Locality Constrained Dictionary Learning for Nonlinear Dimensionality Reduction

Yin Zhou, *Student Member* and Kenneth E. Barner, *Senior Member, IEEE*

***Abstract*—Current nonlinear dimensionality reduction (NLDR) algorithms have quadratic or cubic complexity in the number of data, which limits their ability to process real-world large-scale datasets. Learning over a small set of landmark points can potentially allow much more effective NLDR and make such algorithms scalable to large dataset problems. In this paper, we show that the approximation to an unobservable intrinsic manifold by a few latent points residing on the manifold can be cast in a novel dictionary learning problem over the observation space. This leads to the presented locality constrained dictionary learning (LCDL) algorithm, which effectively learns a compact set of atoms consisting of locality-preserving landmark points on a nonlinear manifold. Experiments comparing state-of-the-art DL algorithms, including K-SVD, LCC and LLC, show that LCDL improves the embedding quality and greatly reduces the complexity of NLDR algorithms.**

***Index Terms*—Dictionary learning, dimensionality reduction, manifold learning, face recognition.**

I. Introduction

MANY computer vision and pattern recognition problems involve high-dimensional large-scale datasets that are computationally expensive to process. Nonlinear dimensionality reduction (NLDR) is an important technique that discovers the most succinct and intrinsic forms of representation of the original high-dimensional data, allowing more effective learning and prediction. Unfortunately, many existing NLDR algorithms are of quadratic or even cubic complexity in the number of data, which diminishes the applicability of these algorithms to real-world large-scale tasks [1]. Efforts have been made on selecting a subset of training data as landmark points on the manifold to improve the efficiency of NLDR algorithms. Landmark points are meaningful points that preserve the local geometric structure of a manifold. In [2], the authors suggest using a subset of randomly selected data points, which, however, may yield a locally optimal solution with poor global performance. Alternatively, [1] proposes utilizing LASSO regression to select landmark points, an approach that has high computational cost due to the required ℓ_1 minimization. The effective learning of landmark points, thus, remains an open challenge.

Sparse representation-based dictionary learning has been proven to be effective in image restoration [3], image denoising [4], [5] and image classification [6]. However, algorithms of this type generally do not ensure locality preservation and thus fails to faithfully depict intrinsic manifold geometry. To address this issue, the approach in [7] approximates nonlinear functions via local coordinate coding. This method is based a modification to ℓ_1 minimization and, as such, has high computational cost. More recently proposed is a locality-constrained linear coding (LLC) approach that favors close samples and suppresses those distant [8]. Moreover, this approach has the advantage of an analytic solution.

In this paper, we show that reconstructing an unobservable intrinsic manifold via a few latent landmark points can be cast, under mild conditions, as a locality constrained dictionary learning problem in the observation space. Utilizing this approach, a novel locality constrained dictionary learning (LCDL) algorithm is introduced. The LCDL algorithm identifies a compact set of landmark points that are simultaneously representational and locality-preserving. Via the landmark points, LCDL naturally embeds training and unseen data onto the intrinsic manifold. Presented results demonstrate that LCDL can significantly improve the performance of NLDR algorithms by yielding a more robust low-dimensional embedding at significantly reduced computational complexity.

II. Problem Formulation

Given an observation set $\{\mathbf{y}_i\}_{i=1}^N$ in $\mathbb{R}^m$, suppose all $\mathbf{y}_i$ reside on a smooth submanifold $\mathcal{M} \subset \mathbb{R}^m$, which is the image of a smooth n-manifold $\mathcal{N}$ under an embedding $f : \mathcal{N} \to \mathbb{R}^m$, where $n \ll m$. f is a diffeomorphism of $\mathcal{N}$ to $\mathcal{M}$ [9].

Let g denote the inverse mapping f^{-1} and let $g(\mathbf{y}_i) \in \mathbb{R}^n$ be the image of $\mathbf{y}_i$ via g located on $\mathcal{N}$. Define $\mathbf{x}_i \in \mathbb{R}^K$ as the local reconstruction code for representing $g(\mathbf{y}_i)$ as a linear combination of K landmarks. Our objective is to learn a codebook of landmark points on $\mathcal{M}$ in the observation space, i.e., $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{m \times K}, (K \ll N)$, such that $g(\mathbf{D})\mathbf{x}_i$ approximates $g(\mathbf{y}_i)$ in terms of ℓ_2 distance, for $i = 1, \ldots, N$. Here $g(\mathbf{D}) = [g(\mathbf{d}_1), \ldots, g(\mathbf{d}_K)] \in \mathbb{R}^{n \times K}$ is a matrix representing the image of the landmarks stored in $\mathbf{D}$ via the inverse mapping g, located on $\mathcal{N}$. Achieving this goal yields much more effective NLDR by learning only $K \ll N$ landmark points, making NLDR algorithms scalable to large dataset problems.

