

一种新的空间数据不确定性重建方法

张 挺¹, 刘金华²

(1. 上海电力学院计算机科学与技术学院, 上海 200090; 2. 浙江传媒学院电子信息工程系, 浙江杭州 310018)

摘 要: 在重建空间数据时, 如果条件数据较少甚至没有任何条件数据, 重建结果常常出现较多的不确定性, 此时适合采用基于统计原理的随机模拟方法重建空间数据. 多点信息统计法 (Multiple-Point Statistics, MPS) 是目前随机模拟的主流方法, 它可以将训练图像中提取的本质特征复制到重建区域. 由于传统采用线性降维的 MPS 无法较好地处理非线性数据, 而局部线性嵌入 (Locally Linear Embedding, LLE) 可以实现对非线性数据的降维, 因此提出 LLE 与 MPS 相结合的空间数据不确定性重建方法. 利用该方法对图像数据进行重建实验, 实验结果证明该方法的有效性.

关键词: 模式; 多点信息统计法; 非线性; 局部线性嵌入; 重建

中图分类号: **文献标识码:** A **文章编号:** 0372-2112 (2018)03-0641-05

电子学报 URL: <http://www.ejournal.org.cn> **DOI:** 10.3969/j.issn.0372-2112.2018.03.019

A New Indefinite Reconstruction Method for Spatial Data

ZHANG Ting¹, LIU Jin-hua²

(1. College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China;

2. Department of Electronic and Information Engineering, Zhejiang Institute of Communication and Media, Hangzhou, Zhejiang 310018, China)

Abstract: When reconstructing spatial data, if conditional data are sparse or even not existent, reconstructed results often show a lot of uncertainties, so it is appropriate to use stochastic simulation based on statistical theories to reconstruct spatial data. As one of the main stochastic simulation methods, multiple-point statistics (MPS) can copy the intrinsic features extracted from training images to the reconstructed regions. Because the traditional MPS methods using linear dimensionality reduction cannot effectively handle nonlinear data but locally linear embedding (LLE) can achieve dimensionality reduction of nonlinear data, an indefinite reconstruction method using LLE and MPS for spatial data is proposed. The experimental results for images show that the proposed method is practical.

Key words: pattern; multiple-point statistics; nonlinear; locally linear embedding; reconstruction

1 引言

数据插值成为重建空间数据的一个主要手段^[1-4]. 插值方法可分为“确定”性插值方法和“不确定”性插值方法. 不确定性插值方法的不确定性主要表现在插值结果的不确定性和多样性, 但是这些结果是对已有数据统计特征的合理反映, 具有较强的指导意义.

不确定性插值方法的一个重要分支是随机模拟方法, 而多点信息统计法 (Multiple-Point Statistics, MPS) 是目前随机模拟的主流^[5,6]. MPS 依赖从训练图像提取的模式特征实现数据重建, 因此训练图像的模式特征决

定了模拟结果, 但这些模式往往维数较高, 数据处理难度较大. 因此, 模式降维问题成为 MPS 中的研究热点. MPS 中的主要方法, 如过滤器随机模拟 (Filter-Based Stochastic Simulation, FILTERSIM)^[5] 和距离模式随机模拟 (Distance-Based Pattern Simulation, DISPAT)^[6], 均利用线性降维处理训练图像, 因此无法较好地处理非线性数据.

局部线性嵌入 (Locally Linear Embedding, LLE) 是一种非线性降维算法, 它的核心是保持降维前后近邻点之间的局部线性结构不变^[7], 因此本文提出一种利用 MPS 和 LLE 重建空间数据的方法, 拟采用 LLE 实现

收稿日期: 2016-10-11; 修回日期: 2017-04-26; 责任编辑: 梅志强

基金项目: 国家自然科学基金项目 (No. 41672114, No. 41702148); 上海市自然科学基金项目 (No. 16ZR1413200); 中石油与中科院重大战略合作项目 (No. 2015A-4812); 中国科学院战略性先导科技专项 (No. XDB10030402); 浙江省科技计划项目 (No. 2017C33163); 中央高校基本科研业务费专项资金资助 (No. WK2090050038)

训练图像的非线性降维. 由于本文方法使用了 MPS, 所以重建结果不唯一, 但是它们能够提供未来数据发展趋势的预测, 具有一定实际意义.

