k / K-Nearest Neighborhood Criterion for Improvement of Locally Linear Embedding

Armin Eftekhari[1], Hamid Abrishami Moghaddam[1], Massoud Babaie-Zadeh[2]

[1] K.N. Toosi University of Technology, Tehran, Iran
[2] Sharif University of Technology, Tehran, Iran
a.eftekhari@ee.kntu.ac.ir, moghaddam@eetd.kntu.ac.ir, mbzadeh@ ee.sharif.edu

Abstract. Spectral manifold learning techniques have recently found extensive applications in machine vision. The common strategy of spectral algorithms for manifold learning is exploiting the local relationships in a symmetric adjacency graph, which is typically constructed using k-nearest neighborhood (k-NN) criterion. In this paper, with our focus on locally linear embedding as a powerful and well-known spectral technique, shortcomings of k-NN for construction of the adjacency graph are first illustrated, and then a new criterion, namely k/K-nearest neighborhood (k/K-NN) is introduced to overcome these drawbacks. The proposed criterion involves finding the sparsest representation of each sample in the dataset, and is realized by modifying Robust-SL0, a recently proposed algorithm for sparse approximate representation. k/K-NN criterion gives rise to a modified spectral manifold learning technique, namely Sparse-LLE, which demonstrates remarkable improvement over conventional LLE through our experiments.

Keywords: Local linear embedding, sparse representation, Robust-SL0

1 Introduction

In the recent years, several algorithms have been developed to perform dimensionality reduction of low-dimensional nonlinear manifolds embedded in a high-dimensional space. In particular, due to technical advantages, local linear embedding (LLE) has found widespread applications in real-world problems [1, 2]. LLE is based on eigen decomposition of a special Gram matrix, which is designed to preserve the local structure of data. This local structure is typically defined using nearest neighborhood criterion in the Euclidean space by constructing a symmetric adjacency graph, in which the nodes represent the training samples and any pair of nodes are connected iff the corresponding data points are adjacent. Indeed, successful recovery of the low-dimensional structure of data highly depends on the construction of an accurate adjacency graph that gives a faithful representation of the local geometry of data [3]. In this regard, though widely used, k-NN criterion suffers from major drawbacks. In fact, since each sample is connected to its k direct nearest neighbors, k-NN rule is generally unable to exclude noisy samples or outliers in the neighborhood. In addition, k-NN criterion considers a fixed neighborhood size about each sample on the manifold. In this paper, with our focus on LLE, k-NN criterion is first represented as an optimization problem, which is then modified to yield k/K-nearest neighborhood (k/K-NN) criterion. As was the case in k-NN, new criterion searches for a small subset of samples in the neighborhood of each data point. However, unlike k-NN, this subset is not limited to k-nearest neighbors of each sample, but instead belongs to a larger neighborhood within the roughly linear patch on the manifold centered at that sample. Furthermore, size of this subset is chosen adaptively to include the minimum required samples among K $(> k)$ nearest neighbors of each data point, which is often believed to give a more reliable representation of the manifold [4]. The proposed criterion involves finding the sparsest approximate representation of a sample in the dataset, and is realized by modifying the recently proposed Robust-SL0 algorithm for sparse approximate representation [5]. The modified spectral method, namely Sparse-LLE is then experimentally validated on several datasets, demonstrating remarkable improvement over the conventional LLE. The rest of this paper is organized as follows. Section 2 is devoted to a review of the LLE. In Section 3, shortcomings of k-NN are studied and k/K-NN criterion is introduced and justified. Implementation details are then discussed in Section 4 and, finally, experimental results are presented in Section 5.

