基于多通道 PCA 模型的手写汉字识别方法

高 学 金连文 尹俊勋
(华南理工大学电子与通信工程系,广州 510641)

摘 要 为了提高手写汉字的识别率和降低训练时间,提出了一种基于多通道 PCA(Principal component analysis)模型的手写汉字识别方法.该方法首先根据汉字的结构特点,将手写汉字分解为"一"、"丨"、"丿"、"乀"4 种方向子模式,然后分别对每个子模式进行主分量分析,最后通过建立起每类汉字的多通道 PCA 模型来进行手写汉字的识别.该方法既兼顾了主分量对手写汉字的描述能力,又有效地降低了建立模型的训练时间.针对 1 034 类别的手写汉字样本的实验结果表明,该汉字识别方法的识别率较欧氏距离分类器提高了 4.4 个百分点,而其训练时间则明显低于直接进行 PCA 重建的识别方法,由此可见,该方法是有效的.

关键词 模式识别(520·2040) 手写汉字识别 主分量分析(PCA) 多通道 PCA 模型
中图法分类号: TP391.43 **文献标识码**: A **文章编号**: 1006-8961(2003)07-0788-04

A New Approach for Handwritten Chinese Character Recognition Based on Multi-Channel PCA Model

GAO xue, JIN Lian-wen, YIN Jun-xun
(*Dept. of Electronics and Communication Engineering, South China University of Technology, Guangzhou* 51064)

Abstract In this paper, a new approach for handwritten Chinese character recognition based on multi-channel PCA (principal component analysis) model is proposed. In terms of the stroke directional characteristics of the handwritten characters, a handwritten Chinese character is decomposed into the four directional sub-patterns at first, namely, horizontal (—), vertical(丨), left up diagonal (丿) and right up diagonal(乀)sub-pattern, each of which could be modeled by its principal components. Then, based on their four sub-pattern PCA models, a multi-channel PCA model for each category of the handwritten Chinese character is constructed respectively, and the model's reconstruction error is used as a matching measure for the handwritten Chinese character recognition. The method can not only exploit principal components' ability for representing the handwritten Chinese character sample set, but also effectively reduce the training time for modeling. Experimental results on 1034 categories of handwritten Chinese characters indicate that, the proposed method can improve recognition rate by 4.4% comparing to the Euclidean distance classifier, while its training time is much lower than that for modeling handwritten Chinese character directly by its PCA model, showing the effectiveness of the proposed approach.

Keywords Handwritten Chinese character recognition, Principal component analysis, Multi-channel PCA model

0 引 言

汉字识别一直是模式识别最重要的研究领域之一.经过多年的研究,已经取得了大量成果[1~3],然而,由于汉字规模大,相似汉字较多,笔画及其书写变形复杂,因此脱机手写体汉字识别至今仍然被认为是文字识别领域最困难的问题之一,也是目前汉字识别研究的热点.

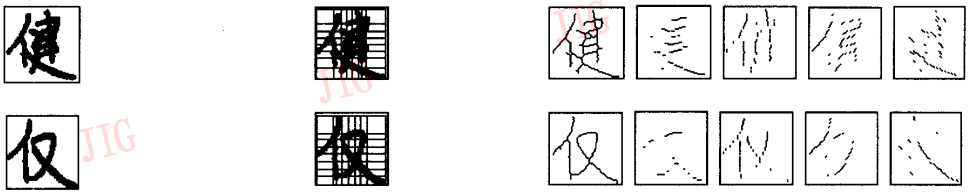
对于手写汉字识别这样一类多类别模式的识别问题,一种方法是根据所有模式类的样本数据及其类间的区分特性来训练分类器;另外一种方法则利用本类样本来建立每类汉字的单字统计或结构模型,即单字模型只与本类样本有关.对于待识别的手写汉字样本,在所定义的匹配度量意义下,通过选取与样本最佳匹配的单字模型类别来作为识别结果.