2 核心理念

设数据模板为 τ_D , 它是由 D 个向量组成的几何形态, $\tau_D = \{\mathbf{h}_\alpha; \alpha = 1, 2, \dots, D\}$. 设数据模板中心位置为 \mathbf{u} , 模板其他位置 $\mathbf{u}_\alpha = \mathbf{u} + \mathbf{h}_\alpha$. 假设变量 S 可以取 K 个状态值, 即有状态值的集合 $\{s_k, k = 1, \dots, K\}$. 令 $S(\mathbf{u}_\alpha)$ 表示在 \mathbf{u}_α 位置的状态值. 选取离向量 \mathbf{u} 最近的 D 个数据作为模拟时的条件数据, D 个条件数据和它们的几何结构组成了“数据事件” $d(\mathbf{u}_\alpha)$.

2.1 建立训练图像的模式库

利用数据模板扫描训练图像来捕获数据事件, 这些数据事件被称为“模式”, 而这些模式包含了训练图像的结构特征. 设 $\mathbf{B}(\mathbf{u})$ 表示以 \mathbf{u} 为中心的模式, $b(\mathbf{u} + \mathbf{h}_\alpha)$ ($\alpha = 1, 2, \dots, D$) 表示 $\mathbf{u} + \mathbf{h}_\alpha$ 位置的状态值, 那么有:

$$\mathbf{B}(\mathbf{u}) = (b(\mathbf{u} + \mathbf{h}_1), b(\mathbf{u} + \mathbf{h}_2), \dots, b(\mathbf{u} + \mathbf{h}_D)) \quad (1)$$

可见模式 $\mathbf{B}(\mathbf{u})$ 和数据模板尺寸相同, 如果将各个模式从训练数据中抽取出来形成模式库, 那么与具体位置 \mathbf{u} 无关, 此时可设模式库中的模式个数为 N_p , 那么第 k 个模式可以表示为:

$$\mathbf{P}_k = (p_k(\mathbf{h}_1), p_k(\mathbf{h}_2), \dots, p_k(\mathbf{h}_D)), k = 1, 2, \dots, N_p \quad (2)$$

其中 $p_k(\mathbf{h}_i)$ 与 $b(\mathbf{u} + \mathbf{h}_i)$ 一一对应. 整个模式库 \mathbf{P} 可以表示为:

$$\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_{N_p}) \quad (3)$$

上式每个 \mathbf{P}_k ($k = 1, 2, \dots, N_p$) 可以视为模式空间中的一个点.

2.2 模式的降维

本文采用 LLE 实现模式的非线性降维. 每个模式 \mathbf{P}_k ($k = 1, 2, \dots, N_p$) 含有 D 个分量, 可以视其维数为 D . 通过 LLE 将 \mathbf{P} 中的各个模式 \mathbf{P}_k 降维到 d ($d \ll D$) 维空间. 降维后得到的低维数据集设为 $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{N_p})$ (其中 $\mathbf{y}_k \in \mathbb{R}^d, k = 1, 2, \dots, N_p$), 即样本个数不变, 每个样本的维数从 D 变为 d . 利用 LLE 降维的步骤如下^[7,8]:

第 1 步, 确定邻近节点. 对于每个模式 \mathbf{P}_i , 通过欧氏距离选择 K 个离其最近的邻近点. 两个模式 \mathbf{P}_i 和 \mathbf{P}_j 的欧氏距离 $d(\mathbf{P}_i, \mathbf{P}_j)$ 定义如下:

$$d(\mathbf{P}_i, \mathbf{P}_j) = \sqrt{\sum_{\alpha=1}^D (p_i(\mathbf{h}_\alpha) - p_j(\mathbf{h}_\alpha))^2}, \quad (4)$$

