基于轮廓追踪的字符识别特征提取

杨 明,刘 强,尹忠科,王建英
YANG Ming,LIU Qiang,YIN Zhong-ke,WANG Jian-ying

西南交通大学 信息科学与技术学院,成都 610031
School of Info Sci.& Tech.,Southwest Jiaotong University,Chengdu 610031,China
E-mail:youngming@mars.swjtu.edu.cn

YANG Ming,LIU Qiang,YIN Zhong-ke,et al.Feature extraction in character recognition based on contour pursuit.Computer Engineering and Applications,2007,43(20):207-209.

Abstract: Character recognition is an important branch of pattern recognition,its key factors are selecting and extracting proper feature vector.Wavelet decomposition and fractal are applied extensively in image processing,and a new feature vector combined their characteristics is proposed in this paper based on contour pursuit.After an input image is preprocessed,contour is extracted.Then edge pixels coordinate sequence is obtained based on it.This method transforms 2-D image data into 1-D data which is decomposed by wavelet to get curves.Afterwards the feature vector is formed by calculating fractal dimension of several segments of curves.Some characters are tested using the method,and the result is satisfied.

Key words: character recognition;feature extraction;wavelet decomposition;fractal dimension

摘 要:字符识别是模式识别的一个重要分支,其关键是特征向量的选择与提取。小波分解和分形在图像处理方面有着广泛的应用,在结合二者特点的基础上提出了一种新的基于轮廓追踪的字符识别特征选取方法。即对于一个输入的字符图像经预处理提取其轮廓,并由轮廓追踪获得边缘点坐标序列,实现了从二维图像数据到一维数据的转化,对得到的一维曲线进行小波分解,计算少数几个分解得到的曲线的分形维数,以它们构成特征向量。并对有关字符做了实验,其效果是令人满意的。

关键词:字符识别;特征提取;小波分解;分形维
文章编号:1002-8331(2007)20-0207-03 **文献标识码**:A **中图分类号**:TP391.43

1 引言

字符识别是模式识别领域的一个古老课题,其历史可追溯到1870年,字符识别一般可以分为两类:联机(On line)字符识别和离线(Off line)字符识别,其中,离线字符识别又称光学字符识别(OCR)。在联机字符识别中,计算机能够通过与计算机相连的手写输入设备获得输入字符笔画的顺序,笔画的方向以及字符的形状等信息;而 OCR 则是要计算机识别那些已经成为字符的东西,例如表格、支票的处理及车牌识别等,由于缺少书写时的动态信息,其识别难度远远大于联机字符识别。

字符识别系统主要由四个部分组成:检测、预处理、特征提取和分类。其中,特征提取是整个字符识别中最为关键的一步,特征选取的好坏将直接影响识别率的高低,所以选取字符特征应满足以下条件:

(1)选取特征必须足以区分各字符,特征应该稳定,受字型影响越小越好。

(2)所选取特征应便于提取,便于在计算机上实现,特征的维数应尽可能少。

目前,字符特征提取方法[1]主要有两种:统计特征方法和结构特征方法。运用于字符识别方面的统计类特征主要有网格特征、穿透特征、投影特征、边缘特征、方向线素特征[2],这类特征本身包含很多信息,识别率均较高,但由于他们都是统计特征,对字形结构描述不足,使得一些在统计特征上差别很小但结构完全不同的字符容易发生混淆[3]。另一类是各类变换系数特征,如 K_L 展开、Fourier 变换、Walsh 变换、Hough 变换、场变换、Zernike 矩及小波变换等[4-7]。以上的变换把二维图像映射到另一个域,把变换的系数作为特征。本文仅讨论用统计方法进行字符识别的有关问题,在文献[8]与[9]提出的基本原则的基础上,利用小波分解及分形维的概念,将字符识别的上述两类特征有机结合,建立了一种新的基于轮廓追踪的字符识别特征向量提取方法,并对有关字符进行了试验。结果表明效果令人满意,即便很相似的字符,也容易将它们区别开来。

2 预处理和轮廓追踪坐标提取

字符识别是模式识别的一个重要分支,具有重大的应用价

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60602043);四川省重点科技计划项目(No.04GG021-020-5,No.03GG006-005-2);四川省应用基础研究项目(No.03JY029-048-2,No.04JY029-059-2,No.2006J13-114)。

