TANGENT SPACE INTRINSIC MANIFOLD REGULARIZATION FOR DATA REPRESENTATION

Shiliang Sun

Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, P. R. China

ABSTRACT

A new regularization method called tangent space intrinsic manifold regularization is presented, which is intrinsic to data manifold and favors linear functions on the manifold. Fundamental elements involved in its formulation are local tangent space representations which we estimate by local principal component analysis, and the connections which relate adjacent tangent spaces. We exhibit its application to data representation where a nonlinear embedding in a low-dimensional space is found by solving an eigen-decomposition problem. Experimental results including comparisons with state-of-the-art techniques show the effectiveness of the proposed method.

***Index Terms*—** Regularization, tangent space, manifold learning, dimensionality reduction, data representation

1. INTRODUCTION

In this paper, we present a new data-dependent regularization method named tangent space intrinsic manifold regularization for function learning, which is motivated largely by the geometry of the data-generating distribution. In particular, data lying in a high-dimensional space are assumed to be intrinsically of low dimensionality. That is, data can be well characterized by far fewer parameters or degrees of freedom than the actual ambient representation. This setting is usually referred to as manifold learning, and the distribution of data is regarded to live on or near a low-dimensional manifold. The validity of manifold learning, especially for high-dimensional image and text data, has already been testified by recent developments [1, 2, 3]. Although our new regularization method has potentials to be applied to a variety of pattern recognition and machine learning problems, here we only consider unsupervised data representation or dimensionality reduction.

The problem of representing data in a low-dimensional space for the sake of data visualization and organization is essentially a dimensionality reduction problem. Given a data set $\{\mathbf{x}_i\}_{i=1}^k$ with $\mathbf{x}_i \in \mathbb{R}^d$, its task is to deliver a data set $\{\mathbf{f}_i\}_{i=1}^k$ where $\mathbf{f}_i \in \mathbb{R}^m$ corresponds to $\mathbf{x}_i$ and $m << d$. Classical linear dimensionality reduction methods include principal component analysis (PCA) and metric multidimensional scaling (MDS) [4]. Recent nonlinear dimensionality reduction methods achieved a great success especially for representing data that obey the manifold assumption, e.g., successfully unfolding the intrinsic degrees of freedom for gradual changes of images and videos. Representative algorithms in this category include locally linear embedding [2], isomap [3], Laplacian eigenmaps [1], Hessian eigenmaps [5], maximum variance unfolding with semidefinite programming [6], and others [7]. Some of the above linear and nonlinear dimensionality reduction methods can be characterized as spectral methods, because computationally they often contain the procedure of eigen-decomposition of an appropriately constructed matrix and then exploit the top or bottom eigenvectors [6].

The principle of regularization has its root in mathematics to solve ill-posed problems, and is widely used in pattern recognition and machine learning. Many well-known algorithms, e.g., SVMs [8, 9], ridge regression and lasso [10, 11], can be interpreted as instantiations of the idea of regularization. The regularization method presented in this paper is intrinsic to data manifold which prefers linear functions on the manifold. Namely, functions with constant manifold derivatives are more appealing. Fundamental elements involved in the regularization formulation are local tangent space representations and the connections which relate adjacent tangent spaces. When applying this regularization principle to nonlinear dimensionality reduction with data modeled as adjacency graphs, it turns out that the resultant problem can be readily solved by eigen-decomposition. We illustrate that this regularization method can obtain good and reasonable data embedding results.

