Discriminative Embedding via Image-to-Class Distances

Xiantong Zhen
zhenxt@gmail.com
Department of Medical Biophysics
The University of Western Ontario
London, ON, Canada

Ling Shao
ling.shao@ieee.org
Department of Electronic and Electrical Engineering
The University of Sheffield,
Sheffield, United Kingdom

Feng Zheng
cip12fz@sheffield.ac.uk
Department of Electronic and Electrical Engineering
The University of Sheffield,
Sheffield, United Kingdom

Abstract

Image-to-Class (I2C) distance firstly proposed in the naive Bayes nearest neighbour (NBNN) classifier 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.

1 Introduction

Local features play a key role in visual recognition, e.g., action recognition. Classification based on local features is still a challenging task due to the large intra-class variance and noisy local features. Widely-used local feature descriptors including SIFT [16], HOG3D [11] and HOG/HOF [13] have shown their effectiveness for both image classification and action recognition. The discriminative ability of local features would greatly influence the performance of later representation and classification [25]. In the last decade, algorithms such as the bag-of-words (BoW) model and sparse coding have been extensively exploited to encode local features as a global representation. The fact is that even images/actions

belonging to the same class would contain quite a large proportion of dissimilar local features, which enlarges the intra-class variance and makes directly comparing local features not optimal for classification.

Recently, a non-parametric approach named naive Bayes nearest neighbour (NBNN) [3] was proposed for image classification, in which the image-to-class (I2C) distance is introduced. Being conceptually simple, NBNN has achieved state-of-the-art performance even comparable with other sophisticated learning algorithms. The success of NBNN is accredited to the employment of the I2C distance, which has been proven to be the optimal distance to use in image classification [3]. It is the I2C distance that effectively deals with the huge intra-class variance of local features. Inspired by the idea of I2C distances, Zhang et al. [27] proposed object-to-class (O2C) distances obtaining state-of-the-art performance for scene classification.

However, the performance of the I2C-based methods highly depends on the effectiveness of local features, because they essentially contribute to the calculation of the I2C distance. The I2C-based methods will be computationally expensive or even intractable with a huge number of local features, especially when the local features are high-dimensional. In addition, the discriminative ability of local features will directly affect the performance of the I2C distance. For instance, the local features with noise or from backgrounds would degenerate the performance of I2C for classification. Therefore, finding a low-dimensional but discriminative space to represent the local features becomes very attractive, especially for action recognition, in which the local features typically amount to tens of thousands and are very high-dimensional.

Dimensionality reduction techniques such as principal component analysis (PCA) can be used to project the features into a low-dimensional space, which has been exploited in [5],[10] for image classification and action recognition. Unfortunately, PCA is an unsupervised feature reduction method treating each local feature equally without considering the label information of images and therefore suffers from being less discriminative in the low-dimensional space. Unsupervised nonlinear dimensionality reduction (manifold learning) methods such as Locally Linear Embedding (LLE) [19], ISOMAP [23], Hessian eigenmaps (HLLE) [6] and Laplacian Eigenmap (LE) [2] suffer from a crucial limitation that the embedding does not generalize well from training to test data. Linearization is a procedure commonly used to construct explicit maps over new samples, e.g., locality preserving projections (LPP) [7] and neighbourhood preserving embedding (NPE) [8].

In addition, some local features could be visually similar or shared by images in different classes which would be misleading for classification. Therefore, the use of conventional discriminative reduction techniques, e.g., linear discriminative analysis (LDA), is suboptimal because LDA, when applied to local features, attempts to minimize the within class variance of different local features and maximize the between-class variance of different local features together. To address the shortcomings of LDA, Sugiyama [22] proposed local Fisher discriminant analysis (LFDA) by combining the ideas of LDA and LPP, which, however, is still not optimal for local feature reduction.

In this paper, 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.

We validate the proposed method for action recognition because the local feature descriptors for actions are always rather lengthy with several hundred even thousand dimensions, e.g., HOG3D [11]. To the best of our knowledge, this is the first work to address the local, discriminative feature reduction for action recognition.

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 remainder of this paper is organized as follows. We review and discuss the related work in Section 2. The details of the proposed method are described in Section 3. We show experiments and results in Section 4 and conclude in Section 5.

