Supervised Local Descriptor Learning for Human Action Recognition

Xiantong Zhen, Feng Zheng, Ling Shao, *Senior Member, IEEE*, Xianbin Cao, *Senior Member, IEEE*, and Dan Xu

***Abstract*—Local features have been widely used in computer vision tasks, e.g., human action recognition, but it tends to be an extremely challenging task to deal with large-scale local features of high dimensionality with redundant information. In this paper, we propose a novel fully supervised local descriptor learning algorithm called discriminative embedding method based on the image-to-class distance (I2CDDE) to learn compact but highly discriminative local feature descriptors for more accurate and efficient action recognition. By leveraging the advantages of the I2C distance, the proposed I2CDDE incorporates class labels to enable fully supervised learning of local feature descriptors, which achieves highly discriminative but compact local descriptors. The objective of our I2CDDE is to minimize the I2C distances from samples to their corresponding classes while maximizing the I2C distances to the other classes in the low-dimensional space. To further improve the performance, we propose incorporating a manifold regularization based on the graph Laplacian into the objective function, which can enhance the smoothness of the embedding by extracting the local intrinsic geometrical structure. The proposed I2CDDE for the first time achieves fully supervised learning of local feature descriptors. It significantly improves the performance of I2C-based methods by increasing the discriminative ability of local features while greatly reducing the computational burden by dimensionality reduction to handle large-scale data. We apply the proposed I2CDDE algorithm to human action recognition on four widely used benchmark datasets. The results have shown that I2CDDE can significantly improve I2C-based classifiers and achieves state-of-the-art performance.**

***Index Terms*—Action recognition, dimensionality reduction, image-to-class distance, large scale local features, manifold regularization, naive Bayes nearest neighbor.**

I. Introduction

RECENTLY, local features [1] have shown great effectiveness and achieved state-of-the-art performance for human action recognition [2]. However, visual recognition based on local feature descriptors is still a challenging task due to the large intra-class variance and the existence of noise and redundant information in local features. Moreover, compared to global representation of actions [3], [4], it tends to be computationally very expensive due to the large scale of local features which are usually of high-dimensional. In computer vision tasks, SIFT [5], [6], HOG3D [7] and HoG/HoF [8] are successfully used and have shown their effectiveness in image classification and human action recognition [9], [10], while their discriminative abilities fundamentally underpin the performance for visual recognition. In the last decade, the bag-of-words (BoW) model [11] have been extensively used to encode local features as a global representation. It has been shown in [12] that the BoW model is a special case of match kernels which actually measure the similarity between two images by directly comparing local features from them. The fact is that even images/actions that belong to the same class would contain quite a large proportion of dissimilar local features, which enlarges the intra-class variance and makes it not optimal to directly compare local features for classification.

In order to avoid the quantization errors in the BoW model, recently, a non-parametric approach named naive Bayes nearest neighbor (NBNN) [13] was proposed for image classification. The core idea of the NBNN classifier is the image-to-class (I2C) distance which shows great effectiveness in handling local features. Although it is conceptually simple, the NBNN classifier 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 [13]. The I2C distance can effectively cope with the large intra-class variance of local features and theoretically avoids the quantization errors in on the BoW model. The NBNN has been extended in [14]–[16] achieve local NBNN and NBNN kernels, which have substantially improved the performance of NBBN and achieve great success in image classification. NBNN has recently been combined with deep convolutional neural networks show great effectiveness for scene classification [17], [18].

The discriminative ability and compactness of local feature descriptors will directly affect the performance of those I2C-based methods for recognition tasks in terms of accuracy and efficiency. For instance, local features of less discrimination with noise and redundant information would severely degenerate the performance of I2C for classification. Moreover, the I2C-based methods would be computationally expensive or even

X. Zhen and F. Zheng are with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: zhenxt@gmail.com; zfeng02@gmail.com).

L. Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K. (e-mail: ling.shao@ieee.org).

X. Cao is with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: xbcao@buaa.edu.cn).

D. Xu is with the Department of Information Engineering and Computer Science, The University of Trento, Trento 38122, Italy (e-mail: dan.xu@unitn.it).

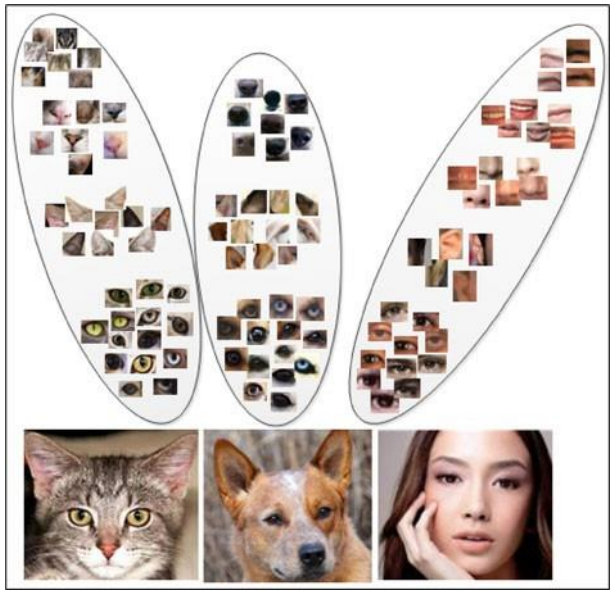


Fig. 1. Toy illustration of local patches from different image categories. The local patches "eyes" from images in different categories can be similar and are close to each other in the feature distribution, while the local patches such as "eyes", "noses", and "ears" are distinctive to each other even though they could be detected from the same image categories.

prohibitive due to the large scale of local features, especially when the local features are high-dimensional, e.g., spatio-temporal local feature descriptors in the video domain for action recognition. It is therefore highly desired and imperative to find a low-dimensional but discriminative space to represent the local features, especially for action recognition [19], in which the local feature descriptors typically amount to tens of thousands and are very high-dimensional. Even more challenges stem from the huge ambiguity of local features. As shown in Fig. 1, some local features in different classes could be visually similar due to the large inter-class ambiguities, which makes it difficult to directly apply existing supervised dimensionality reduction methods, e.g., linear discriminative analysis (LDA). When applied to local features, LDA attempts to minimize the within class variance of different local features and maximize the between-class variance of different local features together.

