Evolutionary Cross-Domain Discriminative Hessian Eigenmaps

Si Si, Dacheng Tao, *Member, IEEE*, and Kwok-Ping Chan, *Member, IEEE*

***Abstract*—Is it possible to train a learning model to separate tigers from elks when we have 1) labeled samples of leopard and zebra and 2) unlabelled samples of tiger and elk at hand? Cross-domain learning algorithms can be used to solve the above problem. However, existing cross-domain algorithms cannot be applied for dimension reduction, which plays a key role in computer vision tasks, e.g., face recognition and web image annotation. This paper envisions the cross-domain discriminative dimension reduction to provide an effective solution for cross-domain dimension reduction. In particular, we propose the cross-domain discriminative Hessian Eigenmaps or CDHE for short. CDHE connects training and test samples by minimizing the quadratic distance between the distribution of the training set and that of the test set. Therefore, a common subspace for data representation can be well preserved. Furthermore, we basically expect the discriminative information used to separate leopards and zebra can be shared to separate tigers and elks, and thus we have a chance to duly address the above question. Margin maximization principle is adopted in CDHE so the discriminative information for separating different classes (e.g., leopard and zebra here) can be well preserved. Finally, CDHE encodes the local geometry of each training class (e.g., leopard and zebra here) in the local tangent space which is locally isometric to the data manifold and thus CDHE preserves the intraclass local geometry. The objective function of CDHE is not convex, so the gradient descent strategy can only find a local optimal solution. In this paper, we carefully design an evolutionary search strategy to find a better solution of CDHE. Experimental evidence on both synthetic and real word image datasets demonstrates the effectiveness of CDHE for cross-domain web image annotation and face recognition.**

***Index Terms*—Cross-domain learning, dimension reduction, evolutionary search, face recognition, manifold learning, web image annotation.**

I. Introduction

RECENTLY, cross-domain learning has attracted more and more attentions for data analysis problems in image or video processing [45] and pattern classification [40]. It deals with the situation where the knowledge in the target domain is insufficient and some auxiliary information can be obtained from other relevant domains to assist solving the problem in the target domain [7]–[9], [11]. A dozen of practical problems fall in this category because data labeling is very expensive [15]. Definitely, it is possible to provide a sufficient number of labeled images of tiger and elk to train a model to separate tigers from elks. This procedure, however, is absolutely expensive. A natural concern is the possibility of utilizing the discriminative information for separating leopards from zebras to classify tiger images and elk images. As shown in Fig. 1, although these four animals are of different categories (i.e., leopard, zebra, tiger, and elk), they share some common discriminative features (e.g., shape). To be specific, tiger versus leopard and elk versus zebra have similar shape. As a consequence, by using the discriminative information (e.g., shape) learned to separate leopards and zebra, cross-domain learning algorithms can classify tigers and elks.

A dozen of cross-domain learning algorithms have been developed in recent years. For example, spectral analysis based cross-domain learning [32] applies the normalized cut to minimize the cut size on the training domain with the least inconsistency, and at the same time maximize the separation of the test domain for text classification. Fei-Fei *et al.* [34] developed a Bayesian learning framework based on representing object categories with probabilistic models to learn new categories with fewer training examples under the help of the prior information from the learned categories. The transfer hidden Markov model (THMM) [33] aims to transfer knowledge from the learned model in one time period to reduce the calibration effort for the current time period for indoor localization, because it is difficult to entirely gather new calibrated data at each new time period. Maximum mean discrepancy embedding (MMDE) [12] tries to find a subspace where training and test samples distribute similarly to solve the sample selection bias problem for text classification in an unsupervised way. All of these algorithms are either classifiers or probabilistic models.

Dimension reduction [14] plays an important role in various tasks in computer vision [3], [5], [10], e.g., face recognition [4], [22], [43] and web image annotation [2]. A dimension reduction algorithm projects the original high-dimensional feature space to a low-dimensional subspace, where specific statistical properties can be well preserved. For example, Fisher's linear discriminative analysis (FLDA) [16], the most traditional supervised dimension reduction algorithm, minimizes the trace ratio between the within class scatter and the between class scatter so that the Gaussian distributed samples can be well separated in the selected subspace; locality preserving projections (LPP) [17] preserves the local geometry of samples by processing an

Manuscript received July 24, 2009; revised October 07, 2009. First published November 03, 2009; current version published March 17, 2010. This work was supported in part by the HKU-SPF Grant (under project number 10400016), Nanyang SUG Grant (under project numberM58020010), in part by Microsoft Operations PTE LTD-NTU Joint R&D (under project number M48020065), and in part by the K. C. Wong Education Foundation Award. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick J. Flynn.

S. Si and K.-P. Chan are with the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: ssi@cs.hku.hk; kpchan@cs.hku.hk).

D. Tao is with the School of Computer Engineering, The Nanyang Technological University, Singapore 639798 (e-mail: dctao@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2035867

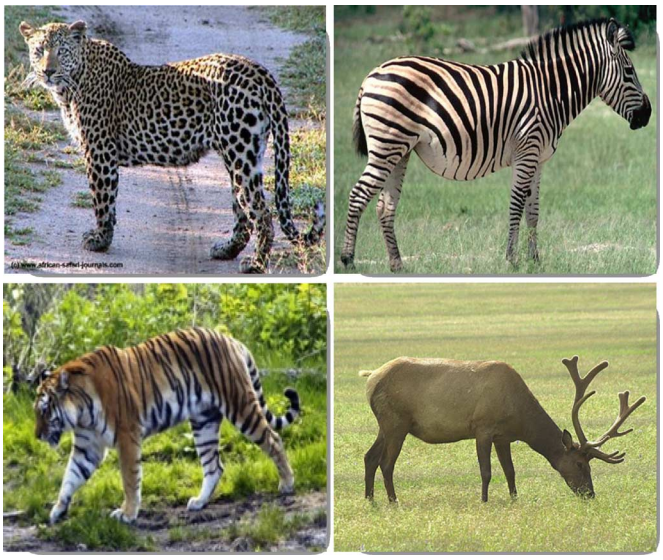


Fig. 1. Training classes are leopard and zebra, while we need to do classification on another two different classes: tiger and elk. These four animals share some common discriminative features, e.g., shape. To be specific, tiger versus leopard and zebra versus elk have the similar shape. As a consequence, it is possible to apply some popular cross-domain learning algorithms to solve the above classification problem.

undirected weighted graph that represents the neighborhood relations of pairwise samples; discriminative locality alignment (DLA) [25] preserves the discriminative information by maximizing the distance among the interclass samples and minimizing the distance among the intraclass samples over local patch of each sample.

The aforementioned dimension reduction algorithms function impressively on both artificial datasets and practical applications, e.g., face recognition. However, they assume that both the training and the test samples are drawn from an identical domain, i.e., in the strict sense, all samples are independent and identically distributed (*i.i.d.*) and this assumption prevents them from many applications, e.g., the cross-domain classification. Therefore, they cannot perform well when the training and the test samples are drawn from different domains.

In this paper, we tackle this problem by searching a shared subspace where training and test samples are distributed in a similar way. In particular, the quadratic distance between the distributions of the training and test domains is minimized in this subspace. However, this subspace could be not optimal for separating samples from different classes. This is because we consider neither the manifold structure of intraclass samples nor the discriminative information of interclass samples, although both of them are important for classification.

By using the patch alignment framework [25], [28], we can model both the intraclass local geometry and the interclass discriminative information conveniently. In particular, for each sample and its associated patch (neighbors of the sample), it is important to consider the following two properties: 1) the intraclass local geometry can be represented by the local tangent space, which is locally isometric to the manifold of the intraclass nearest samples of the patch; and 2) the interclass discriminative information can be represented by the margin between the intraclass neighbor samples and the interclass nearest samples of the patch. Because the method used for local geometry representation is similar to Hessian Eigenmaps [26], the proposed cross-domain dimension reduction algorithm is termed the cross-domain discriminative Hessian Eigenmaps or CDHE for short.

The gradient descent method, the most widely used first-order optimization strategy, can be applied to find a solution or a projection matrix for CDHE. Therefore, the objective function for CDHE is not convex and has several local minima. When optimizing the objective function along the gradient direction by using the gradient descent, it may be trapped into a local minimum, which prevents CDHE from reasonable performance. To solve or at least alleviate this problem, we consider the evolutionary search strategy [20] to find a solution of CDHE. An evolutionary search strategy exploits information of a large number of candidate solutions (known as individuals) in an efficient manner and thus reduces the risk to be trapped into local minima. It operates for searching in a space of individuals and iteratively generating new offspring to update this pool of individuals (called the population) until the best individual is found. Its overall effect is to increase the population's average fitness value, which measures how much the individual will be fit for the optimization problem. As a consequence, along with the evolution, the individual will be more and more suitable for the optimization problem, e.g., the objective of CDHE. In this paper, an evolutionary search strategy is carefully designed to replace the gradient descent strategy for CDHE optimization (E-CDHE) and thorough experimental results show its advantages.