($i, j = 1, 2, \dots, N_p$)

第 2 步, 计算重建权值矩阵 \mathbf{W} . 对于每个模式点, 需要计算与其周围邻近点的重构权值 w_{ij} . 如果两个点不是邻近点, 那么 $w_{ij} = 0$; 否则 $\sum_{j=1}^{N_p} w_{ij} = 1$. 重构权值矩

阵 \mathbf{W} 可以通过对下面式(5)的最小化运算求出:

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{N_p} \left| \mathbf{P}_i - \sum_{j=1}^{N_p} w_{ij} \mathbf{P}_j \right|^2 \quad (5)$$

第 3 步. 计算低维映射 \mathbf{Y} . 定义一个损失函数:

$$\Phi(\mathbf{Y}) = \sum_{i=1}^{N_p} \left| \mathbf{y}_i - \sum_{j=1}^{N_p} w_{ij} \mathbf{y}_j \right|^2 \quad (6)$$

低维映射 \mathbf{Y} 可以通过对式(6)求最小化获得, 两个约束条件是 $\sum_{i=1}^{N_p} \mathbf{y}_i = \mathbf{0}$ 和 $\sum_{i=1}^{N_p} \mathbf{y}_i \mathbf{y}_i^T / N_p = \mathbf{I}$, 这里的 \mathbf{I} 是 $d \times d$ 阶单位阵. 式(6)还可以表述为:

$$\Phi(\mathbf{Y}) = \text{tr}(\mathbf{Y}^T \mathbf{M} \mathbf{Y}),$$

$$\text{其中 } \sum_{i=1}^{N_p} \mathbf{y}_i = \mathbf{0} \text{ 且 } \sum_{i=1}^{N_p} \mathbf{y}_i \mathbf{y}_i^T / N_p = \mathbf{I} \quad (7)$$

上式中 $\mathbf{M} = (\mathbf{I} - \mathbf{W}^T)(\mathbf{I} - \mathbf{W})$, $\text{tr}(\cdot)$ 表示求矩阵的迹. 至此获得 $\mathbf{P} = \{\mathbf{P}_k\}$ ($\mathbf{P}_k \in \mathbb{R}^D, k = 1, 2, \dots, N_p$) 的低维映射 $\mathbf{Y} = \{\mathbf{y}_k\}$ ($\mathbf{y}_k \in \mathbb{R}^d, k = 1, 2, \dots, N_p$), 数据从高维空间 \mathbb{R}^D 映射到低维空间 \mathbb{R}^d . 低维空间的维数 d 可以通过最大似然估计法获得.

2.3 模式的分类

在完成训练图像中模式的降维后, 需要对获取的低维模式进行分类. 对这些模式采用基于密度的聚类方法 (Density-Based Clustering Algorithm, DBSCAN) 进行分类. 在利用 DBSCAN 划分模式空间后, 获得若干“子空间”, 每个子空间命名为 Cell. 对于每个 Cell, 可以定义一个与数据模板相同形状的“平均模板”, 称为 Prototype. Prototype 是属于该 Cell 的所有模式在原高维空间中各节点位置的均值. Prototype 可以视为每个 Cell 中所有模式的均值, 是该 Cell 所有模式的“代表”.

Prototype 的值 $e^{(l)}(\mathbf{h}_\alpha)$ 表示属于一个 Cell 中的所有图案在各个 \mathbf{h}_α 位置的均值, 定义为:

$$e^{(l)}(\mathbf{h}_\alpha) = \frac{1}{c_l} \sum_{i=1}^{c_l} T(\mathbf{u}_i^{(l)} + \mathbf{h}_\alpha), l = 1, \dots, L \quad (8)$$

