第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)18—0219—03 文献标识码：A 中图分类号：TP391.41

基于主分量分析法的脱机手写数字识别

张国华，万钧力
（三峡大学电气信息学院，宜昌 443002）

摘 要：针对手写数字识别研究中统计特征和结构特征融合困难的问题，利用主分量分析法提取数字字符结构特征的统计信息，重建数字模型，并估计重构偏差，同时提取数字的高宽比特征和欧拉特征，通过组合与3种特征相对应的贝叶斯分类器的分类结果实现数字识别。使用该方法对样本库中的样本进行测试，正确识别率为90.73%。

关键词：数字识别；主分量；特征提取；组合分类器

Offline Handwritten Numeral Recognition Based on Principal Component Analysis

ZHANG Guo-hua, WAN Jun-li
(College of Electrical Engineering & Information Technology, China Three Gorges University, Yichang 443002)

【Abstract】 To overcome the difficulty of fusing statistical feature and structural feature in the research on handwritten numeral recognition, principal component analysis is used to reconstruct numeral model and estimate the numeral reconstructive error based on the statistical information of digit structural feature. At the same time, the height-width ratio and Euler value of numeral is extracted. Recognition of digital character is completed through the combination of Bayes classifier corresponding to the three type features. The recognition rate of the method is 90.73% in handwritten numeral database.

【Key words】 numeral recognition; principal component; feature extraction; combining classifier

手写数字识别一直是图像处理和模式识别领域的研究热点，同一个数字在大小、形状、倾斜度和书写风格等方面的无穷变化更增加了这一问题的研究难度。脱机手写数字字符的识别问题至今还没有圆满的解决方案[1]。

根据使用特征的不同，手写数字的识别方法可分为2类：基于字符结构特征的识别方法和基于字符统计特征的识别方法。统计特征和结构特征具有一定的互补性，有效的特征提取方法应能较好地结合字符的结构特征和统计特征，这种方法就是主分量分析法（principal component analysis, PCA）。其充分利用数据中的二阶统计信息进行特征抽取和降维，算法简单、运算量小，是近年来研究较多的一种特征提取方法[2]。本文以PCA为基础，结合数字字符的高宽比特征和欧拉特征研究出一种提高脱机手写数字识别率的新方法。

1 手写数字图像预处理

手写数字是一种特殊的图形图像，对识别有价值的是其中的图形，预处理工作实际上就是从图像中提取图形的过程，即图像向图形的转换。预处理一般包括二值化、字符分割、倾斜校正、归一化等[3,4]。手写数字图形中线的粗细并不重要，重要的是图形由什么样的线组成以及这些线以什么结构组成图形。但线的粗细不一仍会影响识别系统的工作性能，因此，在手写数字预处理过程中，经常要进行笔画粗细的归一化处理。笔画粗细归一化以数学形态学为基础，通过对数字字符进行骨骼化和膨胀运算来实现。

2 手写数字特征提取

2.1 数字的主分量特征及其特征维数选择

PCA的基本思想是：寻找一个最佳子空间，当高维数据x在该子空间进行投影后，所得分量具有最大的方差，同时，在子空间用新分量对原始数据进行重建时，在均方误差最小的意义下逼近效果最优[2]。

本文中经过预处理的手写数字为16×16的点阵图像，将其按行展开为列向量$\boldsymbol{x}^i=[x_1,x_2,...,x_{256}]^{\mathrm{T}}$，其中，$i$=0,1,…,9。用$\omega_i$表示该样本所属的类别，各个数字类的均值$m^i$和协方差矩阵$\boldsymbol{C}^i$分别如下式：

$$m^i=E\{\boldsymbol{x}^i\}\ ;\ \boldsymbol{C}^i=E\left\{(\boldsymbol{x}^i-m^i)(\boldsymbol{x}^i-m^i)^{\mathrm{T}}\right\}$$

协方差矩阵的特征根为$\lambda_1^i\geq\ldots\geq\lambda_{256}^i$，与其相对应的正交归一化特征向量矩阵为$\boldsymbol{U}^i=[u_{i_1},u_{i_2},...,u_{i_{256}}]$。特征向量即为该数字类的主分量，描述了数字的结构特征。向量化的数字$\boldsymbol{x}^i$可以在特征空间$\boldsymbol{U}^i$中进行完全重构，其特征表示为$y(\boldsymbol{x}^i)=[\xi_1,...,\xi_{256}]^{\mathrm{T}}$，即$y(\boldsymbol{x}^i)=\boldsymbol{U}^{i\mathrm{T}}(\boldsymbol{x}^i-m^i)$。