In this paper, we propose a novel fully supervised dimensionality reduction algorithm called Image-to-Class Distances based Discriminative Embedding (I2CDDE) to embed high-dimensional local features into a discriminative low-dimensional space. By taking advantages of the I2C distance to incorporate class labels, the I2CDDE for the first time achieves fully supervised learning of local feature descriptors. The objective of I2CDDE is, in the low-dimensional space, to minimize the I2C distances of images to classes they belong to while maximizing the I2C distances to the classes they do not belong to. To further improve the performance, we propose incorporating a manifold regularization based on the graph Laplacian [20], [21] into the objective function, which can enhance the smoothness of the embeddings by modeling the local geometrical structure and therefore ensure more robust solutions for better performance.

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 class labels of images for classification, which can improve the performance for visual recognition. 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. The newly incorporated manifold regularization term helps extract the intrinsic structure in the lower dimensional space, which can significantly improve the performance of our I2CDDE algorithm [22], [23].

The proposed I2CDDE algorithm can dramatically improve the performance of methods using local features for classification in terms of both computational efficiency and recognition accuracy. To show the effectiveness of I2CDDE in dealing with high-dimensional local descriptors, 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 [7].

The preliminary idea of this work has been presented in [24]. In this work, we have made new contributions in terms of both theoretical novelty and experimental evaluation. We 1) incorporate the graph based manifold regularization into the objective function, which largely improves the performance; 2) provide a more comprehensive study on the proposed algorithm with more experimental evaluation; and 3) investigate the connection to existing important algorithms based on both image-to-class distances and local descriptor learning showing the advantages of our algorithm.

The major contributions of this work can be summarized in the following three aspects.

1) We propose a novel fully supervised learning algorithm for discriminative local feature descriptor learning, which can not only improve the discriminative ability of local features but also reduce the computational cost;
2) We propose incorporating a manifold regularization to extract intrinsic geometrical structure of local features, which ensures smooth and robust solutions to improve the performance;
3) Our algorithm largely speeds up I2C based methods to scale well with a large number of local features and therefore enables its use in real-world applications

The remainder of this paper is organized as follows. We review and discuss the related work in Section II. The details of the proposed method are described in Section III and its connection to existing methods is given in Section IV. We show experimental results in Section V and conclude in Section VI.

II. Related Work

Local feature learning has been widely used for visual recognition tasks. The compactness and discriminative ability of local features play a crucial role in visual recognition and directly affects the performance and computational efficiency. However,

it is still lack of fully supervised learning algorithms for local feature descriptors in that most of exiting algorithms for local feature learning only concern the similarity/dissimilarity without taking into account the class labels for supervised learning. Supervised descriptor learning has recently generated great popularity in both machine learning [25] and computer vision [26].

The I2C distance has recently been proposed in the naive Bayes nearest neighbor (NBNN) classifier showing great advantages [13] over the BoW model for image classification. NBNN is a non-parametric algorithm for image classification based on local features. With the naive Bayes assumption, NBNN is dramatically simple and enjoys many attractive advantages in contrast to parametric learning algorithms. 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 does not suffer from descriptor quantization errors in the BoW model. The core of NBNN is the approximation of the log-likelihood of a local feature by the distance to its nearest neighbor, 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 the NBNN classifier, local features are assumed to be i.i.d. given its class label and the probability density is estimated by the non-parametric Parzen kernel function and can be further approximated by the nearest neighbor 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.* [27] 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.

By incorporating I2C distance measurement into distance metric learning, Wang *et al.* [28] adopted the idea of large margin from SVM and proposed a method named I2C distance metric learning (I2CDML) to learn a distance metric specific to each class. They formulated a convex optimization problem with the constraint that the I2C distance of each training sample to the class it belongs to should be less than those to other classes by a large margin. However, as a conventional distance metric learning algorithm, I2CDML suffers from a major drawback that the number of parameters to be learned grows quadratically with the dimensionality of the data, which tends to be intractable with high-dimensional data.

By combining the ideas of kernels and I2C distances, the NBNN kernel was introduced by Tuytelaars *et al.* [14] which shows 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.

By introducing locality into NBNN, McCann and Lowe [15] developed an improved version of NBNN, named local naive Bayes nearest neighbor (LNBNN), which increases the classification accuracy and scales better with a larger number of classes. The motivation of local NBNN comes 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 neighbor in each of the classes, local NBNN finds in the local neighborhood k nearest neighbors which may only come from some of the classes. The ”localized” idea is shared by the BoW model [29] and sparse coding [30].

Recently, Rematas *et al.* [31] introduce a pooled NBNN kernel to improve the performance of the NBNN kernel. They show that NBNN can be regarded as performing max pooling (finding the nearest neighbor) over the receptive field in the feature space associated with each class, which leads to the 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.

With regard to local feature descriptor learning, prior work in [32]–[34] made attempts to learn discriminative local descriptors. Ke *et al.* [32] proposed the PCA-SIFT which 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 [34] and [33], 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 [33] that the projection directions are the same in their methods, although the approaches used are different. In addition, both need a huge amount of ground truth with matched/unmatched pairs of local feature descriptors for training, which is not applicable in a realistic setting, especially in the spatio-temporal video domain for action recognition.

Local feature learning can also be obtained by dimensionality reduction. The widely used principal component analysis (PCA) can be adopted for image classification and action recognition [32], [35]. Unfortunately, PCA is an unsupervised feature reduction method without taking into consideration the class label information, which results in less discriminative features for classification. Manifold learning methods such as locally linear embedding (LLE) [36], ISOMAP [37], Hessian eigenmaps (HLLE) [38] and Laplacian eigenmap (LE) [22] suffer from a crucial limitation that the embedding does not generalize well from training to test data, namely, the out-of-sample problem. Locality preserving projections (LPP) [39] and neighborhood preserving embedding (NPE) [40] are linear versions of LP and LLE, respectively, which were developed to handle the

out-of-sample problem. A common limitation of the above reduction algorithms is that the discriminative ability is limited due to the fact that class label information is not used. As a consequence, the obtained local descriptors lose the connection to the ultimate goal of classification.