作者简介:杨明,男,硕士研究生,研究方向:信号与信息处理,图像处理与传输等;刘强(1982-),男,硕士研究生,主要研究方向为:信号与信息处理、图像处理与传输;尹忠科(1969-),男,工学博士,教授,主要研究方向为:信号与信息处理、图像处理与传输等;王建英(1972-),女,工学博士,副教授,主要研究方向为:信号与信息处理、图像处理与传输等。

值和理论研究价值。利用计算机进行字符识别,需要对被识别字符进行预处理,例如二值化、位置规格化、大小规格化、轮廓提取等。本文在预处理时,采用 64×64 点阵,而不是传统的 48×48 点阵来进行归一化。实验结果证明,该方法在后续处理过程中并没有降低特征向量的提取效果。图 1 显示了统计方法字符识别系统的四部分,其中特征提取主要是提取待识别字符中较为稳定且能反映字符形状的基本特征作为识别的重要依据,是字符识别的关键部分。

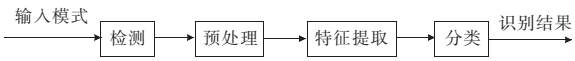

图 1 统计方法字符识别系统简图

轮廓提取的目的是为了获得字符图像的外部轮廓特征以为后面的特征分析和提取服务。图像的轮廓提取有很多方法,如用 Sobel 算子、Robert 算子等方法。本文采用了一种非常简单的方法,即用数学形态学方法来进行字符图像轮廓的提取。因为二值图像的轮廓提取就是掏空其内部点;如果原图像中有一点为黑,且它的 8 个相邻点都是黑色时,则将该点删除。具体方法是用一个九个点的结构元素对图像进行腐蚀,再用原图像减去腐蚀图像。

图 2 预处理后字符图像及其轮廓效果

字符图像在经过预处理之后得到了一系列长度大于一定阈值的封闭轮廓线,对轮廓线上的每一点在轮廓追踪时分别记录下它们的横坐标,这样每个轮廓线就可以得到一个一维数据序列。本文的具体做法是:对字符图像轮廓线进行从上到下由左及右的列扫描,记录下每一黑点的横坐标,将这一坐标序列作为小波变换的原始数据。

3 特征提取的理论基础

3.1 小波分解

设 $L^2(R)$为满足 $\int_{-\infty}^{+\infty}|f(x)|^2\mathrm{d}x<+\infty$,$x\in R$ 的所有 $f(x)$的集合,在空间 $L^2(R)$中任一函数 $f(x)$可以表示为[10]

$$f(x)=\sum_{k,j\in Z}C_{k,j}\psi_{k,j}(x) \tag{1}$$

式中 $\psi_{k,j}(x)=|k|^{-1/2}\psi\left[\frac{x-j}{k}\right]$,$\psi(x)$为小波函数,$C_{k,j}$ 为小波系数。

设 W_j 是集合$\{\psi_{j,k}(x):k\in Z\}$的线性张成在 $L^2(R)$中的闭包,因此 $L^2(R)$能分解成 $W_j:j\in Z$ 的直和,相应子空间为