2 Locally Linear Embedding

LLE, the local properties of the manifold are expressed by writing each sample as a linear combination of its nearest neighbors. LLE then attempts to preserve these local relationships in the low-dimensional space [6]. To be more specific, LLE first constructs the adjacency graph $\mathcal{G}(V,E)$, whose nodes V and edges E represent the data samples and neighborhood relations among samples, respectively. Denoting each sample by $\mathrm{x}_i \in \mathbb{R}^N$, we will use $\mathrm{x}_i \sim \mathrm{x}_j$ to indicate that samples x_i and x_j are adjacent by some criterion, i.e. $\mathrm{x}_i\mathrm{x}_j \in E$. Similarly, $\mathrm{x}_i \nsim \mathrm{x}_j$ will indicate $\mathrm{x}_i\mathrm{x}_j \notin E$. Furthermore, for each sample x_i, the subset of samples x_j satisfying $\mathrm{x}_i \sim \mathrm{x}_j$ will be denoted by $\{\mathrm{x}_{i_j}\}$. In particular, for k-NN criterion, we have $\{\mathrm{x}_{i_j}\} = \mathcal{N}_k(\mathrm{x}_i)$, where $\mathcal{N}_k(\mathrm{x}_i)$ denotes the subset of k-nearest neighbors of x_i. Additionally, $\mathcal{N}_k(i)$ would denote the corresponding subset of indices of $\mathcal{N}_k(\mathrm{x}_i)$.

Once the adjacency graph is constructed, each sample x_i is written as a linear combination of its k nearest neighbors. This is achieved by solving:

$$\boldsymbol{\mathcal{L}}\colon \min_{\mathrm{w}_i} \|\mathrm{x}_i - \mathrm{X}\mathrm{w}_i\|_2^2 \ \text{s.t.}\ \ \mathrm{Supp}(\mathrm{w}_i) \subset \mathcal{N}_k(i), \mathrm{w}_i^{\mathrm{T}}\mathbf{1} = 1 \tag{1}$$

where $\mathrm{w}_i \in \mathbb{R}^N$ contains the reconstruction weights and $\mathrm{Supp}(\eta)$ denotes the support of η, i.e. subset of all indices j, for which η_j is nonzero. In addition, $\mathbf{1} = [1,\dots,1]^{\mathrm{T}} \in \mathbb{R}^N$. The weight matrix $\mathrm{W} = [\mathrm{w}_1,\dots,\mathrm{w}_N] \in \mathbb{R}^{N\times N}$ is then constructed and the embeddings are found by computing the eigenvectors associated with the bottom nonzero eigenvalues of $\mathrm{M} = (\mathrm{I} - \mathrm{W})(\mathrm{I} - \mathrm{W}^{\mathrm{T}})$, where I is the identity matrix. To be more specific, denoting the resulting modal matrix by $\mathrm{V} = [\mathrm{v}_1,\dots,\mathrm{v}_N] \in \mathbb{R}^{N\times N}$, rows of V contain the embeddings $\{\mathrm{y}_i\}_{i=1}^N$. In fact, up to a scaling factor that depends on the algorithm, the embedding of x_i, namely y_i, is a vector with $y_{i,j} = v_{j,i}, j \le m$.

3 k/K-Nearest Neighborhood Criterion

k-NN criterion implies that $\mathrm{x}_i \sim \mathrm{x}_j$ iff $\mathrm{x}_j \in \mathcal{N}_k(\mathrm{x}_i)$, and is justified based on the notion that local geometry of the manifold at x_i is best represented by $\mathcal{N}_k(\mathrm{x}_i)$ rather than by any other subset $V \subset \{\mathrm{x}_i\}_{i=1}^N$ with $\#V = k$. In this section, with our focus on LLE, shortcomings of this notion are discussed. k/K-NN criterion is then introduced, which, to some extent, overcomes the shortcomings of k-NN.

As shown in the [11], (1) is asymptotically equivalent to:

$$\boldsymbol{\mathcal{J}}\colon \min_{\mathrm{w}_i} \lim_{c \to \infty} \|\mathrm{x}_i - \mathrm{X}\mathrm{w}_i\|_2 + c \sum_{j \neq i} u(w_{i,j}) \|\mathrm{x}_i - \mathrm{x}_j\|_2$$
$$\mathrm{s.t.}\ \|\mathrm{w}_i\|_0 = k, \mathrm{w}_i^{\mathrm{T}} \mathbf{1} = 1, \mathrm{w}_{i,i} = 0 \tag{2}$$