Very recently, Simonyan *et al.* [41] proposed learning local feature descriptors using convex optimization. In fact, class labels of images are not used in the learning process, which makes the projections lose connection with classification and are therefore suboptimal. Similar to [33], [34], this method also needs a huge amount of matched/unmatched pairs of local feature descriptors for training with adopting the class labels, which is not applicable in a realistic setting, especially in the video domain for human action action.

In general, most of existing local feature learning algorithms are developed on the similarity/dissimilarity of local features, which not only lacks discriminative ability for recognition tasks, but also limits their use in more challenging tasks due to the need of matched/unmatched local feature pairs. To address the above issues, we propose a novel fully supervised learning algorithm to learn more discriminative but compact local descriptors by leveraging the image-to-class distance to incorporate class labels for supervised local descriptor learning.

III. I2C Distance-Based Discriminative Embedding

The proposed I2CDDE algorithm adopts the image-to-class (I2C) distance and for the first time achieves fully supervised learning of local feature descriptors. The I2C distance provides a direct way to connect local feature descriptors with the class labels, which can be used for supervised descriptor learning. We first revisit the I2C distance based on which we describe our discriminative embedding algorithm.

A. Image-to-Class Distance

By avoiding the quantization errors in the BoW models, the image-to-class (I2C) distance was first introduced in the naive Bayes nearest neighbor (NBNN) classifier which has shown great effectiveness in image classification tasks.

Given an image Q, under the assumption that the class prior $p(C)$ is uniform, the maximum-a-posteriori (MAP) classifier can be simplified as the maximum likelihood (ML) classifier

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

Since the image is represented by a set of local feature descriptors $\{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}$ which are assumed to be i.i.d. given the class C, we therefore have

$$p(Q|C) = p(\mathbf{x}_1, \ldots, \mathbf{x}_N|C) = \prod_{i=1}^{N} p(\mathbf{x}_i|C) \tag{2}$$

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

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

where L is the number of local features from class C and $K(\cdot)$ is the Parzen kernel function. Typically, a Gaussian kernel can be adopted as

$$K(\mathbf{x} - \mathbf{x}_j^C) = \exp\left(-\frac{1}{2h}||\mathbf{x} - \mathbf{x}_j^C||^2\right) \tag{4}$$

where h is the width of the kernel K.

Due to the fact the descriptor distribution is long-tailed, the summation in (3) can be accurately estimation be using only the r largest terms corresponding to the r nearest neighbors of $\mathbf{x}$, which results in

$$p(\mathbf{x}|C) = \frac{1}{L}\sum_{j=1}^{r} K(\mathbf{x} - \mathbf{x}_{NN}^j) \tag{5}$$

where $\mathbf{x}_{NN}^j$ is the j-th nearest neighbor of $\mathbf{x}$.

It is shown in [13] that it can achieve accurate approximation by using the single nearest neighbor, namely, $r = 1$. By further assuming that the kernel bandwidths h in the Parzen function are the equal for all the classes, the likelihood can be simplified using the nearest neighbor, and we can now define the image-to-class (I2C) distance which is the summation of all the distances from the local features of an image to their corresponding nearest neighbors in each class as

$$D_X^c = \sum_{\mathbf{x}\in X} ||\mathbf{x} - NN^c(\mathbf{x})||^2 \tag{6}$$

where NN^c is the nearest neighbor of $\mathbf{x}$ in class c. This results in the NBNN classifier which takes the form

$$\bar{c} = \arg\min_c D_X^c. \tag{7}$$

The NBNN classifier is essentially a lazy learning algorithm [42], which just stores the training sample for testing without any training process. We can observe from the derivation of the NBNN classifier, the main computational buren for testing comes from the exhaustive nearest neighbor search, which is time-consuming, especially when local features are huge and in high-dimensional space. While this is common in action recognition when dense trajectory is used for extracting spatio-temporal interest points (STIPs), and the dimension is always high and up to thousands.

Due to the lack of learning stage in NBNN, the performance is highly dependent on the effectiveness of the raw local feature descriptors. In order to achieve compact but discriminative local feature descriptors, we propose to fully supervised learning of local feature descriptors based on the I2C distance, which can significantly improve the performance while reducing the computational cost.

B. Discriminative Embedding

The I2C distance bridges local feature descriptors and the class labels, which can be used to achieve fully supervised learning of local features. In order to enhance the discriminant abilities of local features, we propose supervised local descriptor learning by incorporating the class labels into the learning process via the image-to-class (I2C) distance.

Recall that given an image X_i, its I2C distance to class c is computed according to (6) as

$$D^c_{X_i} = \sum_{j=1}^{m_i} \|\mathbf{x}_{ij} - \mathbf{x}^c_{ij}\|^2 \quad (8)$$

where $\mathbf{x}^c_{ij}$ is the nearest neighbor in class c. More specifically, we would like to seek a linear projection $W \in R^{D\times d}$ to embed the local features into a lower-dimensional space R^d, where the local feature descriptor is more compact but discriminative for classification.