The rest of the paper is organized as follows. Section II briefs the patch alignment framework to better understand the proposed CDHE and E-CDHE. Section III details the proposed cross-domain discriminative Hessian eigenmaps (CDHE) with the gradient descent based training procedure. To obtain a better solution of CDHE, we carefully design an evolution search strategy for CDHE in Section IV. The cross-domain face recognition is presented in Section V to demonstrate the effectiveness of CDHE and E-CDHE. The cross-domain web image annotation is conducted in Section VI based on two real-world web image datasets: NUS-WIDE and MSRA-MM. In Section VII, we compare our proposed methods with a representative multitask learning method in three databases. Section VIII concludes the paper.

II. Patch Alignment Framework

Patch alignment framework [25] unifies popular dimension reduction algorithms, e.g., Fisher's linear discriminant analysis (FLDA) [16], locally linear embedding (LLE) [30], ISOMap [31], and locality preserving projections (LPP) [17]. It contains two stages: part optimization and whole alignment, and can be applied to design dimension reduction algorithms with specific objectives, e.g., the newly proposed CDHE.

Part optimization—for a given sample x_i in a dataset, based on the labeling information, we can categorize the other samples in this dataset into two groups: 1) samples sharing the same class label with x_i, and 2) samples taking different labels with x_i. Each sample x_i associates with a patch $X_i = [x_i, x_{i^1}, \cdots, x_{i^{k_1}}, x_{i_1}, \cdots, x_{i_{k_2}}]$, wherein $x_{i^1}, \ldots, x_{i^{k_1}}$,

i.e., the k_1 nearest samples of x_i, are from the same class as x_i, and $x_{i_1}, \ldots, x_{i_{k_2}}$, i.e., the other k_2 nearest samples of x_i, are from different classes against x_i. For each patch X_i, the corresponding low-dimensional representation is denoted by $Y_i = [y_i, y_{i^1}, \cdots, y_{i^{k_1}}, y_{i_1}, \cdots, y_{i_{k_2}}]$. In this local patch, specific statistical properties, e.g., discrimination and local geometry can be encoded. For example, the discriminative locality alignment (DLA) [25] encodes the discriminative information over the local patch Y_i by keeping the distances between y_i and its k_1 nearest samples (from the same class as y_i) as small as possible and the distances between y_i and its k_2 nearest samples (from different classes against y_i) as large as possible. The part optimization over the patch Y_i is defined as

$$\arg\min_{Y_i} \operatorname{tr}\left(Y_i L_i Y_i^T\right) \tag{1}$$

where $\operatorname{tr}(.)$ is the trace operator; and L_i varies with different dimension reduction algorithms to encode the discriminative information and the local geometry of the patch.

Whole alignment—each patch Y_i has its own coordinate system and all Y_is can be unified together as a whole one by assuming that the coordinate of the i^{th} patch Y_i is selected from the global coordinate $Y_L = [y_1, y_2, \cdots, y_l]$, i.e., $Y_i = Y_L S_i$, where S_i is the selection matrix. The alignment strategy [29] is adopted to build the global coordinate for all patches as

$$\begin{aligned}\arg\min_{Y} \sum_{i=1}^{N} \operatorname{tr}\left(Y_i L_i Y_i^T\right) &= \arg\min_{Y} \sum_{i=1}^{N} \operatorname{tr}\left(Y_L S_i L_i S_i^T Y_L^T\right) \\ &= \arg\min_{Y} \operatorname{tr}\left(Y_L \sum_{i=1}^{N} S_i L_i S_i^T Y_L^T\right) \\ &= \arg\min_{Y} \operatorname{tr}\left(Y_L L Y_L^T\right) \end{aligned} \tag{2}$$

where $L = \sum_{i=1}^{N} S_i L_i S_i^T$. For linearization, $Y_L = W^T X_L$ is usually considered, where W is the projection matrix. We can impose different constraints, e.g., $Y_L^T Y = I$ or $W^T W = I$, to uniquely determine Y_L. The constraint $W^T W = I$ will be adopted throughout the paper. Under this constraint and $Y_L = W^T X_L$, the solution of (2) can be obtained by using the conventional Lagrangian multiplier method [36] or the generalized eigenvalue decomposition.

There are many dimension reduction approaches taking both local geometric and discriminative information into consideration. A well-known example is Maximum Margin Projection (MMP) [44]. The differences between PAF and MMP are listed as below.

1) MMP and PAF define the 'neighbors' of a sample x_i in different ways. MMP focus on the neighbors of x_i, $N(x_i)$, which includes the k nearest samples to x_i. While for every sample x_i in PAF, it builds a patch $X_i = [x_i, x_{i^1}, \cdots, x_{i^{k_1}}, x_{i_1}, \cdots, x_{i_{k_2}}]$, wherein $x_{i^1}, \ldots, x_{i^{k_1}}$, i.e., the k_1 nearest samples of x_i, are from the same class as x_i, and $x_{i_1}, \ldots, x_{i_{k_2}}$, i.e., the k_2 nearest samples of x_i, are from different classes against x_i. Both k_1 and k_2 can be set manually. Thus, the definitions for $N(x_i)$ in MMP and X_i in PAF are different.
2) MMP is a graph-based approach, so graphs are needed to be built in advance. PAF is not a graph-based method, and thus it is not essential to build a graph to represent the nearby points' relationship. PAF uses the alignment strategy to align all the patches together.
3) Finally, MMP considers the geometry of all the samples including both labelled and unlabelled samples, while PAF focuses on utilizing the geometry of labelled samples.

III. Cross-Domain Discriminative Hessian Eigenmaps

Conventional dimension reduction algorithms assume that training and test samples are drawn from an identical domain. In many practical applications, however, they are actually from different domains. For example, we have a number of duly labeled images of leopard and zebra and we want to train a model to annotate images of tiger and elk. Therefore, these conventional dimension reduction algorithms cannot work well in this scenario. This Section presents the cross-domain discriminative Hessian Eigenmaps or CDHE for short to solve the cross-domain classification tasks. Stemmed from recent results in manifold learning and cross-domain learning, CDHE characterizes three specific properties.

1) The local geometry property—nearby samples in the original Euclidean space are close to each other in the learned subspace.
2) The discriminative property—samples from different classes can be well separated in the learned subspace.
3) The cross-domain property—samples from the training and the test domains are almost independent and identically distributed.

In summary, the discriminative information as well as the local geometry obtained from the training domain will be passed to the test domain by adopting the cross-domain property. Therefore, we can achieve a good classification on the test domain, although the test domain is different from the training domain. To integrate the local geometry and the discriminative information more conveniently, CDHE will be considered under the patch alignment framework [25].

A. Modified Hessian Eigenmaps

Empirically, intraclass geometry is useful for classification. Manifold learning algorithms, e.g., Laplacian Eigenmap (LE) [42] and LPP, recover the low-dimensional manifold structure of samples embedded in a high-dimensional Euclidean space. LE minimizes the average of the Laplacian operator over the manifold, where the Laplacian operator on the mapping function f in the tangent space is defined by $\Delta^{(\tan)}(f) = \sum_{i=1}^{d} \partial^2 f / \partial x_i^2$ and the average function for Laplacian operator over the manifold is $L(f) = \int_M \left(\Delta^{(\tan)}(f)\right)^2 dm$. LPP is a linearization of LE. Hessian Eigenmaps (HE) [26] replaces the Laplacian operator used in LE with the Hessian operator. The Hessian matrix is the square matrix of second-order partial derivatives of a function f, that is, $H(f)_{ij}(x) = \partial^2 f / \partial x_i \partial x_j$. HE recovers the underlying parameterization of a manifold M embedded in a high-dimensional space if the manifold M is locally isometric to an open and connected subset of R^d. Because the parameter space is not necessarily convex in Hessian Eigenmaps, it can model a nonconvex manifold, e.g., an S-curve surface with a hole. Therefore, we adapt Hessian Eigenmaps in CDHE to preserve the local geometry for dimension reduction.

