

文章编号 1004-924X(2009)03-0641-07

应用 SLLE 实现手写体数字识别

杨晓敏, 吴 炜, 何小海, 陈 默

(四川大学 电子信息学院 图像信息研究所, 四川 成都 610064)

摘要: 针对在手写字符识别中由于书写习惯和风格的不同而造成的字符模式不稳定问题, 提出了一种基于流形学习的手写体数字识别方法。在流形学习非监督的基础上引入了监督信息, 从而保证高维到低维的映射在保留流形某些结构的同时也可进一步分离不同类别的流形。算法首先利用基于监督的局部线性嵌入 (SLLE) 对手写体数字图像进行字符特征的降维, 然后再对降维后的特征进行分类识别。对 MINST 库中手写体数字数据库进行了实验, 实验结果表明, 利用 SLLE 降维以后的特征能够有效地区分字符, 识别率可达到 93.27%; 由于具有较好的识别率, 能够发现高维空间的低维嵌入流形。

关键词: 流形学习; 监督局部线性嵌入; 手写字符识别; 非线性降维

中图分类号: TP391.43 **文献标识码:** A

Realization of handwritten numeral character recognition by supervised locally linear embedding

YANG Xiao-min, WU Wei, HE Xiao-hai, CHEN Mo

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China)

Abstract: In order to improve the instability of handwritten character pattern caused by different writing styles, a novel handwritten numeral character recognition approach based on manifold learning is proposed in this paper. Based on non-supervised manifold learning, a supervised information is induced to the algorithm to ensure the map from high dimension to low dimension to retain some manifold structures and also to separate different kinds of manifolds. By proposed method, Supervised Locally Linear Embedding (SLLE) algorithm is used to reduce the dimensionality of input feature. Then, the reduced feature is classified by simple classifier. Finally, the proposed algorithm is tested on the characters in MINST character database. The experimental results demonstrate that the method can effectively improve the recognition rate to 93.27% and can provide a new approach to the research of handwritten numeral character recognition.

Key words: manifold learning; supervised locally linear embedding (SLLE); handwritten character recognition; nonlinear reduction dimensionality

1 引言

机器学习的一大目标是要从数据中学习其相关的规律性。然而,不同的数据往往呈现不同的特点。这使得传统的机器学习方法难以直接用来分析来源于真实世界的高维和非线性数据,如文本、图像、语音等。研究高维非线性数据集面临着诸多困难,这主要是由于高维带来了数据的稀疏和维数灾难^[1-2],非线性使得人们不能有效地利用线性模型进行研究。经研究发现,许多高维采样数据都是由少数隐含变量所决定的,而这些隐含变量是以嵌套在高维欧式空间的组合型非线性流形存在的。例如:人脸表情图像存在内在较低的维数^[3-6]。因此,用基于数据分布的内在维数来分析数据是机器学习和多元数据分析的重要研究方向,流形学习^[3-4]方法为我们提供了一种新的研究途径。

流形学习的数学定义:在 \mathbf{R}^D 空间($D \gg d$)中存在由某个随机过程生成的数据 $\{y_i\} \subset Y$, 经过某个函数 f 可以映射形成 \mathbf{R}^D 空间中的观测数据 $\{x_i = f(y_i)\}$ 。流形学习的目标是要在观测数据 $\{x_i\}$ 中重构 f 和数据 $\{y_i\}$, 以达到数据压缩和降维的目的。与以往的机器学习不同的是流形学习强调整体结构,要通过局部和整体相结合来发现和重建数据的内在规律性。局部线性嵌套(LLE)是流形学习中一种主要的算法,是一种新的针对非线性数据的降维技术。它是将非线性数据局部线性化,找出非线性数据之间的拓扑关系,并成功地将高维数据映射到低维空间中,且不改变原有的相对位置,因此可以广泛地应用于非线性数据的降维。局部线性嵌入本身是一种非监督学习方法,这类方法通常用于数据挖掘和数据可视化,这些过程中数据的类别信息以及各类之间的关系是未知的,并没有考虑到训练样本的类别信息。而基于监督的局部线性嵌入(SLLE)在寻求每个数据点的邻接点过程中利用了样本的类别信息,因此基于监督的局部线性嵌入提高了降维的精度。

