Out-of-Sample Embedding by Sparse Representation

Bogdan Raducanu[1] and Fadi Dornaika[2,3]

[1] Computer Vision Center, 08193 Bellaterra, Spain
[2] University of the Basque Country (UPV/EHU), San Sebastian, Spain
[3] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Abstract. A critical aspect of non-linear dimensionality reduction techniques is represented by the construction of the adjacency graph. The difficulty resides in finding the optimal parameters, a process which, in general, is heuristically driven. Recently, sparse representation has been proposed as a non-parametric solution to overcome this problem. In this paper, we demonstrate that this approach not only serves for the graph construction, but also represents an efficient and accurate alternative for out-of-sample embedding. Considering for a case study the Laplacian Eigenmaps, we applied our method to the face recognition problem. Experimental results conducted on some challenging datasets confirmed the robustness of our approach and its superiority when compared to existing techniques.

1 Introduction

In recent years, a new family of non-linear dimensionality reduction techniques for manifold learning has emerged. The most known ones are: Kernel Principal Component Analysis (KPCA) [1], Locally Linear Embedding (LLE) [2, 3], Isomap [4], Supervised Isomap [5], Laplacian Eigenmaps (LE)[6, 7]. This family of non-linear embedding techniques appeared as an alternative to their linear counterparts which suffer of severe limitation when dealing with real-world data: i) they assume the data lie in an Euclidean space and ii) they may fail to get a faithful representation of data distribution when the number of samples is too small. On the other hand, the non-linear dimensionality techniques are able to discover the intrinsic data structure by exploiting the local topology. In general, they attempt to optimally preserve the local geometry around each data sample while using the rest of the samples to preserve the global structure of the data.

The non-linear embedding approaches model the structure of data by preserving some geometrical property of the underlying manifold. For instance, while the Isomap method attempts to maintain global properties, LE and LLE aim at preserving local geometry which implicitly tends to keep the global layout of the data manifold.

An inherent limitation of these approaches is that they do not provide an explicit mapping function between low and high dimensional spaces. Such function is essential for ensuring continuity of low dimensional representation and projecting data between spaces. This issue has been addressed quite satisfactorily by applying Radial Basis Function network to approximate the optimal mapping function [8]. However, the quality of RBFN relies on the careful selection of a few parameters which are chosen empirically. In [9], the authors cast MDS, ISOMAP, LLE, and LE in a common

framework, in which these methods are seen as learning eigenfunctions of a kernel. The authors try to generalize the dimensionality reduction results for the unseen data samples.

Due to this limitation, the 'out-of-sample' problem (projection of unseen samples on the embedded space) is not a straightforward process and it is less intuitive than in the case of linear manifolds. For this reason, it hasn't received too much attention so far. In this paper, we adopt the sparse representation approach as an optimal solution to the 'out-of-sample' problem. In the past, it was used as an efficient alternative [10] to the parametric construction of the adjacency graph. Without any loss of generality, we chose the Laplacian Eigenmaps as one of the non-linear dimensionality reduction techniques to test our method.

The paper is structured as follows. In section 2, we briefly review the Laplacian Eigenmaps. In section 3, we introduce our proposed approach for the out-of-sample problem based on sparse representation. Section 4 contains the experimental results. We evaluate the performance of proposed out-of-sample method for the face recognition problem. Finally, in section 5 we present our conclusions and provide the guidelines for future work.

2 Review of Laplacian Eigenmaps

Laplacian Eigenmaps is a recent non-linear dimensionality reduction technique that aims to preserve the local structure of data [6]. Using the notion of the Laplacian of the graph, this non-supervised algorithm computes a low-dimensional representation of the data set by optimally preserving local neighborhood information in a certain sense. We assume that we have a set of N samples $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^D$. Let's define a neighborhood graph on these samples, such as a K-nearest-neighbor or ϵ-ball graph, or a full mesh, and weigh each edge $\mathbf{x}_i \sim \mathbf{x}_j$ by a symmetric affinity function $W_{ij} = K(\mathbf{x}_i; \mathbf{x}_j)$, typically Gaussian:

$$W_{ij} = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}) \tag{1}$$

where β is suitable positive scalar. It is usually set to the average of squared distances between all pairs.