In practice, however, it is often infeasible to recover g due to the facts that: 1) the myriad of observed data causes intractable computation complexity and memory consumption; 2) the intrinsic manifold $\mathcal{N}$ is typically unknown. Without knowing g explicitly, even optimizing $\mathbf{D}$ on $\mathcal{N}$ becomes impractical. We therefore need to establish a relationship between the approximation problem among latent variables (i.e., $g(\mathbf{y}_i)$ and $g(\mathbf{D})$) and the approximation problem among observation variables (i.e., $\mathbf{y}_i$ and $\mathbf{D}$).

As noted by [10], $\mathbf{x}_i$ reflects intrinsic geometric properties of each neighborhood on $\mathcal{N}$ and these properties are expected to be equally valid for local patches on $\mathcal{M}$. We can therefore use

Manuscript received November 27, 2012; revised February 01, 2013; accepted February 05, 2013. Date of publication February 11, 2013; date of current version February 22, 2013. This work was supported by NSF under Grant 0812458. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Michael Rabbat.

The authors are with Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: zhouyin@udel.edu; barner@udel.edu.

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/LSP.2013.2246513

the same set of local reconstruction codes to characterize the local geometric relationships between $g(\mathbf{y}_i)$ and $g(\mathbf{D})$ on $\mathcal{N}$ as to characterize those between $\mathbf{y}_i$ and $\mathbf{D}$ on $\mathcal{M}$.

III. Locality Constrained Dictionary Learning

By requiring $g(\mathbf{D})\mathbf{x}_i$ to approximate $g(\mathbf{y}_i)$ in terms of ℓ_2 distance, for $i = 1, \ldots, N$, we essentially obtain a representational $\mathbf{D}$ such that $\sum_{i=1}^{N} \|g(\mathbf{y}_i) - g(\mathbf{D})\mathbf{x}_i\|_2^2$ is minimized. For symmetry, we enforce $\mathbf{1}^{\mathrm{T}}\mathbf{x}_i = 1$ for all i such that the characterization of local geometry by $\mathbf{x}_i$ is invariant to scaling, rotation and shift of the coordinate system [10], where $\mathbf{1}$ is a column vector of all ones.

Lemma 1: Let $\mathcal{M}$, $\mathcal{N}$ and g be as above. Let $\mathbf{p} \in \mathcal{U}_{\mathbf{p}}$ be an open subset of $\mathcal{M}$ with respect to $\mathbf{p}$, such that $\forall \mathbf{q} \in \mathcal{U}_{\mathbf{p}}$, the line segment $\overline{\mathbf{pq}}$ remains in $\mathcal{U}_{\mathbf{p}}$. If $|\partial g^s/\partial q^t| \leq c$, $1 \leq s \leq n$, $1 \leq t \leq m$, at every $\mathbf{q} \in \mathcal{U}_{\mathbf{p}}$, then we have $\forall \mathbf{q} \in \mathcal{U}_{\mathbf{p}}$ [9]:

$$\|g(\mathbf{q}) - g(\mathbf{p})\|_2^2 \leq mnc^2\|\mathbf{q} - \mathbf{p}\|_2^2. \tag{1}$$

The proof can be derived as a generalization of the mean value theorem and as such we omit the steps in this paper (see [9] for details). Lemma 1 indicates that as $\mathcal{U}_{\mathbf{p}}$ shrinks to be a sufficiently small neighborhood of $\mathbf{p}$, $mnc^2\|\mathbf{q} - \mathbf{p}\|_2^2 \longrightarrow \|g(\mathbf{q}) - g(\mathbf{p})\|_2^2$. We use this observation below.