手写体阿拉伯数字字符识别是 OCR(Optical Character Recognition)的一个重要组成部分,尤其是在邮政编码、统计报表、财务报表、银行票据等方面有着广泛的用途,在过去的 30 年里,一直

是图像处理、模式识别领域的研究热点。本文针对手写数字字符书写随意、字符模式不稳定的特点,将 SLLE 算法应用于手写体数字图像特征的降维中,并对降维以后的手写体数字特征进行分类。通过对 MINST 字库的一系列实验表明,该算法可以发现嵌套在字符原始特征中的低维流形,能够达到较高的识别率,具有令人满意的应用效果。

2 LLE 算法介绍

LLE 算法^[7-12]是一种从高维空间非线性映射到低维空间的非监督方法,可以广泛地应用于图像数据的分类与聚类、文字识别、多维数据的可视化、图像分割、检索等领域。LLE 算法的主要思想是:对于一组具有嵌套流形的数据集,在嵌套空间与内在低维空间局部邻域间的点的关系应该不变。即嵌套空间中的每个采样点可以用它的近邻点线性表示,在低维空间中保持每个邻域中的权值不变,重构原数据点,使重构误差最小。图 1 是一个 LLE 用于降维的例子。LLE 将三维空间的数据(图 1(b)),映射到二维空间(图 1(c))中。

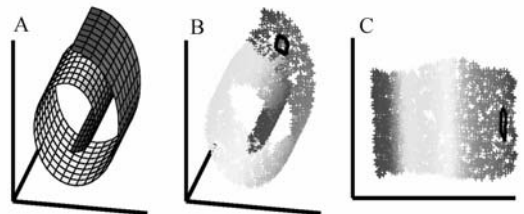


图 1 Swissroll 的 LLE 降维实验
 (a) Swissroll data (b) Sampled data (c) LLE-mapped

图 1 Swissroll 的 LLE 降维实验

Fig. 1 Dimensionality reduction of Swissroll by LLE

LLE 的具体做法为:设在高维欧氏空间 \mathbf{R}^D 中有数据集 $X = \{x_1, x_2, \dots, x_N\}$, 该方法希望将 X 嵌入到一个相对低维的空间 \mathbf{R}^d 中($d \ll D$), 同时尽可能地保持原数据的拓扑结构。

LLE 算法可以归纳为如下 3 个步骤:

步骤 1: 寻找 \mathbf{R}^D 空间中每一个样本点 x_i 欧式距离最近的 K 个近邻点, $i = 1, 2, \dots, N$. K 为预先设定的参数。

步骤 2: 认为空间中的每一个样本点 x_i 可以用它的 K 个近邻线性表示, 即每一个样本点用它

的 K 个近邻点重建。由每个样本点的近邻计算出该样本点的权值矩阵。

定义代价函数:

$$\epsilon(\mathbf{W}) = \sum_i \left| x_i - \sum_j W_{ij} x_{ij} \right|^2, \quad (1)$$

式中 $x_{ij} (j=1, 2, \dots, K)$ 为 x_i 的 K 个近邻点; W_{ij} 为 x_i 和 x_{ij} 的权值关系。 W_{ij} 可以看作是近邻点对重构样本点做出的贡献, 并且权值要满足

$$\sum_{j=1}^K W_{ij} = 1.$$

因此:

$$\epsilon(\mathbf{W}) = \sum_i \left| \sum_j (\mathbf{x}_i - W_{ij} \mathbf{x}_{ij}) \right|^2 = \sum_{j=1}^K \sum_{m=1}^K W_{ij} W_{im} \mathbf{Q}_{jm}^i, \quad (2)$$