Proposition: Define an auxiliary matrix ΔX_{ic} as

$$\Delta X_{ic} = (\Delta \mathbf{x}^c_{i1}, \ldots, \Delta \mathbf{x}^c_{ij}, \ldots, \Delta \mathbf{x}^c_{im_i}) \quad (9)$$

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

$$\hat{D}^c_{X_i} = \mathrm{Tr}(W^* \Delta X_{ic} \Delta X^*_{ic} W) \quad (10)$$

Proof:

$$\begin{aligned}
\hat{D}^c_{X_i} &= \sum_{j=1}^{m_i} \|W^*\mathbf{x}_{ij} - W^*\mathbf{x}^c_{ij}\|^2 \\
&= \sum_{j=1}^{m_i} (W^*\mathbf{x}_{ij} - W^*\mathbf{x}^c_{ij})^*(W^*\mathbf{x}_{ij} - W^*\mathbf{x}^c_{ij}) \\
&= \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^* WW^* (\mathbf{x}_{ij} - \mathbf{x}^c_{ij}) \\
&= \sum_{j=1}^{m_i} \mathrm{Tr}(W^*(\mathbf{x}_{ij} - \mathbf{x}^c_{ij})(\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^* W) \\
&= \mathrm{Tr}(W^* \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})(\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^* W). \quad (11)
\end{aligned}$$

Substituting ΔX_{ic} into (10), we have the I2C distance

$$\hat{D}^c_{X_i} = \mathrm{Tr}(W^* \Delta X_{ic} \Delta X^*_{ic} W). \quad (12)$$

□

In contrast to the methods in [33], [34] which only concern similarity/dissimilarity of local features without taking into the class label information, our objective 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. As a consequence, the discriminative abilities of local feature descriptors are directly related to the class labels of images from which local features are extracted.

Specifically, the objective function takes the following form:

$$W^* = \arg\max_W \frac{\mathrm{Tr}(W^* \sum_{n=1}^{N_i} \sum_i \Delta X_{in} \Delta X^*_{in} W)}{\mathrm{Tr}(W^* \sum_i \Delta X_{iP} \Delta X^*_{iP} W)}$$

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 W^* to maximize the ratio in (13). The above equation can be rewritten in terms of covariance matrices as

$$W^* = \arg\max_W \frac{\mathrm{Tr}(W^* C_N W)}{\mathrm{Tr}(W^* C_P W)} \quad (14)$$

where

$$\begin{aligned}
C_N &= \sum_{n=1}^{N_i} \sum_i \Delta X_{in} \Delta X^*_{in} \\
C_P &= \sum_i \Delta X_{iP} \Delta X^*_{iP}. \quad (16)
\end{aligned}$$

In practical implementation, due to the fact that local features can be extracted from backgrounds and shared by similar actions of different categories, the I2C distance using the nearest neighbor (NN) would not be always reliable. To make the I2C distance more robust and insensitive to noisy features, we further improve the algorithm by incorporating locality (using r nearest neighbors) in the objective function, which could, to some extent, preserve the local structure of features in the reduced space [43]. With the neighborhood relaxation, the $D^c_{X_i}$ in (8) is replaced by

$$D^c_{X_i,K} = \sum_{k=1}^{r} \sum_{j=1}^{m_i} \|\mathbf{x}_{ij} - \mathbf{x}^c_{ij,k}\|^2 \quad (17)$$

where $\mathbf{x}_{ij,k}$ is the k-th nearest neighbor of $\mathbf{x}_{ij}$ in the c-th class and k is the number of nearest neighbors. The objective function in (14) needs also to be updated accordingly.

C. Manifold Regularization

To further improve the performance of the proposed algorithm, we consider incorporating a manifold regularization [23], [44] based on the graph Laplacian to model the local geometrical structure of data points. It has been shown that learning performance can be significantly enhanced if the geometrical structure is exploited and the local invariance is considered. Many well-known manifold learning algorithms, such as LLE, ISOMAP, and LE, use the so-called locally invariant idea [45], i.e., the nearby points are likely to have similar embeddings, to detect the underlying manifold structure. This will make the solutions more robust to noisy and outlier features and therefore achieve better performance.

A natural and intuitive assumption is that if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the intrinsic geometry of the data distribution, and then their low-dimensional representations $W^*\mathbf{x}_i$ and $W^*\mathbf{x}_j$ with respect to the projection W are also close to

with respect to the projection W are also close to

$$= \arg\max_{W} \frac{\mathrm{Tr}(W^{*}(\sum_{n=1}^{N_i}\sum_{i} \Delta X_{in}\Delta X^{*}_{in})W)}{\mathrm{Tr}(W^{*}(\sum_{i}\Delta X_{iP}\Delta X^{*}_{iP})W)} \quad (13)$$

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 each other. This assumption, also known as local invariance assumption [22], plays a fundamental role in the development of sumption [22], plays a fundamental role in the development of various kinds of algorithms, including dimensionality reduction algorithms [22], [39] and semi-supervised learning algorithms [44], [46]. It has also been shown in spectral graph theory [47]

and manifold learning theory [22] that the local geometric structure can be effectively modeled through a nearest neighbor graph on a scatter of data points [20].

To this end, we first construct a weighted graph $tt = (V, E)$ [39], where V and E respectively represent L vertices and edges between vertices. We denote $A \in \mathbb{R}^{N \times N}$ as the symmetric similarity matrix with non-negative elements corresponding to the edge weight of the graph tt, where each element A_{ij} is computed by a heat kernel with parameter σ as

$$A_{ij} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \tag{18}$$

where $i, j = 1, \ldots, N$. We set the diagonal elements of A to be zeros, i.e., $A_{ii} = 0$.

In the low-dimensional space, we would like to minimize the following term:

$$\sum_{i,j} \|W^*\mathbf{x}_i - W^*\mathbf{x}_j\|^2 A_{ij}. \tag{19}$$

Since the similarity matrix A characterizes the manifold structure of the local feature space, in the lower-dimensional space, $\{W^*\mathbf{x}_i\}_{i=1}^{L}$ preserve the intrinsic local geometrical structure of data distribution. An intuitive consequence of minimizing the regularization term is that, in the low-dimensional space, data points close to each other in the original space are forced to be close while those far away from each other in the original space tend to be far apart.

