Neurocomputing 99 (2013) 448–457

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Enhanced and parameterless Locality Preserving Projections for face recognition

Fadi Dornaika [a,b,*], Ammar Assoum [c]

[a] Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Spain
[b] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[c] LaMA Laboratory, Lebanese University, Tripoli, Lebanon

ARTICLE INFO

*Article history:*
Received 25 April 2011
Received in revised form
30 April 2012
Accepted 18 July 2012
Communicated by M.S. Bartlett
Available online 13 August 2012

*Keywords:*
Dimensionality reduction
Graph-based linear embedding
Locality Preserving Projections
Face recognition

ABSTRACT

In this paper, we address the graph-based linear manifold learning method for object recognition. The proposed method is called enhanced Locality Preserving Projections. The main contribution is a parameterless computation of the affinity matrix that draws on the notion of meaningful and adaptive neighbors. It integrates two interesting properties: (i) being entirely parameter-free and (ii) the mapped data are uncorrelated. The proposed method has been integrated in the framework of three graph-based embedding techniques: Locality Preserving Projections (LPP), Orthogonal Locality Preserving Projections (OLPP), and supervised LPP (SLPP). Recognition tasks on six public face databases show an improvement over the results of LPP, OLPP, and SLPP. The proposed approach could also be applied to other category of objects.



## 1. Introduction

In most computer vision and pattern recognition problems, the large number of sensory inputs, such as images and videos, are computationally challenging to analyze. In such cases it is desirable to reduce the dimensionality of the data while preserving the original information in the data distribution, allowing for more efficient learning and inference. If the variance of the multivariate data is faithfully represented as a set of parameters, the data can be considered as a set of geometrically related points lying on a smooth low-dimensional manifold. The fundamental issue in dimensionality reduction is how to model the geometry structure of the manifold and produce a faithful embedding for data projection. During the last few years, a large number of approaches have been proposed for computing the embedding. We categorize these methods by their linearity. The linear methods, such as principal component analysis (PCA) [14], multidimensional scaling (MDS) [2], are evidently effective in observing the Euclidean structure. Unlike PCA which is unsupervised, linear discriminant analysis (LDA) [4] is a supervised technique. One limitation of PCA and LDA is that they effectively see only the linear global Euclidean structure. However, some recent research shows that the samples may reside on a nonlinear submanifold, which makes PCA and LDA inefficient.

The nonlinear methods such as locally linear embedding (LLE) [10], Laplacian eigenmaps [1] focus on preserving the local structures. Isomap [13] attempts to preserve the geodesic distances on the manifold. Principal component analysis (PCA) projects the samples along the directions of maximal variances and aims to preserve the Euclidean distances between the samples. There is considerable interest in geometrically motivated approaches to visual analysis. Various researchers (see [10,13,12]) have considered the case when the data lives on or close to a low dimensional submanifold of the high dimensional ambient space. One hopes then to estimate geometrical and topological properties of the submanifold from random points lying on this unknown submanifold. Maximum variance unfolding (MVU) [15] is a global algorithm for nonlinear dimensionality reduction, in which all the data pairs, nearby and far, are considered. MVU attempts to unfold a data set by pulling the input patterns as far apart as possible subject to the constraints that distances and angles between neighboring points are strictly preserved.

Linear dimensionality reduction (LDR) techniques have been increasingly important in pattern recognition [9,17] since they permit a relatively simple mapping of data onto a lower-dimensional subspace, leading to simple and computationally efficient classification strategies. The main advantage of the linear methods over the non-linear ones is that the embedding function of the linear techniques is defined everywhere in the input space, while

* Corresponding author at: Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Spain.
*E-mail address:* fdornaika@hotmail.fr (F. Dornaika).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.07.016

for non-linear embedding techniques, it is only defined for a set of data samples. Many linear dimensionality reduction techniques can be derived from a graph whose nodes represent the data samples and whose edges quantify the similarity among pairs of samples [11]. Recently, a graph-based method was proposed for linear dimensionality reduction (LDR). It is based on Locality Preserving Projections (LPP) [5,8,18]. LPP is a typical linear graph-based dimensionality reduction (DR) method that has been successfully applied in many practical problems such as face recognition. LPP is essentially a linearized version of Laplacian Eigenmaps [1]. The main applications of LPP are data visualization, data dimensionality reduction, and object recognition. When dealing with face recognition problems, LPP is preceded by a principal component analysis (PCA) step in order to avoid possible singularities.

In the literature, many extensions have been proposed. In [16], schemes have been introduced to improve the original LPP by (i) introducing linear transforms before performing the LPP and (ii) changing the quotient criterion to a difference criterion. While LPP is a linear unsupervised technique, it preserves the locality structures of data better than PCA. Furthermore, it is shown that in some cases it can give better results than the supervised technique LDA [6]. Orthogonal LPP (OLPP) [3] was recently proposed as an extension of LPP. OLPP provides orthogonal projection directions using the LPP criterion. Ref. [8] is to solve the singularity of the matrix $XDX^T$ of LPP.

Although the LPP framework can effectively preserve the manifold structure of the input data, its discriminability between different classes is little because the label information is neglected during the estimation of the linear transform. Therefore, a supervised LPP (SLPP) is proposed to overcome this limitation [19]. In SLPP, the affinity matrix $\mathbf{A}$ is computed with the constraint that each point's K nearest neighbors must be chosen from the samples with the same class label as its.