其中:

$$\mathbf{Q}_{jm}^i = (\mathbf{x}_i - \mathbf{x}_{ij})^T (\mathbf{x}_i - \mathbf{x}_m). \quad (3)$$

\mathbf{Q}_{jm}^i 也可以写成:

$$\mathbf{Q}_{jm}^i = \frac{1}{2} (D_{i,N(j)} + D_{i,N(m)} - D_{N(j),N(m)}). \quad (4)$$

求最优权值就是对于公式(1)在约束条件下求解最小二乘问题。利用拉格朗日乘子法, 即可求出局部最优重建权值矩阵:

$$W_{ij} = \frac{\sum_{m=1}^K R_{jm}^i}{\sum_{p=1}^K \sum_{q=1}^K R_{pq}^i}, \quad (5)$$

其中:

$$\mathbf{R}^i = (\mathbf{Q}^i)^{-1}. \quad (6)$$

步骤3: 保持权值不变, 在低维空间 \mathbf{R}^d 中 ($d \ll D$) 中对原数据点重构。将所有的样本点映射为低维空间中的数据点, 并使输出数据在低维空间中保持原有的拓扑结构。设低维空间的数据点为 \mathbf{y}_i , 可以通过求最小的代价函数(7)得到。

$$\Phi(\mathbf{y}) = \sum_i \left| \mathbf{y}_i - \sum_j W_{ij} \mathbf{y}_j \right|^2. \quad (7)$$

式(7)的最优解要满足如下条件:

$$\sum_i \mathbf{y}_i = 0, \quad (8)$$

$$\frac{1}{N} \sum_i \mathbf{y}_i \mathbf{y}_i^T = \mathbf{I}. \quad (9)$$

式(8)的目的是保持平移不变性, 式(9)的目的是避免产生退化解。式(7)可以重写成:

$$\Phi(\mathbf{y}) = \sum_{i,j} M_{i,j} \mathbf{y}_i^T \mathbf{y}_j. \quad (10)$$

式(10)中 \mathbf{M} 是一个 $n \times n$ 的对称矩阵

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W}). \quad (11)$$

由式(10)可知, 要使 $\Phi(\mathbf{y})$ 值达到最小, 则取 \mathbf{y} 为 \mathbf{M} 的最小 d 个非零特征值所对应的特征向量。

3 SLLE 算法介绍

局部线性嵌入本身是一种非监督学习方法, 这类方法通常用于数据挖掘和数据可视化, 在这些过程中数据的类别信息以及各类之间的关系是未知的, 并没有考虑到训练样本的类别信息。因此可以加入样本的类别信息。基于监督的局部线性嵌入(SLLE)^[9]在寻求每个数据点的邻接点过程中利用了样本的类别信息。即一个给定的数据点的邻接点仅仅是从与其类别相同的点中选取。达到这一目的的方法之一就是增大不同类样本 x_i 和 x_j 的距离:

$$D_{i,j}' = D_{i,j} + \alpha \text{Max}(D_{i,j}) \Delta_{i,j}, \quad (12)$$

其中:

$$\Delta_{i,j} = \begin{cases} 1 & \omega_j \neq \omega_i \\ 0 & \text{otherwise} \end{cases}. \quad (13)$$

公式(12)中, $D_{i,j}$ 表示原始的距离, $\text{Max}(D_{i,j})$ 是训练样本集合中的最大距离。 $\alpha \in (0, 1)$ 表示类别信息对数据集合的贡献。事实上 SLLE 可以看做是一个非线性的 Fisher 映射, 而 α 表示监督的程度。对于 $\alpha \in (0, 1)$, 从高维到低维的映射在保留流形某些结构的同时, 也进一步分离了不同类别的流形。