where $u(\cdot)$ is the step function and $\|\eta\|_0$ is the ℓ^0-norm of the vector η, i.e. number of nonzero elements of η. It is observed that solving $\boldsymbol{\mathcal{J}}$ primarily minimizes the second term of the functional by choosing $\{\mathrm{x}_{i_j}\} = \mathcal{N}_k(\mathrm{x}_i)$. Then, keeping $\{\mathrm{x}_{i_j}\}$ fixed, $\boldsymbol{\mathcal{J}}$ minimizes the reconstruction error $\|\mathrm{x}_i - \mathrm{X}\mathrm{w}_i\|_2$ by solving the linear system $\mathrm{x}_i^{\perp} = \mathrm{X}\mathrm{w}_i$ subject to $\mathrm{Supp}(\mathrm{w}_i) = \mathcal{N}_k(i)$, where $\mathrm{x}_i^{\perp}$ is the projection of x_i onto $\mathrm{Span}(\{\mathrm{x}_{i_j}\})$. It is observed that, despite its importance, minimizing the reconstruction error does not contribute to the choice of $\{\mathrm{x}_{i_j}\}$ in $\boldsymbol{\mathcal{J}}$. Furthermore, $\#\{\mathrm{x}_j\}$ is fixed to k in $\mathcal{J}$, while it is generally better to let the algorithm automatically decide on $\#\{\mathrm{x}_j\}$ by selecting only necessary samples for representation of x_i [4]. To overcome these drawbacks, the following optimization problem is introduced:

$$\min_{\mathrm{w}_i} \|\mathrm{x}_i - \mathrm{X}\mathrm{w}_i\|_2 + c_1 \sum_{j \neq i} u(w_{i,j}) \|\mathrm{x}_i - \mathrm{x}_j\|_2 + c_2 \|\mathrm{w}_i\|_0$$
$$\mathrm{s.t.}\ \mathrm{w}_i^{\mathrm{T}} \mathbf{1} = 1, w_{i,i} = 0 \tag{3}$$

where c_1, c_2 are finite positive scalars. By choosing $c_1 < \infty$, we ensure that, in contrast to $\mathcal{J}$, minimizing the reconstruction error $\|\mathrm{x}_i - \mathrm{X}\mathrm{w}_i\|_2$ contributes to our choice of $\{\mathrm{x}_{i_j}\}$. Moreover, (3) uses the minimum required number of samples to best represent x_i, and hence adaptively selects $\#\{\mathrm{x}_{i_j}\}$ on the manifold. Note that for every pair c_1 and c_2, there always exist a pair ϵ_1 and ϵ_2, for which (3) is equivalent to:

$$\min_{\mathrm{w}_i} \|\mathrm{w}_i\|_0$$
$$\mathrm{s.t.} \|\mathrm{x}_i - \mathrm{X}\mathrm{w}_i\|_2 \leq \epsilon_1, \textstyle\sum_{j \neq i} u(w_{i,j}) \|\mathrm{x}_i - \mathrm{x}_j\|_2 \leq \epsilon_2, \mathrm{w}_i^{\mathrm{T}} \mathbf{1} = 0, w_{i,i} = 0 \tag{4}$$

Furthermore, we notice that $\sum_{j \neq i} u(\mathrm{w}_{i,j}) \|\mathrm{x}_i - \mathrm{x}_j\|_2 \leq \epsilon_2$ sets an upper limit on $\|\mathrm{x}_i - \mathrm{x}_j\|_2$ for $\mathrm{x}_j \in \{\mathrm{x}_{i_j}\}$ and hence there exists some $\epsilon_3 > 0$, for which the second constraint in (4) can be safely replaced by $\|\mathrm{x}_i - \mathrm{x}_j\|_2 \leq \epsilon_3$, $\forall \mathrm{x}_j \in \{\mathrm{x}_{i_j}\}$ [7]. A closer look reveals that this in turn could be safely replaced by $\mathrm{Supp}(\mathrm{w}_i) \subset \mathcal{N}_K(i)$, for some integer K. Therefore, we can rewrite (4) as follows:

$$\boldsymbol{\mathcal{J}\mathcal{J}}\colon \min_{\mathrm{w}_i} \|\mathrm{w}_i\|_0$$
$$\mathrm{s.t.} \|\mathrm{x}_i - \mathrm{X}\mathrm{w}_i\|_2 \leq \epsilon_1, \mathrm{Supp}(\mathrm{w}_i) \subset \mathcal{N}_K(i), \mathrm{w}_i^{\mathrm{T}} \mathbf{1} = 1 \tag{5}$$