$$V_j=\cdots+W_{j-2}+W_{j-1},j\in Z \tag{2}$$

则有

$$L^2(R)=\sum_{k\in Z}W_k=\cdots+W_{-1}+W_0+W_1+\cdots \tag{3}$$

$$f(x)=\sum_{k\in Z}g_k(x)=\cdots+g_{-1}(x)+g_0(x)+g_1(x)+\cdots \tag{4}$$

对于某一固定 $N\in Z$ 和任意正整数 M,有

$$L^2(R)=V_{N-M}+W_{N-1}+W_{N-2}+\cdots+W_{N-M} \tag{5}$$

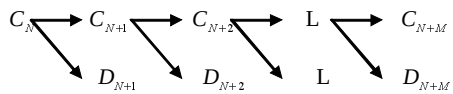

图 3 Mallat 分解算法

$$f_N(x)=f_{N-M}(x)+g_{N-1}(x)+g_{N-2}(x)+\cdots+g_{N-M}(x) \tag{6}$$

式中:$f_k(x)=\sum_{j=-\infty}^{+\infty}C_{k,j}\Phi(2^kx-j)g_k(x)=\sum_{j=-\infty}^{+\infty}D_{k,j}\psi(2^kx-j)$。其中,$C_N$ 代表第 N 层小波分解后的低频部分;很明显 C_0 代表原始待分解数据,本算法 C_0 中是轮廓追踪提取的坐标序列;D_N 代表第 N 层小波分解后的高频细节部分。经过一层小波分解后,C_1 是代表了 C_0 中的低频部分,表征数据的轮廓信息,而 D_1 则代表了 C_0 中的高频部分,表征数据中的细节信息。

3.2 分形维数

分形是研究自然界自相似现象的有力数学工具[11,12],自相似现象产生的动力学基础是混沌吸引子。设 $f(t)$是一个一维信号,以 t 为 X 轴,$f(t)$为 Y 轴,得到的二维图形记为 $F(t,f(t))$。若 F 满足自相似条件 $f(t)=a^{-H}f(at)$,其中:a 为尺度,H 为常数,则称 F 是一个二维分形图形。

分形维数又称 Hausdorff 维数,是描述分形体的一个重要特征量。它不同于经典几何学中的整数型的欧几里德维数,而是建立在 Hausdorff 测度下的一种分数型的维数。实际应用中,分形体的 Hausdorff 维数一般是无法直接计算得到的,而是计算其近似值。下面仅介绍本文所使用的分形维数的近似值——计盒维数。

定义 对于分形图形 F,用边长为 ε 的正方形网格覆盖 F,设与分形图形 F 相交的正方形个数为 $N(\varepsilon)$,则 F 的计盒维数 $D_B(F)$为 $D_B(F)=\lim_{\varepsilon\to 0}\frac{\ln(N(\varepsilon))}{\ln(1/\varepsilon)}$。

4 实验结果

本文对字符“A”、“H”、“己”、“已”等做了具体实验。实验过程中对轮廓追踪坐标序列进行两层 db5 小波分解,得到 5 条曲线,再对每条曲线求其分形维。在小波分解中,横坐标序列 C_0 与第一层分解中 C_1(C_0 的低频部分),D_1(C_0 的高频部分)是非自相似的,而 C_1 与第二层分解中的 C_2,D_2 也是非自相似的,所以计算所得的各曲线的分形维数它们之间是不相关的,可以以这些分形维数构成特征向量,进而组成训练集。由于训练集中的样本是 5 维向量,维数低,分类识别时便于待识别字符的特征向量与样本的相关性度量准则的建立。

通过表 1 和表 2 的对比可以看出,本文算法在特征向量

表 1 字符 A、H、己、已的特征向量(本文算法)

字符	特征向量				
A	1.490 4	1.431 0	1.477 9	1.303 1	1.394 4
H	1.573 5	1.421 8	1.292 7	1.281 2	1.308 6
己	1.616 3	1.379 5	1.532 0	1.249 0	1.314 7
已	1.611 2	1.515 8	1.354 0	1.238 6	1.221 7

表 2 字符 A、H、己、已的特征向量(环形统计法)

字符	特征向量				
A	1.095 2	1.086 6	1.080 7	1.119 0	1.106 1
H	1.182 7	1.174 2	1.140 5	1.134 8	1.191 9
己	1.151 6	1.153 8	1.150 5	1.108 7	1.099 3
已	1.191 9	1.164 2	1.160 1	1.215 7	1.159 0