基金项目:国家自然科学基金(60275005);广东省自然科学基金(011611、020828);Motorola 研究基金
收稿日期:2002-06-21;**改回日期**:2003-02-28

本文采用后一种方法. 由于单字模型的建立只与本类样本有关,因此使得训练时间和内存开销大为降低,特别是对于大类别的手写汉字识别,其优越性更为明显. 同时,增加新的模式类也无需重新训练已完成的汉字模型. PCA 是数据降秩的有效方法,由于用该方法确定的主分量对应于数据的均方重建误差曲面的最小点[4],因而对数据具有较强的描述能力. 主分量分析(PCA)最初应用于图象的压缩和重建[5],最近也被应用于人脸识别[6,7]和人造物体的识别[8],并取得了较好的效果. 本文将结合主分量分析,研究手写汉字的建模和识别方法,然而,PCA 的缺点是其确定主分量的训练时间较长,特别是对于特征维数和类别数都比较大的手写汉字识别问题. 为了有效地降低训练时间和充分利用 PCA 描述数据的能力,本文提出了一种基于多通道 PCA 模型的手写汉字识别方法. 该方法首先根据汉字的基本结构特点(即每个汉字都是由"一"、"丨"、"丿"、"乀"4 种基本笔画划组成),将手写汉字图象分解为相应的 4 种方向子模式,并分别提取相应的特征作为 PCA 重建的对象,这样,每种子模式的特征向量维数就相应降低;然后,针对每一子模式,可分别通过主分量分析来建立手写汉字的多通道 PCA 模型;最后,根据待识别样本的重建误差来进行分类.

1　方向分解与特征提取

与其他模式识别问题一样,特征提取也是手写汉字识别的一项关键技术,它将直接影响整个识别系统的性能. 为了建立有效的 PCA 模型,本文将汉字图象进行方向分解,并分别提取弹性网格特征作为 PCA 重建的对象. 根据汉字的结构特点,可将手写汉字分解为"一"、"丨"、"丿"、"乀"4 种方向子模式,由于其可以有效地提取包含汉字笔画信息的特征,从而可改进分类器的识别性能,而弹性网格特征则能够适应笔画的不规则书写变形,是手写汉字识别中较为有效的特征之一[9]. 汉字图象经二值化和归一化处理后,即可首先划分成基于笔画像素分布均衡的弹性网格,并进行了细化;然后,根据细化图象的 8 邻域像素分布情况,采用文献[9]提出的"OR"分解策略,将每个笔画像素点分解到 4 种方向子模式中,实验证明,"OR"分解策略在性能上优于其他分解方法;最后,根据所划分的弹性网格计算笔画分解子模式的弹性网格特征. 对于 8×8 的网格,特征向量的总维数为 256,而每个笔画分解子模式特征向量的维数则为 64. 图 1 为汉字的弹性网格划分及笔画方向分解结果.



(a) 归一化后的汉字点阵　(b) 8×8 弹性网格　(c) 4 种笔画方向分解子模式

图 1　汉字的弹性网格划分及笔画方向分解

2　多通道 PCA 重建与识别

给定一个由 n 维向量组成的数据集合 $\boldsymbol{P}$,则由 PCA 可以确定 d 个正交的向量($d<n$),而由此向量张成的子空间就构成了原数据集合的一个近似描述. 设集合 $\boldsymbol{P}=\{\boldsymbol{x}_i\in\mathbf{R}^n \mid i=1,2,\cdots,s\}$,其均值向量与协方差矩阵分别为 $\bar{\boldsymbol{x}}$ 和 $\boldsymbol{C}$,即

$$\bar{\boldsymbol{x}}=\frac{1}{s}\sum_{i=1}^{s}\boldsymbol{x}_i \tag{1}$$

$$\boldsymbol{C}=\frac{1}{s}\sum_{i=1}^{s}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} \tag{2}$$

