting of small number of local features. Thus we define a threshold N_{th} for N_k, and set z_k to be $\mathbf{0}$ if $N_k < N_{\text{th}}$.

Now we have K features $z_1, z_2, \ldots, z_K$ for each image. They are fused into one vector that is employed as input data to a classifier. By virtue of LDA, we can adopt a simple averaging scheme for this purpose:

$$\bar{z} = \frac{1}{K} \sum_{k=1}^{K} z_k. \qquad (9)$$

As a result, each image is represented by a single C-dimensional feature vector $\bar{z}$. It is noteworthy that the dimensionality of this feature is considerably smaller than those of the conventional methods (e.g., the Fisher Kernel method typically employs $2KD$-dimensional feature vector).

3. Experiment

We evaluate the performance of the proposed image classification method. The results are compared with those obtained by the Fisher kernel method [5].

3.1 Experimental Procedure

We conducted the experiments on the following two datasets.

PASCAL-VOC2007 [9]: This dataset consists of 9,963 images of 20 different object categories. According to the prespecified partition, 5,011 images (trainval) were used for learning and 4,952 images (test) were used for testing. The classification performance was measured by the mean of average precision (mAP) across the 20 categories.

Caltech-256 [10]: This dataset contains 30,607 images of 256 object categories and one background (clutter) category. According to the standard experimental protocol for this dataset, the images of each category (excluding the background category) were randomly divided into two sets: 30 images for learning and the remainings for testing. The classification performance was evaluated using the average of the correct classification rate for each category. We repeated each run 5 times with different learning and test splits.

For describing local image features, we used only the 128-dimensional SIFT descriptor [1]; hence, $D = 128$. The descriptors were computed at dense spatial grid points with 4-pixel spacing in three scales, 16, 24, and 32 pixels. We used the VLFeat library [18]. It must be noted that we did not adopt the spatial pooling technique [19] in all our experiments, though it is applicable to our method.

The descriptors extracted from learning images were employed for training our classification system. First, K visual words were acquired by using the standard k-means clustering. Next, the D-dimensional descriptors were transformed into D'-dimensional vectors by the whitening transformation with dimensionality reduction for each cluster. The transformation matrices were computed by applying PCA to each clustered descriptors. We substituted these D'-dimensional features for raw descriptors in the subsequent processes. We have chosen K and D' to be 32 and 64, respectively. The prototype subspaces were learned by applying the proposed algorithm to each of K clustered data. The learning was terminated after 10 iterations. In order to investigate the influence of the choice of the number of prototypes L and the dimensionality of the prototypes H, we repeated the experiments with different settings for these parameters. The threshold N_{th} was set to $D'/2 = 32$. The feature vectors $\bar{z}$ were classified by linear SVM [20], [21], [22]. The soft margin parameters were chosen via 5-fold cross validation.

We compare the classification performance of the proposed method with the Fisher kernel method. In the case of the Fisher kernel method, following to the setup described in Ref. [5], PCA was first applied to reduce the dimensionality of the descriptors to D'. Then, these D'-dimensional data were used for fitting the diagonal-covariance GMM with K mixture components. In addition, the power normalization with $\alpha = 0.5$ and the L2-normalization were also applied to the resulting Fisher vectors. The parameters D' and K were set to 64 and 256, respectively.

3.2 Results

Table 1 shows the classification accuracy of the proposed method and the Fisher Kernel method. The rows of the table labeled as "L-dim." display the values of the mAP (VOC2007) and the correct recognition rate (Caltech-256) obtained by the proposed method using the feature vectors computed from Eq. (3), while the rows labeled as "LH-dim." display those obtained by using the feature vectors computed from Eq. (4). In the former

Table 1 Classification accuracy on PASCAL-VOC2007 and Caltech-256 datasets. FK: Fisher Kernel method.