Hessian Eignmaps finds the $(d+1)$-dimensional null-space of $H(f)$, where $H(f)$ is the Hessian matrix of a smooth mapping f, i.e., $f: M \mapsto R$. This $H(f)$ can be calculated by using $H(f) = \int_M \|H_f(x_i)\|_F^2 dx$ wherein $H_f(x_i)$ is the Hessian of f on the patch $X_{H(i)} = [x_i, x_{i^1}, \cdots, x_{i^{k_1}}]$ and the corresponding output in low-dimensional space is $Y_{H(i)} = [y_i, y_{i^1}, \cdots, y_{i^{k_1}}]$. The tangent plane $T_{x_i}(M)$, a Euclidean space tangential to M at x_i, is an orthogonal coordinate system. In order to estimate $H_f(x_i)$, we calculate the local coordinate system of $X_{H(i)}$ and each sample in $X_{H(i)}$ has its own local coordinate Π_i on the tangent plane $T_{x_i}(M)$. Afterwards, this $H_f(x_i)$ can be estimated by using Π_i.

However, Hessian Eigenmaps cannot be applied to many practical applications, e.g., face recognition and image annotation because it requires that $k_1 > d$ where k_1 is the number of the neighboring samples and d is the dimension of the subspace. It is difficult to guarantee this condition because we have a limited number of samples. We propose to overcome this problem by performing PCA on M at x_i and orthnormalizing the d-dimensional representation to obtain the tangent coordinate in $T_{x_i}(M)$. The following steps for the modified Hessian Eigenmaps are similar to those in Hessian Eigenmaps.

Under the patch alignment framework, the objective function for the modified Hessian Eigenmaps to preserve the local geometry on a local patch $Y_{H(i)}$ can be written as

$$\begin{aligned} H(y_i) &= \operatorname{tr}\left(Y_{H(i)} H_f(x_i) H_f^T(x_i) Y_{H(i)}^T\right) \\ &= \operatorname{tr}\left(Y_{H(i)} L_{H(i)} Y_{H(i)}^T\right) \end{aligned} \tag{3}$$

where $L_{H(i)} = H_f(x_i) H_f^T(x_i)$ encodes the local geometry information of the patch $X_{H(i)}$ and $H(y_i)$ is the local geometry representation. Under the help of $L_{H(i)}$, local geometric information can be further preserved.

B. Margin Maximization

As for classification, however, it is insufficient to only retain the local geometry, because no labeling information is taken into account. To further exploit the discriminative power, like the definition of the local geometry, we can define a new margin maximization based scheme for discriminative information preservation over patches [35]. In particular, for each sample x_i associated with a patch $X_{M(i)} = [x_i, x_{i^1}, \cdots, x_{i^{k_1}}, x_{i_1}, \cdots, x_{i_{k_2}}]$, wherein $x_{i^1}, \ldots, x_{i^{k_1}}$, i.e., the k_1 nearest samples of x_i, are from the same class as x_i, and $x_{i_1}, \ldots, x_{i_{k_2}}$, i.e., the other k_2 nearest samples of x_i, are from different classes against x_i, we define the margin as the average difference between two kinds of distances on this patch. One is called interclass distance, that is, the distance between x_i and samples taking different labels, i.e., $x_{i_1}, \ldots, x_{i_{k_2}}$; the other is called intraclass distance, that is, the distance between x_i and samples sharing the same label, i.e., $x_{i^1}, \ldots, x_{i^{k_1}}$. Basically, in the patch $X_{M(i)}$'s low-dimensional representation $Y_{M(i)} = [y_i, y_{i^1}, \cdots, y_{i^{k_1}}, y_{i_1}, \cdots, y_{i_{k_2}}]$, we expect the margin between intraclass and interclass samples will be maximized as large as possible, i.e.,

$$\sum_{p=1}^{k_2} \left\|y_i - y_{i_p}\right\|^2 \frac{1}{k_2} - \sum_{j=1}^{k_1} \|y_i - y_{i^j}\|^2 \frac{1}{k_1}. \tag{4}$$

On the other hand, based on (4), we try to minimize the bellowing objective function:

$$\begin{aligned} M(y_i) &= \sum_{j=1}^{k_1} \|y_i - y_{i^j}\|^2 \frac{1}{k_1} - \sum_{p=1}^{k_2} \left\|y_i - y_{i_p}\right\|^2 \frac{1}{k_2} \\ &= \operatorname{tr}\left(Y_{M(i)} \begin{bmatrix} -e_{k_1+k_2}^T \\ I_{k_1+k_2} \end{bmatrix} \operatorname{diag}(w_i) \right. \\ &\qquad \left. \times [-e_{k_1+k_2}, I_{k_1+k_2}] Y_{M(i)}^T\right) \\ &= \operatorname{tr}\left(Y_{M(i)} L_{M(i)} Y_{M(i)}^T\right) \end{aligned} \tag{5}$$

where $w_i = [\overbrace{1/k_1, \ldots, 1/k_1}^{k_1}, \overbrace{-1/\mathrm{k}_2, \ldots, -1/\mathrm{k}_2}^{k_2}]^T$; $I_{k_1+k_2}$ is the $(k_1+k_2)\times(k_1+k_2)$ identity matrix; $e_{k_1+k_2} = [1, \ldots, 1]^T \in R^{k_1+k_2}$; $L_{M(i)} = \begin{bmatrix} \sum_{j=1}^{k_1+k_2} (w_i)_j & -w_i^T \\ -w_i & \operatorname{diag}(w_i) \end{bmatrix}$ and $M(y_i)$ is the margin information representation.

C. Cross-Domain Parser

If samples from training and test domains are *i.i.d.*, both the local geometry and the discriminative information can be well passed from the training domain to the test domain. However, in the cross-domain setting, the training and the test samples are distributed differently in the original high-dimensional space. Therefore, it is essential to find a subspace so that 1) the training and the test samples are distributed similarly and 2) the local geometry and the discriminative information obtained from the training domain can be passed to the test domain.

The subspace can be obtained by minimizing a distance between the distribution of the training samples and that of the test samples. Given a dataset set $X = [x_1, x_2, \cdots, x_l, x_{l+1}, \cdots, x_{l+u}]$, suppose the first l samples are from the training set with the labeling information and the rest u samples are from the test set without the labeling information. Its low-dimensional representation is $Y = [y_1, y_2, \cdots, y_l, y_{l+1}, \cdots, y_{l+u}]$. To provide a computational and tractable method to measure the distance between $p_L(y)$ (the distribution of training samples in the low-dimensional subspace) and $p_U(y)$ (the distribution of test samples in the low-dimensional subspace), the quadratic distance is applied here

$$\begin{aligned} Q_W(P_L \| P_U) &= \int \left(p_L(y) - p_U(y)\right)^2 dy \\ &= \int \left(p_L(y)^2 - 2 p_L(y) p_U(y) + p_U(y)^2\right) dy. \end{aligned} \tag{6}$$

In order to estimate the distribution $p_L(y)$ and $p_U(y)$ in the projected subspace, we apply the kernel density estimation (KDE) [6], [24], i.e., $p(y) = (1/n)\sum_{i=1}^{n} G_{\sum}(y - y_i)$. Here, n is the number of samples, and $G_{\sum}(y)$ is the d-dimensional Gaussian kernel with the covariance matrix $\sum$. By introducing the estimated distributions based on KDE to (6), we have

$$\begin{aligned} &Q_W(P_L\|P_U) \\ &\quad= \int \left(\frac{1}{l}\sum_{i=1}^{l} G_{\sum_1}(y-y_i)\right)^2 dy \\ &\qquad+ \int \left(\frac{1}{u}\sum_{j=l+1}^{l+u} G_{\sum_2}(y-y_j)\right)^2 dy \\ &\qquad- \int \frac{2}{lu}\sum_{i=1}^{l}\sum_{j=l+1}^{l+u} G_{\sum_1}(y-y_i)G_{\sum_2}(y-y_j)dy. \quad (7) \end{aligned}$$

Because $\int G_{\sum_1}(y-y_s)G_{\sum_2}(y-y_t)dy = G_{\sum_1+\sum_2}(y_s-y_t)$ holds for two arbitrary Gaussian kernels, we get a discrete form of (7) as

$$\begin{aligned} Q_W(P_L\|P_U) = &\frac{1}{l^2}\sum_{s=1}^{l}\sum_{t=1}^{l} G_{\sum_{11}}(y_t-y_s) \\ &+ \frac{1}{u^2}\sum_{s=l+1}^{l+u}\sum_{t=l+1}^{l+u} G_{\sum_{22}}(y_t-y_s) \\ &- \frac{2}{lu}\sum_{s=1}^{l}\sum_{t=l+1}^{l+\mathrm{u}} G_{\sum_{12}}(y_t-y_s) \qquad (8) \end{aligned}$$

where $\sum_{11} = \sum_1 + \sum_1$, $\sum_{12} = \sum_1 + \sum_2$ and $\sum_{22} = \sum_2 + \sum_2$. The quadratic distance $Q_W(P_L\|P_U)$ serves as a bridge to pass the local geometry and the discriminative information from the training domain to the test domain.