In this paper, we propose a novel LPP that integrates two interesting properties: (i) being entirely parameter-free and (ii) the mapped data are uncorrelated. The paper is organized as follows. Section 1 summarizes the linear mapping of Locality Preserving Projections (LPP). Section 2 describes the proposed parameterless and enhanced LPP. Section 3 presents experimental results for face recognition using five face databases.

## 2. Locality Preserving Projections

We assume that we have a set of $N$ samples $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^D$. Define a neighborhood graph on these data, such as a K-nearest-neighbor or $\epsilon$-ball graph, or a full mesh, and weigh each edge $\mathbf{x}_i \sim \mathbf{x}_j$ by a symmetric affinity function $A_{ij} = K(\mathbf{x}_i; \mathbf{x}_j)$, typically Gaussian, i.e.,

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

where $\beta$ is usually set to the average of squared distances between all pairs. Let $\mathbf{A}$ denotes the symmetric affinity matrix whose elements are defined by $A_{ij}$.

We seek latent points $\{\mathbf{y}_i\}_{i=1}^N \subset \mathbb{R}^L$ that minimize $\frac{1}{2}\sum_{i,j}\|\mathbf{y}_i - \mathbf{y}_j\|^2 A_{ij}$, which discourages placing far apart latent points that correspond to similar observed points. For the purpose of presentation simplicity, we present the one dimensional mapping case in which the original data set $\{\mathbf{x}_i\}_{i=1}^N$ is mapped to a line.

Let $\mathbf{z} = (y_1, y_2, \ldots, y_N)^T$ be such a map (a column vector). Note that here every data sample is mapped to a real value. A reasonable criterion for choosing a "good" map is to optimize the following objective function under some constraints:

$$\min \frac{1}{2}\sum_{i,j}(y_i - y_j)^2 A_{ij} \tag{1}$$

Minimizing function (1) imposes a heavy penalty if neighboring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. By simple algebra formulation, function (1) can be written as

$$\frac{1}{2}\sum_{i,j}(y_i - y_j)^2 A_{ij} = \mathbf{z}^T\mathbf{D}\mathbf{z} - \mathbf{z}^T\mathbf{A}\mathbf{z} = \mathbf{z}^T\mathbf{L}\mathbf{z} \tag{2}$$

where $\mathbf{D}$ is the diagonal weight matrix, whose entries are column (or row, since $\mathbf{A}$ is symmetric) sums of $\mathbf{A}$, and $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the Laplacian matrix.

In the LPP formulation, the latent data are simply given by a linear mapping of the original data. Thus, the one dimensional map $\mathbf{z}$ (column vector) is giving by

$$\mathbf{z} = \mathbf{X}^T\mathbf{w} \tag{3}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ is the data matrix, and $\mathbf{w}$ is a projection direction. Finally, by combining (3) and (2) and by imposing the constraint $\mathbf{z}^T\mathbf{D}\mathbf{z} = 1$ for setting an arbitrary scale, the minimization problem reduces to the finding:

$$\min_{\mathbf{w}} \mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} \quad \text{s.t. } \mathbf{w}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{w} = 1 \tag{4}$$

The transformation vector $\mathbf{w}$ that minimizes the objective function is given by the minimum eigenvalue solution to the generalized eigenvalue problem:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{w} \tag{5}$$

For a multidimensional mapping, each data sample $\mathbf{x}_i$ is mapped into a vector $\mathbf{y}_i$. The aim is to compute the projection directions ($\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_L)$). These vectors are given by the generalized eigenvectors of (5), ordered according to their eigenvalues, $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_L$. Then, the mapping of $\mathbf{x}$ is given by $\mathbf{y} = \mathbf{W}^T\mathbf{x}$.

In many real world problems such as face recognition, the dimensionality of the sample, $D$, is usually larger than the number of the samples, $N$, and the generalized eigen-equation cannot be directly solved due to the matrix singularity problem. In such cases both matrices $\mathbf{X}\mathbf{D}\mathbf{X}^T$ and $\mathbf{X}\mathbf{L}\mathbf{X}^T$ are singular. This problem is also referred to as the Small Sample Size (SSS) problem.

To overcome the complication of singular matrices, original data are first projected to a PCA subspace or a random orthogonal space so that the resulting matrix $\mathbf{X}\mathbf{D}\mathbf{X}^T$ is non-singular. The global transform is given by $\mathbf{W} = \mathbf{W}_{PCA}\mathbf{W}_{LPP}$ or $\mathbf{W} = \mathbf{W}_{rand}\mathbf{W}_{LPP}$ where $\mathbf{W}_{rand}$ is an orthogonal random matrix with enough columns. The use of random projection has been proved to give equivalent performance to PCA yet with the obvious advantage that the corresponding transform does not need a training set.

*Shortcomings of LPP*: The classical LPP has two shortcomings. The first shortcoming concerns the selection of parameters. Indeed, the computation of the affinity matrix $\mathbf{A}$ needs the setting of two parameters: (i) the width of the Gaussian Kernel and (ii) the size of neighborhood for non-full mesh graphs. In practice, the width parameter, $\beta$, is set to the average of pairwise distances over the training set. However, this heuristic does not necessarily provide the optimal width value for a given data set. On the other hand, the neighborhood size has an impact on the learning of the manifold, and was usually set in advance to the same value for all samples. The second shortcoming concerns the eigenvectors that are solved from the generalized eigenvalue problem of (5), these ones are generally statistically correlated and contain some redundancy information.