LE seeks latent points $\{\mathbf{y}_i\}_{i=1}^N \subset \mathbb{R}^L$ that minimize $\frac{1}{2}\sum_{i,j}\|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij}$, which discourages placing far apart latent points that correspond to similar observed points. If $\mathbf{W} \equiv W_{ij}$ denotes the symmetric affinity matrix and $\mathbf{D}$ is the diagonal weight matrix, whose entries are column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$, then the Laplacian matrix is given $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The objective function can also be written as:

$$\frac{1}{2}\sum_{i,j}\|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} = tr(\mathbf{Z}^T\,\mathbf{L}\,\mathbf{Z}) \tag{2}$$

where $\mathbf{Z}^T = \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ is the $L \times N$ embedding matrix and $tr(.)$ denotes the trace of a matrix. The i^{th} row of the matrix $\mathbf{Z}$ provides the vector $\mathbf{y}_i$—the embedding coordinates of the sample $\mathbf{x}_i$.

The embedding matrix $\mathbf{Z}$ is the solution of the optimization problem:

$$\min_{\mathbf{Z}} tr(\mathbf{Z}^T \mathbf{L} \mathbf{Z}) \;\; s.t. \;\; \mathbf{Z}^T \mathbf{D} \mathbf{Z} = \mathbf{I}, \;\; \mathbf{Z}^T \mathbf{L} \mathbf{1} = \mathbf{0} \tag{3}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1} = (1, \ldots, 1)^T$. The first constraint eliminates the trivial solution $\mathbf{Z} = \mathbf{0}$ (by setting an arbitrary scale) and the second constraint eliminates the trivial solution $\mathbf{1}$ (all samples are mapped to the same point). Standard methods show that the embedding matrix is provided by the matrix of eigenvectors corresponding to the smallest eigenvalues of the generalized eigenvector problem,

$$\mathbf{L} \mathbf{z} = \lambda \mathbf{D} \mathbf{z} \tag{4}$$

Let the column vectors $\mathbf{z}_0, \ldots, \mathbf{z}_{N-1}$ be the solutions of (4), ordered according to their eigenvalues, $\lambda_0 = 0 \leq \lambda_1 \leq \ldots \leq \lambda_{N-1}$. The eigenvector corresponding to eigenvalue 0 is left out and only the next eigenvectors for embedding are used. The embedding of the original samples is given by the row vectors of the matrix $\mathbf{Z}$, that is, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] = \mathbf{Z}^T$.

$$\mathbf{x}_i \longrightarrow \mathbf{y}_i = (z_1(i), \ldots, z_L(i))^T \tag{5}$$

where $L < N$ is the dimension of the new space.

From equation (4), we can observe that the dimensionality of the subspace obtained by LE is limited by the number of samples N.

3 Proposed out-of-sample embedding

3.1 Projection of new samples

Assume we have obtained a LE embedding $\mathbf{Y}_s = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ of seen samples $\mathbf{X}_s = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and consider unseen (out-of-sample) sample in observed space $\mathbf{x}_{N+1}$. The natural way to embed the new sample would be to recompute the whole embedding $(\mathbf{Y}_s, \mathbf{y}_{N+1})$ for $(\mathbf{X}_s, \mathbf{x}_{N+1})$ from Eq. (3). This is computationally costly and does not lead to defining a mapping for new samples; we seek a way of keeping the old embedding fixed and embed new sample based on that. Then, the next most natural way is to recompute the embedding but keeping the old embedded samples fixed and imposing that the embedding of the new sample (vector $\mathbf{y}_{N+1}$) should minimize the following target function:

$$\sum_{i=1}^{N} \|\mathbf{y}_{N+1} - \mathbf{y}_i\|^2 \, W_{(N+1)i} \tag{6}$$

$$\sum_{i=1}^{N} (\mathbf{y}_{N+1} - \mathbf{y}_i)^T (\mathbf{y}_{N+1} - \mathbf{y}_i) W_{(N+1)i} \tag{7}$$

The above should correspond to a minimum, and thus the derivative with respect to $\mathbf{y}_{N+1}$ of the target function should vanish:

$$2\sum_{i=1}^{N}(\mathbf{y}_{N+1}-\mathbf{y}_i)W_{(N+1)i}=0 \tag{8}$$

From the above, we can conclude that the embedding $\mathbf{y}_{N+1}$ is given by:

$$\mathbf{y}_{N+1}=\frac{\sum_{i=1}^{N}W_{(N+1)i}\,\mathbf{y}_i}{\sum_{i=1}^{N}W_{(N+1)i}} \tag{9}$$

The above formula stipulates that the embedding of an unseen sample is simply the linear combination of all fixed embedded samples where the linear coefficients are set to the similarity between the unseen sample and the existing sample.