D. Cross-Domain Discriminative Hessian Eigenmaps

By using the results obtained from the previous subsections, we can obtain the optimization framework to learn the projection matrix W, which can pass both the local geometry and the discriminative information from the training domain to the test domain. Because the margin representation $M(y_i)$ and the local geometry representation $H(y_i)$ are defined over patches, and each patch has its own coordinate system, alignment strategy is adopted here to build a global coordinate for all patches defined for the training samples. As a consequence, the objective function to solve the cross-domain dimension reduction is given by

$$W = \arg\min_{W\in R^{D\times d}} \sum_{i=1}^{l} (M(y_i) + \beta H(y_i)) + \lambda Q_W(P_L\|P_U) \quad (9)$$

where λ and β are two tuning parameters. If we define two selection matrixes $S_{H(i)}$ and $S_{M(i)}$, which select samples in the i^{th} patch from all the training samples $Y_L = [y_1, y_2, \cdots, y_l]$ for constructing $M(y_i)$ and $H(y_i)$, respectively. Therefore, $Y_{H(i)} = Y_L S_{H(i)}$ and $Y_{M(i)} = Y_L S_{M(i)}$ with $Y_{H(i)}$ representing the patch for the local geometry preservation and $Y_{M(i)}$ denoting the patch for margin maximization. After plugging (3), (5) and $Y_L = W^T X_L$, the objective function in (9) will turn to

$$\begin{aligned} W = &\arg\min_{W\in R^{D\times d}} \sum_{i=1}^{l} \Big(\mathrm{tr}\left(Y_{M(i)} L_{M(i)} Y_{M(i)}^T\right) \\ &\qquad +\beta \mathrm{tr}\left(Y_{H(i)} L_{H(i)} Y_{H(i)}^T\right)\Big) \\ &+ \lambda Q_W(P_L\|P_U) \\ = &\arg\min_{W\in R^{D\times d}} \sum_{i=1}^{l} \Big(\mathrm{tr}\left(Y_L S_{M(i)} L_{M(i)} \left(Y_L S_{M(i)}\right)^T\right) \\ &\qquad +\beta \mathrm{tr}\left(Y_L S_{H(i)} L_{H(i)} \left(Y_L S_{H(i)}\right)^T\right)\Big) \\ &+ \lambda Q_W(P_L\|P_U) \\ = &\arg\min_{W\in R^{D\times d}} \mathrm{tr}\Bigg(Y_L \sum_{i=1}^{l} \Big(S_{M(i)} L_{M(i)} S_{M(i)}^T \\ &\qquad +\beta S_{H(i)} L_{H(i)} S_{H(i)}^T\Big) Y_L^T\Bigg) \\ &+ \lambda Q_W(P_L\|P_U) \\ = &\arg\min_{W\in R^{D\times d}} \mathrm{tr}\left(Y_L L Y_L^T\right) + \lambda Q_W(P_L\|P_U) \\ = &\arg\min_{W\in R^{D\times d}} \mathrm{tr}\left(W^T X_L L X_L^T W\right) + \lambda Q_W(P_L\|P_U) \quad (10) \end{aligned}$$

where $L = \sum_{i=1}^{l}(S_{M(i)} L_{M(i)} S_{M(i)}^T + \beta S_{H(i)} L_{H(i)} S_{H(i)}^T)$ is the alignment matrix encoding both the local geometry and the discriminative information, and X_L represents all training samples.

E. Gradient Descent Based Strategy for Optimization

To obtain a possible solution W of (10), a direct method is to optimize (10) with respect to W iteratively by adopting the gradient descent technique. For the $k+1^{th}$ iteration, according to (10), the update rule for solving W is

$$\begin{aligned} W_{k+1} = W_k - \eta(k)\Bigg(&\frac{\partial \mathrm{tr}\left(W^T X_L L X_L^T W\right)}{\partial W} \\ &+\lambda \frac{\partial Q_W(P_L\|P_U)}{\partial W}\Bigg) \quad (11) \end{aligned}$$

where $\eta(k)$ is the learning rate factor at the iteration k, and it controls the gradient step size for the k^{th} iteration. Gradients of $\mathrm{tr}(W^T X_L L X_L^T W)$ and $Q_W(P_L\|P_U)$ can be easily calculated through their original functions.

Based on (10), it is obvious to generate the derivative of $\mathrm{tr}(W^T X_L L X_L^T W)$ with respect to W as

$$\frac{\partial \mathrm{tr}\left(W^T X_L L X_L^T W\right)}{\partial W} = \left(X_L L X_L^T + \left(X_L L X_L^T\right)^T\right) W. \quad (12)$$

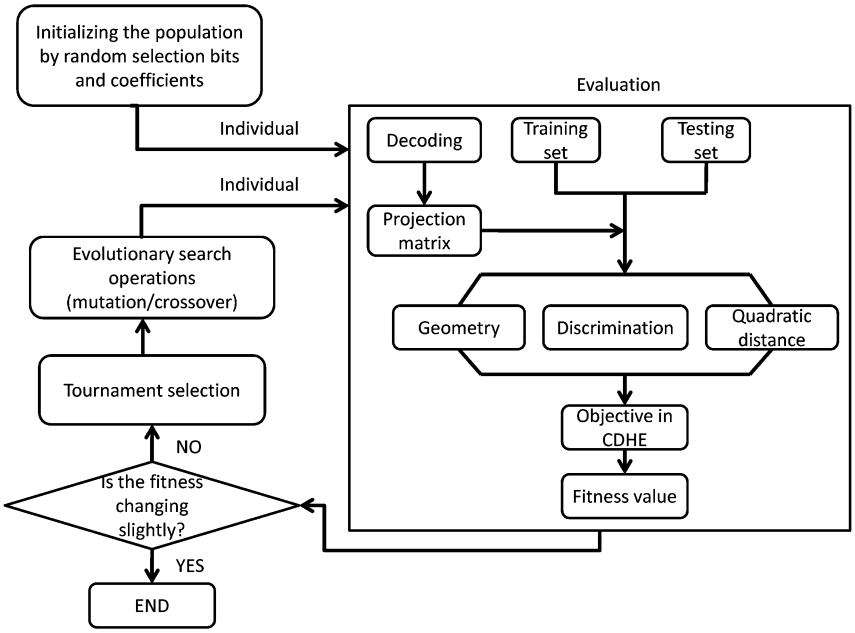


Fig. 2. Flowchart of E-CDHE.

In addition, according to (8), the derivative of $Q_W(P_L\|P_U)$ with respect to W is

$$\begin{aligned}
&\frac{Q_W(P_L\|P_U)}{\partial W} \\
&= \sum_{i=1}^{l+u} \frac{Q_W(P_L\|P_U)}{\partial y_i}\frac{\partial y_i}{\partial W} \\
&= \sum_{i=1}^{l} \frac{Q_W(P_L\|P_U)}{\partial y_i}x_i^T + \sum_{i=l+1}^{l+u} \frac{Q_W(P_L\|P_U)}{\partial y_i}x_i^T \\
&= \frac{2}{l^2}\sum_{i=1}^{l}\sum_{t=1}^{l} G_{\sum_{11}}(y_i - y_t)\Sigma_{11}^{-1}(y_t - y_i)x_i^T \\
&\quad - \frac{2}{lu}\sum_{i=1}^{l}\sum_{t=l+1}^{l+u} G_{\sum_{12}}(y_t - y_i)\Sigma_{12}^{-1}(y_t - y_i)x_i^T \\
&\quad + \frac{2}{u^2}\sum_{i=l+1}^{l+u}\sum_{t=l+1}^{l+u} G_{\sum_{22}}(y_i - y_t)\Sigma_{22}^{-1}(y_t - y_i)x_i^T \\
&\quad - \frac{2}{lu}\sum_{i=l+1}^{l+u}\sum_{t=1}^{l} G_{\sum_{12}}(y_t - y_i)\Sigma_{12}^{-1}(y_t - y_i)x_i^T. \quad (13)
\end{aligned}$$

Based on (11), (12), and (13), we can obtain a solution of CDHE iteratively by imposing $W^TW = I$.