Let w_i^* and $\{\mathrm{x}_{i_j}^*\}$ denote the solution of $\boldsymbol{\mathcal{JJ}}$ and the subset of samples corresponding to the nonzero elements of w_i^*, respectively. Note that, as a result of the second constraint in $\boldsymbol{\mathcal{JJ}}$, $\{\mathrm{x}_{i_j}^*\} \subset \mathcal{N}_K(x_i)$. In order to preserve the computational advantages of working with highly sparse matrices, we further limit $\#\{\mathrm{x}_{i_j}^*\}$ to k, for some integer $k < K$. This is achieved by keeping at most k top nonzero elements of w_i^* and setting others (if any) to zero. $\{\mathrm{x}_{i_j}^*\}$ is also modified by discarding the corresponding samples. The new criterion will be referred to as k/K-NN rule and is summarized in Fig. 1. Notice that, in k-NN, $\{\mathrm{x}_{i_j}^*\}$ is the subset of first k nearest neighbors of x_i, whereas in $\mathcal{JJ}$, $\{\mathrm{x}_{i_j}^*\}$ is the best subset $V \subset \mathcal{N}_K(\mathrm{x}_i)$ with $\#V \leq k$, that contains the minimum required samples to achieve a reconstruction error less than the error tolerance ϵ_1. When compared to k-NN, k/K-NN criterion is able to exclude noisy neighbors and outliers, which is achieved by the constraint on the reconstruction error in $\boldsymbol{\mathcal{JJ}}$. On the other hand, when compared to K-NN, k/K-NN criterion adaptively selects $\#\{\mathrm{x}_{i_j}^*\}$ $(\leq k)$ to best represent x_i with the minimum required number of samples. Now, using k/K-NN criterion to construct the adjacency graph, LLE is modified to obtain an improved spectral algorithm, dubbed Sparse-LLE. Note that the only difference between LLE and Sparse-LLE lies in the construction of the adjacency graph.

Given integers k and K, with $k < K$, solve $\boldsymbol{\mathcal{JJ}}$ for each sample $\mathrm{x}_i \in \{\mathrm{x}_i\}_{i=1}^N$, and denote the answer by w_i^*. Then, nodes i and j in the adjacency graph $\mathcal{G}$ are connected iff $w_{i,j}^*$ is among the top k nonzero elements of w_i^*.

Figure 1. k/K-NN criterion for construction of the adjacency graph.

4 Implementation

In Section 3, k/K-NN criterion for construction of the adjacency graph was introduced and justified. In order to apply this criterion, we shall study the following optimization problem: $\mathcal{P}_{0,\epsilon,S}\colon \min\|\mathrm{s}\|_0 \ \mathrm{s.t.}\ \|\mathrm{b} - \mathrm{As}\|_2 \leq \epsilon$ and $\mathrm{Supp}(\mathrm{s}) \in S$, where $\mathrm{b} \in \mathbb{R}^n$ and S is a given subset of indices of $\mathrm{s} = [s_1, \ldots, s_m]^{\mathrm{T}} \in \mathbb{R}^m$. Our implementation assumes $m > n$, which fairly happens almost always in real-world situations. As the starting point, we first consider the well-known sparse approximate representation problem $\mathcal{P}_{0,\epsilon}\colon \min\|\mathrm{s}\|_0 \ \mathrm{s.t.}\ \|\mathrm{b} - \mathrm{As}\|_2 \leq \epsilon$. Among available approaches, we opt for the recently proposed Robust-SL0 as a fast and accurate algorithm [5]. Briefly speaking, Robust-SL0 solves a sequence of problems of the form $\mathcal{Q}_{\epsilon,\sigma}\colon \max_{\mathrm{s}} \sum_{i=1}^m e^{-s_i^2/2\sigma^2}$ s.t. $\|\mathrm{b} - \mathrm{As}\|_2 \leq \epsilon$, decreasing σ at each step, and initializing the next step at the maximizer of the previous (larger) value of σ. Each $\mathcal{Q}_{\epsilon,\sigma}$ is solved approximately by few iterations of gradient ascent. Convergence analysis of Robust-SL0 has been thoroughly considered in [5] and it was shown that, under some mild conditions, the sequence of maximizers of $\mathcal{Q}_{\epsilon,\sigma}$ indeed converges to the unique minimizer of $\mathcal{P}_{0,\epsilon}$, whenever such answer exists. Moreover, Robust-SL0 runs significantly faster than the competing algorithms, while producing answers with the same or better accuracy [5]. The idea is now to modify $\mathcal{P}_{0,\epsilon,S}$ in a way that enables using Robust-SL0 algorithm to solve $\mathcal{P}_{0,\epsilon,S}$. This necessitates proper modification of the second constraint in $\mathcal{P}_{0,\epsilon,S}$, i.e. $\mathrm{Supp(s)} \in S$. While this may be achieved by, for instance, setting $s_i = 0$ for $i \notin S$ at each iteration, we prefer to preserve the studied convergence properties of Robust-SL0 by replacing $\mathrm{Supp(s)} \in S$ with a term in functional that smoothly favors small values for s_i when $i \notin S$. Therefore, $\mathcal{P}_{0,\epsilon,S}$ is modified to:

$$\lim_{\sigma \to 0} \max_{\mathrm{s}} \sum_{i \in S} (1-\alpha) e^{-s_i^2/2\sigma^2} + \sum_{i \notin S} (1+\beta) e^{-s_i^2/2\sigma^2}$$
$$\text{s.t. } \|\mathrm{b} - \mathrm{As}\|_2 \le \epsilon \tag{6}$$

where we take $0 \le \alpha, \beta < 1$. Convergence properties of (6) are obtained by minimal modifications in the proof presented in [5]. Note that Robust-SL0 algorithm is now applicable to (6) by merely using the gradient of the functional of (6) in the algorithm. The interested reader is referred to [5] for details.

5 Experiments

The objective of this section is to experimentally assess the merits of the proposed k/K-NN criterion for construction of the adjacency graph. To this end, the performance of LLE and Sparse-LLE are compared on several datasets. In each experiment, k (and if available K) are experimentally tuned for the best results. Other parameters of Sparse-LLE are fixed to: $\epsilon = 0.05, \alpha = \beta = 0.9$. As our first experiment, we compare the performance of LLE and Sparse-LLE for visualizing the Frey face dataset, which consists of 1965 gray-level images of a single individual acquired under different expression and pose conditions [2]. Few images in this dataset are depicted in Fig. 2(a). Fig. 2(b) depicts the first two components of these images discovered by LLE. Depicted in Fig. 2(c) are the visualization results obtained by Sparse-LLE, which may be interpreted as follows. We can recognize four pair of opposite branches in the embedded space, labeled from 1 to 4. It is observed that the main trend in branches 1 and 2 includes left pose or slightly left pose images, whereas images in branches 3 and 4 are mainly either right pose or slightly right pose. In particular, while containing opposite poses, both branches 1 and 3 are similar in that one of their ends includes happy faces and the other end includes either sad faces or faces with visible tongue. The main trend of images in each branch is represented in Fig. 3. It is observed that the main trend in branches 1 and 2 includes left pose or slightly left pose images, whereas images in branches 3 and 4 are mainly either right pose or slightly right pose. In particular, while containing opposite poses, both branches 1 and 3 are similar in that one of their ends includes happy faces and the other end includes either sad faces or faces with visible tongue.

As our second experiment, the performance of LLE and Sparse-LLE is compared in face recognition task on the extended Yale face database. The dataset includes 2432 cropped frontal images of 38 individuals under expression and illumination variation [8], where the first 16 images of each individual are considered in this experiment. After vectoriation, using LLE and Sparse- LLE, dimension of image data is reduced to 10. Subsequently, motivated by the well-designed experimental setup in

[6], quality of the resulting low-dimensional representations is evaluated by measuring the classification errors of 1 nearest neighbor classifiers trained on the low-dimensional representations using leave-one-out cross-validation. In other words, class of each sample is predicted by its nearest neighbor in the embedded space and the overall classification error is reported in Table 1.

Retinal biometrics refers to identity verification of individuals based on their retinal vessel tree pattern. Our third experiment is conducted on VARIA database containing 153 (multiple) retinal images of 59 individuals [9]. To compensate for the variations in the location of optic disc (OD) in retinal images, a ring-shaped region of interest (ROI) in the vicinity of OD is used to construct the feature matrix. To extract the ROI, using the technique presented in [10], OD and vessel tree are extracted. Then, a ring-shaped mask with proper radii centered at OD is used to form the feature vectors $X \in \mathbb{R}^{6\times 8}$ by collecting the pixels along 8 beams of length 6 originating from OD. A special case is depicted in Fig. 4. After vectorization of feature matrices, dimension is reduced to 10 using LLE and Sparse-LLE. The performance of the resulting low- dimensional representations is then evaluated similar to the second experiment (Table 1).