其中 L 是所有非空 Cell 数目; c_l 表示属于第 l 个 Cell 的所有模式数目; $\mathbf{u}_i^{(l)}$ ($i = 1, \dots, c_l$) 是属于第 l 个 Cell 中的各个模式所对应数据模板的中心位置; T 表示训练图像; $T(\mathbf{u}_i^{(l)} + \mathbf{h}_\alpha)$ 是训练图像在 $\mathbf{u}_i^{(l)} + \mathbf{h}_\alpha$ 位置的状态值. 第 l 个 Prototype 表示为:

$$\mathbf{E}_l = (e^{(l)}(\mathbf{h}_1), e^{(l)}(\mathbf{h}_2), \dots, e^{(l)}(\mathbf{h}_D)), l = 1, \dots, L \quad (9)$$

\mathbf{E}_l 可以视为第 l 个 Cell 对应的平均模式.

2.4 模式的提取

利用扫描训练图像时使用的数据模板扫描待重建区域, 检索以未知点 \mathbf{u} 为中心的数据事件内的所有已知节点, 这些点视为条件数据. 设条件数据点数目为 n' ($\leq D$). 将数据模板内的节点分为两个部分: inner 部分和 outer 部分. 它们可以作为其他后继节点重建时的条

件数据.在完成一次模拟之后,inner 部分会被“固定”为重建结果,不会再对其进行模拟;而模板的 outer 部分不仅作为其他节点模拟的条件数据,而且该部分的所有节点会被当作未知点重新模拟.

设 t 表示数据类型,数据模板捕获的重建区域内的数据事件设为 patch,则 patch 中一共可能有 3 种已知数据:

(1) $t=1$:原始条件数据,它们被分配到各节点位置上;

(2) $t=2$:已经模拟过的节点,它们被固定为条件数据,即 patch 的 inner 部分(不包含 inner 部分内的原始条件数据);

(3) $t=3$:通过“粘贴”Cell 中的 patch 而已知的节点,但这些点会被重新模拟,即 patch 的 outer 部分(不包含 outer 部分内的原始条件数据).

另外定义“距离函数”如下:

$$D(d(\mathbf{u}_\alpha), E_l) = \sum_{\alpha=1}^d \omega(t) |d(\mathbf{u} + \mathbf{h}_\alpha) - e^{(l)}(\mathbf{h}_\alpha)| \quad (10)$$

上式中 $t=1, 2, 3$. $D(d(\mathbf{u}_\alpha), E_l)$ 表示求取数据事件 $d(\mathbf{u}_\alpha)$ 和 E_l 中对应的已知节点间的距离. 每种节点会根据其类型而给定一个权值 $\omega(t)$, 表示其在求取距离函数中的重要性, 注意有 $\omega(3) \leq \omega(2) \leq \omega(1)$. 可见原始条件数据点在距离函数中的作用最大.

令 D_l 表示 $d(\mathbf{u}_\alpha)$ 与各个 E_l 之间的距离, 即:

$$D_l = D(d(\mathbf{u}_\alpha), E_l), l=1, \dots, L \quad (11)$$

上式中 $D(\cdot)$ 表示利用式(10)求距离, 可得距离向量:

$$\mathbf{U}_L = (D_1, D_2, \dots, D_L) \quad (12)$$

在获取 \mathbf{U}_L 之后, 从 D_1, D_2, \dots, D_L 中寻找最小值, 假设该最小值对应序号为 R , 则从 E_R 对应的 Cell 中随机提取出一个模式, 然后将其复制到以当前模拟点 \mathbf{u} 为中心的待重建区域; 再选择该模式的部分节点作为 outer 部分用于后继模拟, 同时“固定”该模式的其他节点作为 inner 部分的节点. 至此完成一个模式的提取.

3 重建方法的完整流程

根据初始条件数据的不同, 节点 \mathbf{u} 的重建可以分为以下 2 种情况: (1) 无条件数据模拟(称为 C1 情况); (2) 有条件数据的模拟(称为 C2 情况). 第 2 节对本方法各核心部分作了介绍, 将上述部分整合起来, 形成本方法的完整流程:

- ①、利用数据模板扫描训练图像, 构建模式库 P .
- ②、利用 LLE 对模式库中的模式进行降维.
- ③、采用 DBSCAN 对降维后的模式进行分类.
- ④、定义一条随机访问路径, 对重建区域内的未知节点进行访问.