Whenever $W_{(N+1)i}$ is set to a Kernel function (i.e., $W_{(N+1)i}=K(\mathbf{x}_{N+1},\mathbf{x}_i)$, Eq. (9) is equivalent to the Laplacian Eigenmaps Latent Variable Model (LELVM) introduced in [11].

3.2 Computation of the similarity coefficients via Sparse Representation

The problem of out-of-sample embedding boils down to the estimation of the similarities $W_{(N+1)i}, i=1,\ldots,N$. In [11], these $W_{(N+1)i}$ were computed using a K nearest neighbor and a Heat Kernel. However, it is well known that the neighborhood size as well as the Kernel parameter may affect the embedding process. We will bypass this limitation by using the coding provided by sparse representation.

In traditional graph construction process, the graph adjacency structure and the graph weights are derived separately. It was argued that the graph adjacency structure and the graph weights are interrelated and should not be separated. Thus it is desired to develop a procedure which can simultaneously completes these two tasks within one step. In [10], the authors proposed to simultaneously build the adjacency graph and its weights. To this end, they used the sparse representation of each training sample as a linear superposition of basis functions (rest of the training samples) plus the noise.

We apply the sparse coding/representation principle for computing the set of coefficients $W_{(N+1)i}$. Let the vector $\mathbf{a}=(W_{(N+1)1},W_{(N+1)2},\ldots,W_{(N+1)N})^T$. Thus, the objective is to compute the vector $\mathbf{a}$ given the unseen sample and the training data. Based on sparse coding, the unseen sample $\mathbf{x}_{N+1}$ can be written as

$$\mathbf{x}_{N+1}=\sum_{i=1}^{N}a_i\,\mathbf{x}_i+\mathbf{e}=\mathbf{X}\,\mathbf{a}+\mathbf{e} \tag{10}$$

The goal is to minimize both the reconstruction error and the L_1 norm of the vector $\mathbf{a}$:

$$\min_{\mathbf{a},\mathbf{e}}\left(\|\mathbf{a}\|_{L_1}+\|\mathbf{e}\|_{L_1}\right)\quad s.t.\quad \mathbf{x}_{N+1}=\mathbf{X}\,\mathbf{a}+\mathbf{e} \tag{11}$$

Let $\mathbf{a}'$ denote the vector $\mathbf{a}'=(\mathbf{a}^T,\mathbf{e}^T)^T$ and $\mathbf{I}$ denote the $D\times D$ identity matrix, then the objective function (11) can be written as:

$$\min\|\mathbf{a}'\|_{L_1}\quad s.t.\quad [\mathbf{X}\ \mathbf{I}]\,\mathbf{a}'=\mathbf{x}_{N+1} \tag{12}$$

Although no sparse priors are imposed, the sparse property of the coefficient vector **a** is generated naturally by the L_1 optimization. Once the vector $(\mathbf{a}^T, \mathbf{e}^T)^T$ is computed, the similarity coefficients $W_{(N+1)i}$ are set to:

$$W_{(N+1)i} = |a_i|, i = 1, \ldots, N$$

3.3 Advantages of the proposed out-of-sample embedding scheme

Although our proposed out-of-sample formula (Eq. (9)) is similar to that of the Latent Variable Model [11], it has the two following interesting differences and advantages:

1. For the LVM scheme, the neighborhood size must be set manually, and the optimal setting may be different for different data sets. In our scheme, the computation of similarity coefficients adapts to the dataset through the use of sparse coding. No parameter is required.
2. There have been many ways to compute the similarity coefficients and the most popular one among them is the typical Heat Kernel (Gaussian weighting function) described in Eq.(1). However, the Gaussian aperture may affect the final classification results significantly, and how to optimally determine this parameter is still an open problem. Our scheme get rid of this since we exploit the sparseness property of the deduced coefficients in order to express both adjacency structure and the associated weights without any predefined parameter.