4 手写体数字识别过程

由于构成样本的原始图像尺寸往往有多种, 因此首先要把原始图像归一化成一个标准尺寸为 $m \times n$ 的图像, 如图2所示, 则每一幅归一化以后



图2 部分实验字库样本

Fig. 2 Part of samples of handwritten numeral characters

的图像都可以看作是 $m \times n$ 维特征向量。根据第 1, 2 部分的论述, 可以看到 SLLE 的计算与样本点(训练样本即原始图像)的个数相关, 当样本点的个数较大时, 求解过程的计算量就会增加。因此本文提出在原始高维空间对每一类数字字符 c ($c=0, 1, 2, \dots, 9$) 的训练样本进行类内聚类^[13-15]以减少求解过程的计算量。将聚类结束后的每个聚类的均值作为数字字符的最终训练样本。

聚类算法是以整个数据集内部存在若干“分组”或“聚类”为出发点产生的一种数据描述方法, 每个子集中的点都有高度的相似性。类内聚类以后每一类的训练样本数目减少, 进而减少了 SLLE 求解过程的计算量。类内聚类以后得到高维欧氏空间 \mathbf{R}^D 中的最终训练样本集 $X = \{x_1, x_2, \dots, x_N\}$, 并构成训练样本矩阵 $\mathbf{X}_{D \times N}$ 。其中 x_i 为归一化以后的数字图像的二值化特征, $D = m \times n$, 为训练样本的输入维数, N 为训练样本的个数。训练样本在低维空间 \mathbf{R}^d 的输出为 $Y = \{y_1, y_2, \dots, y_N\}$, 构成输出矩阵 $\mathbf{Y}_{d \times N}$, d 为训练样本的输出维数。

4.1 训练过程

(1) 根据式(12)寻找 \mathbf{R}^D 空间中每一个最终训练样本点 x_i 的 K 个近邻点。

(2) 由每个最终训练样本点的近邻点计算出该样本点的局部重建权值矩阵。

(3) 由该最终训练样本点的局部重建权值矩阵和其近邻点计算出低维的空间 \mathbf{R}^d 中训练样本的输出为 $Y = \{y_1, y_2, \dots, y_N\}$ 。

4.2 识别过程

假设待识别样本 x_{new} 与训练样本是同分布的。采用如下算法得到测试样本的输出结果。算法步骤:

(1) 将 x_{new} 加入 \mathbf{X} 矩阵中, 则 \mathbf{X} 变为 $D \times (N + 1)$ 的矩阵。在训练样本中找 x_{new} 的 K 个近邻点。

(2) 根据式(12)分别求 x_{new} 在 $x_{\text{new}} \in c$ ($c=0, 1, 2, \dots, 9$) 假设条件下与其 K 个近邻点的权值系数, 且满足式(14)最小

$$\epsilon(W_{x_{\text{new}}}) = \left| x_{\text{new}} - \sum_j W_{\text{new},j} x_{\text{new},j} \right|^2, \quad (14)$$

式中: $\sum_j W_{\text{new},j} = 1$; $x_{\text{new},j}$ ($j=1, 2, \dots, k$) 为 x_{new} 的 K 个近邻点; $W_{\text{new},j}$ 为 x_{new} 与其近邻点 $x_{\text{new},j}$ 间的权值。

(3) 在 $x_{\text{new}} \in c$ ($c=0, 1, 2, \dots, 9$) 假设条件下分别计算 x_{new} 的低维输出向量。

$$y_{\text{new}} = \sum_j W_{\text{new},j} y_{\text{new},j}, \quad (15)$$

式中, $y_{\text{new},j}$ 为 $x_{\text{new},j}$ 对应的低维输出向量。

(4) 在 $x_{\text{new}} \in c$ ($c=0, 1, 2, \dots, 9$) 假设条件下分别计算求出 y_{new} 与低维的空间 \mathbf{R}^d 中训练样本 y_1, y_2, \dots, y_N 的最近距离