2. THE PROPOSED REGULARIZATION

We are interested in estimating a function $f(\mathbf{x})$ defined on $\mathcal{M} \subset \mathbb{R}^d$, where $\mathcal{M}$ is a smooth manifold on $\mathbb{R}^d$. We assume that $f(\mathbf{x})$ can be well approximated by a linear function with respect to the manifold $\mathcal{M}$. Let m be the dimensionality of $\mathcal{M}$. At each point $\mathbf{z} \in \mathcal{M}$, $f(\mathbf{x})$ can be represented as a linear function $f(\mathbf{x}) \approx b_{\mathbf{z}} + \mathbf{w}_{\mathbf{z}}^\top \mathbf{u}_{\mathbf{z}}(\mathbf{x}) + o(\|\mathbf{x} - \mathbf{z}\|^2)$ lo-

Thanks to NSFC Project 61075005 and Shanghai Knowledge Service Platform Project (No. ZF1213) for funding, and to Prof. Tong Zhang from Rutgers University for helpful discussions. Email: slsun@cs.ecnu.edu.cn

cally around $\mathbf{z}$, where $\mathbf{u}_{\mathbf{z}}(\mathbf{x}) = T_{\mathbf{z}}(\mathbf{x}-\mathbf{z})$ is an m-dimensional vector representing $\mathbf{x}$ in the tangent space around $\mathbf{z}$, and $T_{\mathbf{z}}$ is an $m \times d$ matrix that projects $\mathbf{x}$ around $\mathbf{z}$ to a representation in the tangent space of $\mathcal{M}$ at $\mathbf{z}$. Note that in this paper the basis for $T_{\mathbf{z}}$ is computed using local PCA for its simplicity and wide applicability. In particular, the point $\mathbf{z}$ and its neighbors are sent over to the regular PCA procedure and the top m eigenvectors are returned back as rows of matrix $T_{\mathbf{z}}$. The weight vector $\mathbf{w}_{\mathbf{z}} \in \mathbb{R}^m$ is an m-dimensional vector, and it is also the manifold-derivative of $f(\mathbf{x})$ at $\mathbf{z}$ with respect to the $\mathbf{u}_{\mathbf{z}}(\cdot)$ representation on the manifold, which we write as $\nabla_T f(\mathbf{x})|_{\mathbf{x}=\mathbf{z}} = \mathbf{w}_{\mathbf{z}}$.

Mathematically, a linear function with respect to the manifold $\mathcal{M}$, which is not necessarily a globally linear function in $\mathbb{R}^d$, is a function that has constant manifold derivative. However, this does not mean $\mathbf{w}_{\mathbf{z}}$ is a constant function of $\mathbf{u}$ due to the different coordinate systems when the "anchor point" $\mathbf{z}$ changes from one point to another. This needs to be compensated using "connections" that map a coordinate representation $\mathbf{u}_{\mathbf{z}'}$ to $\mathbf{u}_{\mathbf{z}}$ for any $\mathbf{z}'$ near $\mathbf{z}$.

To see how our approach works, we assume for simplicity that $T_{\mathbf{z}}$ is an orthogonal matrix for all $\mathbf{z}$: $T_{\mathbf{z}} T_{\mathbf{z}}^\top = I_{(m\times m)}$. This means that if $\mathbf{x} \in \mathcal{M}$ is close to $\mathbf{z} \in \mathcal{M}$, then $\mathbf{x} - \mathbf{z} \approx T_{\mathbf{z}}^\top T_{\mathbf{z}}(\mathbf{x}-\mathbf{z}) + O(\|\mathbf{x}-\mathbf{z}\|^2)$. Now consider $\mathbf{x}$ that is close to both $\mathbf{z}$ and $\mathbf{z}'$. We can express $f(\mathbf{x})$ both in the tangent space representation at $\mathbf{z}$ and $\mathbf{z}'$, which gives

$$b_{\mathbf{z}} + \mathbf{w}_{\mathbf{z}}^\top \mathbf{u}_{\mathbf{z}}(\mathbf{x}) \approx b_{\mathbf{z}'} + \mathbf{w}_{\mathbf{z}'}^\top \mathbf{u}_{\mathbf{z}'}(\mathbf{x}) + O(\|\mathbf{x}-\mathbf{z}'\|^2 + \|\mathbf{x}-\mathbf{z}\|^2).$$