协方差矩阵 $\boldsymbol{C}$ 的 n 个特征值按大小递减排列为 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n$,其对应的特征向量为 $\mu_1,\mu_2,\cdots,\mu_n$,若前 d 个正交特征向量对应于集合 $\boldsymbol{P}$ 中数据投影方差最大的 d 个方向,则称其为数据集合的 d 个主分量. 如果数据在某些主分量方向的投影方差远大于其他方向的方差,则利用这些向量描述和重建原始数据,可以获得较小的期望重建误差. 对于手写汉字样本,尽管不同的书写风格具有不同的笔画变形,由于汉字本身的间架结构的约束,笔画变形在统计上表现出一定的规律性. 例如对不同书写者,某些局部笔画自由度比较大,而其他一些局部笔画则相对稳定. 对手写汉字特征向量的主分量分析也证实了

这一点. 图 2 所示为“啊”字 300 套手写样本特征向量集合的前 100 个主分量方向的方差曲线. 由图中可以看出,其最大与最小方差相差比较大. (图中最大方差为 12.224,最小为 0.294,相差 40.6 倍.)

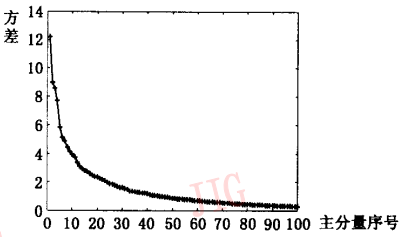


图 2 手写汉字样本“啊”的方差曲线

因此,对于每类手写汉字,可以应用其训练样本集的前 d 个主分量来建立其 PCA 模型

$$\tilde{\boldsymbol{x}} = \sum_{j=1}^{d} \mu_j^{\mathrm{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}})\mu_j + \bar{\boldsymbol{x}} \tag{3}$$

$$E(\boldsymbol{x}) = \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2 = \left\| \sum_{j=1}^{d} \mu_j^{\mathrm{T}}(\boldsymbol{x} - \bar{\boldsymbol{x}})\mu_j - (\boldsymbol{x} - \bar{\boldsymbol{x}}) \right\|^2 \tag{4}$$

其中,$\tilde{\boldsymbol{x}}$ 为原数据 $\boldsymbol{x}$ 的重建,$E(\boldsymbol{x})$ 为 $\boldsymbol{x}$ 的重建误差. 当 d 为零时,重建误差 $E(\boldsymbol{x})$ 就对应输入样本 $\boldsymbol{x}$ 与其类均值的欧氏距离.

数据的相对均方重建误差为

$$\xi = \sum_{j=d+1}^{n} \lambda_j \Big/ \sum_{j=1}^{n} \lambda_j \tag{5}$$

然而,当原始数据的维数 n 较大时,则确定协方差矩阵的特征值与特征向量的训练时间较长,特别是对于手写汉字的识别问题,其特征向量的维数和需要建模的类别数都比较大. 为了有效地降低特征数据的维数,以减少训练时间,可根据汉字的结构特点,将每个汉字图象分解为“一”、“丨”、“丿”、“乀”4 种方向子模式. 这样,由于每个子模式的特征向量的维数可相应降低,因此可以有效地减少训练时间;然后,根据式(3)分别对每个子模式进行建模,并由此来进行手写汉字的多通道重建,此即称为多通道 PCA 模型.

设 $\mu^{(l,k)}$ 和 $\bar{\boldsymbol{x}}^{(l,k)}$ ($l=1,2,\cdots,N;k=1,\cdots,4$)分别为第 l 类汉字的第 k 个子模式的主分量和均值向量,$\boldsymbol{x}^{(l,k)}$ 为样本 $\boldsymbol{x}$ 第 k 个子模式的特征向量,其中 N 为待建模汉字的类数. 由式(3)可得第 l 类汉字的多通道 PCA 模型

$$\tilde{\boldsymbol{x}}^{(l,k)} = \sum_{i=1}^{d} (\mu_i^{(l,k)})^{\mathrm{T}}(\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}^{(l,k)})\mu_i^{(l,k)} + \bar{\boldsymbol{x}}^{(l,k)}, k = 1,\cdots,4 \tag{6}$$

其中,$\tilde{\boldsymbol{x}}^{(l,k)}$ 为 $\boldsymbol{x}^{(k)}$ 由第 l 类汉字模型的重建向量.