2 Related work

The image-to-class (I2C) distance was first introduced by Bioman et al. [3] in the naive Bayes nearest neighbour (NBNN) classifier. NBNN is a non-parametric algorithm for image classification based on local features. With the naive Bayes assumption, NBNN is dramatically simple, while in contrast to parametric learning algorithms, NBNN enjoys many attractive advantages. It requires no training stage and can naturally deal with a huge number of classes. Due to the use of the I2C distance calculated on original local features, NBNN can get rid of descriptor quantization errors. The core of NBNN is the approximation of the log-likelihood of a local feature by the distance to its nearest neighbour, which brings about the image-to-class (I2C) distance. Taking advantage of the I2C distance, several variants of NBNN have been proposed in the past few years to improve the generalization ability of NBNN.

In NBNN, local features are assumed to be i.i.d. given the class labels and the probability density is estimated by the non-parametric Parzen kernel function and can be further approximated by the nearest neighbour under the assumption that the normalization factor in the kernel function is class-independent. However, this assumption is too strict and restricts its generalization on multiple features. Towards an optimal NBNN by relaxing the assumption, Behmo et al. [1] addressed this problem by learning parameters specific to each class via hinge-loss minimization. The optimal NBNN demonstrates good generalization on combining multiple feature channels.

A kernelized version of NBNN, termed the NBNN kernel, was introduced by Tuytelaars et al. [24]. It was shown in their work that the NBNN kernel is complementary to the bag-of-features kernel. By preserving the core idea of the NBNN algorithm, for each image, the I2C distances to all classes are computed. Instead of directly classifying the image as the class with the minimum I2C distance, they concatenated all the I2C distances as a vector, which can be regarded as a high-level image representation. A linear support vector machine (SVM) is employed for image classification. The success of the NBNN kernel is largely attributed to the discriminative representation of an image by the I2C distances to its own class but also to classes it does not belong to. This representation gains more discriminative information in contrast to directly using the absolute I2C distance measurement.

Recently, McCann and Lowe [17] developed an improved version of NBNN, named local naive Bayes nearest neighbour (LNBNN), which increases the classification accuracy and scales better with a larger number of classes. The motivation of local NBNN stems from the observation that only the classes represented in the local neighbourhood of a descriptor contribute significantly and reliably to their posterior probability estimation. Specifically, instead of finding the nearest neighbour in each of the classes, local NBNN finds in the local neighbourhood k nearest neighbours which may only come from some of the classes. The "localized" idea is shared with localized soft assignment coding (LSC) [15] in the BoW model and locality-constrained linear coding (LLC) [26] in sparse coding.

A pooled NBNN kernel has also been introduced by Rematas et al. [18] to enhance the NBNN kernel. NBNN can be regarded as performing max pooling (finding the nearest neighbour) over the receptive field in the feature space associated with each class, which leads to the image-to-class (I2C) distance. Based on this understanding, they generalized the max pooling in NBNN to propose the image-to-subclass and image-to-word distances, which improves both the image-to-image and image-to-class baselines.

In terms of local feature reduction, our method is closely related to the work in [10], [4], [9]. PCA-SIFT by Ke et al. [10] is the first attempt to address the dimensionality reduction for local features. PCA was applied to project the gradient image vector of a patch to obtain a more compact feature vector, which is significantly shorter than the standard SIFT descriptor. Discriminative local feature reduction has been explored in [9] and [4], both of which use the same covariance matrices of pairwise matched distances and pairwise unmatched feature distances to find the linear projection. It is demonstrated in [4] that the projection directions are the same in their methods, although the approaches used are different.

3 I2C distance-based discriminative embedding

We first revisit the image-to-class (I2C) distance based on which our algorithm is built, and then describe the proposed I2C-based discriminative embedding (I2CDDE) in details.

3.1 Revisit of I2C distances

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. Taking the assumption that the class prior $p(C)$ is uniform, MAP can be simplified as the maximum likelihood (ML) classifier:

$$\hat{C} = arg\max_C p(C|Q) = arg\max_C p(Q|C). \tag{1}$$

Under the naive-Bayes assumption that $\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots,\mathbf{x}_N$ are i.i.d. given the class C, we have:

$$p(Q|C) = p(\mathbf{x}_1,\ldots,\mathbf{x}_N|C) = \prod_{i=i}^{N} p(\mathbf{x}_i|C), \tag{2}$$

where $p(\mathbf{x}_i|C)$ can be approximated using the non-parametric Parzen density estimation.