That is, $b_{\mathbf{z}} + \mathbf{w}_{\mathbf{z}}^\top \mathbf{u}_{\mathbf{z}}(\mathbf{x}) \approx b_{\mathbf{z}'} + \mathbf{w}_{\mathbf{z}'}^\top \mathbf{u}_{\mathbf{z}'}(\mathbf{x})$. This means that

$$b_{\mathbf{z}} + \mathbf{w}_{\mathbf{z}}^\top T_{\mathbf{z}}(\mathbf{x}-\mathbf{z}) \approx b_{\mathbf{z}'} + \mathbf{w}_{\mathbf{z}'}^\top T_{\mathbf{z}'}(\mathbf{x}-\mathbf{z}').$$

Setting $\mathbf{x} = \mathbf{z}$, we obtain $b_{\mathbf{z}} \approx b_{\mathbf{z}'} + \mathbf{w}_{\mathbf{z}'}^\top T_{\mathbf{z}'}(\mathbf{z}-\mathbf{z}')$, and

$$b_{\mathbf{z}'} + \mathbf{w}_{\mathbf{z}'}^\top T_{\mathbf{z}'}(\mathbf{z}-\mathbf{z}') + \mathbf{w}_{\mathbf{z}}^\top T_{\mathbf{z}}(\mathbf{x}-\mathbf{z}) \approx b_{\mathbf{z}'} + \mathbf{w}_{\mathbf{z}'}^\top T_{\mathbf{z}'}(\mathbf{x}-\mathbf{z}').$$

This implies that

$$\begin{aligned} &\mathbf{w}_{\mathbf{z}}^\top T_{\mathbf{z}}(\mathbf{x}-\mathbf{z}) \approx \mathbf{w}_{\mathbf{z}'}^\top T_{\mathbf{z}'}(\mathbf{x}-\mathbf{z}) \\ \approx &\mathbf{w}_{\mathbf{z}'}^\top T_{\mathbf{z}'} T_{\mathbf{z}}^\top T_{\mathbf{z}}(\mathbf{x}-\mathbf{z}) + O(\|\mathbf{x}-\mathbf{z}'\|^2 + \|\mathbf{x}-\mathbf{z}\|^2). \quad (1) \end{aligned}$$

Since (1) holds for arbitrary $\mathbf{x} \in \mathcal{M}$ close to $\mathbf{z} \in \mathcal{M}$, it follows that, $\mathbf{w}_{\mathbf{z}}^\top \approx \mathbf{w}_{\mathbf{z}'}^\top T_{\mathbf{z}'} T_{\mathbf{z}}^\top + O(\|\mathbf{z}-\mathbf{z}'\|)$ or $\mathbf{w}_{\mathbf{z}} \approx T_{\mathbf{z}} T_{\mathbf{z}'}^\top \mathbf{w}_{\mathbf{z}'} + O(\|\mathbf{z}-\mathbf{z}'\|)$.

This means that if we expand at points $\mathbf{z}_1, \ldots, \mathbf{z}_k \in Z$, and denote neighbors of $\mathbf{z}_j$ as $\mathcal{N}(\mathbf{z}_j)$, then the correct regularizer will be

$$\begin{aligned} R(\{b_{\mathbf{z}}, \mathbf{w}_{\mathbf{z}}\}_{\mathbf{z}\in Z}) = \sum_{i=1}^{k} \sum_{j\in\mathcal{N}(\mathbf{z}_i)} \big[\big(b_{\mathbf{z}_i} - b_{\mathbf{z}_j} - \\ \mathbf{w}_{\mathbf{z}_j}^\top T_{\mathbf{z}_j}(\mathbf{z}_i - \mathbf{z}_j)\big)^2 + \gamma\|\mathbf{w}_{\mathbf{z}_i} - T_{\mathbf{z}_i} T_{\mathbf{z}_j}^\top \mathbf{w}_{\mathbf{z}_j}\|_2^2\big]. \quad (2) \end{aligned}$$

With $\mathbf{z}(\mathbf{x}) = \arg\min_{\mathbf{z}\in Z} \|\mathbf{x}-\mathbf{z}\|_2$, the function $f(\mathbf{x})$ is approximated as

$$f(\mathbf{x}) = b_{\mathbf{z}(\mathbf{x})} + \mathbf{w}_{\mathbf{z}(\mathbf{x})}^\top T_{\mathbf{z}(\mathbf{x})}\big(\mathbf{x} - \mathbf{z}(\mathbf{x})\big), \quad (3)$$

which is a very natural formulation for out-of-example extensions.