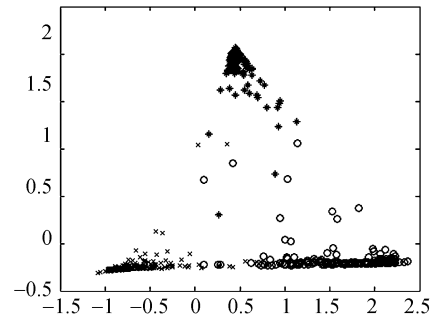
$$\text{dis}_c = \min\{\text{dis} \| y_{\text{new}} - y_i \| \mid i=1 \dots N\}. \quad (16)$$

(5) 求 x_{new} 的类别 c :

$$c = \arg \min_c \text{dis}_c. \quad (17)$$

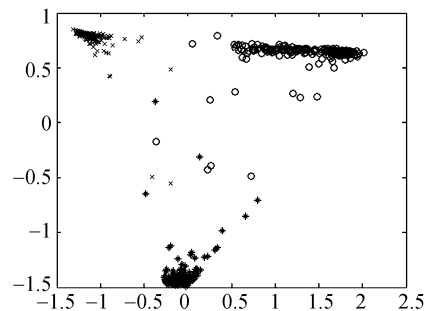
5 实验结果与分析

实验使用 MINST 数据库中的手写数字字符。从中选取 1 800 个训练样本和 450 个测试样本, 每个样本都只经过简单的二值化并归一化为 17×33 ($=561$) 像素的图像。图 2 为部分字符的图像, 可以看出该字库中各字符笔画的形态、粗细和灰度等级的差别都很明显。图 3 为手写体数字



(a) LLE 算法

(a) LLE algorithm



(b) SLLE 算法

(b) SLLE algorithm

图 3 手写体数字‘0’, ‘2’, ‘6’低维空间的可视化图

Fig. 3 Visualization results of characters ‘0’, ‘2’, ‘6’ in low dimension

‘0’, ‘2’, ‘6’分别采用 LLE 和 SLLE 降维为 3 维时的可视化效果图。其中‘x’代表数字‘0’, ‘o’代表‘6’, ‘*’代表‘2’。可以看出通过 SLLE 降维以后,可以更容易地将它们分开。

SLLE 算法需要设置邻近点的个数和输出维数 d 。 K 的选取在算法中起到关键因素,如果 K 取值太大, LLE 不能体现局部特性;反之, LLE 不能保持样本点在低维空间中的拓扑结构。图 4 反映了当 $d=18$ 时,近邻数对手写体数字识别率的影响。

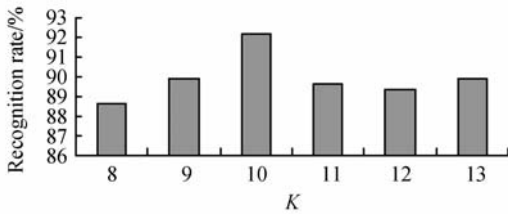


图 4 近邻数 K 对识别率的影响

Fig. 4 Influence of K on recognition rate

如果输出维数 d 选得太高,输出数据会受到噪声的影响;反之,则不能正确地提取样本数据的固有特征。图 5 为当 $K=10$ 时,输出维数 d 对识别率的影响。

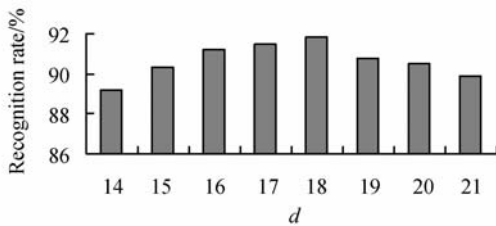


图 5 d 对识别率的影响

Fig. 5 Influence of d on recognition rate

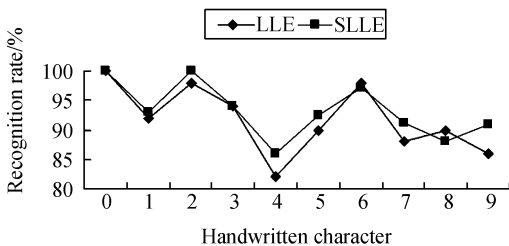


图 6 SLLE 与 LLE 识别率的比较

Fig. 6 Comparison of SLLE and LLE in recognition rates

别率的影响。

图 6 为 MINST 字符库下, SLLE 与 LLE 算法识别率的比较。证明了 SLLE 从高维到低维的映射在保留流形某些结构的同时,也进一步分离了不同类别的流形。