D. Manifold Regularized Discriminative Embedding

The manifold regularization term in (19) can be rewritten in terms of the graph Laplacian as

$$\mathrm{Tr}(W^*XLX^*W) \tag{20}$$

where $L = D - A$ is the graph Laplacian and D is a diagonal matrix whose entries are column/row sums of A, i.e., $D = \sum_j A_{i,j}$. If we define

$$C_M = XLX^* \tag{21}$$

where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_L]$ is a matrix of the local feature descriptors from the training set, we have

$$\mathrm{Tr}(W^*C_M W). \tag{22}$$

In order not to forego the convenience of solving the problem in (14), we incorporate the manifold regularization term into the denominator of (14) to build our final objective function as follows:

$$W^* = \arg\max_W \frac{\mathrm{Tr}(W^*C_N W)}{\mathrm{Tr}(W^*(C_P + \beta C_M)W)} \tag{23}$$

where $\beta \in (0, \infty)$ is a free parameter which can be experimentally obtained by cross validation.

The objective function in (23) is a trace ratio optimization problem which can be efficiently solved [48].

E. Computational Complexity

The proposed I2CDDE algorithm can significantly reduce the complexity of algorithms based local features. A key deficit in I2C-based methods is the heavy computational burden resulting from the nearest neighbor search, which is extremely expensive especially when local features are high-dimensional. I2CDDE can greatly decrease 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 $O(NMD^2)$, 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 $O(NMd^2)$, where d ($d < D$) is the dimensionality of local features in the embedded space. Taking 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.

IV. Relations to Existing Methods

The proposed I2CDDE algorithm is the first fully supervised dimensionality reduction of local features by explicitly incorporating the class labels in the feature learning process. We provide a description to connect the proposed I2CDDE with existing important methods, which shows the advantages of I2CDDE as a first fully supervised local descriptor learning algorithm.

A. Difference From LDP

Our I2CDDE is closely connected to, while essentially different from linear discriminant projection (LDP) [33], [34], as both address the dimensionality reduction of local features. In LDP, the objective function is to maximize the ratio of the variance of differently labeled points (unmatched points) to that of same-labeled points (matched points). The matched and unmatched features vary with different applications. For instance, in image/object classification, matched features could be the points on the objects that are visually similar [33]. Our I2CDDE is fundamentally different from LDP in multiple asepects.

1) LDP deals with the relationship between local features instead of images, which does not secure the discriminative ability of local features for classification due to the loss of link to class labels.
2) I2CDDE treats local features from each image/sequence as a whole and copes with the relationship between images/videos and classes. By differentiating the I2C distances to the same class and to different classes, I2CDDE makes the local features globally discriminative on an image/video level and can naturally benefit classification.
3) LDP requires extra ground truth of matched and unmatch local features, which are hard to obtain for spatio-temporal local features in action recognition. I2CDDE directly used

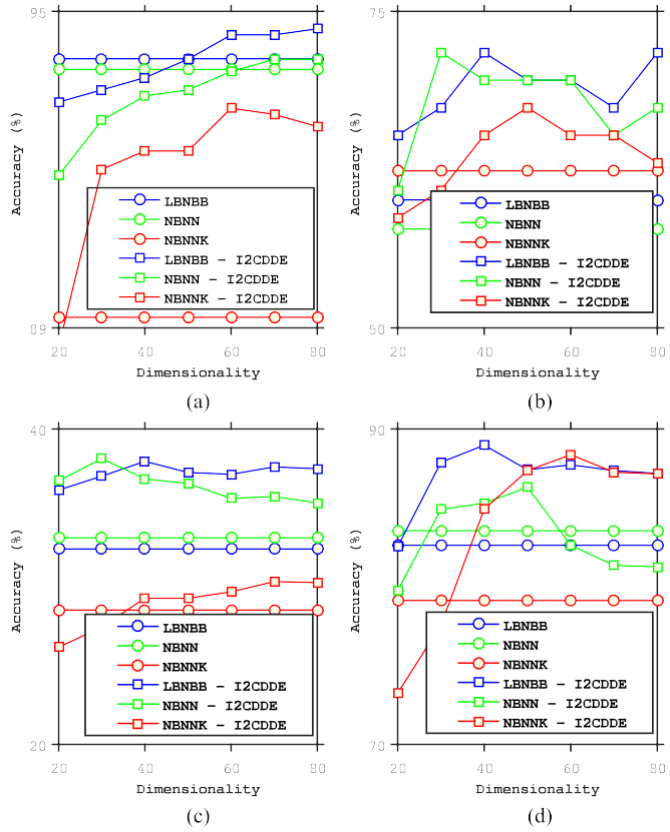


Fig. 2. Performance of NBNN (green), local NBNN (blue), and the NBNN kernel (red) with different dimensions on the four datasets. Lines with Q and S denote the performance before and after dimensionality reduction by I2CDDE. (a) KTH, (b) UCF YouTube, (c) HMDB51, and (d) UCF 101.

image class labels and is therefore more applicable for local feature reduction.

4) I2CDDE takes into consideration the intrinsic structure of local features by imposing a manifold regularization, which provides more smooth and robust solutions.

B. Difference From I2CDML

In the image-to-class distance metric learning (I2CDML) algorithm [28], the squared Euclidean distance in (8) is replaced with the parametric Mahalanobis distance which is to be learned. The I2C distance becomes

$$D^c_{X_i} = \sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^* M_c (\mathbf{x}_{ij} - \mathbf{x}^c_{ij}) \qquad (24)$$

where M_c is the distance metric learned in [28].