3. GENERALIZATION AND REFORMULATION

Relating data with a discrete weighted graph is popular especially in graph-based machine learning methods. It also makes sense for us to generalize the regularizer in (2) using a symmetric weight matrix W constructed from the above data collection Z.

Entries in W characterize the closeness of different points where the points are often called nodes in the terminology of graphs. Usually there are two steps involved in constructing a weighted graph. The first step builds an adjacency graph by putting an edge between two "close" points. People can choose to use parameter $\epsilon \in \mathbb{R}$ or parameter $n \in \mathbb{N}$ to determine close points, which means that two nodes would be connected if their Euclidean distance is within ϵ or either node is among the n nearest neighbors of the other as indicated by the Euclidean distance. The second step calculates weights on the edges of the graph with a certain similarity measure. For example, the heat-kernel method computes weight W_{ij} for two connected nodes i and j by $W_{ij} = \exp^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/t}$ with parameter $t > 0$, while for nodes not directly connected the weights would be zero [1].

Therefore, the generalization of the tangent space intrinsic manifold regularizer turns out to be

$$\begin{aligned} R(\{b_{\mathbf{z}}, \mathbf{w}_{\mathbf{z}}\}_{\mathbf{z}\in Z}) = \sum_{i=1}^{k} \sum_{j=1}^{k} W_{ij} \big[\big(b_{\mathbf{z}_i} - b_{\mathbf{z}_j} - \\ \mathbf{w}_{\mathbf{z}_j}^\top T_{\mathbf{z}_j}(\mathbf{z}_i - \mathbf{z}_j)\big)^2 + \gamma\|\mathbf{w}_{\mathbf{z}_i} - T_{\mathbf{z}_i} T_{\mathbf{z}_j}^\top \mathbf{w}_{\mathbf{z}_j}\|_2^2\big]. \quad (4) \end{aligned}$$

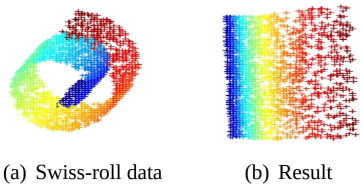
Now we reformulate the regularizer (4) into a canonical matrix quadratic form to facilitate subsequent formulations on data representation. In particular, we would like to rewrite the regularizer as a quadratic form in terms of a symmetric matrix S as follows,

$$R(\{b_{\mathbf{z}}, \mathbf{w}_{\mathbf{z}}\}_{\mathbf{z}\in Z}) = \begin{pmatrix} b_{\mathbf{z}_1} \\ \vdots \\ b_{\mathbf{z}_k} \\ \mathbf{w}_{\mathbf{z}_1} \\ \vdots \\ \mathbf{w}_{\mathbf{z}_k} \end{pmatrix}^\top \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix} \begin{pmatrix} b_{\mathbf{z}_1} \\ \vdots \\ b_{\mathbf{z}_k} \\ \mathbf{w}_{\mathbf{z}_1} \\ \vdots \\ \mathbf{w}_{\mathbf{z}_k} \end{pmatrix}, \quad (5)$$

where $\begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix}$ is a block matrix representation of S and the size of S_1 is $k \times k$. In this paper, we suppose that $\mathbf{w}_{\mathbf{z}_i}$ $(i = 1, \ldots, k)$ is an m-dimensional vector. Thus, the size of S is $(k + mk) \times (k + mk)$. Due to space limit, we omit the detailed derivation for S_1, S_2 and S_3, which will be made available on the internet.