Theorem 1: Let $g(\mathbf{y}_i)$, $\mathbf{y}_i$, $g(\mathbf{D})$, $\mathbf{D}$ and g be as above. Let $\mathbf{y}_i \in \mathcal{U}_{\mathbf{y}_i}$ and $\mathbf{D}\mathbf{x}_i \in \mathcal{U}_{\mathbf{D}\mathbf{x}_i}$ be open sets as in Lemma 1, that also satisfy $\mathbf{D}\mathbf{x}_i \in \mathcal{U}_{\mathbf{y}_i}$ and $\{\mathbf{d}_j | x_{ji} \neq 0, \forall j\} \subset \mathcal{U}_{\mathbf{D}\mathbf{x}_i}$ $\forall i$. If $\mathbf{1}^{\mathrm{T}}\mathbf{x}_i = 1$ and $\|\mathbf{x}_i\|_0 = \tau$ ($\tau \ll K$) for all i, then the following inequality holds:

$$\sum_{i=1}^{N} \|g(\mathbf{y}_i) - g(\mathbf{D})\mathbf{x}_i\|_2^2 \leq \alpha \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \beta \sum_{i=1}^{N}\sum_{j=1}^{K} \left[x_{ji}^2 \|\mathbf{D}\mathbf{x}_i - \mathbf{d}_j\|_2^2\right] \tag{2}$$

where x_{ji} is the j-th element in vector $\mathbf{x}_i$, $\tau \in \mathbb{Z}^+$, and $\alpha = 2c_1$, $\beta = 2\tau c_2$, with $c_1 = \sup(\{|\partial g^s/\partial q^t| \mid \mathbf{q} \in \mathcal{U}_{\mathbf{y}_i}, \forall i, s, t\})$ and $c_2 = \sup(\{|\partial g^s/\partial q^t| \mid \mathbf{q} \in \mathcal{U}_{\mathbf{D}\mathbf{x}_i}, \forall i, s, t\})$. Note that i exclusively represents the indexes of $\mathbf{y}_i$ and its code $\mathbf{x}_i$ while j only denotes the j-th element in $\mathbf{x}_i$.

Proof: Denote by $\mathbf{Y} \in \mathbb{R}^{m \times N}$ the matrix containing all $\mathbf{y}_i$ and let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{K \times N}$ be the matrix containing N local reconstruction codes. We have

$$\begin{aligned}
&\sum_{i=1}^{N} \|g(\mathbf{y}_i) - g(\mathbf{D})\mathbf{x}_i\|_2^2 \\
&= \|g(\mathbf{Y}) - g(\mathbf{D})\mathbf{X}\|_F^2 \\
&\overset{(a)}{=} \|g(\mathbf{Y}) - g(\mathbf{D}\mathbf{X}) + g(\mathbf{D}\mathbf{X}) - g(\mathbf{D})\mathbf{X}\|_F^2 \\
&\overset{(b)}{\leq} 2\|g(\mathbf{Y}) - g(\mathbf{D}\mathbf{X})\|_F^2 + 2\|g(\mathbf{D}\mathbf{X}) - g(\mathbf{D})\mathbf{X}\|_F^2 \\
&\overset{(c)}{=} 2\sum_{i=1}^{N} \|g(\mathbf{y}_i) - g(\mathbf{D}\mathbf{x}_i)\|_2^2 + \|g(\mathbf{D}\mathbf{x}_i) - g(\mathbf{D})\mathbf{x}_i\|_2^2
\end{aligned} \tag{3}$$

where in (a) $g(\mathbf{D}\mathbf{X}) \in \mathbb{R}^{n \times N}$ is a matrix representing the image of the reconstructed signals $\mathbf{D}\mathbf{X}$ via g; (b) is from Cauchy-Schwarz inequality; in (c) $g(\mathbf{D}\mathbf{x}_i) \in \mathbb{R}^n$ is the i-th column in $g(\mathbf{D}\mathbf{X})$. Since $\mathbf{1}^{\mathrm{T}}\mathbf{x}_i = \sum_{j=1}^{K} x_{ji} = 1$ and $\|\mathbf{x}_i\|_0 = \tau$ for all i, (3) can be written as:

$$\begin{aligned}
&\sum_{i=1}^{N} \|g(\mathbf{y}_i) - g(\mathbf{D})\mathbf{x}_i\|_2^2 \\
&\leq 2\sum_{i=1}^{N} \|g(\mathbf{y}_i) - g(\mathbf{D}\mathbf{x}_i)\|_2^2 + \left\| \sum_{j=1}^{K} x_{ji}[g(\mathbf{D}\mathbf{x}_i) - g(\mathbf{d}_j)] \right\|_2^2 \\
&\leq 2\sum_{i=1}^{N} \|g(\mathbf{y}_i) - g(\mathbf{D}\mathbf{x}_i)\|_2^2 \\
&\quad + 2\tau \sum_{i=1}^{N}\sum_{j=1}^{K} \left[x_{ji}^2 \|g(\mathbf{D}\mathbf{x}_i) - g(\mathbf{d}_j)\|_2^2\right]
\end{aligned} \tag{4}$$

