Discriminative Embedding via Image-to-Class Distances

Xiantong Zhen
zhenxt@gmail.com

Ling Shao
ling.shao@ieee.org

Feng Zheng
cip12fz@sheffield.ac.uk

Department of Medical Biophysics
The University of Western Ontario
London, ON, Canada

Department of Electronic and Electrical Engineering
The University of Sheffield

Department of Electronic and Electrical Engineering
The University of Sheffield

Image-to-Class (I2C) distance firstly proposed in the naive Bayes nearest neighbour (NBNN) classifier [1, 5, 6] has shown its effectiveness in image classification. However, due to the large number of nearest-neighbour search, I2C-based methods are extremely time-consuming, especially with high-dimensional local features. In this paper, with the aim to improve and speed up I2C-based methods, we propose a novel discriminative embedding method based on I2C for local feature dimensionality reduction. Our method **1)** greatly reduces the computational burden and improves the performance of I2C-based methods after reduction; **2)** can well preserve the discriminative ability of local features, thanks to the use of I2C distances; and **3)** provides an efficient closed-form solution by formulating the objective function as an eigenvector decomposition problem. We apply the proposed method to action recognition showing that it can significantly improve I2C-based classifiers.

We incorporate the I2C distance to propose a novel dimensionality reduction method to embed high-dimensional local features into a discriminative low-dimensional space. The use of the I2C distance benefits in two aspects. On the one hand, local features from one image are treated as a whole and class labels can be directly used for supervised learning. This increases the discriminative capacity of local features. On the other hand, it provides an intuitive and effective venue to couple local feature reduction with classification, which can improve the performance of classification. In the low-dimensional space, local features from each image are aligned according to the I2C distances and the I2C distance to its own class is minimized and the I2C distances to other classes are maximized.

Our work contributes in the following aspects: **1)** a novel discriminative subspace learning algorithm based on the I2C distances is proposed for the dimensionality reduction of local features; **2)** after embedding, I2C-based methods are remarkably speeded up and scale well with a large number of local features and therefore become more attractive in real-world applications; and **3)** we formulate the method as an eigenvector decomposition problem, which is efficient with a closed-form solution.

The image-to-class (I2C) distance was first defined in the naive Bayes nearest neighbour (NBNN) classifier. NBNN is an approximation of the optimal MAP naive-Bayes classifier under some assumptions. Given an image Q represented as a set of local features, $\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots,\mathbf{x}_N$, where $\mathbf{x}_i \in R^D$ and D is the dimensionality of local features. The summation of all the distances from the local features of an image to their corresponding nearest neighbours in each class is defined as the Image-To-Class (I2C) distance, which can be calculated by:

$$D_X^c = \sum_{\mathbf{x}\in X} ||\mathbf{x} - NN^c(\mathbf{x})||^2, \tag{1}$$

where NN^c is the nearest neighbour of $\mathbf{x}$ in class c. The resulting classifier takes the form as:

$$\bar{c} = arg\min_c D_X^c, \tag{2}$$

Our task is to classify a collection of videos $\{X_i\}$, each of which is represented by a set of local features: $\{\mathbf{x}_{i1},\ldots,\mathbf{x}_{ij},\ldots,\mathbf{x}_{im_i}\}$, e.g., HOG3D [4], where m_i is the number of local features from image X_i. Given an image/video X_i, its I2C distance to class c is computed according to Eq. 1 as:

$$D_{X_i}^c = \sum_{j=1}^{m_i} ||\mathbf{x}_{ij} - \mathbf{x}_{ij}^c||^2, \tag{3}$$

where $\mathbf{x}_{ij}^c$ is the nearest neighbour in class c.

We aim to find a linear projection $\mathbf{W} \in R^{D\times d}$ to embed the local features into a lower-dimensional space R^d. Unlike the methods in [3], [2],

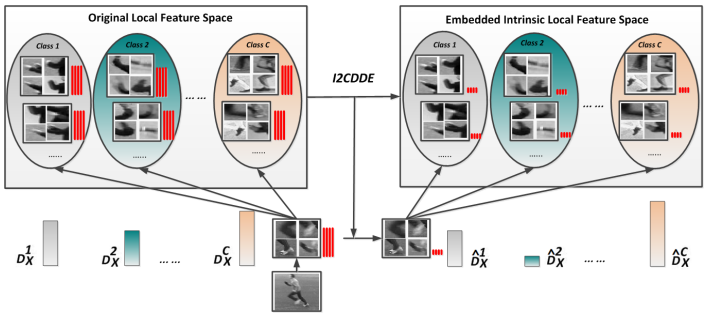

Figure 1: Illustration of the discriminative embedding based on the I2C distance. Action classes are represented by the ellipses in which the rectangles denote local patches from frames (Classes 1, 2 and c represent '*Boxing*', '*Handwaving*' and '*Running*' from the KTH dataset, respectively). The length of the red bars indicates the dimensionality of the local features. The color bars are the I2C distances. D_X^c is the I2C distance from the action X to class c. $\hat{D}_X^c$ is the I2C distance in the embedded space.

our aim in the embedded space is to minimize the I2C distances from images to the classes they belong to while simultaneously maximizing the I2C distances to the classes they do not belong to. The objective function we used takes the form as:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}(\sum_{n=1}^{N_i}\sum_i \Delta X_{in}\Delta X_{in}^{\mathrm{T}})\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}(\sum_i \Delta X_{iP}\Delta X_{iP}^{\mathrm{T}})\mathbf{W})}, \tag{4}$$

where ΔX_{iP} is the auxiliary matrix associated with the class (positive class) that image X_i belongs to and ΔX_{in} is with the class (negative class) that image X_i does not belong to. Note that, given a dataset, the number of negative classes N_i is the same for all images in the dataset.

We can now seek the embedding $\mathbf{W}^*$ to maximize the ratio in Eq. 4. The above equation can be rewritten in terms of covariance matrices as:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_N\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_P\mathbf{W})}, \tag{5}$$

where $\mathbf{C}_N = \sum_{n=1}^{N_i}\sum_i \Delta X_{in}\Delta X_{in}^{\mathrm{T}}$, and $\mathbf{C}_P = \sum_i \Delta X_{iP}\Delta X_{iP}^{\mathrm{T}}$.

It can be seen that maximizing the objective function in Eq. 5 is a well-known eigensystem problem [2]:

$$\mathbf{C}_N\mathbf{W} = \lambda\mathbf{C}_P\mathbf{W} \tag{6}$$

The linear projection is composed of d eigenvectors corresponding to the d largest eigenvalues $\lambda_1,\ldots,\lambda_d$. The whole procedure of the embedding is illustrated in Fig 1.

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[2] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *IEEE TPAMI*, 33(2):338–352, 2011.

[3] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. In *ICCV*, 2007.

[4] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[5] S. McCann and D.G. Lowe. Local naive bayes nearest neighbor for image classification. In *CVPR*, 2012.

[6] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The nbnn kernel. In *ICCV*, 2011.