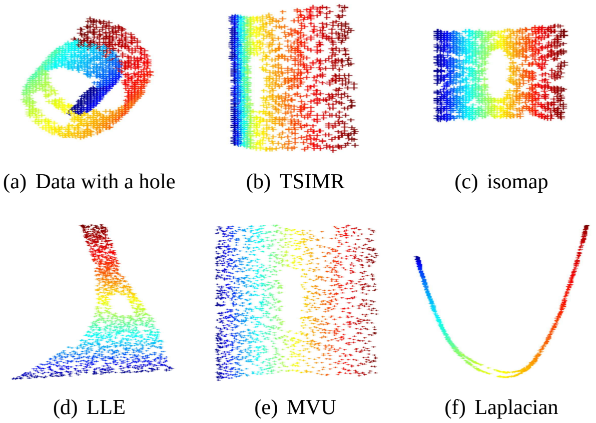
4. APPLICATION TO DATA REPRESENTATION

Data representation or dimensionality reduction is intrinsically an ill-posed problem since this is an unsupervised task



(a) Swiss-roll data (b) Result

Fig. 1. Embedding results of the TSIMR on the Swiss roll.



(a) Data with a hole (b) TSIMR (c) isomap

(d) LLE (e) MVU (f) Laplacian

Fig. 2. Embedding results on the Swiss roll with a hole.

and there can be multiple different solutions depending on the preferences or specific objectives defined by users. When the tangent space intrinsic manifold regularization method is used, we actually prefer embedding functions with constant manifold derivatives.

Define $\mathbf{w} = (\mathbf{w}_{\mathbf{z}_1}^\top, \mathbf{w}_{\mathbf{z}_2}^\top, \dots, \mathbf{w}_{\mathbf{z}_k}^\top)^\top$. Suppose vector $\mathbf{f} = (b_{\mathbf{z}_1}, b_{\mathbf{z}_2}, \dots, b_{\mathbf{z}_k})^\top$ is an embedding or representation of points $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k$ in a line. Suppose the whole graph is connected (otherwise we can do embedding for each connected component). A reasonable criterion for finding a good embedding under the principle of the tangent space intrinsic manifold regularization is to minimize the objective $\begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix}^\top S \begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix}$ under appropriate constraints.

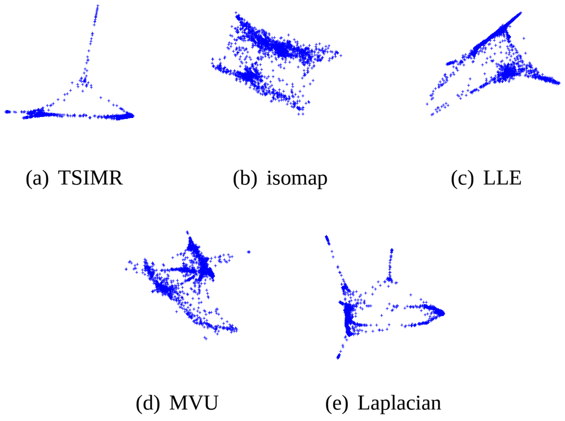
To remove an arbitrary scaling factor in both $\mathbf{f}$ and $\mathbf{w}$, we take into account the constraint $\begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix}^\top \begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix} = 1$. Therefore, the optimization problem becomes

$$\min_{\mathbf{f},\mathbf{w}} \begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix}^\top S \begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix}, \quad \text{s.t.} \begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix}^\top \begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix} = 1. \tag{6}$$

It is easy to show that the solution is given by the eigenvector corresponding to the minimal eigenvalue of the eigendecomposition $S\begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix} = \lambda\begin{pmatrix}\mathbf{f}\\ \mathbf{w}\end{pmatrix}$. Since matrix S is positive semidefinite, all its eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_{k+mk}\}$, which we suppose are sorted in ascending order, are non-negative. Hence, the embedding $\mathbf{f}$ can be found as the first k entries of the eigenvector corresponding to the least eigenvalue. Meanwhile we also obtain the $\mathbf{w}$ that matches the embedding $\mathbf{f}$.