method	K	L	H	VOC2007	Caltech-256
proposed (L-dim.)	32	128	1	53.0	38.5 ±0.19
			2	53.4	39.5 ±0.46
			4	53.2	39.9 ±0.25
			8	52.9	39.6 ±0.44
		256	1	55.5	40.8 ±0.38
			2	56.0	41.3 ±0.51
			4	55.9	41.5 ±0.47
			8	55.5	41.4 ±0.49
proposed (LH-dim.)	32	128	1	53.0	38.5 ±0.19
			2	55.3	41.3 ±0.41
			4	57.6	42.4 ±0.50
			8	58.8	42.8 ±0.42
		256	1	55.5	40.8 ±0.38
			2	57.6	42.3 ±0.33
			4	59.2	**42.9** ±0.24
			8	**59.8**	42.8 ±0.42
FK	256	–	–	57.1	39.7 ±0.17

case little or no accuracy improvement is seen as the subspace dimensionality H is increased, whereas significant improvement is observed in the latter case. We can also see that the accuracy values are about the same with the identical feature dimensionality LH. The proposed method attains the highest accuracy when $K = 32$, $L = 256$, and $H = 4$ or 8. In these conditions, the sum of the dimensions of $\boldsymbol{y}_k$, KLH, is equal to 32,768 or 65,536, which is comparable to the dimension of the Fisher vector, $2KD' = 32{,}768$. These results imply that the proposed method can perform more efficient feature encoding in comparison to the Fisher kernel method.

4. Discussion

The above experimental results suggest that the classification accuracy of the proposed image classification method is comparable to that of the well-established state-of-the-art methods. In order to confirm the validity of our method, however, it is necessary to conduct more extensive experiments employing various datasets and to compare with some more recent studies, for instance, Refs. [23], [24]. It is also an open question whether this method is efficient in terms of the computational cost, although the computational complexity of the proposed method appears not to be high since almost all operations are arithmetic.

On the other hand, the proposed feature extraction method is applicable to other recognition problems in which data is composed of a set of multiple features. In future, we will investigate how to apply this method to such recognition problems as face recognition from image sequence [25], [26].

References

[1] Lowe, D.G.: Distinctive image features from scale-invariant key points, *Int'l. J. Computer Vision*, Vol.60, No.2, pp.91–110 (2004).

[2] Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *Proc. ICCV '03*, Vol.2, pp.1470–1477 (2003).

[3] Csurka, G., Dance, C.R., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *In Workshop on Statistical Learning in Computer Vision*, *ECCV*, pp.1–22 (2004).

[4] Perronnin, F. and Dance, C.: Fisher Kernels on Visual Vocabularies for Image Categorization, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*) (2007).

[5] Perronnin, F., Sanchezz, J. and Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification, *Proc. European Conference on Computer Vision* (*ECCV*), pp.143–156 (2010).

[6] Zhou, X., Yu, K., Zhang, T. and Huang, T.S.: Image Classification using Super-Vector Coding of Local Image Descriptors, *Proc. European Conference on Computer Vision* (*ECCV*) (2010).

[7] Tipping, M.E. and Bishop, C.M.: Probabilistic Principal Component Analysis, *Journal of the Royal Statistical Society, Series B*, Vol.21, pp.611–622 (1999).

[8] Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer (2006).

[9] Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J. and Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, available from ⟨http://www.pascal-network.org/challenges/VOC/voc2007/⟩.

[10] Griffin, G., Holub, A. and Perona, P.: Caltech-256 Object Category Dataset, California Institute of Technology, available from ⟨http://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001⟩.

[11] 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* (*BMVC*) (2011).

[12] Tipping, M.E. and Bishop, C.M.: Mixtures of Probabilistic Principal Component Analyzers, *Neural Computation*, Vol.11, No.2, pp.443–482 (1999).

[13] Kambhatla, N. and Leen, T.K.: Dimension Reduction by Local Principal Component Analysis, *Neural Computation*, Vol.9, No.7, pp.1493–1516 (1997).

[14] Ho, J., Yang, M.-H., Lim, J., Lee, K.-C. and Kriegman, D.: Clustering appearances of objects under varying illumination conditions, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), pp.11–18 (2003).

[15] Vidal, R.: Subspace Clustering, *IEEE Signal Processing Magazine*, Vol.28, pp.52–68 (2011).