Applying Lemma 1 to each $\|g(\mathbf{y}_i) - g(\mathbf{D}\mathbf{x}_i)\|_2^2$ and to each $[x_{ji}^2\|g(\mathbf{D}\mathbf{x}_i) - g(\mathbf{d}_j)\|_2^2]$ in (4), $\exists$ $c_1 = \sup(\{|\partial g^s/\partial q^t| \mid \mathbf{q} \in \mathcal{U}_{\mathbf{y}_i}, \forall i, s, t\})$ and $c_2 = \sup(\{|\partial g^s/\partial q^t| \mid \mathbf{q} \in \mathcal{U}_{\mathbf{D}\mathbf{x}_i}, \forall i, s, t\})$ such that $2\|g(\mathbf{y}_i) - g(\mathbf{D}\mathbf{x}_i)\|_2^2 \leq 2c_1\|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2$, $\forall i$ and $2\tau[x_{ji}^2\|g(\mathbf{D}\mathbf{x}_i) - g(\mathbf{d}_j)\|_2^2] \leq 2\tau c_2[x_{ji}^2\|\mathbf{D}\mathbf{x}_i - \mathbf{d}_j\|_2^2]$, $\forall i, j$. Letting $\alpha = 2c_1$ and $\beta = 2\tau c_2$ completes the result. ∎

Theorem 1 establishes a relationship between the latent variables and the observation variables by upper-bounding the approximation error on the intrinsic manifold $\mathcal{N}$ (LHS) in terms of the approximation error on $\mathcal{M}$ in the observation space (RHS). As indicated by Lemma 1, when $\mathbf{D}\mathbf{x}_i$ and all $\mathbf{d}_j \in \{\mathbf{d}_j | x_{ji} \neq 0, \forall j\}$ lie within a sufficiently small neighborhood of $\mathbf{y}_i$, the RHS of (2) $\longrightarrow$ the LHS of (2).

By minimizing the RHS of (2) with respect to $\mathbf{D}$ and $\mathbf{x}_i$ for all i, we achieve faithful approximation (1-st term) and secure compact localization (2-nd term), i.e., all $\mathbf{d}_j \in \{\mathbf{d}_j | x_{ji} \neq 0, \forall j\} \longrightarrow \mathbf{D}\mathbf{x}_i \longrightarrow \mathbf{y}_i$, indicating that $\beta\sum_{i=1}^{N}\sum_{j=1}^{K}[x_{ji}^2\|\mathbf{D}\mathbf{x}_i - \mathbf{d}_j\|_2^2] \approx \beta\sum_{i=1}^{N}\sum_{j=1}^{K}[x_{ji}^2\|\mathbf{y}_i - \mathbf{d}_j\|_2^2]$. We therefore formulate the practical LCDL optimization problem as:

$$\min_{\mathbf{D},\mathbf{X}} \quad \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda \sum_{i=1}^{N}\sum_{j=1}^{K} \left[x_{ji}^2 \|\mathbf{y}_i - \mathbf{d}_j\|_2^2\right] + \mu\|\mathbf{X}\|_F^2 \tag{5}$$

$$\text{s.t.} \quad \begin{cases} \mathbf{1}^{\mathrm{T}}\mathbf{x}_i = 1 & \forall i \qquad (*) \\ x_{ji} = 0 & \text{if } \mathbf{d}_j \notin \Omega_\tau(\mathbf{y}_i) \quad \forall i, j \qquad (**) \end{cases}$$

where $\Omega_\tau(\mathbf{y}_i)$ is defined as the τ-neighborhood containing τ nearest neighbors of $\mathbf{y}_i$, and λ, μ are positive regularization constants. $\mu\|\mathbf{X}\|_F^2$ is included for numerical stability of the least—squares solution. The sum-to-one constraint $(*)$ follows from the symmetry requirement, while the locality constraint $(**)$ ensures that $\mathbf{y}_i$ is reconstructed by atoms belonging to its τ-neighborhood, allowing $\mathbf{x}_i$ to characterize the intrinsic local geometry.