TSIMR	isomap	LLE	MVU	Laplacian
0.6950	0.9281	0.9247	0.9308	0.7991

Table 1. Residual variances on the Swiss roll with a hole.



(a) TSIMR (b) isomap (c) LLE

(d) MVU (e) Laplacian

Fig. 3. Embedding results for the face images.

In order to find an embedding in $\mathbb{R}^m$ where m is the dimensionality of manifold $\mathcal{M}$, we definite a matrix $F = (\mathbf{f}_1, \dots, \mathbf{f}_m)$ where $\mathbf{f}_1, \dots, \mathbf{f}_m$ are the first k entries of the eigenvectors in accordance with the m least eigenvalues, respectively. Then, the embedding of $\mathbf{z}_i$ ($i = 1, \dots, k$) in $\mathbb{R}^m$ would be the i-th row of matrix F.

5. EXPERIMENTS

We evaluated the tangent space intrinsic manifold regularization (TSIMR) for data representation with synthetic and real-world data sets. The parameter γ is set to 1.

5.1. Swiss roll without and with a hole

The "Swiss roll" is a 2-dimensional manifold embedded in a 3-dimensional ambient space. For the first experiment we uniformly sampled 2000 points from the manifold, which are depicted in Fig. 1(a). The construction of the adjacency graph uses 10 nearest neighbors, and the heat kernel is employed to assign weights to the edges of the graph. Specifically, the kernel parameter t is fixed as the average of the squared distances between all points and their most nearest neighbors. For data embedding in a 2-dimensional space, the result is shown in Fig. 1(b), which precisely reflects the intrinsic degrees of freedom of the manifold.

Moreover, we show that our method still works well even if there is a hole existing in the Swiss roll. This case resembles the situation in practice that data are not sufficiently sampled from a certain region. We shoveled a rectangular hole

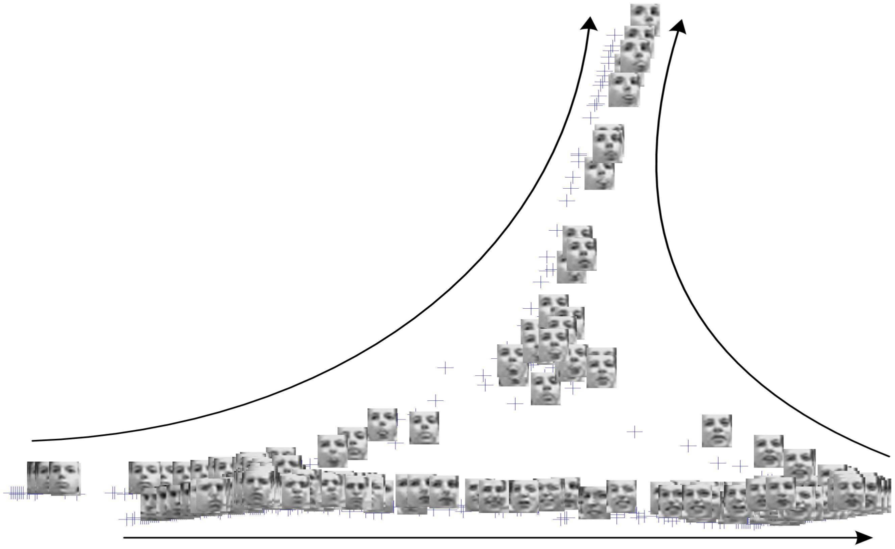


Fig. 4. Embedding results with the corresponding face images.

through the manifold, and the corresponding sample is given in Fig. 2(a). The data representation in a 2-dimensional space by the TSIMR is shown in Fig. 2(b), which faithfully reflects the existence and shape of the rectangular hole. The embedding results using isomap, LLE (locally linear embedding), MVU (maximum variance unfolding), and Laplacian regularization are also included in Fig. 2.