The Parzen likelihood estimation of the probability of $\mathbf{x}$ from class C is:

$$\hat{p}(\mathbf{x}|C) = \frac{1}{L}\sum_{j=1}^{L} K(\mathbf{x}-\mathbf{x}_j^C), \tag{3}$$

where L is the number of local features from class C.

By further assuming that the kernel bandwidths in the Parzen function are the same for all the classes, the likelihood can be simplified using the nearest neighbour. 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:

$$D_X^c = \sum_{\mathbf{x}\in X} ||\mathbf{x} - NN^c(\mathbf{x})||^2, \tag{4}$$

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

$$\bar{c} = arg\min_c D_X^c, \tag{5}$$

3.2 Discriminative embedding

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

$$D_{X_i}^c = \sum_{j=1}^{m_i} ||\mathbf{x}_{ij} - \mathbf{x}_{ij}^c||^2, \tag{6}$$

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.

Proposition. *Define an auxiliary matrix ΔX_{ic} as:*

$$\Delta X_{ic} = (\Delta \boldsymbol{x}_{i1}^c,\ldots,\Delta \boldsymbol{x}_{ij}^c,\ldots,\Delta \boldsymbol{x}_{im_i}^c), \tag{7}$$

where $\Delta \boldsymbol{x}_{ij}^c = \boldsymbol{x}_{ij} - \boldsymbol{x}_{ij}^c$, therefore the I2C distance in the low dimensional space projected by $\boldsymbol{W}$ becomes:

$$\hat{D}_{X_i}^c = \mathrm{Tr}(\boldsymbol{W}^{\mathrm{T}}\Delta X_{ic}\Delta X_{ic}^{\mathrm{T}}\boldsymbol{W}), \tag{8}$$

The proof of this proposition will be given in the Appendix.

Unlike the methods in [9], [4], 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:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\sum_{n=1}^{N_i}\sum_i \mathbf{W}^{\mathrm{T}}\Delta X_{in}\Delta X_{in}^{\mathrm{T}}\mathbf{W})}{\mathrm{Tr}(\sum_i \mathbf{W}^{\mathrm{T}}\Delta X_{iP}\Delta X_{iP}^{\mathrm{T}}\mathbf{W})} = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}(\sum_{n=1}^{N_i}\sum_i \Delta X_{in}\Delta X_{in}^{\mathrm{T}})\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}(\sum_i \Delta X_{iP}\Delta X_{iP}^{\mathrm{T}})\mathbf{W})}, \tag{9}$$

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.

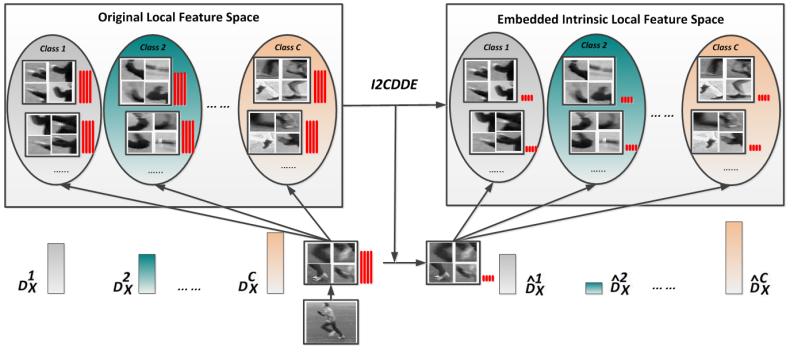

Figure 1: Illustration of the discriminative embedding based on the I2C distance. Action classes are represented by the ellipses in which the rectangles denote local patches from frames (Classes 1, 2 and c represent '*Boxing*', '*Handwaving*' and '*Running*' from the KTH dataset, respectively). The length 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.

We can now seek the embedding $\mathbf{W}^*$ to maximize the ratio in Eq. 9. The above equation can be rewritten in terms of covariance matrices as:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_N\mathbf{W})}{\mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\mathbf{C}_P\mathbf{W})}, \tag{10}$$

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. 10 is a well-known eigensystem problem [4]:

$$\mathbf{C}_N\mathbf{W} = \lambda\mathbf{C}_P\mathbf{W} \tag{11}$$

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.