给定手写汉字的输入样本 $\boldsymbol{x}=(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\boldsymbol{x}^{(3)},\boldsymbol{x}^{(4)})$,则由第 l 类汉字的多通道 PCA 模型重建的误差 $E^{(l)}(\boldsymbol{x})$ 为

$$E^{(l)}(\boldsymbol{x}) = \sum_{k=1}^{4} \|\tilde{\boldsymbol{x}}^{(l,k)} - \boldsymbol{x}^{(k)}\|^2$$

$$= \sum_{k=1}^{4} \left\| \sum_{j=1}^{d} (\mu_j^{(l,k)})^{\mathrm{T}}(\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}^{(l,k)})\mu_j^{(l,k)} - (\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}^{(l,k)}) \right\|^2 \tag{7}$$

对于手写汉字的分类问题,本文采用重建误差作为待识别样本与汉字模型匹配程度的度量. 给定待识别样本 $\boldsymbol{x}$,如果 r 满足

$$E^{(r)}(\boldsymbol{x}) = \min_{l=1,\cdots,N} \{E^{(l)}(\boldsymbol{x})\} \tag{8}$$

则输入样本应属于第 r 类模式.

3 实验结果

为了检验基于多通道 PCA 模型的手写汉字识别方法的有效性,随机地从 863 手写体汉字样本库 HCL2000 中取出 350 套样本,每套样本取国标 16～26 区的 1 034 个汉字,作为训练和测试样本. 其中 300 套样本用于建立手写汉字的多通道 PCA 模型,其余 50 套用于测试系统的识别性能. 实验所采用的弹性网格大小为 8×8,样本特征的总维数为 256,每个子模式的特征向量维数为 64.

在第 1 个实验中,分别测试了基于 PCA 模型及多通道 PCA 模型的手写汉字识别性能. 为了提高分类器的识别速度,在进行样本的重建和分类之前,首先应用欧氏距离分类器确定待识别样本 m 的个候选字类别集合. 表 1 给出了欧氏距离分类器的前 m 个候选字的识别率.

表 1 欧氏距离分类器的前 m 个候选字的识别率

候选字数	1	5	10	20	50	90
识别率(%)	91.95	97.18	98.24	98.87	99.45	99.63

注:候选字数等于 1 对应表 2 中主分量个数为 0 的情况.

由表 1 可以看出,当候选字个数 m 取 90 时,识别率为 99.63%,这已经满足候选分类器的要求.

然后,待识别样本与手写汉字的 PCA 及多通道 PCA 模型的匹配(本文分别称为方法 1 和方法 2)只在候选类别集合中进行. 实验系统如图 3 所示.

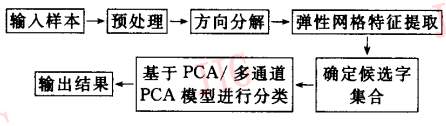


图 3 实验系统流程

表 2 给出了两种方法的识别结果.

表 2　方法 1 与方法 2 的识别率(1 034 类汉字)

主分量个数 d	0	1	3	5	10	15
方法 1(%)	91.95	94.36	96.04	96.51	96.79	96.79
方法 2(%)	91.95	94.98	96.09	96.35	95.97	95.34

注：主分量个数为 0 对应欧氏距离分类器

由表 2 可以看出，基于 PCA 模型及基于多通道 PCA 模型的手写汉字识别方法，其识别率均高于欧氏距离分类器，最好识别率分别提高 4.4 和 4.84 个百分点，证明了 PCA 模型对手写汉字样本具有较强的描述能力. 而方法 1 与方法 2 的识别性能虽没有明显的区别(最好识别率仅相差 0.44 个百分点)，但方法 2 可大大缩短分类器的训练时间，这表明基于多通道 PCA 模型的手写汉字识别方法识别速度较快.