IV. Optimization

An iterative process is employed for LCDL optimization. That is, the local reconstruction code $\mathbf{X}$ is optimized first, followed by $\mathbf{D}$. The iterations are repeated, with one aspect held fixed while the other is optimized. The repetition is terminated once either objective function is below a preset threshold or a maximum number of iterations is reached.

A. Solving for Local Reconstruction Codes

Fixing $\mathbf{D}$, which is initialized or set from previous iteration, the i-th column $\mathbf{x}_i \in \mathbf{X}$ is obtained by solving:

$$\min_{\mathbf{x}_i} \quad \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda \sum_{j=1}^{K} \left[x_{ji}^2 \|\mathbf{y}_i - \mathbf{d}_j\|_2^2 \right] + \mu \|\mathbf{x}_i\|_2^2 \tag{6}$$

$$\text{s.t.} \quad \begin{cases} \mathbf{1}^{\mathrm{T}}\mathbf{x}_i = 1 \\ x_{ji} = 0 \qquad \text{if } \mathbf{d}_j \notin \Omega_\tau(\mathbf{y}_i) \quad \forall j \end{cases}$$

Taking both of the constraints into consideration, and using Lagrange multiplier, we obtain

$$\mathcal{L}(\hat{\mathbf{x}}_i, \eta) = \|\mathbf{y}_i - \Omega_\tau \hat{\mathbf{x}}_i\|_2^2 + \lambda \sum_{\mathbf{d}_j \in \Omega_\tau} \left[x_{ji}^2 \|\mathbf{y}_i - \mathbf{d}_j\|_2^2 \right] + \mu \|\hat{\mathbf{x}}_i\|_2^2 + \eta(\mathbf{1}^{\mathrm{T}}\hat{\mathbf{x}}_i - 1) \tag{7}$$

where for simplicity we express $\Omega_\tau(\mathbf{y}_i)$ as $\Omega_\tau \in \mathbb{R}^{m \times \tau}$ and $\hat{\mathbf{x}}_i$ as a succinct subvector containing only τ nonzero elements for those $\mathbf{d}_j \in \Omega_\tau$. Denote as $\mathbf{G} = (\Omega_\tau - \mathbf{y}_j\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}(\Omega_\tau - \mathbf{y}_j\mathbf{1}^{\mathrm{T}})$ the local covariance matrix. Define $\delta(\cdot)$ as the operator that preserves only the diagonal of a square matrix and sets the remaining elements to zero. Thus $\delta(\mathbf{G})$ is a diagonal matrix of size τ-by-τ. For mathematical simplicity, we impose $\mathbf{1}^{\mathrm{T}}\hat{\mathbf{x}}_i = 1$ onto the 1-st term of (7) and get:

$$\mathcal{L}^*(\hat{\mathbf{x}}_i, \eta) = \hat{\mathbf{x}}_i^{\mathrm{T}}(\mathbf{G} + \lambda\delta(\mathbf{G}) + \mu\mathbf{I})\hat{\mathbf{x}}_i + \eta(\mathbf{1}^{\mathrm{T}}\hat{\mathbf{x}}_i - 1), \tag{8}$$

where $\mathbf{I}$ is the identity matrix. Setting $\nabla_{\hat{\mathbf{x}}_i}\mathcal{L}^*(\hat{\mathbf{x}}_i, \eta)$ and $\nabla_\eta \mathcal{L}^*(\hat{\mathbf{x}}_i, \eta)$ to zero, we obtain the solution as

$$\hat{\mathbf{x}}_i = \frac{(\mathbf{G} + \lambda\delta(\mathbf{G}) + \mu\mathbf{I})^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}(\mathbf{G} + \lambda\delta(\mathbf{G}) + \mu\mathbf{I})^{-1}\mathbf{1}} \tag{9}$$

Although the formulations are different, we can still adopt the strategy in [10] to compute $\hat{\mathbf{x}}_i$ efficiently, i.e., first solving the linear system of equations $(\mathbf{G} + \lambda\delta(\mathbf{G}) + \mu\mathbf{I})\hat{\mathbf{x}}_i = \mathbf{1}$ and then normalizing $\hat{\mathbf{x}}_i$ to satisfy the sum-to-one constraint. Note that in contrast with sparse coding algorithms, the proposed coding scheme has an analytic solution and thus is of substantially lower computational complexity.