[16] Hariharan, B., Malik, J. and Ramanan, D.: Discriminative Decorrelation for Clustering and Classification, *Proc. European Conference on Computer Vision* (*ECCV*) (2012).

[17] Ledoit, O. and Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis*, Vol.88, No.2, pp.365–411 (2004).

[18] Vedaldi, A. and Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008), available from ⟨http://www.vlfeat.org/⟩.

[19] Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*) (2006).

[20] Vapnik, V.: *Statistical Learning Theory*, Wiley (1998).

[21] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research*, Vol.9, pp.1871–1874 (2008).

[22] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Trans. Intelligent Systems and Technology*, Vol.2, pp.27:1–27:27 (2011).

[23] Kobayashi, T.: BoF meets HOG: Feature Extraction based on Histograms of Oriented p.d.f. Gradients for Image Classification, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*) (2013).

[24] Harada, T. and Kuniyoshi, Y.: Graphical Gaussian Vector for Image Categorization, *Neural Info. Proc. Systems* (*NIPS*) (2012).

[25] Yamaguchi, O., Fukui, K. and Maeda, K.: Face Recognition Using Temporal Image Sequence, *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, pp.318–323 (1998).

[26] Sakano, H. and Mukawa, N.: Kernel mutual subspace method for robust facial image recognition, *Proc. 4th Int'l. Conf. Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, Vol.1, pp.245–248 (2000).

Appendix

A.1 Derivation of the Similarity $\widetilde{S}(X_k, P_\ell)$

In the PPCA model, the covariance matrix $\boldsymbol{\Sigma}$ is given as $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}$, where $\boldsymbol{W}$ is a $D \times H$ matrix ($H < D$) and $\sigma^2 > 0$ (subscripts are omitted). If we estimate the parameters of PPCA model by maximum likelihood approach, the optimal solution for $\boldsymbol{\mu}$ is simply the sample mean, while those of $\boldsymbol{W}$ and σ^2 are known to have the following forms [7], [8]:

$$\boldsymbol{W} = \boldsymbol{U}(\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I})^{\frac{1}{2}}\boldsymbol{R} \tag{A.1}$$

$$\sigma^2 = \frac{1}{D-H}\sum_{d=H+1}^{D}\lambda_d, \tag{A.2}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$ are the eigenvalues of the sample covariance matrix, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_H)$, $\boldsymbol{U}$ is the $D \times H$ matrix composed of the H eigenvectors corresponding to the H largest eigenvalues, and $\boldsymbol{R}$ is an arbitrary $H \times H$ orthogonal matrix. The model parameters do not depend on the choice of $\boldsymbol{R}$; therefore, we assume that $\boldsymbol{R} = \boldsymbol{I}$.

When Eqs. (A.1) and (A.2) hold, $\mathcal{L} = \log \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written as follows:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{2}\left(D\log 2\pi + \log|\boldsymbol{\Sigma}|\right)\\ &-\frac{1}{2\sigma^2}\left(\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 - \|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{W}^\top(\boldsymbol{x}-\boldsymbol{\mu})\|^2\right).\end{aligned} \tag{A.3}$$

Here let us assume that

- $\boldsymbol{\mu} = \boldsymbol{0}$,

- $|\boldsymbol{\Sigma}|$ and σ^2 take identical values, respectively, for every prototype,
- $\boldsymbol{W}\boldsymbol{\Lambda}^{-\frac{1}{2}} = \boldsymbol{U}\mathrm{diag}(\sqrt{1-\sigma^2/\lambda_h}) \approx \boldsymbol{U}$.

Then,

$$\mathcal{L} = a(\|\boldsymbol{U}_\ell^\top \boldsymbol{x}\|^2 - \|\boldsymbol{x}\|^2 + b) \tag{A.4}$$

approximately holds for any $\ell = 1, 2, \ldots, L$, where a and b are common positive constants. The term $\|\boldsymbol{x}\|^2$ can also be omitted since it takes an equal value for every prototype. Hence, by omitting the constant terms, we can derive the approximated similarity $\widetilde{S}(X_k, P_\ell)$ as Eq. (2).

(Communicated by *Mark S. Nixon*)