在第 2 个实验中，PCA 模型与多通道 PCA 模型的训练时间比较见图 4(主机为 P Ⅲ/733).

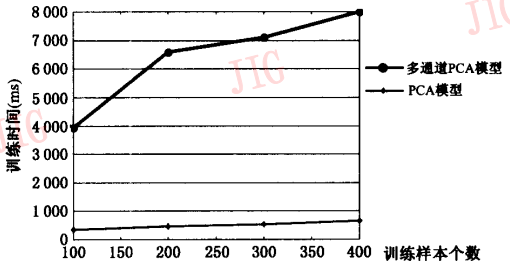


图 4　各类汉字的 PCA 与多通道 PCA 模型的平均训练时间曲线

由图 4 可以看出，多通道 PCA 模型在训练时间上具有明显的优势.

4　结　论

手写汉字识别一直是文字识别领域最困难的问题之一. 针对汉字的结构特点，本文提出了一种基于多通道 PCA 模型的手写汉字识别方法. 该方法首先将手写汉字进行方向分解，然后通过主分量分析来分别建立每类汉字的多通道 PCA 模型，并将其用于手写汉字的识别. 这样既利用了主分量对手写汉字数据的描述能力，又能有效地降低其训练时间. 从实验结果可知，基于多通道 PCA 模型的手写汉字识别方法，其识别率要高于欧氏距离分类器，而训练时间则较直接进行 PCA 重建的识别方法有明显降低，由此可见，该方法是行之有效的.

参 考 文 献

1 Hildebrand T H, Liu W. Optical recognition of handwritten Chinese characters: advances since 1980 [J]. Pattern Recognition, 1993, 26(2): 205～225.

2 Tsukumo J. Handprinted Kanji OCR development-what was solved in handprinted Kanji character recognition? [J]. Institute of Electronics Information and Communication Engineers Transactions on Information and Systems. 1996, E79-D: 411～416.

3 Kimura Y, Wakahara T. Toward robust handwritten Kanji character recognition [J]. Pattern Recognition Letters, 1999, 20(10): 979～990.

4 Xu Lei. Theories of unsupervised learning, PCA and its nonlinear extensions[A]. In: Proceedings of IEEE International Conference on Neural Networks'94 [C], Orlando, Florida, USA, 1994: 1254～1257.

5 Kirby M, Sirovich L. Application of the Karhunen-Loève procedure for the characterization of human faces [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(1): 103～108.

6 Pentland A, Moghaddam B, Starner T. View-based and modular eigenspaces for face recognition[A]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition [C], Seattle, WA, USA, 1994: 84～91.

7 Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711～720.

8 Murase H, Nayar S. Visual learning and recognition of 3D objects from appearance[J]. Intelernational Journal of Computer Vision, 1995, 14(1): 5～24.

9 Jin Lian-wen. Handwritten Chinese character recognition with directional decomposition cellular features [J]. Journal of Circuits, System, and Computers, 1998, 8(4): 517～524.

高　学　1967 年生，2000 年获北京工业大学自动控制专业硕士学位，现为华南理工大学电子与通信工程系 2000 级博士生. 主要研究领域为汉字识别、图象处理、遗传算法. 发表论文 8 篇.

金连文　1968 年生，1991 年毕业于中国科技大学电子工程系，1996 年获华南理工大学电子与通信工程系博士学位，现为华南理工大学电子信息学院副教授，IEEE 会员. 目前主要研究方向为文字识别、图象处理等. 发表论文 20 多篇.

尹俊勋　1942 年生，华南理工大学电子信息学院教授、博士生导师. 主要研究方向为信号处理、模式识别、通信编码等. 在信号处理及通信领域已发表学术论文 40 余篇.