B. Dictionary Optimization

Having obtained the optimal $\mathbf{X} \in \mathbb{R}^{K \times N}$, we now consider this term fixed and present a procedure for individually optimizing each atom of $\mathbf{D}$. Let $\mathbf{d}_j \in \mathbb{R}^m$ be the j-th atom in $\mathbf{D}$ and define row vector $\mathbf{x}_{j*} \in \mathbb{R}^{1 \times N}$ as the j-th row of $\mathbf{X}$. With $\mathbf{X}$ and all other atoms fixed, we rewrite (5) and cast the optimization problem as

$$\min_{\mathbf{d}_j} \mathcal{H}(\mathbf{d}_j) = \left\| \mathbf{Y} - \sum_{k \neq j} \mathbf{d}_k \mathbf{x}_{k*} - \mathbf{d}_j \mathbf{x}_{j*} \right\|_F^2 + \lambda \left\{ \sum_{i=1}^{N} \left[x_{ji}^2 \|\mathbf{y}_i - \mathbf{d}_j\|_2^2 \right] + \sum_{i=1}^{N} \sum_{k \neq j}^{K} \left[x_{ki}^2 \|\mathbf{y}_i - \mathbf{d}_k\|_2^2 \right] \right\} \tag{10}$$

Setting $\mathbf{E} = \mathbf{Y} - \sum_{k \neq j} \mathbf{d}_k \mathbf{x}_{k*}$ and eliminating irrelevant terms, (10) is simplified to

$$\min_{\mathbf{d}_j} \quad \mathcal{H}(\mathbf{d}_j) = \mathrm{Tr}\left\{ (\mathbf{E} - \mathbf{d}_j\mathbf{x}_{j*})(\mathbf{E} - \mathbf{d}_j\mathbf{x}_{j*})^{\mathrm{T}} \right\} + \lambda \sum_{i=1}^{N} \left[x_{ji}^2 (\mathbf{y}_i - \mathbf{d}_j)^{\mathrm{T}}(\mathbf{y}_i - \mathbf{d}_j) \right] \tag{11}$$

Since $\mathcal{H}(\mathbf{d}_j)$ is convex, setting the gradient of $\mathcal{H}(\mathbf{d}_j)$ with respect to $\mathbf{d}_j$ to zero yields the optimal solution

$$\mathbf{d}_j = \frac{1}{(1+\lambda)\left(\mathbf{x}_{j*}\mathbf{x}_{j*}^{\mathrm{T}}\right)} \left(\mathbf{E}\mathbf{x}_{j*}^{\mathrm{T}} + \lambda \mathbf{Y}\alpha \right) \tag{12}$$

where $\alpha = [x_{j1}^2, \ldots, x_{jN}^2]^{\mathrm{T}} \in \mathbb{R}^N$ is a column vector with terms the squared values of those in $\mathbf{x}_{j*}^{\mathrm{T}}$.

Discussion: In the framework of LCDL, the mapping g can be found by any NLDR algorithm. NLDR algorithms, however, require complexity $O(mN^2)$ or $O(mN^3)$ in time (depending on the specific formulation utilized) and $O(N^2)$ in space, when operating on the full set of data. Utilizing a landmark points approach greatly reduces the NLDR complexity to $O(mK^2)$ or $O(mK^3)$ in time, again depending on the formulation utilized, and $O(K^2)$ in space, where $K \ll N$. The LCDL time complexity, for computing a single $\mathbf{x}_i$, is $O(\tau mK) + O(m\tau^3)$, which is dominated by $O(\tau mK)$ as $\tau^3 \ll K$. Additionally, the LCDL optimizing, for each $\mathbf{d}_j$, has time complexity $O(mN)$. The overall asymptotic complexity of LCDL is $O(\tau mKN) + O(mNK)$ per iteration. Though the convergence speed is task-dependent, the convergence to a local minimum is guaranteed and in our experiments 15 iterations are typically sufficient to achieve satisfactory results. When N is large, the LCDL complexity is negligible compared to that of NLDR. Thus LCDL, by efficiently establishing a faithful embedding and reconstruction representations, can significantly reduce the time and space complexity of NLDR algorithms, especially for large-scale datasets.