3.3 Neighbourhood relaxation

Due to the noisy local features, e.g., local features from backgrounds and shared by similar actions, the I2C distance using the nearest neighbour (NN) would not be reliable and the assumption that the nearest neighbor is preserved during the projection will be too strict. To relax this, we incorporate locality (using K nearest neighbours) in the objective function, which is to preserve the local structure of features in the reduced space. We will show experimentally that this modification can improve the performance especially on more complex datasets, e.g., HMDB51, in which the backgrounds are quite complicated and local features are extremely noisy. With the neighbourhood relaxation, the $D_{X_i}^c$ in Eq. 6 is replaced by:

$$D_{X_i,K}^c = \sum_{k=1}^{K}\sum_{j=1}^{m_i} ||\mathbf{x}_{ij} - \mathbf{x}_{ij,k}^c||^2, \tag{12}$$

where $\mathbf{x}_{ij,k}^c$ is the k-th nearest neighbour of $\mathbf{x}_{ij}^c$ in the c-th class and K is the number of neighbours. The objective function in Eq. 10 needs also to be updated accordingly.

3.4 Computational complexity

A key deficit in I2C-based methods is the heavy computational burden resulting from the nearest neighbour search, which is extremely expensive especially when local features are high-dimensional. I2CDDE can greatly reduce the computational cost and at the same time even enhance the discriminative ability of local features. At the test stage, the computational complexity in the original space is $\mathcal{O}(NMD^2)$, where N is the number of local features from a test sample, M is the total number of local features in the training set and D is the dimensionality of local features in the original space. After the embedding, the computational complexity is reduced to $\mathcal{O}(NMd^2)$, where d ($d \ll D$) is the dimensionality of local features in the embedded space. Take the local descriptor in action recognition for instance, we use the HOG3D descriptor. The dimensionality in the original space is 1000 while in the embedded space it is only tens of dimensions. The computational complexity in the reduced space is $d^2/D^2 = 10^2/1000^2 = 1/10000$ of that in the original space.

4 Experiments and results

Although our method can be used for both image classification and action recognition, we choose to validate our method for action recognition because local features used in action sequences are of much higher-dimensional than those, e.g., SIFT, in the image domain. We comprehensively evaluate I2CDDE for action recognition. Experiments are conducted on the benchmark KTH dataset, the realistic UCF YouTube and HMDB51 datasets. We compare the performance of I2CDDE with typical dimensionality reduction methods including PCA, LDA, LFDA, LPP and NPE, and also show the improvement of I2C-based methods including NBNN, local NBNN and the NBNN kernel. LDP is not included for comparison due to the unavailability of ground truth (matched and unmatched local features) for action datasets.

4.1 Datasets and settings

The KTH dataset [20] is a commonly used benchmark action dataset with 2391 video clips and six human action classes including walking, jogging, running, boxing, hand waving and hand clapping, performed by 25 subjects. We follow the experimental setup [20].

The UCF YouTube dataset [14] is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination condition. This dataset contains a total of 1168 sequences with 11 action categories. We follow the experimental settings in [14].

The HMDB51 dataset [12] contains 51 distinct categories with at least 101 clips in each for a total of 6766 video clips extracted from a wide range of sources. It is a challenging and realistic dataset for action recognition. As in [21, 28, 29], we test our algorithm on a subset of this dataset, i.e. the general body movements with 19 action categories. We follow the experimental setting in the original work [12] using three training/test splits.

We utilize Dollar's periodic detector [5] to detect spatio-temporal interest points (STIPs) and the three-dimensional histograms of oriented gradients (HOG3D) [11], which is descriptive and relatively compact with 1000 dimensions, is used to describe STIPs.