从韩国延世大学手写数字样本库[5]中选取400个数字“0”的样本，采用主分量分析法获得的主分量所恢复成的特征子图如图1所示，其中，第1、第2行为主分量u_1~u_8所对应的特征子图，第3行为主分量u_{253}~u_{256}所对应的特征子图。从图1中可以看出：前8个特征子图对数字“0”的结构特征描述能力较强，而最后的4个特征子图对特征描述几乎毫无作用。由于特征值反映了相应的特征子图在组成数字类图像时的贡献大小情况，因此在不影响特征向量对字符类别信息描述能

基金项目：湖北省教育厅科学技术基金资助项目(2003A002)
作者简介：张国华(1982－)，男，硕士，主研方向：模式识别；万钧力，教授
收稿日期：2006-11-23 **E-mail：**wjlem@163.com

力的前提下可对主分量进行适当截取。

图 1　数字“ 0 ”的特征子图

因为协方差矩阵$\boldsymbol{C}^i$特征根的大小反映了相应的主分量在该类数字的重构过程中贡献的大小，所以可根据特征根来进行数字类特征维数n_i的选择，即

$$\alpha(n_i)=\frac{\sum_{j=n_i+1}^{256}\lambda_j^i}{\sum_{j=1}^{256}\lambda_j^i}\le a$$

其中，$\alpha(n_i)$的大小反映了由于截断而使类别信息丢失的程度；a为一个固定值，且$a\in(0,1)$。

2.2　数字字符的高宽比特征

主分量特征没有反映每个数字类的高宽比特征，因此常导致“ 1 ”和其他数字字符混淆。由于未经倾斜校正的手写数字字符的高宽比 r 受字符倾斜程度的影响很大，不能正确反映数字样本的类别信息，因此需要对每个数字类经过倾斜校正后的字符高宽比 r 进行统计。设经过倾斜校正的数字字符的高和宽分别为 h 和 w，则字符的高宽比特征 r 为 $r=h/w$。

在高宽比特征类条件概率密度函数具体形式未知的情况下，非参数的窗函数估计法能够直接使用训练集中的样本对各个数字类条件概率密度进行估计[6]。正态Parzen窗函数为

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}u^2\right\}$$

则数字类 ω_i 高宽比特征 r 的类条件概率密度函数为

$$p(r|\omega_i)=\frac{1}{400}\sum_{j=1}^{400}\varphi(r-r_j^i)=\frac{1}{400}\sum_{j=1}^{400}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(r-r_j^i)^2\right\}$$

估计结果表明[5]：除“ 6 ”、“ 9 ”的高宽比特征相似难以区分外，其他数字都具有明显的高宽比特征，尤其是数字“ 1 ”，与其他数字的高宽比特征明显不同。

2.3　数字字符的欧拉特征

图像的拓扑性质可用于描述图像平面区域的形状，只要图形不出现撕裂或粘连，其拓扑性质不受形变的影响。图像的欧拉数是图像的一种拓扑度量。欧拉数 eu 等于图像中所有对象的总数 c 减去这些对象中孔洞的数目 h。即 $eu=c-h$。

对于手写数字字符而言，字体本身是一个完整的对象，其欧拉数 eu 取决于图形中所具有的孔洞数 h。在一般情况下，手写数字中的字符“ 1 ”～“ 5 ”和“ 7 ”的欧拉数为 1，字符“ 0 ”、“ 6 ”、“ 9 ”的欧拉数为 0，字符“ 8 ”的欧拉数为-1。但由于手写数字图像预处理中的二值化以及个人书写习惯的差异，数字的字体经常出现错误的断笔或多余的孔洞，使得数字字符的欧拉数与实际情况不符。数字字符的欧拉数 eu 只能为离散的整数。因此，只能根据训练样本集估计数字类 ω_i 中各个欧拉数 eu 的条件概率 $P(eu|\omega_i)$，i=0,…,9，即

$$P(eu|\omega_i)=\frac{n_{eu}^i}{N^i},\quad eu=-5,\ldots,5$$

其中，N^i为训练样本集中 ω_i 类数字样本的总数；n_{eu}^i 为N^i个 ω_i 类的样本中欧拉数为eu的样本总数，且 $\sum_{eu=-5}^{eu=5}P(eu|\omega_i)=1$。统计分析表明，欧拉特征可以有效区分数字“ 3 ”、“ 8 ”、“ 9 ”。

3　分类器设计

3.1　基于主分量的最小重构偏差分类器

采用主分量分析法求得的各个数字类 ω_i 的协方差矩阵的特征向量矩阵$\boldsymbol{U}^i$构成了一个标准正交基，任意一个向量化的数字$\boldsymbol{x}$都可以在该特征空间内进行完全的重构。但是，当选取的基向量个数m小于 256 时，则每个数字类 ω_i 有固定个数的基向量对该类数字样本的重构能力最强。因此，对于一个向量化的数字$\boldsymbol{x}$，当其选择固定的特征维数$m(m<256)$分别在各个数字类的特征空间内进行重构时，其在所属类别 ω_i 的特征空间内的重构结果 $\hat{\boldsymbol{x}}$ 与自身$\boldsymbol{x}$的偏差最小。