As shown in [49], the Mahalanobis distance metric learning can be considered as learning a linear transformation of the data and measuring the squared Euclidean distance in the transformed space after applying the linear transformation. This can be shown by factorizing the distance matrix M_c in (24) as: $M_c = ttt^*$, where tt is the linear transformation to be learned. The I2C distance in (24) becomes

$$D^c_{X_i} = \sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^* ttt^* (\mathbf{x}_{ij} - \mathbf{x}^c_{ij}) \qquad (25)$$

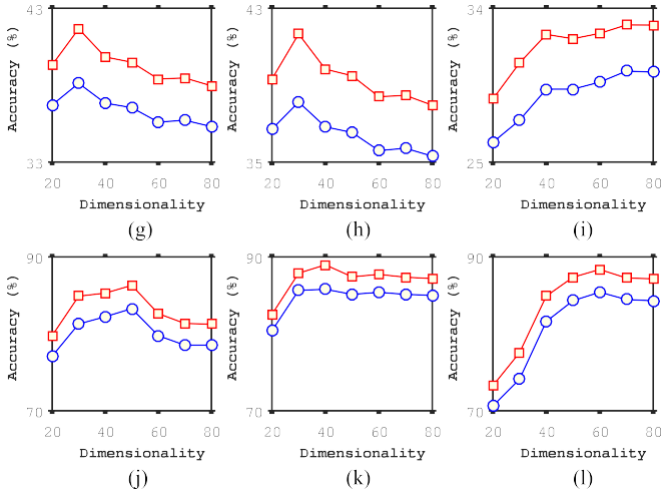


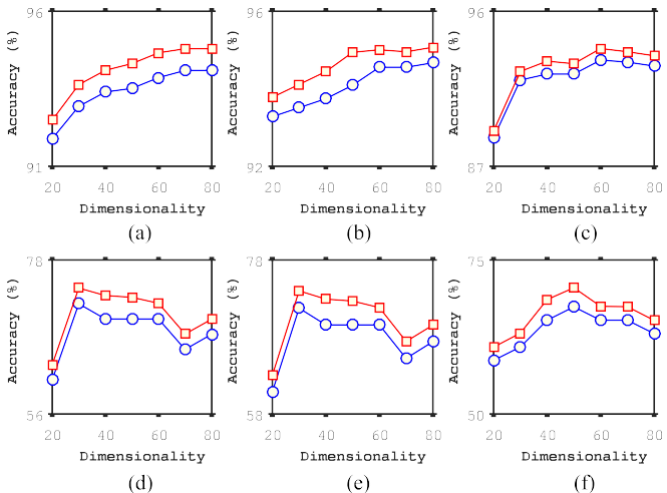
Fig. 3. Performance of I2CDDE with/without manifold regularization on the KTH (the top row), UCF YouTube (the second row), HMDB51 (the third row), and UCF 101 (bottom) datasets. Q and S denote I2CDDE with and without manifold regularization, respectively. (a) NBNN, (b) LBNBB, (c) NBNNK, (d) NBNN, (e) LBNBB, (f) NBNNK, (g) NBNN, (h) LBNBB, (i) NBNNK, (j) NBNN, (k) LBNBB, and (l) NBNNK.

We can see that (25) is equivalent to (10) in terms of linear transformations. The main differences between I2CDDE and I2CDML are summarized as follows.

1) I2CDML adopts the large margin framework from an SVM in the objective function which is solved by the gradient descent algorithm, while I2CDDE can be efficiently solved via a well-studied trace ratio optimization.
2) I2CDML learns multiple distance metrics for all the classes leading to a high computational cost in the high-dimensional space, especially with a huge amount of classes, while I2CDDE learns a unified linear projection, which alleviates the computational burden without compromising the discriminative ability.

V. Experiments and Results

We have conducted extensive experiments on the commonly used benchmark KTH dataset [50], the challenging realistic UCF YouTube [51], HMDB51 [52] and UCF 101 datasets



$j=1$

TABLE I
COMPARISON OF I2CDDE WITH OTHER REDUCTION METHODS BY ACCURACY IN PERCENTAGE (%)

	KTH			HMDB51			YouTube			UCF101		
	NBNN	LNBNN	NBNNK	NBNN	LNBNN	NBNNK	NBNN	LNBNN	NBNNK	NBNN	LNBNN	NBNNK
I2CDDE	**93.6**	**94.1**	**92.5**	**39.7**	**41.7**	**30.7**	**68.8**	**74.7**	**63.1**	**86.3**	**88.9**	**88.3**
PCA	91.7	91.8	89.8	35.6	35.7	25.8	58.6	58.7	53.6	71.4	81.3	79.3
LDA	82.9	83.3	18.3	31.6	31.4	13.1	54.3	56.5	23.9	63.2	64.5	61.2
LFDA	86.6	86.8	67.4	29.6	28.5	10.2	63.1	71.7	23.9	66.5	85.9	74.6
LPP	92.8	93.3	91.0	34.4	35.2	28.3	56.8	60.9	58.7	76.5	86.2	73.5
NPE	91.9	92.6	91.0	34.8	34.9	27.9	55.6	60.9	57.4	74.4	86.7	71.0
Baseline*	**93.9**	**94.1**	89.2	31.8	33.1	29.8	57.8	60.1	62.4	63.9	65.9	64.2

*The baseline results are obtained by HOG3D descriptors of 1000 dimensions without dimensionality reduction.

[53]. We have compared with popular dimensionality reduction methods including PCA, LDA, LFDA, LPP and NPE, and also showed the improvement of I2C-based methods including NBNN, local NBNN and the NBNN kernel. Since LDP is trained on matched and unmatched local features, which are not available for action datasets, we do not include it for comparison.

A. Experimental Settings

The proposed I2CDDE algorithm can work with any raw local feature descriptors to improve the performance, and to benchmark with existing algorithms, we use the three-dimensional histogram of oriented gradients (HOG3D) [7] descriptor which is descriptive and relatively compact with 1000 dimensions is used to describe spatio-temporal interest points (STIPs). We use the computational efficient HOG3D descriptors to demonstrate the effectiveness of the proposed I2CDDE for compact feature descriptor learning rather than beating state-of-the-art algorithms. The I2CDDE can seamlessly work with recently advanced convolutional neural networks to achieve state-of-the-art performance [54]–[56]. We adopt Dollar's periodic detector [35] to detect STIPs. This method can detect a high number of space-time interest points, was proven to be faster, simpler, more precise and gives better performance, even though only one scale is used [57]. Roughly, up to 150 STIPs have been detected for each video clip. Note that our I2CDDE algorithm can be applied to improve the performance of any local feature descriptors [58]. By using dense sampling or dense trajectory based local features, the overall performance can be further improved to achieve state-of-the-art performance [18].