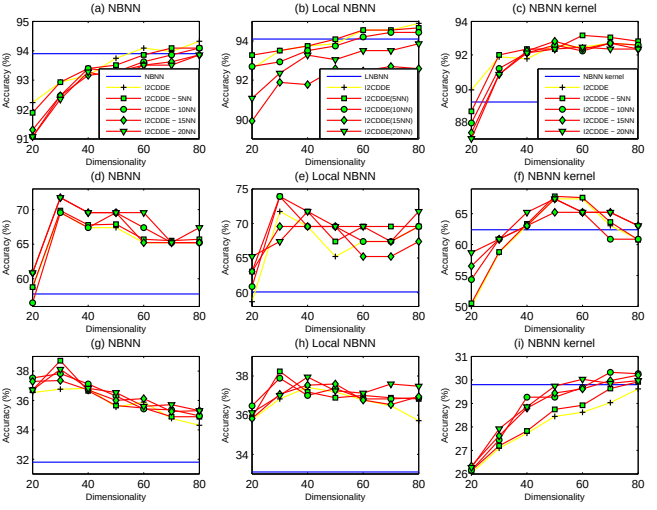

Figure 2: The performance of I2CDDE with different numbers of nearest neighbours on the KTH (the top row), UCF YouTube (the middle row) and HMDB51 (the bottom row) datasets. Blue lines are the baselines of NBNN, local NBNN and the NBNN kernel without dimensionality reduction, and yellow lines are I2CDDE with the nearest neighbour (1NN).

4.2 Results

The performance of I2CDDE for action recognition with different dimensions on the KTH, UCF YouTube and HMDB51 datasets are plotted in Fig. 2. On all the three datasets, we observe that the performance of NBNN, local NBNN and the NBNN kernel has been dramatically improved. On the KTH dataset, the increase on the NBNN kernel is more significant than NBNN and local NBNN, while on the UCF YouTube and HMDB51 datasets, the improvement over NBNN and local NBNN is much more remarkable than that over the NBNN kernel. Note that the superior performance of I2CDDE can be achieved with the local features of less than 60 dimensions, which manifests the effectiveness of I2CDDE for dimensionality reduction of local features.

We have also investigated the effects of different numbers of nearest neighbours on the performance of the neighbourhood embedding. As shown in Fig. 2, on the KTH dataset, the performance of the neighbourhood embedding is comparable with the baseline I2CDDE with the nearest neighbour. On the realistic datasets including UCF YouTube and HMDB51, the benefit of incorporating neighbourhood turns to be more significant, especially on HMDB51. This is expected and reasonable because the KTH is relatively easy with simple actions and clear backgrounds, while HMDB51 contains rather complicated actions and clutters in background. Note that NBNN, local NBNN and the NBNN kernel with neighbourhood embedding are all largely improved over the baseline with the nearest neighbour.

Methods	NBNN	Local NBNN	NBNN Kernel
No Reduction	16.4s	8.4s	22685.3s
Reduction	0.9s	0.6s	365.4s

Table 1: The run time before and after applying I2CDDE (d=30).

		KTH	HMDB51	YouTube
NBNN	I2CDDE	**92.9**	**38.7**	**71.7**
	PCA	91.7	35.6	58.6
	LDA	82.9	31.6	54.3
	LFDA	86.6	29.6	63.1
	LPP	92.8	34.4	56.8
	NPE	91.9	34.8	55.6
	Original	**93.9**	31.8	57.8
LNBNN	I2CDDE	**93.5**	**38.2**	**73.9**
	PCA	91.8	35.7	58.7
	LDA	83.3	31.4	56.5
	LFDA	86.8	28.5	71.7
	LPP	93.3	35.2	60.9
	NPE	92.6	34.9	60.9
	Original	**94.1**	33.1	60.1
NBNN Kernel	I2CDDE	**92.0**	**30.2**	**60.9**
	PCA	89.8	25.8	53.6
	LDA	18.3	13.1	23.9
	LFDA	67.4	10.2	23.9
	LPP	91.0	28.3	58.7
	NPE	91.0	27.9	57.4
	Original	89.2	29.8	**62.4**

Table 2: The comparison of I2CDDE with other reduction methods. Note that the results listed in the table are the accuracies (%) achieved by the methods with **30 dimensions** (except for LDA and LFDA).

4.3 Run time

Since one of the key contributions of I2CDDE is to speed up the I2C-based methods including NBNN, local NBNN and the NBNN kernel, we have compared the run time (in seconds) to classify a test sample before and after using I2CDDE, which is shown in Table 1. The I2C-based methods are dramatically faster after dimensionality reduction. The run time after reduction is calculated by setting reduced dimensionality as 30 for each method and experiments are conducted on the KTH dataset.