V. Experimental Results

A. Synthetic Datasets

The proposed LCDL is evaluated by measuring the root mean square error (RMSE) introduced through the reconstruction of an intrinsic manifold $\mathcal{N}$, i.e., $\|g(\mathbf{Y}) - g(\mathbf{D})\mathbf{X}\|_F/\sqrt{N}$. LCDL is compared with three state-of-the-art DL algorithms, K-SVD [4], LCC [7] and the recently proposed LLC [8]. Note that $g(\mathbf{Y})$ and $g(\mathbf{D})$ are the low-dimensional embedding of training data and landmark points, respectively, computed via the NLDR algorithm, where $g(\mathbf{Y})$ is employed as the ground truth. Also, $\mathbf{X}$ is computed according to (9). For each synthetic dataset, $N = 3000$ training data are randomly generated, among which K samples are randomly selected for initialization. We set $K = 500, 200$, and 100 for the Swiss roll, Punctured sphere and Gaussian manifold, respectively. The NLDR algorithms are Hessian LLE [11], Laplacian Eigenmap [12], and LLE [10] for these three manifolds. We restrict training samples to be reconstructed by 2 atoms, as the intrinsic manifolds are 2D. The visualization and RMSE of the reconstructed low-dimensional manifolds are illustrated in Fig. 1. LCDL outperforms other competitive methods, yielding the closet approximation to the ground truth in all cases.

B. Face Recognition

Consider next the effectiveness in classification of the reconstructed low-dimensional manifolds produced by LCDL, with effectiveness determined through comparisons to the aforenoted methods. Though this paper only reports classification results using LLE, drawn conclusions can be generalized to other NLDR algorithms.

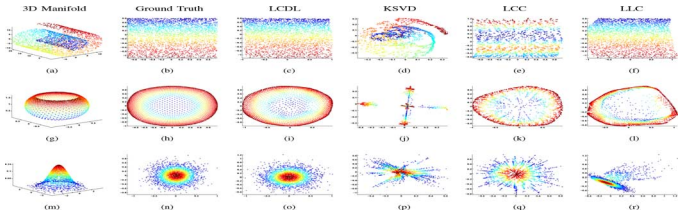

Fig. 1. Low-dimensional embedding reconstruction comparison on Swiss roll (1st row), Punctured sphere (2nd row) and Gaussian (3rd row). Ground truth is the low-dimensional embedding obtained from all training samples. The NLDR nearest neighbor $k = 6$. The RMSE values are (c) 0.0299, (d) 0.7409, (e) 0.0666, (f) 0.0535, (i) 0.0705, (j) 0.8664, (k) 0.1060, (l) 0.1743, (o) 0.0076, (p) 0.2774, (q) 0.0727, (r) 0.0380

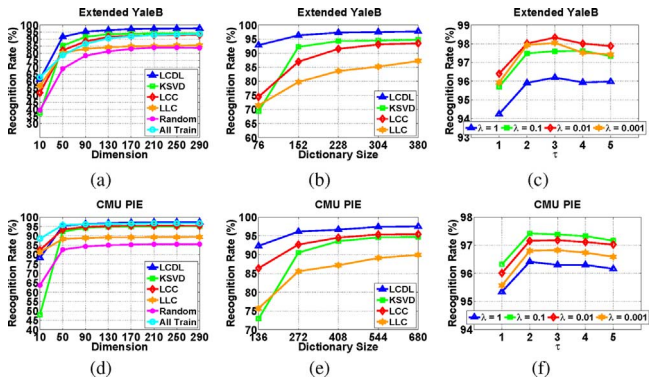

Fig. 2. Classification results over two face databases. The parameter k of LLE is set to 60 for both Extended YaleB and CMU PIE.

The Extended YaleB Database [13] contains 2414 face images of 38 subjects. For each subject, we randomly select half of the images (about 32 per person) for training and the other half for testing. As in [14], we use a subset of the CMU PIE Database [15], i.e., C05, C07, C09, C27, and C29, which yields a total 11554 images of 68 subjects. Following [14], a random selection of 130 images per person is employed to form the training set and the rest of the database is designated for testing. All images are normalized to 32×32 pixels and preprocessed by histogram equalization. The nearest neighbor classifier is employed and the results averaged over 10 repetitions are reported. The LCDL parameters are set as $\lambda = 0.1$, $\mu = 0.001$, and $\tau = 2$.