IV. Evolutionary CDHE

However, the high-dimensional structure of (10) is not convex and it has several local minima, and thus the gradient descent strategy for solving (10) often traps into one of these local minima. To overcome this problem, the evolutionary search strategy, an alternative to gradient descent when the optimization problem is not convex, is carefully designed to solve CDHE. We term it the evolutionary search for CDHE (E-CDHE).

Evolutionary search [20] is a generic population-based metaheuristic optimization strategy that mimics the metaphor of natural biological evolution. In analogy to natural genetics, at each generation, a new set of population (a number of potential solutions) is created by the processes of the selection, crossover, and mutation operations. Evolutionary search operates on this population of potential solutions and applies the principal of survival of the fittest to produce a better approximation to the solution until the best solution is found.

Fig. 2 illuminates the detailed process of E-CDHE. First, a population is randomly initialized in the search space. The population includes a pool of individuals represented by binary strings to save the space. Because every individual from the population has been represented in a binary string, next for every individual, we decode this individual into a projection matrix and calculate its fitness value through this projection matrix in the evaluation process. Its fitness value includes three parts: the intraclass local geometry of the training samples, the interclass discriminative information of the training samples and the quadratic distance between training and test sets in the projected subspace. When fitness values of all individuals are ready, we check whether the mean of all fitness values in this population is unchanged or changed slightly comparing against the anterior generation. If unchanged or changed slightly, the individual with the best fitness value is recorded and output. Otherwise, we randomly select two individuals through tournament selection and undertake mutation or crossover operations under certain probability to generate a new individual. Hereby, a new population is generated after these three operations and the whole process will be repeated many times until convergence.

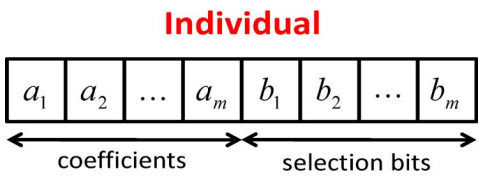


Fig. 3. Illumination of an individual.

A. Constructing Projection Matrix in the Search Space

To avoid searching blindly in the original high-dimensional space and increase the searching efficiency, we propose to construct the projection matrix by using the linear combination of the basis of the search space. In E-CDHE, the search space contains two parts: 1) the eigenvectors of $X_L L X_L^T$ corresponding to the leading eigenvalues and 2) the projection vectors from CDHE which include discriminative information for cross-domain learning.

The first part contains rich discriminative information in the training set. Using only discriminative components from the training set does not necessarily lead to good classification performance in the test set, because they are not perfect for the cross-domain learning. Therefore, on the other hand, to make the search space suitable for cross-domain learning, we take the projection vectors from CDHE into consideration.

Let $\Phi = \{\alpha_i \in \Re^D | i = 1, 2, \cdots, m\}$ be the search space discussed above. E-CDHE will search among different combinations of basis vectors in Φ, i.e., a new projection matrix W can be constructed by linearly combining the vectors from the basis vectors in Φ, and then orthnormalizing these vectors.

According to the above method of generating a projection matrix W, the information encoded in an individual should include m selection bits and m combination coefficients. Each selection bit represents whether the corresponding basis vector in Φ is selected to produce W, and its coefficient is represented by 10 bits in a binary decimal. Fig. 3 shows the structure of an individual, in which the selection bits $b_1, b_2, \cdots, b_m$ taking the value of "0" or "1," indicating whether the corresponding basis vector in Φ is selected to construct the projection matrix and each combination coefficient a_i will take 10 bits to represent the coefficient of the corresponding vector to construct W. An individual under such definition takes $11m$ bits and thus E-CDHE achieves a low space complexity.

B. Evaluating Individuals

The fitness value evaluates how good the solution will be for the optimization problem in E-CDHE. It influences how E-CDHE will choose offspring from the current generation. The objective function characterizes the optimality of a candidate solution, so the fitness function of E-CDHE will be the versa of the objective function of CDHE in (10), i.e.,

$$Fitness(W) = -\left(tr\left(W^T X_L L X_L^T W\right) + \lambda Q_W(P_L || P_U)\right) \tag{14}$$

where $Fitness(W)$ is the fitness function of E-CDHE and W is the candidate solution. When evaluating the individual's fitness value, we first decode the individual into the projection matrix W according to the basis vectors in Φ and then calculate its fitness value in (14).

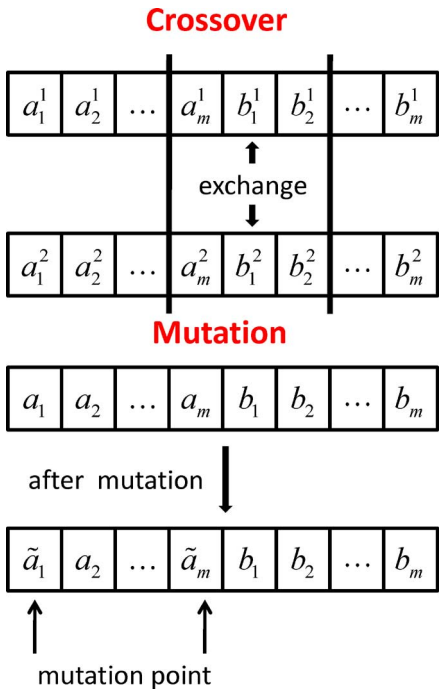


Fig. 4. Illumination of two operations: Crossover and mutation in E-CDHE.

C. Generating New Individuals

During each successive generation, E-CDHE will breed offspring from the current generation at a certain proportion. E-CDHE generates new solutions via three operations on the individuals based on the fitness function defined in (14). The three operations include selection, crossover, and mutation. In this paper, we use the following specific strategies.

1) Tournament selection [19]—the selection of suitable solutions simulates the action of tournament. It operates under a fitness-based process, where fitter solutions (as measured by the fitness value) are typically more likely to be selected, because the possibility of an individual to be a winner of the tournament selection is directly related to its fitness value. This will help keep the diversity of the population, and preventing premature convergence on poor solutions.
2) Two-point crossover [39]—after the tournament selection of two individuals from the population, we randomly select two crossover points and implement an exchange procedure between these two individuals. Fig. 4 shows that the middle segment of two individuals, i.e., a_m^1, b_1^1, b_2^1 and a_m^2, b_1^2, b_2^2 are exchanged and hereby two new individuals are generated. The exchange procedure is not simply swapping the segments between two crossover points. It is essential to guarantee the number of 1s in the selection bits unchanged after exchange the segmentation, because the number of 1s in the selection bits represents the dimension of the subspace. If after the exchange the number of 1s in the selection bits changes, we will randomly add some 1s in the selection bits if the number of 1s decreases or vice versa.

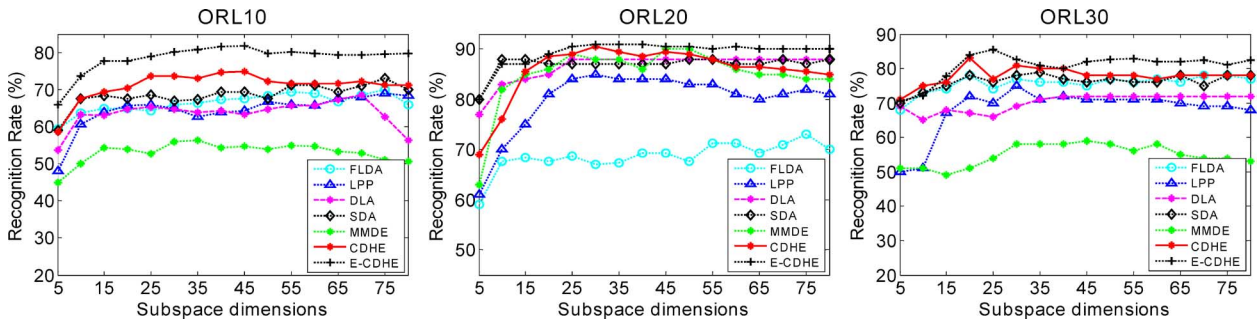


Fig. 5. Recognition rates versus subspace dimensions under the ORL10, ORL20, ORL30 (from left to right) experimental settings.

3) Probability mutation—every bit in an individual is subjective to mutation from 0 to 1 or vice versa under a certain probability. Fig. 4 shows the bits $\tilde{a}_1$ and $\tilde{a}_m$ will be subjective to mutation from 0 to 1 or vice verse and a new individual will be generated. It is worth emphasizing that if there is a change in the selection bits, another randomly selected bit should be changed accordingly to make the total number of 1s in the selected bits unchanged.