4.4 Comparison with other dimension reduction methods

We have also compared I2CDDE with widely used linear dimensionality reduction methods including PCA, LDA, LFDA, LPP and NPE, in Table 2. As expected, I2CDDE uniformly outperforms the compared methods. PCA, LPP and NPE are unsupervised without using the label information and therefore tend to be less discriminative for classification. LDA and LFDA discriminatively learn the projections by labeling the local features with the label of the image that it belongs to, which, however, could mislead the classifier as discussed in

Section 1. We can see that for the NBNN kernel, they even fail to produce reasonable results for all the three datasets. In I2CDDE, the I2C distance actually creates a bridge between the class labels and local features (by using I2C distance), providing an effective and intuitive venue to impose the discriminative information on local features, and therefore can improve the performance of classification.

5 Conclusion

In this paper, we have proposed a method named image-to-class distance-based embedding (I2CDDE) for dimensionality reduction of local features. The experimental results on the KTH, UCF YouTube and HMDB51 datasets have demonstrated that I2CDDE can significantly improve the performance of previously proposed I2C-based methods including NBNN, local NBNN and the NBNN kernel. More importantly, I2CDDE dramatically speeds up these methods, which could boost I2C-based methods for large-scale applications. In addition, I2CDDE uniformly outperforms the classical linear dimensionality reduction techniques such as PCA, LDA, LFDA, LPP and NPE, which further validates the effectiveness of I2CDDE.

6 Appendix

Proof.

$$\begin{aligned}\hat{D}^c_{X_i} &= \sum_{j=1}^{m_i} ||\mathbf{W}^{\mathrm{T}}\mathbf{x}_{ij} - \mathbf{W}^{\mathrm{T}}\mathbf{x}^c_{ij}||^2 = \sum_{j=1}^{m_i} (\mathbf{W}^{\mathrm{T}}\mathbf{x}_{ij} - \mathbf{W}^{\mathrm{T}}\mathbf{x}^c_{ij})^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{x}_{ij} - \mathbf{W}^{\mathrm{T}}\mathbf{x}^c_{ij}) \\ &= \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_{ij} - \mathbf{x}^c_{ij}) = \sum_{j=1}^{m_i} \mathrm{Tr}(\mathbf{W}^{\mathrm{T}}(\mathbf{x}_{ij} - \mathbf{x}^c_{ij})(\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^{\mathrm{T}}\mathbf{W}) \\ &= \mathrm{Tr}(\mathbf{W}^{\mathrm{T}} \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})(\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^{\mathrm{T}}\mathbf{W}). \end{aligned} \tag{13}$$

Substitute ΔX_{ic} into Eq. (13), we have the I2C distance: $\hat{D}^c_{X_i} = \mathrm{Tr}(\mathbf{W}^{\mathrm{T}}\Delta X_{ic}\Delta X_{ic}^{\mathrm{T}}\mathbf{W})$.

□

References

[1] R. Behmo, P. Marcombes, A. Dalalyan, and V. Prinet. Towards optimal naive bayes nearest neighbor. In *ECCV*. 2010.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.

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

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

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[6] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *PNAS*, 100(10):5591–5596, 2003.

[7] X. He and X Niyogi. Locality preserving projections. In *NIPS*, 2004.

[8] X. He, D. Cai, S. Yan, and H. Zhang. Neighborhood preserving embedding. In *ICCV*, 2005.

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

[10] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *CVPR*, 2004.

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

[12] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[14] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos ąřin the wildąś. In *CVPR*, pages 1996–2003, 2009.

[15] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*, 2011.

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

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

[18] K. Rematas, M. Fritz, and T. Tuytelaars. The pooled nbnn kernel: Beyond image-to-class and image-to-image. In *ACCV*, 2012.

[19] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[21] L. Shao, X. Zhen, D. Tao, and X. Li. Spatio-temporal laplacian pyramid coding for action recognition. *IEEE TCYB*, 44(6):817, 2014.

[22] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *JMLR*, 8:1027–1061, 2007.

[23] J. B Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

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

[25] H. Wang, H. Huang, and C. Ding. Discriminant laplacian embedding. In *AAAI*, 2010.

[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[27] L. Zhang, X. Zhen, and L. Shao. Learning object-to-class kernels for scene classification. *IEEE TIP*, 23(8):3241–3253, 2014.

[28] X. Zhen and L. Shao. A local descriptor based on laplacian pyramid coding for action recognition. *Pattern Recognition Letters*, 34(15):1899–1905, 2013.

[29] X. Zhen, L. Shao, and X. Li. Action recognition by spatio-temporal oriented energies. *Information Sciences*, 281(10):295–309, 2014.