B. Performance on Action Recognition

The proposed I2CDDE shows great effectiveness in improving I2C based methods for human action recognition on the four datasets. The proposed I2CDDE algorithm can greatly enhance the performance of NBNN, local NBNN and the NBNN kernel even with very low dimensionality on all the four datasets. The performance of I2CDDE for action recognition with different dimensions on the KTH, UCF YouTube, HMDB51 and UCF 101 datasets are plotted in Fig. 2(a), 2(b), 2(c) and 2(d) respectively. On the KTH dataset, the increase on the NBNN kernel is more significant than NBNN and local NBNN, while on more challenging UCF YouTube, HMDB51 and UCF 101 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 compared to the original HOG3D features of 1000 dimensions. This demonstrates that the proposed I2CDDE can effectively extract the most discriminative features and achieves compact local feature descriptors

The incorporated manifold regularization can largely boost the performance of I2CDDE. The comparison results with and without the manifold regularization on the four datasets are shown in Fig. 3. On the KTH dataset, the performance with manifold regularization outperforms the baseline I2CDDE with a large margin. On the realistic datasets including UCF YouTube, HMDB51 and UCF 101, the benefit of incorporating the manifold regularization term turns to be more significant, especially on HMDB51 and UCF 101. This is expected and reasonable because the KTH is relatively easy with simple actions and clear backgrounds, while HMDB51 and UCF 101 contain rather complicated actions and clutters in background. The incorporated manifold regularization makes the solutions robust to noisy features. It is worth mentioning that even with simple handcrafted HOG3D features, our I2CDDE can achieve competitive performance with the state-of-the-art algorithms. The larger improvement on the challenging realistic datasets demonstrates the great effectiveness of our I2CDDE for compact feature descriptor learning.

To choose the best dimensions of the learned descriptors for a specific dataset, we use the cross validation, which is simple but effective although more sophisticated technologies can also be used. Roughly, on more challenging datasets, larger number of dimensions are usually required to achieve better performance due to the greater variability of realistic datasets.

C. Comparison With Representative Reduction Methods

The effectiveness of the proposed I2CDDE algorithm has been demonstrated by the advantages over other existing dimensionality reduction algorithms. We have conducted comprehensive comparison with widely used dimensionality reduction algorithms including PCA, LDA, LFDA, LPP and NPE. As shown in Table I, the proposed I2CDDE consistently 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 introduction. We can see that for the NBNN kernel with the reduction of LDA and LFDA, they even fail to produce reasonable results for all the four datasets due to the huge information loss. The proposed I2CDDE incorporates the class labels into dimensionality reduction of 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. The proposed I2CDDE provides the first fully supervised learning to achieve compact yet highly discriminative local feature descriptors.

VI. Conclusion

In this paper, we have proposed a novel supervised learning algorithm called discriminative embedding based on image-to-class distances (I2CDDE) for large scale local feature descriptor learning. The proposed I2CDDE leverages the strengths of the I2C distance, which is for the first time introduced for local feature descriptor learning. Graph-based manifold learning has also been incorporated as a regularization, which further improves the performance of the learned local descriptors. The proposed I2CDDE for the first time achieves fully supervised learning of local feature descriptors. We apply I2CDDE to action recognition tasks which contain large scale spatio-temporal local feature descriptors of high dimensionality. The experimental results on four widely used benchmark datasets: KTH, UCF YouTube, HMDB51 and UCF 101 have demonstrated that I2CDDE can consistently improve the performance and surpass the widely used dimensionality reduction algorithms. More importantly, I2CDDE dramatically speeds up these methods, which enables I2C-based methods to be used for large-scale multimedia datasets.

References

[1] M. Yu, L. Shao, X. Zhen, and X. He, "Local feature discriminant projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1908–1914, Sep. 2016.

[2] X. Zhen and L. Shao, *Introduction to Human Action Recognition*. Hoboken, NJ, USA: Wiley, 2015.

[3] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[4] X. Zhen, L. Shao, and X. Li, "Action recognition by spatio-temporal oriented energies," *Inf. Sci.*, vol. 281, pp. 295–309, 2014.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[6] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3425–3440, Nov. 2015.

[7] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proc. Brit. Machi. Vis. Conf.*, 2008.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[9] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[10] X. Zhen and L. Shao, "Action recognition via spatio-temporal local features: A comprehensive study," *Image Vis. Comput.*, vol. 50, pp. 1–13, 2016.

[11] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, vol. 2, no. 1470, pp. 1470–1477.

[12] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, vol. 2, no. 3, pp. 135–143.

[13] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[14] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell, "The NBNN kernel," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1824–1831.

[15] S. McCann and D. Lowe, "Local naive Bayes nearest neighbor for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3650–3656.

[16] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug. 2014.

[17] M. Fornoni and B. Caputo, "Scene recognition with naive Bayes non-linear learning," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 3404–3409.

[18] I. Kuzborskij, F. M. Carlucci, and B. Caputo, "When naive Bayes nearest neighbors meet convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 2100–2109.

[19] X. Zhen, L. Shao, S. Maybank, and R. Chellappa, "Handcrafted vs. learned representations for human action recognition [editorial]," *Image Vis. Comput.*, vol. 55, no. 2, pp. 39–41, 2016.

[20] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[21] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[22] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, vol. 14, pp. 585–591.

[23] P. Niyogi, "Manifold regularization and semi-supervised learning: Some theoretical analyses," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1229–1250, 2013.

[24] X. Zhen, L. Shao, and F. Zheng, "Discriminative embedding via image-to-class distances," in *Proc. Br. Mach. Vis. Conf.*, 2014.

[25] X. Zhen *et al.*, "Descriptor learning via supervised manifold regularization for multioutput regression," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published. doi: 10.1109/TNNLS.2016.2573260.

[26] X. Zhen, Z. Wang, M. Yu, and S. Li, "Supervised descriptor learning for multi-output regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1211–1218.