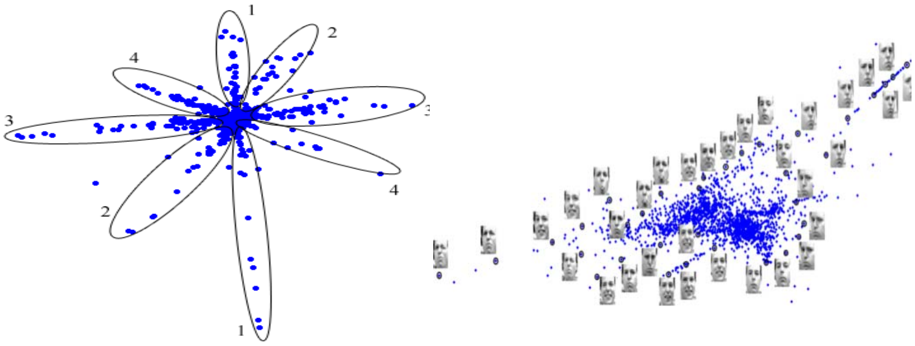
Conclusions

LLE is a well-known and powerful spectral dimension reduction algorithm. For successful recovery of the low-dimensional structure of data, however, LLE requires an adjacency graph, which is typically constructed using k-NN criterion. In this paper, deficiencies of k-NN for construction of the adjacency graph were first studied and k/K-NN criterion was then introduced to overcome the drawbacks. Implementation of k/K-NN involved a variant of Robust-SL0 algorithm for sparse approximate representation. The modified spectral method, namely Sparse-LLE, is experimentally validated on several datasets, demonstrating remarkable improvement over the conventional LLE.

Table 1. Generalization errors of 1-NN classifiers for different dimension reduction algorithms.

	Yale face database		VARIA database	
Algorithm	Parameters	Generalization error of 1-NN	Parameters	Generalization error of 1-NN
PCA	-	35.5263	-	59.4771
LLE	$k = 12$	29.9342	$k = 4$	61.4379
Sparse-LLE	$k = 5$ $K = 12$	23.5197	$k = 5$ $K = 7$	56.2092

(a)



(c) (b)

Figure 2. A few samples of Frey dataset used in the first experiment (a). Images of faces mapped into the embedding space described by the first two coordinates of LLE with $k = 12$ (b), and sparse-LLE with $k = 5$ and $K = 12$ (c).