4 Performance evaluation

To validate the effectiveness of our proposed approach, we applied it to the face recognition problem.

4.1 Data sets

We considered in our experiments four public face data sets. All these databases are characterized by a large variation in face appearance.

1. **Yale**[1]**:** The YALE face data set contains 165 images of 15 persons. Each individual has 11 images. The images demonstrate variations in lighting condition, facial expression. Each image is resized to 32×32 pixels.
2. **ORL**[2]**:**. There are 10 images for each of the 40 human subjects, which were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The images were taken with a tolerance for some tilting and rotation of the face up to 20^o.
3. **UMIST**[3]**:**. The UMIST data set contains 575 gray images of 20 different people. The images depict variations in head pose.

[1] $http://see.xidian.edu.cn/vipsl/database\_Face.html$
[2] $http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html$
[3] $http://www.shef.ac.uk/eee/research/vie/research/face.html$

4. **Extended Yale - part B[4]:**. It contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. In our study, a subset of 1800 images has been used. Figure 1 shows some face samples in the extended Yale Face Database B.

Fig. 1. Some samples in Extended Yale data set.

4.2 Experimental results

To make the computation of the embedding process more efficient, the dimensionality of the original face samples was reduced by applying random projections [12]. It has a similar role to that of PCA yet with the obvious advantage that random projections do not need any training data.

We have compared our method with other two approaches. One of them is the Latent Variable Model (LVM), proposed in [11]. The other one, is a linearization of the existing mapping $\mathbf{X}_s \rightarrow \mathbf{Y}_s$. To this end, we use simple linear regression in order to infer a linear matrix transform $\mathbf{A}$ that best approximates the existing mapping through the linear equation $\mathbf{Y}_s = \mathbf{A}^T \mathbf{X}_s$. We stress the fact the linearization has not been thoroughly tested as an out-of-sample method. Instead, this linearization was used for spectral regression (e.g., [13]).

For each face data set and for every method, we conducted three groups of experiments for which the percentage of training samples was set to 30%, 50% and 70% of the whole data set. The remaining data was used for testing. Here, the testing implies: (i) the out-of-sample embedding of the unseen sample (face), and (ii) recognizing it through the use of the Nearest Neighbor classifier in the embedded space. The partition of the data set was done randomly. For a given embedding method, the recognition rate was computed for several dimensions belonging to $[5, L_{max}]$, where L_{max} is a parameter directly related with the number of training samples.

In figures 2, 3, and 4 we show the average recognition rates for all 4 datasets, based on the average of 10 random splits.

In table 1, we present the best (average) performance obtained by each 'out-of-sample' method, based on 10 random splits. For the case of LVM method, the ϵ parameter corresponds to the number of neighbors used to approximate the unseen sample. We could appreciate that the smaller this number is, the better the result.

The above results confirm the superiority of our approach when compared with existing ones. We can observe that this superiority was obtained for all data sets and for all dimensions tested for the obtained embedding space. We can also observe that