The above generation process is repeated until the fitness value is unchanged or changed only slightly. Because we are searching in a constraint search space but not the original high-dimensional space, it will converge to an optimal solution efficiently.

V. Cross-Domain Face Recognition

In this section, we justify the effectiveness of the proposed two cross-domain dimension reduction methods, i.e., CDHE and E-CDHE for the application of cross-domain face recognition on two cross-domain face databases in comparison with four classical dimension reduction algorithms, e.g., FLDA [16], LPP [17], DLA [18] and the semi-supervised discriminate analysis (SDA) [13]. In addition, we also compare CDHE and E-CDHE with the maximum mean discrepancy embedding (MMDE) [12] which is an unsupervised cross-domain learning algorithm to demonstrate the effectiveness of CDHE and E-CDHE in cross-domain setting. Furthermore, we also show that E-CDHE performs better than CDHE.

A. ORL Test

To justify the effectiveness of the proposed CDHE and E-CDHE, we compare them against FLDA, LPP, DLA, SDA and MMDE on the ORL [37] face database. The ORL database contains 400 face images with 40 distinct subjects and each subject has 10 images taken at different times, lighting conditions, facial expressions (open/closed eyes, smiling/not smiling), and facial details (glasses/no glasses) against a dark homogeneous background. The images used for our experiments are of size 32 $\times$ 32 in raw pixel. There is no public face dataset constructed in the cross-domain setting, so we build a new dataset based on ORL face database by randomly selecting $n(= 10, 20, 30)$ subjects as the training classes and utilize the rest subjects as the test classes to test the effectiveness of CDHE and E-CDHE for cross-domain learning. In these three settings (ORL10, ORL20, and ORL30), images in training and test sets come from different subjects and thus training and test domains are distinct, which is suitable for cross-domain learning.

TABLE I
Best Recognition Rates (%) of Seven Algorithms on the Experiment of Cross-Domain Face Recognition on the Settings of ORL10, ORL20, ORL30 (the Numbers in the Parentheses are the Optimal Subspace Dimensions)

	ORL10	ORL20	ORL30
FLDA	70.00(75)	73.00(75)	78.00(70)
LPP	69.00(75)	85.00(30)	75.00(30)
DLA	68.67(70)	88.00(25)	72.00(40)
SDA	73.00 (75)	88.00(75)	79.00(35)
MMDE	56.33(35)	90.00(45)	59.00(45)
CDHE	**75.00(45)**	**90.05(30)**	**83.00(20)**
E-CDHE	**81.91(45)**	**91.00(30)**	**85.42(25)**

In the test stage, we select one reference image from each class in the test domain and then apply the nearest-neighbor rule to predict labels of the remaining test images in the selected subspace. It is worth emphasizing that the sample's labeling information from the reference images is inaccessible to all the dimension reduction algorithms in the training stage. Table I reports the highest recognition rate of each algorithm with respect to the corresponding optimal subspace dimension. Fig. 5 shows the results of CDHE and E-CDHE against FLDA, LPP, DLA, SDA, and MMDE with regard to recognition accuracy. From this figure, we can see CDHE and E-CDHE consistently and significantly outperform the other dimension reduction algorithms. Because CDHE and E-CDHE do not have the *i.i.d.* assumption and can better utilize the labeling information in the training set. Furthermore, E-CDHE can make the improvement over CDHE because it has less risk to be trapped in the local minima.

B. Yale2PIE and PIE2Yale Test

In this section, we will further investigate the effectiveness of CDHE and E-CDHE and also compare them with other five algorithms, i.e., FLDA, LPP, DLA, SDA and MMDE for cross-domain face recognition between two quite different databases. There is no public face dataset constructed for the application of cross-domain face recognition between two databases, so we build a new dataset by combining the YALE dataset [1] and the CMU-PIE dataset [23]. The YALE dataset consists of 165 images with 15 individuals and each with 11 images captured under different facial expressions and configurations. The CMU-PIE dataset contains 41,368 images of 68 people under 13 different poses, 43 different illumination

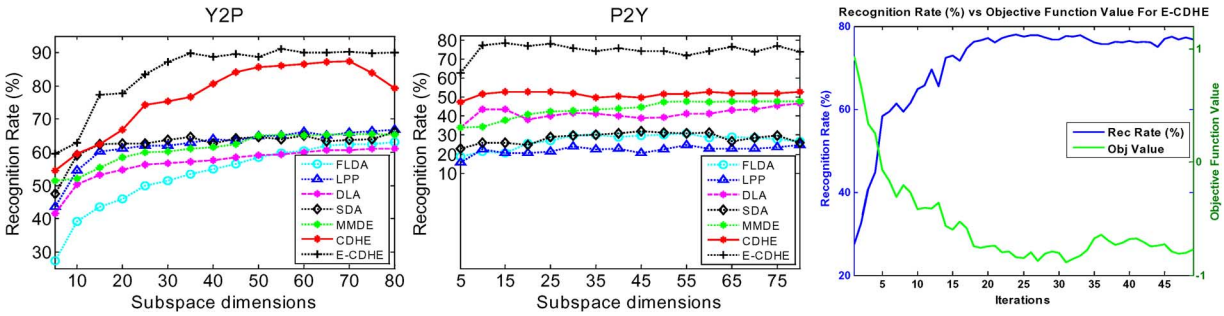


Fig. 6. Recognition rates versus subspace dimensions under the Y2P and P2Y (from left to right) experimental settings. The most right one is the objective function value and the recognition rates versus iterations.

conditions, and 4 different expressions. The images from YALE and PIE used for our experiments are of size 32×32 in raw pixel. All images in YALE are used in our experiments, but we randomly select 10 images per individual in the CMU-PIE dataset. As a consequence, the combined dataset has two domains: one from YALE and the other from CMU-PIE. Based on the dataset, two experiments can be conducted: 1) Y2P: the training set is YALE and the test set is CMU-PIE; 2) P2Y: the training set is CMU-PIE and the test set is YALE.

In the test stage, we select one reference image from each class in the test domain and then apply the nearest-neighbor rule to predict labels of the remaining test images in the selected subspace. It is worth emphasizing that the sample's labeling information from the reference images is inaccessible to all the dimension reduction algorithms in the training stage.

The face recognition rates versus subspace dimensions on the databases of Y2P and P2Y are presented in Fig. 6. Table II reports the best recognition rate of each algorithm with respect to the corresponding optimal subspace dimension. It is shown that CDHE and E-CDHE significantly outperform the other five dimension reduction algorithms. Both experiments show conventional dimension reduction algorithms, e.g., FLDA, LPP, DLA, and SDA, are unsuitable for cross-domain tasks because they assume that both the training and the test samples are drawn from the same distribution. MMDE considers the distribution bias between the training and the test sets, but it ignores both the local geometry of the intraclass samples and the discriminative information of the interclass samples. Therefore, it cannot work as well as CDHE and E-CDHE. CDHE and E-CDHE perform consistently and significantly better than others, because they successfully pass both the local geometry and the discriminative information from the training set to the test set. Furthermore, E-CDHE performs better than CDHE because it can search for a better solution of the optimization problem defined in (10).

Fig. 6 (right most) presents the trend of E-CDHE's objective function values and its recognition rates with respect to the training iterations to examine the convergence property of E-CDHE. Since different dimensions on different datasets share similar convergence properties in the training stage, we only show analyses of E-CDHE on the database of Y2P at dimension 20 in the Fig. 6. In this figure, $\lambda = 1$ [in (10)] and the size of the population is 50. This figure shows the objective values and its variance are decreasing with the increasing of the training iterations, i.e., evolutionary search can find a solution for E-CDHE. This is consistent with the discussions in Section IV.