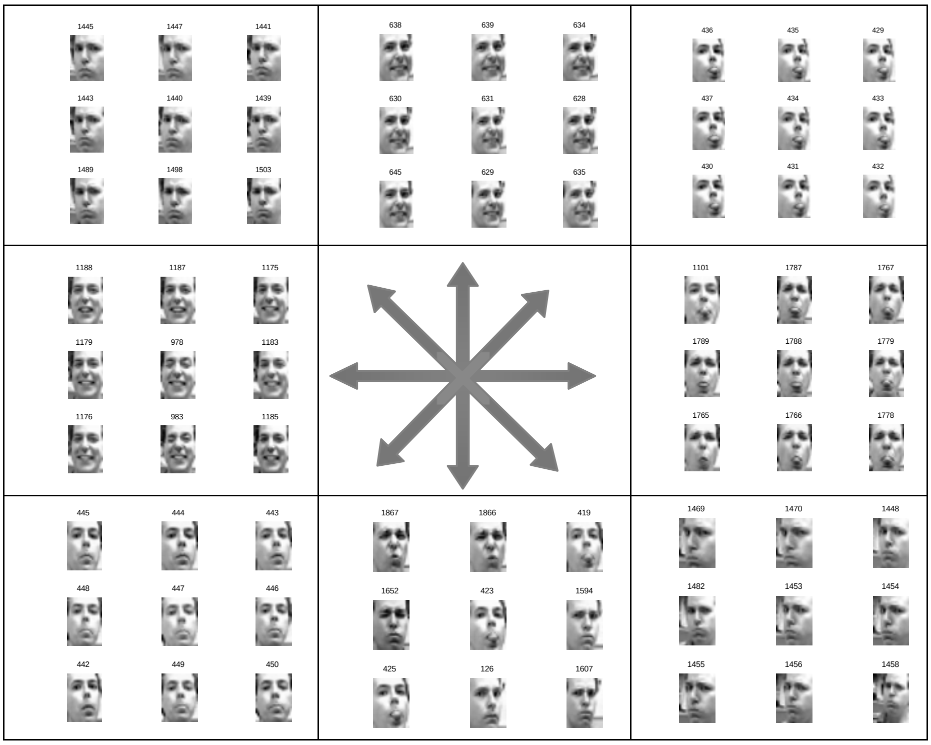


Figure 3. Further study of the embedded space obtained by sparse-LLE in Fig. 2(c). Outer boxes are positioned similar to the distribution of branches in Fig. 2(c), where label of corresponding branches are indicated by the arrows. Each outer box contains few samples of the corresponding branch, which are selected to represent the main trend of the images inside the branch.

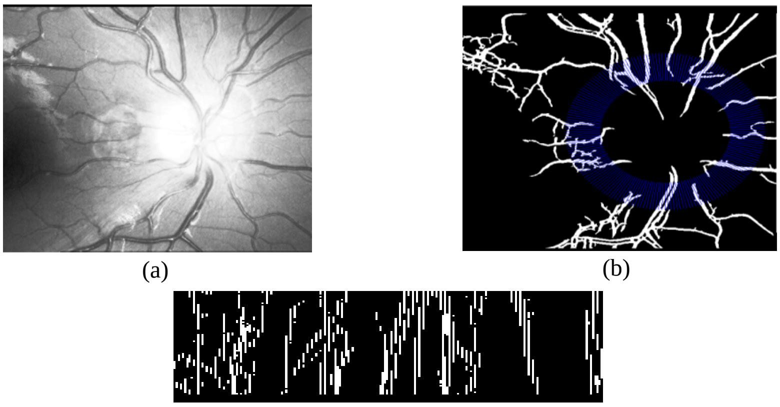


Figure 4. (a) Retinal image; bright area is OD. (b) Vessel tree (in white) and mask (in blue). (c) Feature matrix obtained from 300 beams of length 100 pixels (images (a) and (b) are cropped).

References

1. L.K. Saul, K.Q. Weinberger, and J.H. Ham, F. Sha, and D.D. Lee, “Spectral methods for dimensionality reduction,” in Semisupervised Learning, O. Chapelle, B. Schölkopf, and A. Zien (eds), MIT Press, 2006.
2. S. Roweis, and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” Science, vol. 290, pp. 2323-2326, 2000.
3. G. Lebanon, “Riemannian geometry and statistical machine learning,” Doctoral Thesis, School of Computer Science, Carnegie Mellon University, 2005.
4. T. Lin, H. Zha, and S.U. Lee, “Riemannian manifold learning for nonlinear dimensionality reduction,” in Proc. ECCV (1), pp. 44-55, 2006.
5. A. Eftekhari, M. Babaie-Zadeh, C. Jutten, H. Abrishami Moghaddam, “Robust-SL0 for stable sparse representation in noisy settings,” Int. Conf. Acoustics, Speech, and Signal Proc., 2009, Accepted.
6. L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik, “Dimensionality reduction: a comparative review,” submitted to Neurocomputing, 2009.
7. M. Bernstein, V. de Silva, J.C. Langford, and J. B. Tenenbaum, “Graph approximations to geodesics on embedded manifolds,” Technical Report, Stanford University, 2000.
8. A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, “From few to many: illumination cone models for face recognition under variable lighting and pose,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, pp. 643-660, 2001.
9. VARIA database, available online at http://www.varpa.es/varia.html
10. Farzin, H., Abrishami, H.: A novel retinal identification system. EURASIP Jr. on Advances in Signal Proc., 2008.
11. A. Eftekhari, H. Abrishami Mohgaddam, "k/K-nearest neighborhood criterion for improvement of locally linear embedding", Technical report, available online at http://nasim.kntu.ac.ir/MS/a_eftekhari.