[27] R. Behmo, P. Marcombes, A. Dalalyan, and V. Prinet, "Towards optimal naive Bayes nearest neighbor," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 171–184.

[28] Z. Wang, Y. Hu, and L.-T. Chia, "Image-to-class distance metric learning for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 706–719.

[29] X. Zhen and L. Shao, "A local descriptor based on Laplacian pyramid coding for action recognition," *Pattern Recog. Lett.*, vol. 34, no. 15, pp. 1899–1905, 2013.

[30] J. Wang *et al.*, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3360–3367.

[31] K. Rematas, M. Fritz, and T. Tuytelaars, "The pooled NBNN kernel: Beyond image-to-class and image-to-image," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 176–189.

[32] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2004, pp. 506–513.

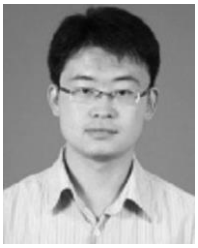
[33] H. Cai, K. Mikolajczyk, and J. Matas, "Learning linear discriminant projections for dimensionality reduction of image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 338–352, Feb. 2011.

[34] G. Hua, M. Brown, and S. Winder, "Discriminant embedding for local image descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[35] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 65–72.

[36] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

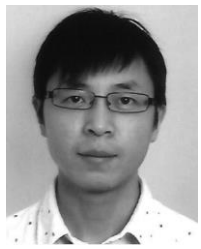
[37] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[38] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci.*, vol. 100, no. 10, pp. 5591–5596, 2003.
[39] X. He and X. Niyogi, "Locality preserving projections," in *Proc. Neural Inf. Process. Syst.*, 2004, vol. 16, pp. 153–160.
[40] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1208–1213.
[41] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1573–1585, 2014.
[42] E. K. Garcia, S. Feldman, M. R. Gupta, and S. Srivastava, "Completely lazy learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 9, pp. 1274–1285, Sep. 2010.
[43] Y. Wang and Y. Wu, "Complete neighborhood preserving embedding for face recognition," *Pattern Recog.*, vol. 43, no. 3, pp. 1008–1015, 2010.
[44] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.
[45] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, vol. 2, 2006, pp. 1735–1742.
[46] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.* 2004, vol. 16, no. 16, pp. 321–328.
[47] F. R. Chung, *Spectral Graph Theory*, vol. 92. Providence, RI, USA: Amer. Math. Soc., 1997.
[48] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[49] P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon, "Metric and kernel learning using a linear transformation," *J. Mach. Learn. Res.*, vol. 13, pp. 519–547, 2012.
[50] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recog.*, 2004, pp. 32–36.
[51] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1996–2003.
[52] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.
[53] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.
[54] J. Song, L. Gao, F. Zou, Y. Yan, and N. Sebe, "Deep and fast: Deep learning hashing with semi-supervised graph construction," *Image Vis. Comput.*, vol. 55, pp. 101–108, 2016.
[55] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image Vis. Comput.*, vol. 55, pp. 42–52, 2016.
[56] J. Xie, G. Dai, F. Zhu, E. Wong, and Y. Fang, "DeepShape: Deep-learned shape descriptor for 3d shape retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1335–1345, Jul. 2017, doi: 10.1109/TPAMI.2016.2596722.
[57] L. Shao and R. Mattivi, "Feature detector and descriptor evaluation in human action recognition," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2010, pp. 477–484.
[58] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1080–1088.



Xiantong Zhen received the B.S. and M.E. degrees from Lanzhou University, Lanzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., in 2013.

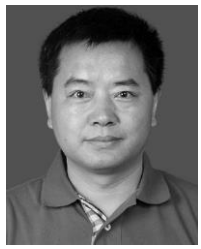
He is currently a Postdoctoral Fellow with the University of Texas at Arlington, Arlington, TX, USA. His research interests include machine learning, computer vision, and medical image analysis.

Feng Zheng received the B.S. and M.S. degrees in applied mathematics from Hubei University, Wuhan, China, in 2006 and 2009, respectively, and the Ph.D. degree in electronic and electrical engineering from the University of Sheffield, Sheffield, U.K., in 2016. From 2009 to 2012, he was an Assistant Re- search Professor (2011–2012) and Assistant Re- searcher (2009–2011) with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China. Currently, he is a Research Fellow with the University of Texas at Arlington, Arlington, TX, USA. His research interests include computer vision, machine learning, and human–computer interaction.



Ling Shao (M'09–SM'10) is a Professor with the School of Computing Sciences, University of East Anglia, Norwich, U.K. Previously, he was a Professor (2014–2016) with Northumbria University, Newcastle upon Tyne, U.K., a Senior Lecturer (2009–2014) with the University of Sheffield, Sheffield, U.K., and a Senior Scientist (2005–2009) with Philips Research, Eindhoven, The Netherlands. His research interests include computer vision, image/video processing, and machine learning.

Prof. Shao is a Fellow of the British Computer Society and the Institution of Engineering and Technology. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals.



Xianbin Cao (M'08–SM'10) received the Ph.D. degree in information science from the University of Science and Technology of China, Beijing, China, in 1996.

He is currently a Professor with the School of Electronic and Information Engineering, Beihang University, Beijing, China. He is also the Director of the Laboratory of Intelligent Transportation System. His current research interests include intelligent transportation systems, airspace transportation management, and intelligent computation.

Dan Xu is currently working toward the Ph.D. degree in information engineering and computer science at The University of Trento, Trento, Italy.

He is a member of Multimedia and Human Understanding Group led by Prof. Nicu Sebe at the University of Trento. He was a Research Assistant with the Department of Mechanical and Automation Engineering and a Visiting Scholar of Multimedia Laboratory, Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, China. His research focuses on computer vision, multimedia, and machine learning. Specifically, he is interested in deep learning and its applications to a variety of topics such as pixel-level prediction and reconstruction, cross-domain retrieval, and video activity analysis.

Mr. Xu was a recipient of a Best Scientific Paper Award at ICPR 2016.