Now we use the metric of residual variance, namely $1 - R^2(D_1, D_2)$ [3], to measure the embedding performance of different methods. A low residual variance reflects a good data representation. From the results reported in Table 1, we see that the TSIMR method gives the best data representation.

5.2. Face images

This data set contains 1965 face images taken from sequential frames of a small video [2]. The size of each image is 20×28. However, since the face images mainly include varying pose and expression, they are believed to reside on a low-dimensional manifold.

The construction of the adjacency graph uses 12 nearest neighbors, which is identical to the setup in [2]. The heat kernel is adopted to assign weights to the edges of the graph, where the parameter t is set to $5d_{av}$ with d_{av} being the average of the squared distances between all points and their most nearest neighbors. For data embedding in a 2-dimensional space, the TSIMR method provides a very tidy and compact representation as shown in Fig. 3(a). Embedding results using the other methods with the same setting of 12 nearest neighbors are given in Fig. 3(b)$\sim$3(e). The outcome of our TSIMR is more concise and orderly than the other results.

For interpretation of the embedding results, in Fig. 4 we randomly select and render a half of the original face pictures in the low-dimensional space. We identify three tendencies of variances of pose and expression, which are indicated with curves. The left curse shows a gradual change of pose from left to right, and expression from calm to grimace. The right curve corresponds to the change of expression which is first from happy to happy and grimace, and then to calm and grimace. The bottom straight line shows an interleaving between pose and expression: first, under the sad expression, pose changes from left to right; then under the happy expression, pose repeats the same "from left to right" pattern. The outcome from the TSIMR illustrates a well interpretable property on the intrinsic degrees of freedom of the face images.

To further understand the behavior of the TSIMR, we examine the locally linear representation given in (3). Vector $\mathbf{w}_{\mathbf{z}(x)}$ and matrix $T_{\mathbf{z}(x)}$ jointly determine a projection vector $\mathbf{w}_{\mathbf{z}(x)}^\top T_{\mathbf{z}(x)}$, whose effect is to perform an inner product with $\mathbf{x}$ in the original coordinate system. Consequently, similar to the eigenface representation of major eigenvectors derived from PCA, we can visualize $\mathbf{w}_{\mathbf{z}(x)}^\top T_{\mathbf{z}(x)}$ as a tangent space intrinsic manifold face whose visualization is omitted here due to space limit. Note that $\mathbf{w}_{\mathbf{z}(x)}^\top T_{\mathbf{z}(x)}$ actually represents a linear combination of the eigenvectors given by local PCA.

6. CONCLUSION

In this paper, we have proposed a new regularization method called tangent space intrinsic manifold regularization, which favors linear functions on the manifold and can perform direct out-of-sample extensions. We derived the corresponding matrix quadratic form representation and further proposed a new method for data representation whose solution tends out to be an eigen-decomposition problem. Experimental results including comparisons with other methods have shown the effectiveness of the proposed method.

7. REFERENCES

[1] M. Belkin and P. Niyogi, "Lapalcian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[2] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[3] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[4] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley, New York, 2000.

[5] D. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," in *Proceedings of the National Academy of Arts and Sciences*, 2003, pp. 5591–5596.

[6] K. Weinberger and L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *International Journal of Computer Vision*, vol. 70, pp. 77–90, 2006.

[7] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction by local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, pp. 313–338, 2004.

[8] J. Shawe-Taylor and S. Sun, "A review of optimization methodologies in support vector machines," *Neurocomputing*, vol. 74, pp. 3609–3618, 2011.

[9] V.N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[10] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *Journal of Transportation Engineering*, vol. 138, pp. 1358–1367, 2012.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, pp. 267–288, 1996.