⑤、检查访问路径上的待模拟节点 \mathbf{u} 是否已知. 如果是已知的条件数据或已模拟节点, 则对随机路径上的下个节点进行判定; 否则转向步骤⑥;

⑥、检索以 \mathbf{u} 为中心的数据模板内的条件数据(设数目为 n'). 根据条件数据的不同, 提取模式的方法分为以下 2 种情况:

(1) 如果条件数据属于 C1 情况(即 $n'=0$, 无条件数据), 那么从训练图像模式库 P 中随机提取一个模式, 然后将该模式直接复制到以当前模拟节点 \mathbf{u} 为中心的重建区域.

(2) 如果条件数据属于 C2 情况(即 $n'>0$), 即以 \mathbf{u} 为中心的数据事件非空, 那么就在训练图像所有的 E_l 中利用式(10)寻找与当前数据事件最接近的 E_l . 一旦该 E_l 被确定, 则从与之对应的 Cell 中随机提取一个模式, 然后将它复制到以当前模拟节点 \mathbf{u} 为中心的重建区域.

⑦、将模式复制到重建区域后, 再将模式的部分节点固定为 inner 部分, 该模式的其他节点作为 outer 部分.

⑧、重复步骤⑤到⑦, 继续对其他节点进行模拟, 直到随机路径上的所有节点被模拟完毕.

4 实验结果及分析

4.1 重建质量的评价标准

本文评价重建结果时并不直接比较重建结果和训练图像是否相同, 而是通过多个重建结果的均值, 以及变差函数(variogram)来衡量. 变差函数能够反映变量在某个方向上空间结构变化的相关性和变异性, 定义如下:

$$\gamma(\mathbf{h}) = \frac{1}{2} E[Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})]^2 \quad (13)$$

其中 $Z(\mathbf{u})$ 和 $Z(\mathbf{u} + \mathbf{h})$ 分别表示位置 \mathbf{u} 和 $\mathbf{u} + \mathbf{h}$ 处的变量状态值. \mathbf{h} 表示距离向量. E 表示求数学期望值.

4.2 重建实验

为了验证本方法有效性, 分别对四幅图像 Lena、Baboon、Boat 和 Pepper 进行重建实验. 四幅图像如图 1 所示, 大小均为 $256 * 256$ 像素. 所有实验都是在 Intel core i3-4160T (3.1 GHz CPU) 和 4GB 内存环境下运行的. 公式(10)中的权值 $\omega(1) = 0.6$, $\omega(2) = 0.25$, $\omega(3) = 0.15$. 为了简便, 本文提出的方法简称为 LLESIM.

分别从图 1(a) ~ (d) 中随机抽取 1000 个像素作为重建时的条件数据, 即占每幅图像像素点的 1.526%. 基于上述训练图像和条件数据, 利用 FILTERSIM、DISPAT 和 LLESIM 分别重建 100 幅的 Lena、Baboon、Boat 和 Pepper 图像, 重建区域均为 $256 * 256$ 像素. 根据最大似然估计法, LLESIM 和 DISPAT 的低维空间维数 d 设置

为 4. FILTERSIM 的低维维数为 6. 模板尺寸为 13×13 像素, inner 部分设置为 7×7 像素. LLESIM 方法重建 Lena、Baboon、Boat 和 Pepper 图像的重建结果各取 1 幅,

见图 2. 利用 FILTERSIM 和 DISPAT 生成的重建图像见图 3 和图 4. 可见 LLESIM 重建图像较为清晰.