[4] $http://vision.ucsd.edu/\sim leekc/ExtYaleDatabase/ExtYaleB.html$

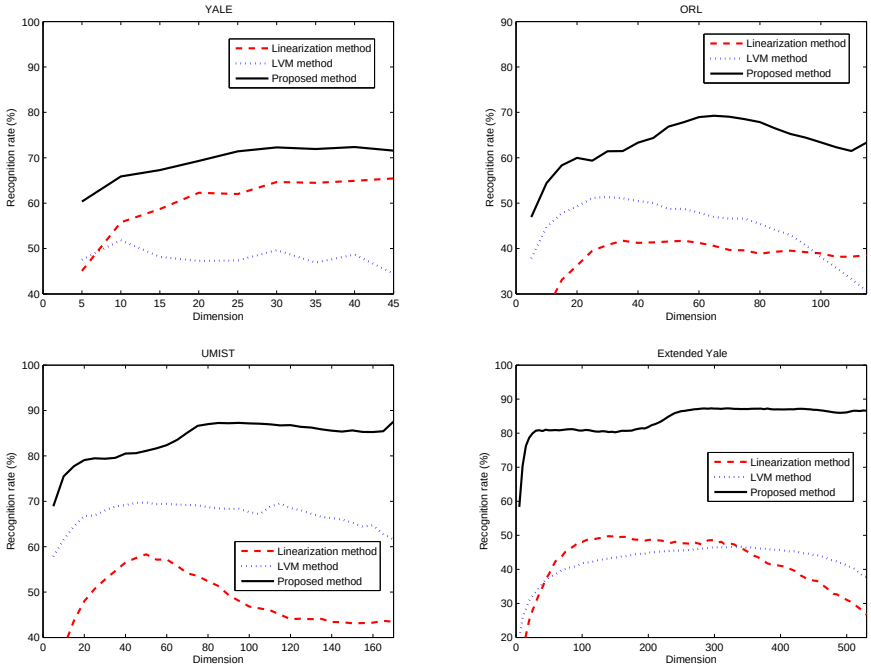

Fig. 2. Experimental results on all 4 datasets for the 30-70 modality.

Dataset \ Method	Sparse Rep.	LVM			Linearization
30%-70%		$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 7$	
YALE	**72.36%**	51.84%	41.66%	33.15%	65.43%
ORL	**69.25%**	51.35%	37.71%	30.25%	41.71%
UMIST	**87.56%**	69.72%	60.49%	52.65%	58.31%
Ext. Yale	**87.29%**	46.66%	31.33%	24.25%	49.90%
50%-50%		$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 7$	
YALE	**81.85%**	70.12%	61.60%	52.09%	68.14%
ORL	**82.50%**	72.05%	60.35%	49.25%	46.38%
UMIST	**95.03%**	85.90%	76.04%	70.03%	76.25%
Ext. Yale	**91.46%**	61.09%	46.85%	39.03%	53.14%
70%-30%		$\epsilon = 3$	$\epsilon = 5$	$\epsilon = 7$	
YALE	**86.73%**	77.15%	73.87%	67.95%	75.51%
ORL	**88.75%**	82.16%	73.66%	65.41%	53.25%
UMIST	**97.74%**	93.06%	85.20%	79.94%	80.52%
Ext. Yale	**92.12%**	70.97%	58.36%	48.74%	57.14%

Table 1. Maximum average recognition rate.

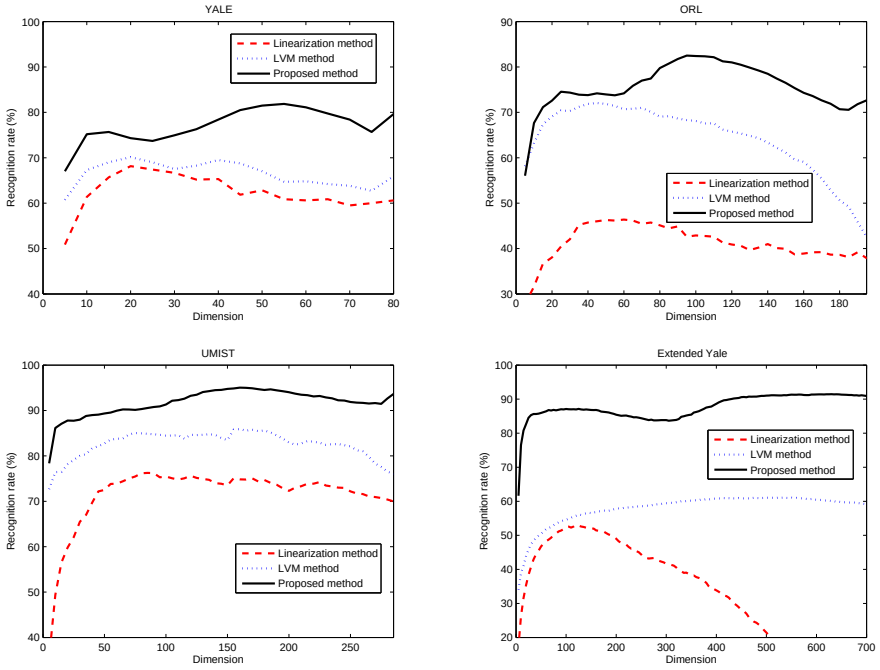

Fig. 3. Experimental results on all 4 datasets for the 50-50 modality.

the linearization method provided the poorest results, which can be explained by the fact that the linear method is global and does not take into account the local adjacency information.