在特征空间$\boldsymbol{U}^i$内进行固定维数重构所引起的偏差　$e^i(\boldsymbol{x})$为

$$e^i(\boldsymbol{x})=\left\|\boldsymbol{x}-\left(\sum_{j=1}^{m}u_j^{iT}(\boldsymbol{x}-m^i)u_j^i+m^i\right)\right\|_2$$

数字的分类规则为

$$e^i(\boldsymbol{x})=\min_{j=0,\ldots,9}\{e^j(\boldsymbol{x})\},\ \boldsymbol{x}\in\omega_i$$

根据上式的分类规则分别采用不同的主分量个数对样本库中的 3 000 个样本进行分类识别，其结果如表 1 所示。从表 1 中可以看出，当只选取前 3 个主分量进行重构时，正确识别率就已经达到 84.57%，这说明由每个数字类的前 3 个主分量构成的特征空间已能描述数字类的大部分特征；当主分量维数从 6 增加到 16 时，正确识别率只有小幅度的提高；继续提高主分量个数，则正确识别率开始下降，这说明数字类的主要信息集中在前十几个主分量中；数字“ 0 ”的识别效果随着主分量个数的增加反而下降，这是由于该数字的字形结构简单，类别信息主要集中在前几个主分量中。

表 1　基于主分量重构偏差的识别结果

选取PC维数	正确识别样本总数										正确识别字符总数	正确识别率/%
	数字0	数字1	数字2	数字3	数字4	数字5	数字6	数字7	数字8	数字9		
3	299	233	238	251	258	253	270	228	249	258	2 537	84.57
6	298	232	249	263	258	262	281	236	269	262	2 610	87.00
10	295	235	252	269	263	261	286	242	275	251	2 629	87.63
14	295	237	264	270	264	266	281	244	268	244	2 633	87.77
16	293	241	267	270	266	273	278	238	269	242	2 637	87.90
20	291	236	268	273	255	273	265	235	255	242	2 593	86.43

3.2　组合分类器

对于手写数字的主分量特征以神经网络为分类器模型，使用样本库中的数字分别进行训练和测试，可达到的最高识别率为 85.6%。即使采用不同的方法对独立训练神经网络分类器的结果进行组合，识别率也只有小幅的提高。由于不同类型的特征反映了数字不同方面的特性，在一种特征空间很难区分的 2 种数字可能在另一个特征空间很容易区分[5]。高宽比特征和欧拉特征的类条件概率密度函数或类条件概率容易估计，因此可以根据贝叶斯公式求得特征值的后验概率。

测试样本高宽比特征 r 对于数字类 ω_i 的后验概率为

$$P(\omega_i|r)=\frac{P(r|\omega_i)P(\omega_i)}{\sum_{j=0}^{9}P(r|\omega_j)P(\omega_j)}$$

$$p_r=\left[P(\omega_0|r),\ldots,P(\omega_9|r)\right]^{\mathrm{T}}$$

测试样本欧拉特征 eu 对于数字类 ω_i 的后验概率为

$$P(\omega_i|eu)=\frac{P(eu|\omega_i)P(\omega_i)}{\sum_{j=0}^{9}P(eu|\omega_j)P(\omega_j)}$$

$$p_{eu}=\left[P(\omega_0|eu),\ldots,P(\omega_9|eu)\right]^{\mathrm{T}}$$

组合分类器的输出为

$$C = Combine(p, p_r, p_{eu}) = \begin{bmatrix} P(\omega_0|\boldsymbol{x})P(\omega_0|r)P(\omega_0|eu) \\ \vdots \\ P(\omega_9|\boldsymbol{x})P(\omega_9|r)P(\omega_9|eu) \end{bmatrix}$$

组合分类器的分类规则为

$c_j = \max_{i=0,\ldots,9}\{c_i\}$，$\boldsymbol{x} \in \omega_j$

从样本库中选出 3 000 个样本用组合分类器进行分类，其结果如表 2 所示。

表 2　组合分类器的识别结果

样本类别	识别结果									
	0	1	2	3	4	5	6	7	8	9
0	294	0	0	2	1	0	2	1	0	0
1	0	273	15	0	0	7	3	2	0	0
2	1	0	281	7	5	5	0	1	0	0
3	1	0	2	280	1	8	0	0	8	0
4	3	3	11	5	282	4	3	3	3	7
5	4	3	7	16	10	262	6	4	4	8
6	3	1	3	6	3	18	268	1	3	2
7	1	0	51	7	1	1	0	235	0	4
8	5	0	0	7	4	6	1	0	274	3
9	1	1	8	6	2	3	0	3	3	273
测试样本总数：3 000 正确识别样本总数：2 722　　正确识别率：90.73% 错误识别样本总数：278　　错误识别率：9.27%										