TABLE II
BEST RECOGNITION RATES (%) OF SEVEN ALGORITHMS ON THE EXPERIMENT OF CROSS-DOMAIN FACE RECOGNITION ON THE SETTINGS OF Y2P AND P2Y (THE NUMBERS IN THE PARENTHESES ARE THE OPTIMAL SUBSPACE DIMENSIONS)

	Y2P	P2Y
FLDA	63.08 (80)	30.90 (55)
LPP	66.91 (80)	24.84 (55)
DLA	61.17 (80)	46.66 (55)
SDA	66.17 (80)	32.12 (45)
MMDE	65.59 (70)	47.94 (20)
CDHE	**87.35 (70)**	**52.73 (25)**
E-CDHE	**91.15 (55)**	**78.50 (15)**

VI. CROSS-DOMAIN WEB IMAGE ANNOTATION

We also conducted experiments on two real-world image annotation databases: NUS-WIDE [21] and MSRA-MM [38]. The NUS-WIDE database contains 269,648 labeled web images with 81 concepts and MSRA-MM database consists of 65,443 labeled web images with 68 concepts collected from the Internet by using Microsoft Live Search. The features used in the experiment for NUS-WIDE are 500-D bag of visual words [27]. For representing images in the MSRA-MM, the dimension of features is 899-D, including (1) 225D block-wise color moment; (2) 64D HSV color histogram; (3) 256D RGB color histogram; (4) 144D color correlogram; (5) 75D edge distribution histogram; (6) 128D wavelet texture; and (7) 7D face features. Example web images from the NUS-WIDE database and MSRA-MM database are shown in Fig. 7 and Fig. 8, respectively. In this experiment, we evaluated the effectiveness of CDHE and E-CDHE for cross-domain web image annotation by comparing against five representative dimension reduction algorithms, i.e., FLDA, LPP, DLA, SDA, and MMDE. We required that training and test samples shared some common properties, or nothing useful could be passed from the training set to the test set. In the experiment on the NUS-WIDE database, animal was applied as the main concept. We selected 12 categories, including bear, bird, cat, cow, dog,

Fig. 7. Sample images under the 'animal' concept (including 12 kinds of animals) from the NUS-WIDE database.

Fig. 8. Sample images from five training classes (tree, waterpark, baseball, party, and military) and five testing classes (plant, hotel, football, medical, and war) in the MSRA-MM database.

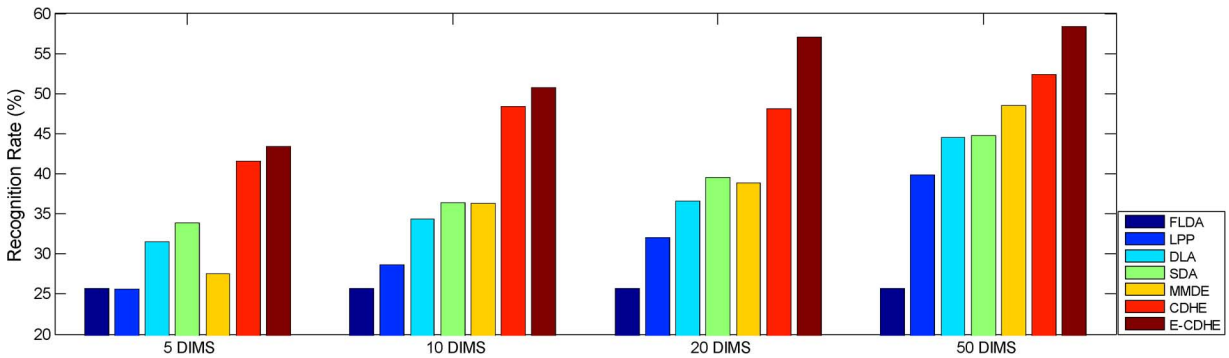


Fig. 9. Recognition rates versus different dimension reduction algorithms under the 5, 10, 20, and 50 dimensions on the NUS-WIDE database.

elk, fish, fox, horse, tiger, whale and zebra. To test the effectiveness of CDHE and E-CDHE for cross-domain learning, we randomly selected six animals for training and used the rest six for testing (for five times). For each animal, 100 images were randomly selected from the dataset. For the experiment on the MSRA-MM database, we used one concept versus one concept method. Specifically, we considered five concepts: tree, waterpark, baseball, party and military as the training classes and utilized the other five relevant concepts: plant, hotel, football, medical and war for testing. For each concept, 100 images were randomly selected from the dataset.

Like the testing stage of the cross-domain face recognition, in our annotation test, we selected a reference image randomly from each category and then apply the nearest-neighbor rule to predict the labels of the remaining test images in the selected subspace. In the training stage of all the dimension reduction algorithms, the labeling information from the reference images was inaccessible.

Fig. 9 compares CDHE and E-CDHE against the other five dimension reduction algorithms on NUS-WIDE database under 4 different dimensions. It uses the boxplot to describe the comparison results. It has four groups, each of which stands for one dimension, i.e., 5, 10, 20, and 50. Each group contains seven boxes, where boxes from left to right are the average accuracies of FLDA, LPP, DLA, SDA, MMDE, CDHE, and E-CDHE, respectively. The figure shows that CDHE and E-CDHE consistently and significantly outperform other dimension reduction algorithms. Furthermore, E-CDHE is more effective than CDHE because the evolutionary search can find a better solution than the gradient descent used in CDHE.

Fig. 10 compares CDHE and E-CDHE against the other five dimension reduction algorithms on MSRA-MM database under four different dimensions. In this figure, we have four groups, which indicate 5, 10, 20, and 50 dimensions. Each group contains seven boxes, where boxes from left to right show the annotation accuracy of FLDA, LPP, DLA, SDA, MMDE, CDHE, and E-CDHE, respectively. This figure shows that CDHE and E-CDHE consistently and significantly outperform other dimension reduction algorithms. In addition, E-CDHE is more effective than CDHE.

VII. Comparison Against Multitask Learning

In this Section, we compare the proposed CDHE and E-CDHE against LatentMTL [41], which is a representative

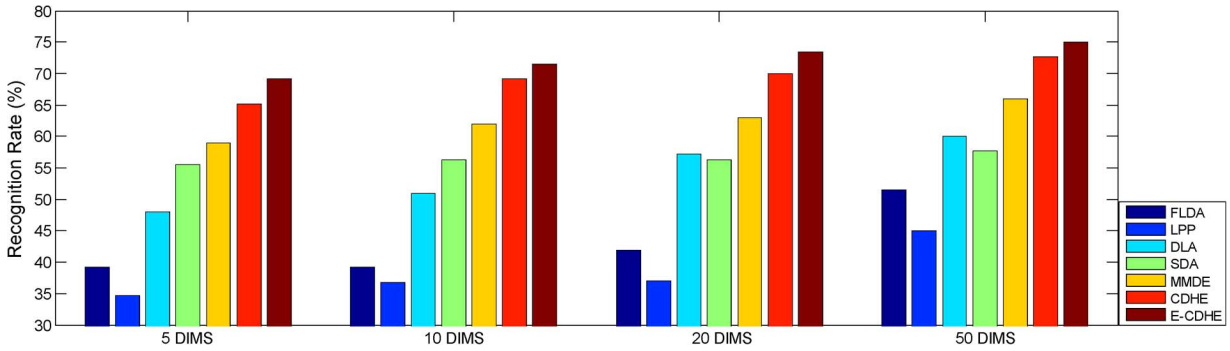


Fig. 10. Recognition rates versus different dimension reduction algorithms under the 5, 10, 20, and 50 dimensions on the MSRA-MM database.

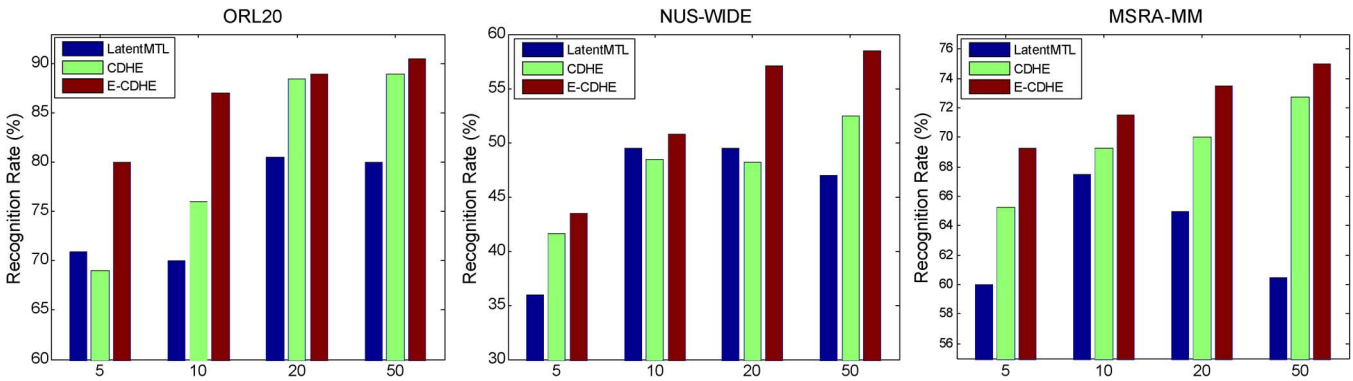


Fig. 11. Recognition rates versus different dimension reduction algorithms under the 5, 10, 20, and 50 dimensions on the ORL 20, NUS-WIDE, and MSRA-MM databases (from left to right) respectively.

multitask learning algorithm. LatentMTL accomplishes the multitask classification in a latent feature space, but it ignores the distribution difference between training and testing domains. Since this method can only handle the situation when the number of training classes is equivalent to that of the testing classes, we compared it against the proposed algorithms on three of the seven databases mentioned above, which are ORL20, NUS-WIDE, and MSRA-MM. We selected a reference image randomly from each class and then apply the nearest-neighbor rule to predict the labels of the remaining test images in the selected subspace. In the training stage of all the dimension reduction algorithms, the labeling information from the reference images was unavailable. Fig. 11 shows that both CDHE and E-CDHE consistently outperform LatentMTL. In addition, E-CDHE is more effective than CDHE.