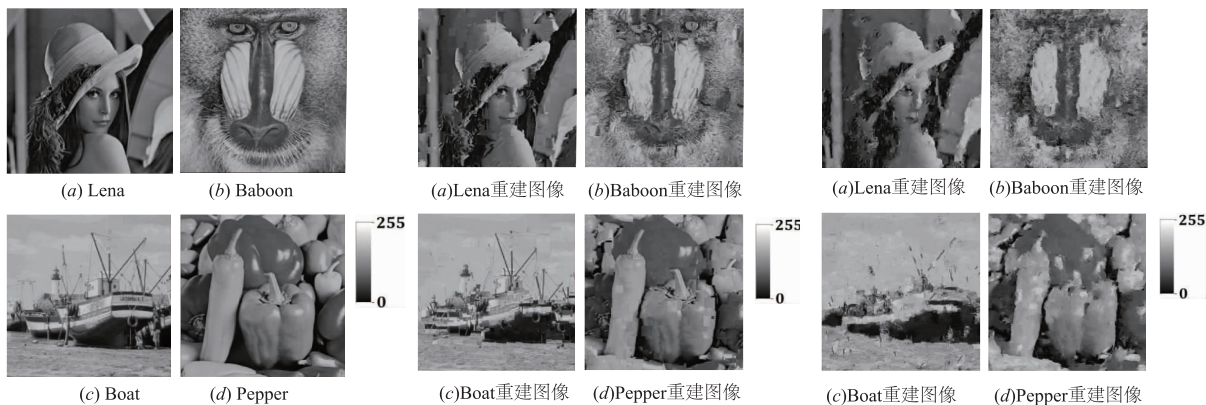


图1 四幅训练图像

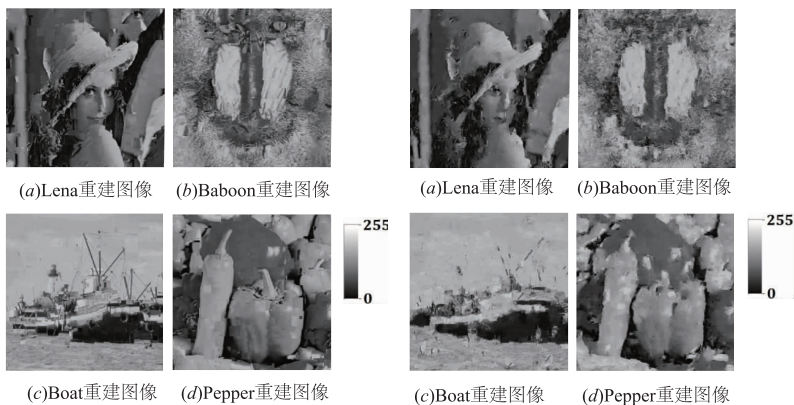


图2 利用LLESIM重建的Lena、Baboon、Boat和Pepper图像

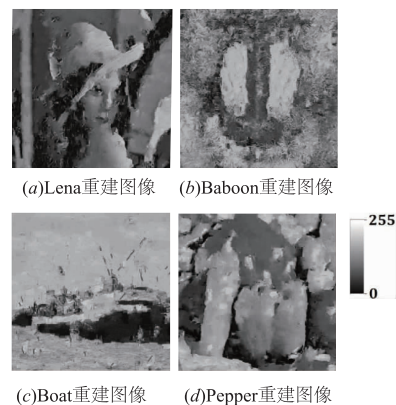


图3 利用FILTERSIM重建的Lena、Baboon、Boat和Pepper图像

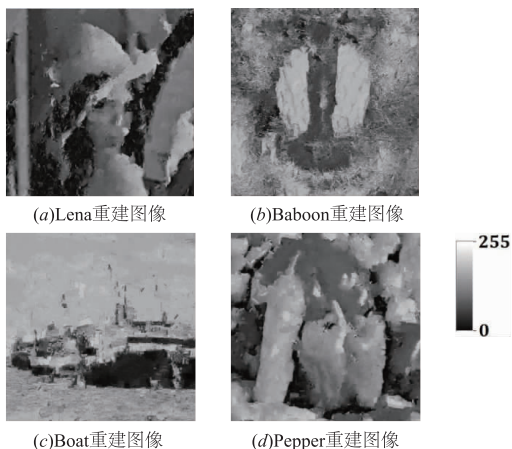
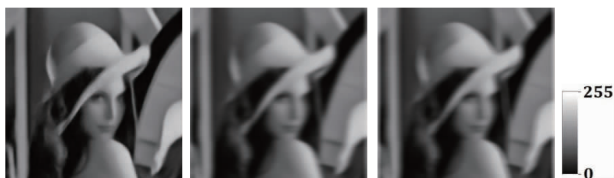