6 结 论

手写体数字字符在现实生活中的应用极为广泛,因此其识别问题一直是图像处理与模式识别领域的热点。SLLE 是一种基于监督的流形学习方法,它将非线性数据局部线性化,找出非线性数据之间的拓扑关系,并成功地将高维数据映射到低维空间中,且不改变原有的相对位置。与局部线性嵌入相比,基于监督的局部线性嵌入 (SLLE) 在寻求每个数据点的邻接点过程中利用了样本的类别信息。即一个给定的数据点的邻接点仅仅是从与其类别相同的点中选取。因此在保留流形某些结构的同时,也进一步分离了不同类别的流形。

实验中未对 MINST 字符做任何预处理,仅将字符的二值信息作为原始特征进行 SLLE 降维,为了证明 SLLE 降维效果,分类器只选用简单的最近邻分类器。实验证明了本文算法具有良好的降维效果,较高的识别率,识别率可达到 93.27%。在进一步的研究中将考虑结合其它的特征提取方式得到字符图像的原始特征,并结合 SVM 分类器进行分类以进一步提高字符的识别率。

本文基于 SLLE 的识别算法,不仅适用于手写体数字字符的识别,也可以应用于其它字符和工程图形符号的识别。

参考文献:

[1] PANG Y W, LIU ZH K, YU N H. A new nonlinear

feature extraction method for face recognition[J]. *Neurocomputing*, 2006, 69: 949-953.

[2] YEASIN M, BULLOT B. Comparison of linear and

- non-linear data projection techniques in recognizing universal facial expressions[J]. *Neural Networks*, 2005(5):3087-3092.
- [3] WU Y M, CHAN K L, WANG L. Face recognition based on discriminative manifold learnings[J]. *Pattern Recognition*, 2004(4):171-174.
- [4] 聂祥飞, 郭军. 利用 Gabor 小波变换解决人脸识别中的小样本问题[J]. *光学 精密工程*, 2007, 15(6): 973-977.
- NIE X F, GUO J. Solution of small sample size problem in face recognition using Gabor wavelet transform[J]. *Opt. Precision Eng.*, 2007, 15(6): 973-977. (in Chinese)
- [5] 李粉兰, 唐文彦, 段海峰, 等. 分数次幂多项式核函数在核直接判别式分析中的应用[J]. *光学 精密工程*, 2007, 15(9):1410-1414.
- LI F L, TANG W Y, DUAN H F, *et al.*. Application of fractional power polynomial kernel function to kernel direct discriminant analysis[J]. *Opt. Precision Eng.*, 2007, 15(9):1410-1414. (in Chinese)
- [6] SOUVENIR R, PLESS R. Manifold clustering [C]. *In: Proc of the 21st Int'1 Conf on Computer Vision (ICCV'05), Los Alamitos: IEEE Computer Society Press*, 2005:648-653.
- [7] ROWEIS S T. Nonlinear dimensionality reduction by locally linear embedding [J]. *Science*, 2000, 290(5500):2323-2326.
- [8] SAUL L K, ROWEIS S T. Think globally, fit locally:unsupervised learning of nonlinear manifolds[J]. *Journal of Machine Learning Research*, 2003, 4(6):119-155.
- [9] DICK D R, ROBERT P W. Locally linear embedding for classification [EB/OL]. <http://www.ph.tn.tudelft.nl/~dick/ph 2002 01.pdf>, 2002-1
- [10] LAWRENCE K S, SAM T R. An introduction to locally linear embedding [EB/OL]. <http://www.cs.toronto/~roweis/lle/>, 2001-06-10.
- [11] 徐志节, 杨杰, 王猛. 一种新的彩色图像降维方法 [J]. *上海交通大学学报*, 2004, 38(12):2063-2067.
- XU ZH J, YANG J, WANG M. A new nonlinear dimensionality reduction for color image[J]. *Journal of Shanghai Jiaotong University*, 2004, 38(12):2063-2067. (in Chinese)
- [12] 吴炜, 杨晓敏, 陈默, 等. 一种新颖的人脸图像超分辨率技术 [J]. *光学 精密工程*, 2008, 16(5):815-821.
- WU W, YANG X M, CHEN M, *et al.*. Novel method of face hallucination [J]. *Opt. Precision Eng.*, 2008, 16(5):815-821. (in Chinese)
- [13] DUDA R O, HART P E, STORK D G. *Pattern Classification (Second Edition)* [M]. New York: John Wiley & Sons, 2001.
- [14] 姜铮钊, 丁晓青. 基于 MQDF 的英文 OCR 多模板分类器 [J]. *计算机工程*. 2005, 31(15):56-58.
- JIANG ZH Y, DING X Q. English OCR multi-template classification based on modified quadratic discriminant function [J]. *Computer Engineering*, 2005, 31(15):56-58. (in Chinese)
- [15] 王和勇, 郑杰, 姚正安, 等. 基于聚类和改进距离的 LLE 方法在数据降维中的应用 [J]. *计算机研究与发展*. 2006, 43(8):1485-1490.
- WANG H Y, ZHENG J, YAO ZH A, *et al.*. Application of dimension reduction on using improved LLE based on clustering [J]. *Journal of Computer Research and Development*, 2006, 43(8):1485-1490. (in Chinese)