从表 2 中可以看出，分类的识别率已经达到了 90%以上。但是分类器对数字“ 7 ”的识别效果依然不佳。为了满足实际应用中系统对高识别精度的要求，当样本的类别属性不明确时，可采取拒绝作出分类判决的办法，即

$$P(\omega_i|\boldsymbol{x}) = \max_{j=0,\ldots,9} P(\omega_j|\boldsymbol{x}), P(\omega_i|\boldsymbol{x}) \geq t,\ \boldsymbol{x} \in \omega_i$$

其中，参数 $t(t \in [0.1,1])$为拒绝门限值。当由分类器所得的最大后验概率估计值 $P(\omega_i|\boldsymbol{x})$小于拒绝门限值 t 时，系统采取拒绝分类的办法。经过测试，系统可达到的最高识别精度为 97.64%。

4　结论

手写数字的主分量特征描述了数字字符结构特征的统计信息，同时为了弥补其对有些数字之间的差异描述能力不强的缺点，提取了数字字符的高宽比特征和欧拉特征。将 3 种特征均用于手写数字的分类识别过程中，取得了不错的实验效果。将不同类型的特征融合应用于分类识别的过程是一种很有前景的方法。

参考文献

1 Plamondon R. On-line and Off-line Handwriting Recognition: A Comprehensive Survey[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(1): 63-84.

2 芮　挺, 沈春林, 丁　健, 等. 基于主分量分析的手写数字字符识别[J]. 小型微型计算机系统, 2005, 26(2): 289-292.

3 Liu Chenglin, Nakashima K, Sako H, et al. Handwritten Digit Recognition: Investigation of Normalization and Feature Extraction Techniques[J]. Pattern Recognition, 2004, 37(2): 265-279.

4 王有伟, 刘　捷. 手写数字识别中一种新的倾斜校正的方法[J]. 计算机工程, 2004, 30 (11): 128-137.

5 张国华. 基于主分量分析和多分类器组合的手写数字识别技术研究[D]. 宜昌: 三峡大学, 2006.

6 Girolami M, He Chao. Probability Density Estimation from Optimally Condensed Data Samples[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003, 25(10): 1253-1264.

(上接第 218 页)

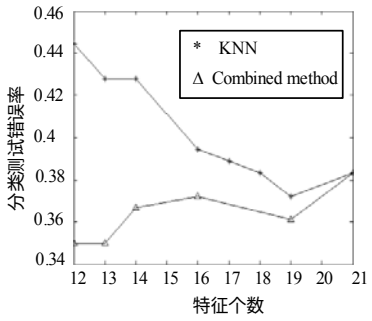

图 3　KNN 及其组合算法在 waveform 数据集上的测试结果

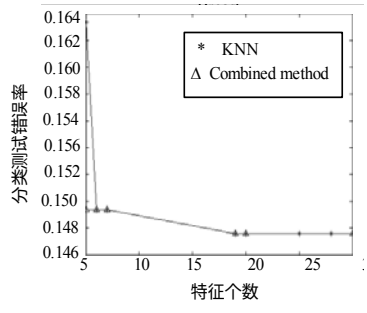

图 4　KNN 及其组合算法在 breastcancer 数据集上的测试结果

5　结束语

由仿真结果可知，基于 KNN 的组合式算法与 KNN 算法都能有效进行特征选择，减小特征数目，确保分类错误率没有明显的提高。 甚至在一定的情况下，还能降低特征选择后样本的分类错误率。基于 KNN 的组合式算法与 KNN 算法相比，其样本分类错误率较小，效果更好。

参考文献

1 Siedlecki W, Sklansky J. On Automatic Feature Selection[J]. Int'l Journal of Pattern Recognition and Artificial Intelligence, 1988, 2(2): 197-220.

2 Glover F. Future Paths for Integer Programming and Links to Artificial Intellifence[J]. Comput.& Ops., 1986, 13(5): 533-549.

3 Pudil P, Novovicova, Kittler J. Floating Search Methods in Feature Selection[J]. Pattern Recognition Letters, 1994, 15(11): 1119-1125.

4 Narendra P M, Fukunaga K. A Branch and Bound Algorithm for Feature Selection[J]. IEEE Transaction on Computers, 1977, 26(9): 917-922.

5 Mitra P, Murthy C A, Pal S K. Unsupervised Feature Selection Using Feature Similarity[J]. IEEE Trans. on Pattern Recognition and Machine Intelligence, 2002, 24(3).

6 孙即祥. 现代模式识别[M]. 长沙：国防科技大学出版社, 2002.