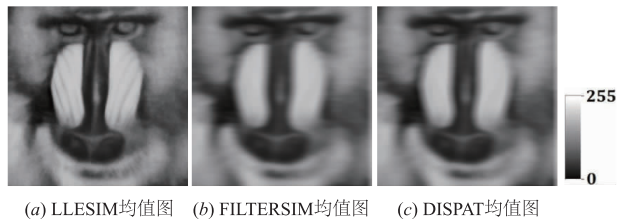
图4 利用DISPAT重建的Lena、Baboon、Boat和Pepper图像

对 LLESIM, FILTERSIM 和 DISPAT 分别生成的 100 幅重建图像求均值, 见图 5 ~ 图 8. LLESIM 生成的均值图比 FILTERSIM 和 DISPAT 的均值图更加清晰, 与训练图像更为相似, 说明 LLESIM 各重建图像间的差异性较小, 因此重建质量较高. LLESIM, FILTERSIM 和 DISPAT 生成的重建图像和训练图像在 X 和 Y 方向的变差函数均值如图 9 ~ 图 12 所示. 可见 LLESIM 重建图像和训练图像的变差函数最为接近.



(a) LLESIM均值图 (b) FILTERSIM均值图 (c) DISPAT均值图

图5 利用三种方法生成的100幅Lena重建图像的均值图



(a) LLESIM均值图 (b) FILTERSIM均值图 (c) DISPAT均值图

图6 利用三种方法生成的100幅Baboon重建图像的均值图



(a) LLESIM均值图 (b) FILTERSIM均值图 (c) DISPAT均值图

图7 利用三种方法生成的100幅Boat重建图像的均值图



(a) LLESIM均值图 (b) FILTERSIM均值图 (c) DISPAT均值图

图8 利用三种方法生成的100幅Pepper重建图像的均值图

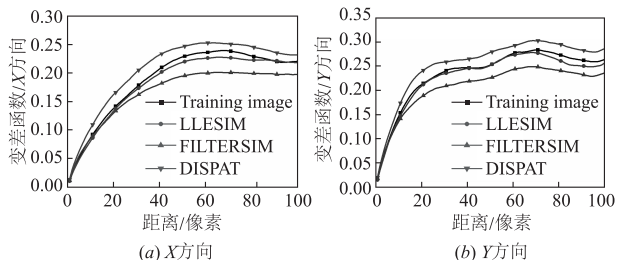


图9 Lena重建图像和训练图像在 X 和 Y 方向的变差函数均值

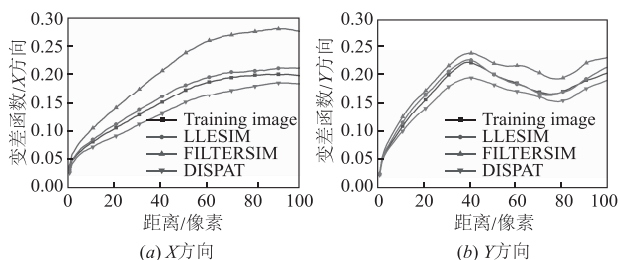


图10 Baboon重建图像和训练图像在X和Y方向的变差函数均值

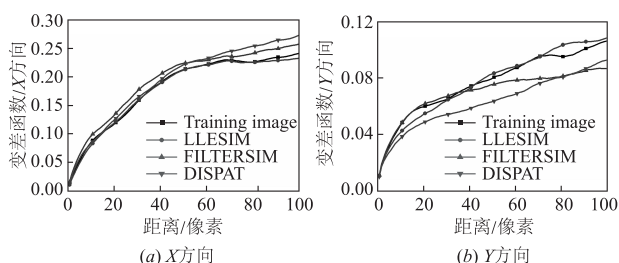


图11 Boat重建图像和训练图像在X和Y方向的变差函数均值

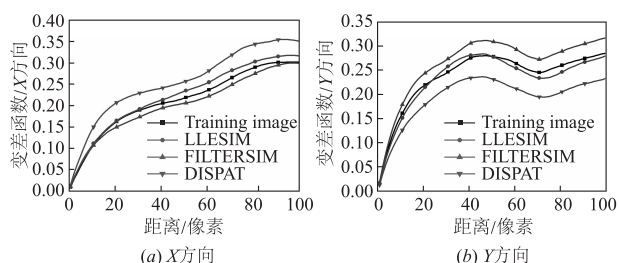


图12 Pepper重建图像和训练图像在X和Y方向的变差函数均值

5 总结

在传统多点信息统计法对训练图像数据降维的过程中,主要面向线性数据进行处理,而对于非线性数据没有提出处理方法,这会导致降维效果较差.一些经典的MPS方法(如DISPAT和FILTERSIM)只是利用线性降维方法处理训练图像数据,这将极大地影响MPS重建空间数据的效果.

本文结合LLE和MPS对训练图像中的高维模式进行降维.由于LLE是一种经典的非线性降维方法,适用于非线性数据的降维,因此可以有效降低训练图像中非线性数据的维度.与传统采用线性降维的DISPAT和FILTERSIM实验比较,可以看出本方法的优势,并且本方法重建的多种基于统计原则的不确定性结果可以作为风险评测和分析决策的指导工具.

参考文献

- [1] 张立峰,刘昭麟,田沛.基于压缩感知的电容层析成像图像重建算法[J].电子学报,2017,45(2):353-358.
Zhang Li-feng, Liu Zhao-lin, Tian Pei. Image reconstruction algorithm for electrical capacitance tomography based on