作者简介:



杨晓敏(1980—),女,四川广安人,讲师,博士,主要从事图像处理和图像通信方面的研究。E-mail: arielyang2006@163.com



吴炜(1975—),男,四川宜宾人,讲师,博士,主要从事图像处理和模式识别方面的研究。E-mail: wuwei75@yeah.net



何小海(1964—),男,四川盐亭人,教授、博士生导师,主要从事图像处理和图像通信方面的研究。E-mail: chenmo1984@gmail.com



陈默(1984—),男,四川成都人,博士研究生,主要从事人工智能方面的研究。E-mail: cias@263.net

●下期预告

基于父系数及邻域系数的 DTCWT 图像去噪

卢刚¹, 闫敬文², 寇业泉¹, 张建中¹

(1. 厦门大学 通信工程系, 福建 厦门 361005;

2. 汕头大学 电子工程系, 广东 汕头 515063)

目前小波变换(DWT)在图像去噪中的应用取得了较好的效果,但 DWT 不具有位移不变性和良好的方向性。二维双树复数小波变换(DTCWT)由于具有良好的平移不变性和方向选择性,比传统的二维离散小波变换具有更好的图像去噪能力。根据基于当前系数与父系数及邻域系数间的关系,本文构造了 DTCWT 图像去噪阈值计算公式,提出了一种去噪声新方法 PNDTCWT(Parental and Neighboring Coefficients of DTCWT)。该方法在对图像进行二维 DTCWT 变换后,利用阈值公式根据当前系数和父系数及相邻系数计算收缩阈值,对当前系数进行去噪处理。最后经过二维 DTCWT 反变换,得到去噪结果。实验结果表明,PNDTCWT 的噪声抑制效果明显优于各种基于 DWT 的去噪方法和其他 DTCWT 去噪方法。和基于父系数的 DTCWT 去噪方法相比,PNDTCW 的 PSNR 平均提高了 0.5 dB 左右。从视觉效果来看,PNDTCW 在去除噪声的同时,能较好地保留图像细节,物体轮廓显得比较平滑,不存在传统 DWT 算法中的混淆现象。