5 Conclusion

In this paper, we demonstrated that sparse representation can serve as an efficient and accurate alternative for out-of-sample embedding. Considering for a case study the Laplacian Eigenmaps, we applied our method to the face recognition problem. The experimental results demonstrate that our algorithm can maintain an accurate low-dimensional representation of the data without any parameter tuning. A natural extension of our approach is its application to online learning.

References

1. B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998) 1299–1319.
2. S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

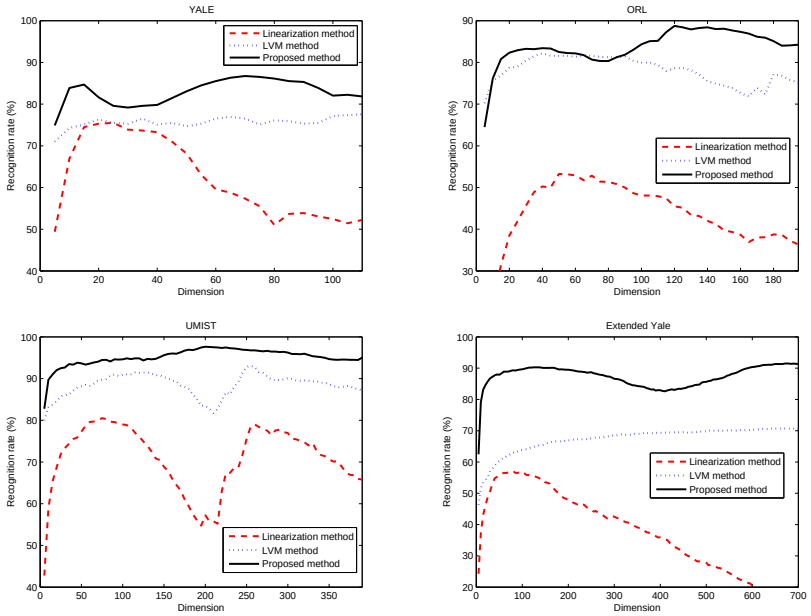

Fig. 4. Experimental results on all 4 datasets for the 70-30 modality.

3. L. K. Saul, S. T. Roweis, Y. Singer, Think globally, fit locally: Unsupervised learning of low dimensional manifolds, Journal of Machine Learning Research 4 (2003) 119–155.
4. J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
5. X. Geng, D. Zhan, Z. Zhou, Supervised nonlinear dimensionality reduction for visualization and classification, IEEE Transactions on systems, man, and cybernetics-part B: cybernetics 35 (2005) 1098–1107.
6. M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.
7. P. Jia, J. Yin, X. Huang, D. Hu., Incremental Laplacian Eigenmaps by preserving adjacent information between data points, Pattern Recognition Letters 30 (16) (2009) 1457–1463.
8. A. Elgammal, C. Lee, Non-linear manifold learning for dynamic shape and dynamic appearance, Computer Vision and Image Understanding 106 (1) (2007) 31–46.
9. Y. Bengio, J. Paiement, P. Vincent, Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps and spectral clustering, in: Advances in Neural Information Processing, 2004.
10. S. Yan, H. Wang, Semi-supervised learning by sparse representation, in: SIAM International Conference on Data Mining, 2009.
11. M. A. Carreira-Perpinan, Z. Lu, The Laplacian Eigenmaps latent variable model, Journal of Machine Learning Research 2 (2007) 59–66.
12. N. Goel, G. Bebis, A. Nefian, Face recognition experiments with random projections, in: SPIE Conference on Biometric Technology for Human Identification, 2005.
13. D. Cai, X. He, J. Han, Spectral regression for efficient regularized subspace learning, in: Proc. Int. Conf. Computer Vision (ICCV'07), 2007.