compressed sensing[J]. Acta Electronica Sinica, 2017, 45(2):353-358. (in Chinese)

- [2] Vedadi F, Shirani S. A map-based image interpolation method via Viterbi decoding of Markov chains of interpolation functions[J]. IEEE Transactions on Image Processing, 2014, 23(1):425-438.
- [3] 詹毅,李梦.非局部特征方向图像插值方法研究[J].电子学报,2016,44(5):1064-1070.
Zhan Yi, Li Meng. Research on image interpolation with non-local feature directions[J]. Acta Electronica Sinica, 2016, 44(5):1064-1070. (in Chinese)
- [4] 叶程,刘真,吴明光.基于球坐标及三角形插值的颜色信号色域映射算法[J].电子学报,2015,43(11):2180-2186.
Ye Cheng, Liu Zhen, Wu Ming-guang. Gamut mapping algorithm for color signal based on spherical coordinates and triangular interpolation[J]. Acta Electronica Sinica, 2015, 43(11):2180-2186. (in Chinese)
- [5] Zhang T F. Filter-based training pattern classification for spatial pattern simulation[D]. Palo Alto; Stanford University, 2006.
- [6] Honarkhah M. Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling[D]. Palo Alto; Stanford University, 2011.
- [7] Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(22):2323-2326.
- [8] Liu X, Tosun D, Weiner M, et al. Locally linear embedding (LLE) for MRI based Alzheimer's disease classification[J]. NeuroImage, 2013, 83(6):148-157.

作者简介



张挺男,1979年9月生,安徽安庆人.2009年获得中国科学技术大学博士学位,现为上海电力学院计算机科学与技术学院副教授,主要研究方向为图像重建.
E-mail: tingzh@shiep.edu.cn



刘金华(通讯作者)女,1972年11月生,黑龙江佳木斯人,2008年获得杭州电子科技大学硕士学位,现为浙江传媒学院电子信息学院实验师,主要研究方向为应用电子、图像处理.
E-mail: liujinh@163.com