For all algorithms, a structured dictionary is learned as $\mathbf{D} = [\mathbf{D}_1|\mathbf{D}_2|\ldots|\mathbf{D}_C]$, where $\mathbf{D}_i$ is the sub-dictionary for class i. As in [6], we fix the number of atoms per class to be 8, yielding a dictionary of 304 atoms for the Extended YaleB Database and a dictionary of 544 atoms for the CMU PIE Database. The sparsity factor is set, through exhaustive search optimization, to 24 and 32 for K-SVD over the two databases, respectively. All Train is selected as the baseline method, which represents the results obtained in performing LLE on the entire training set. Random means employing randomly selected training samples as the dictionary. The recognition rates versus dimension for all methods are illustrated in Fig. 2(a) and (d). The proposed LCDL achieves the highest accuracies, 97.5% on the Extended YaleB Database and 97.4% on the CMU PIE Database. LCDL outperforms All Train since the optimization ameliorates the noise and outlier effects within the training data, which leads to more robust dimensionality reduction.

In addition, we evaluate the proposed approach by fixing the dimension and varying the number of atoms per class from 2 to 10. As shown in Fig. 2(b) and (e), LCDL consistently produces higher accuracy than competing algorithms across a range of dictionary sizes. This results from the fact that LCDL establishes a dictionary that is both representational and locality preserving. Moreover, we examine the impact τ and λ have on LCDL performance. As shown in Fig. 2(c) and (f), LCDL maintains higher recognition rate than other methods over a wide range of τ and λ, indicating that performance is relatively robust to parameter value selections.

TABLE I
THE OVERALL TIME (SECONDS) INCLUDES DICTIONARY LEARNING AND TRAINING DATA EMBEDDING. NOTE THE TIME MEASUREMENT MAY VARY BASED ON DIFFERENT IMPLEMENTATIONS

	Extended YaleB		CMU PIE	
	Overall Time	Speedup	Overall Time	Speedup
All Train	22.1577 s	No	11807.3121 s	No
K-SVD [4]	71.2387 s	No	2751.2620 s	4.3x
LCC [7]	38.7172 s	No	1299.7146 s	9.1x
LLC [8]	11.6593 s	1.9x	69.2321 s	170.5x
LCDL	7.1001 s	3.1x	45.8025 s	257.8x

Finally, we evaluate the implementation efficiency of LCDL by measuring the speedup in terms of the overall training time compared to the All Train baseline, Table I. The results show that LCDL is more efficient than comparison methods and significantly improves the learning efficiency of LLE by more than 2 orders of magnitude over the CMU PIE Database.

VI. CONCLUSION AND FUTURE WORK

We propose a novel algorithm, LCDL, that learns dictionary atoms as landmark points which are simultaneously representational and locality preserving. Experiments demonstrate that LCDL is superior to existing dictionary learning algorithms in terms of yielding more meaningful atoms for NLDR algorithms with greatly reduced computational complexity. Future research will consider incorporating a sparse outlier term to improve robustness and testing over additional datasets.

REFERENCES

[1] J. Silva, J. Marques, and J. Lemos, "Selecting landmark points for sparse manifold learning," in *NIPS*, 2006, vol. 18, 1.
[2] V. de Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *NIPS*, 2002, 1.
[3] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, pp. 53–69, 2008, 1.
[4] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionarys for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, 2006, 1, 3, 4.
[5] K. Engan, S. Aase, and J. Husoy, "Frame based signal compression using method of optimal directions (mod)," in *ISCAS'99*.
[6] Q. Zhang and B. Li, "Discriminative k-SVD for dictionary learning in face recognition," in *CVPR*, 2010, 1, 4.
[7] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *NIPS09*, 1, 3, 4.
[8] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, 1, 3, 4.
[9] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd ed. New York, NY, USA: Academic, 2003.
[10] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000, 1, 2, 3.
[11] D. L. Donoho and C. Grimes, "Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-Dimensional Data," 2003, Stanford Univ., Stanford, CA.
[12] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.* 2003.
[13] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Patt Anal. Mach. Intell.*, 2005.
[14] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *ICCV 2007*, 2007, pp. 1–7, 3.
[15] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Trans. Patt Anal. Mach. Intell.*, pp. 1615–1618, 2003.