VIII. Conclusion

In this paper, we have studied the problem of dimension reduction under the cross-domain setting and developed a novel dimension reduction algorithm, termed the cross-domain discriminative Hessian Eigenmaps (CDHE). It passes the discriminative information and the local geometry from the training samples to the test samples by considering the difference between the distribution of the training samples and that of the test samples. Therefore, CDHE reduces the limitation of the traditional dimension reduction algorithms in data distribution assumption and well transfers the discriminative information from the training to the testing domain. To further improve the optimization technique in CDHE (gradient descent) which is susceptible to be trapped in local minima, we have proposed E-CDHE that optimizes the cross-domain learning problem by using a newly proposed evolutionary search strategy. Cross-domain experimental results on face recognition and real-world web image annotation have demonstrated the superiority of the proposed CDHE and E-CDHE in comparison with popular dimension reduction and cross-domain learning algorithms.

In the future, we will exploit the local geometry of unlabelled samples, because the geometry of the unlabelled samples is helpful for learning the optimal shared subspace. In addition, we plan to apply other distance measures to further reduce the computational cost of the proposed CDHE and E-CDHE.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "A unified framework for image retrieval using keyword and visual features," *IEEE Trans. Image Process.*, vol. 14, no. 7, pp. 979–989, Jul. 2005.

[3] X. He, M. Ji, and H. Bao, "Graph embedding with constraints," in *Proc. 21st Int. J. Conf. Artificial Intelligence*, 2009, pp. 1065–1070.

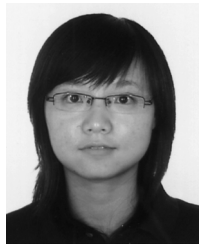
[4] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal Laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.

[5] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, to be published.

[6] M. Wang, X.-S. Hua, T. Mei, R. Hong, G.-J. Qi, Y. Song, and L.-R. Dai, "Semi-supervised kernel density estimation for video annotation," *Comput. Vis. Image Understand.*, vol. 113, no. 3, pp. 384–396, Mar. 2009.

[7] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jan. 1997.

[8] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 862–871.
[9] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multi-level image annotation," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 407–426, Mar. 2008.
[10] X. Li, S. Lin, S. Yan, and D. Xu, "Discriminant locally linear embedding with high-order tensor data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 2, pp. 342–352, Feb. 2008.
[11] S. Si, D. Tao, and K. P. Chan, "Transfer discriminative logmaps," in *Proc. 10th IEEE Int. Conf. Pacific-Rim Conference on Multimedia*, 2009, pp. 131–143.
[12] J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. 23rd AAAI Conf. Artificial Intelligence*, 2008, pp. 677–682.
[13] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE Int. Conf. Computer Vision*, 2007, pp. 1–7.
[14] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Ed. Psych.*, vol. 24, no. 2, pp. 417–441, Jun. 1933.
[15] S. Si, D. Tao, and K. P. Chan, "Cross-domain web image annotation," Int. Workshop on Internet Multimedia Mining.
[16] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Feb. 1936.
[17] X. He and P. Niyogi, "Locality preserving projections," *Adv. Neural Inf. Process. Syst.*, vol. 16, pp. 1–8, 2003.
[18] T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *Proc. 10th Eur. Conf. Computer Vision*, 2008, pp. 725–738.
[19] B. L. Miller and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Complex Syst.*, vol. 9, no. 3, pp. 193–212, Mar. 1995.
[20] T. Bäck and H.-P. Schwefel, "An overview of evolutionary algorithms for parameter optimization," *Evol. Comput.*, vol. 1, pp. 1–23, Jan. 1993.
[21] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image and Video Retrieval*, 2009, pp. 1–8.
[22] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General averaged divergence analysis," in *Proc. 7th IEEE Int. Conf. Data Mining*, 2007, pp. 302–311.
[23] T. Sim, S. Baker, and M. Bsat, The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces, Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep., CMU-RI-TR-01-02, 2001.
[24] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Mar. 1962.
[25] T. Zhang, D. Tao, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.
[26] D. L. Donoho and C. Grimes, "Hessian Eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Arts Sci.*, vol. 100, no. 10, pp. 5591–5596, Oct. 2003.
[27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Feb. 2004.
[28] T. Zhang, D. Tao, X. Li, and T. Yang, "A unifying framework for spectral analysis based dimensionality reduction," in *Proc. IEEE Int. Joint Conf. Neural Networks*, Jun. 2008, pp. 1670–1677.
[29] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, Jan. 2002.
[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, Dec. 2000.
[31] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[32] X. Ling, W. Dai, G. Xue, Q. Yang, and Y. Yu, "Spectral domain-transfer learning," in *Proc. 14th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2008, pp. 488–496.
[33] W. Zheng, W. Xiang, Q. Yang, and D. Shen, "Transferring localization models over time," in *Proc. 23rd AAAI Conf. Artificial Intelligence*, 2008, pp. 1421–1426.
[34] F. Li, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.
[35] W. Liu, D. Tao, and J. Liu, "Transductive component analysis," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 433–442.
[36] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier (Optimization and Neural Computation Series)*. New York: Athena Scientific, 1996.
[37] F. Samaria and A. Harter, "Parameterisation of a Stochastic model for human face identification," in *Proc. 2nd IEEE Workshop on Applications of Computer Vision*, Dec. 1994, pp. 138–142.
[38] H. Li, M. Wang, and X.-S. Hua, "MSRA-MM 2.0: A large-scale web multimedia dataset," Int. Workshop on Internet Multimedia Mining, in Association With ICDM.
[39] T. D. Gwiazda, *Genetic Algorithms Reference Vol.1 Crossover for Single-Objective Numerical Optimization Problems*. Lomianki: Tomasz Gwiazda, 2006.
[40] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, May 2004.
[41] W. Zheng, J. Pan, Q. Yang, and J. J. Pan, "Transferring multi-device localization models using latent multi-task Learning," in *Proc. 23rd AAAI Conf. Artificial Intelligence*, 2008, pp. 1427–1432.
[42] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, June 2003.
[43] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.
[44] X. He, D. Cai, and J. Han, "Learning a maximum margin subspace for image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 2, pp. 189–201, Feb. 2008.
[45] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 17–26.



Si Si received the B.Eng. degree from the University of Science and Technology of China (USTC). She is currently pursuing the M.Phil. degree in the Department of Computer Science, University of Hong Kong.

She is currently an exchange student with the School of Computer Engineering, Nanyang Technological University, Singapore. Her research interests include computer vision and machine learning.

Dacheng Tao (M'07) received the B.Eng. degree from the University of Science and Technology of China, the M.Phil. degree from the Chinese University of Hong Kong, and the Ph.D. degree from the University of London, London, U.K.

Currently, he is a Nanyang Assistant Professor with the School of Computer Engineering, Nanyang Technological University. His research is mainly on applying statistics and mathematics for data analysis problems in computer vision, data mining, machine learning, multimedia, and video surveillance. He has published more than 100 scientific articles extensively in the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, CVPR, ECCV, ICDM, ACM, TKDD, Multimedia, KDD, etc.

Dr. Tao is the recipient of best paper awards and finalists. He holds the K. C. Wong Education Foundation Award of the Chinese Academy of Sciences.

Kwok-Ping Chan (M'95) received the B.Sc. (Eng.) and Ph.D. degrees from the University of Hong Kong.

He is currently an Associate Professor in the Department of Computer Science, University of Hong Kong. His research interest is in Chinese computing, pattern recognition, and machine learning.