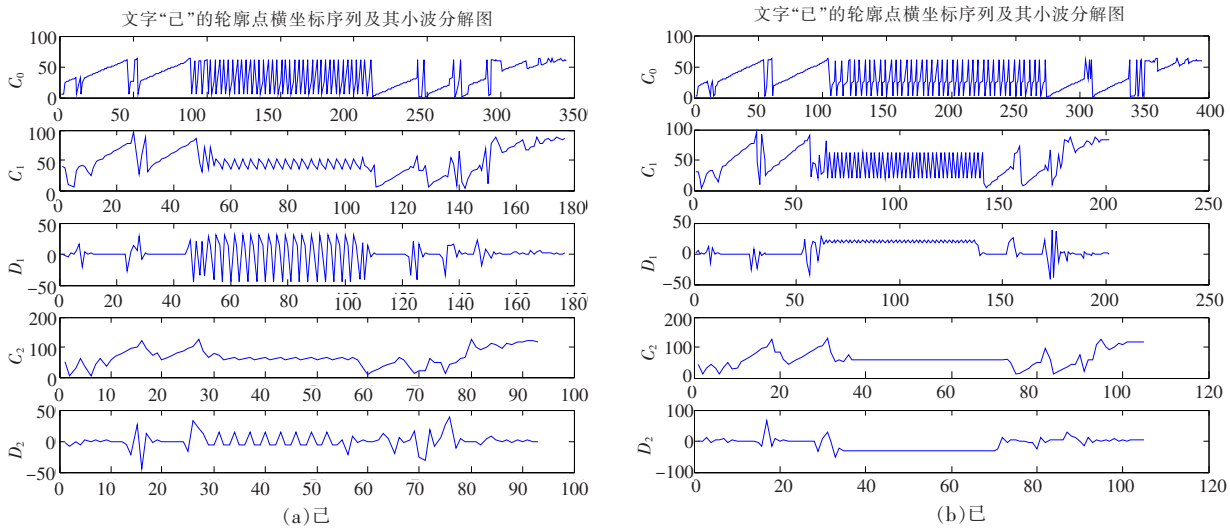

图 4　文字“己”和“已”的轮廓点横坐标序列及其小波分解图

维数一致的情况下，较之文献[13]的环形统计法具有更佳的特征提取效果。字符“A”、“H”和“己”、“已”的特征向量对应元素的最大距离分别由 0.087 6 和 0.107 0 提高到 0.185 2 和 0.178 0，而最小距离近乎不变。

5　结论

通过以上具体的字符识别实验可以看出，基于轮廓追踪的字符识别特征选取方法是有效的。即便字符间差距很小，例如“己”和“已”，所得特征向量之间的差异仍较大，保证了识别系统的准确率。其优点是特征向量维数低，识别率高，但其抗干扰性较差，对旋转字符识别存在问题，需要在预处理过程中加以解决。（收稿日期：2007 年 1 月）

参考文献：

[1] 戚飞虎.模式识别与图像处理[M].上海：上海交通大学出版社，1989.

[2] 郝红卫.集成手写汉字识别方法与系统[D].中科院自动化所，1996-06.

[3] 林晓帆，丁晓青，吴右寿.基于置信度分析和多信息融合的高性能手写数字识别[J].清华大学学报：自然科学版，1998，38(9)：47-50.

[4] Chao Kan，Mandyam D Srinath.Invariant character recognition with zernike and orthogonal fourier-mellin moments[J].Pattern Recognition，2000，28.

[5] Lin C S，Hwang C L.New forms of shape invariance from elliptic fourier descriptors[J].Pattern Recognition，1990，23(11)：1155-1166.

[6] Lei Huang，Xiao Huang.Multiresolution recognition of offline hand-written Chinese characters with wavelet transform[C]//International Conference on Documents Analysis and Reference，USA，2001.

[7] Plamondon R，Srihari S N.On line and off line handwriting recognition：a comprehensive survey[J].IEEE Trans Pattern Analysis and Machine Intelligence，2000，22(1)：63-84.

[8] 程正兴，林勇平.小波分析在图像处理中的应用[J].工程数学学报，2001，18(5)：57-86.

[9] Tang Y Y，Yang L H，Liu J，et al.Wavelet theory and its application to pattern recognition[M].London：World Scientific Singapore，2000.

[10] 程正兴.小波分析算法与应用[M].西安：西安交通大学出版社，1998.

[11] Mandelbort B B.The fractal geometry of nature[M].New York：Freeman，1982.

[12] Peitgen O，Jurgens H，Saupe D.Chaos and fractals [M].New York：Springer-Verlag，1992.

[13] 张国华，穆静，吴琼.小波分析在文字识别中的应用[J].西安工业学院学报，2004，24(3)：226-230.

（上接 183 页）

参考文献：

[1] 王伟凝，余英林，张创超.基于线条方向直方图的图像情感语义分类[J].计算机工程，2005，31(10)：7-9.

[2] 陈希超，夏顺仁.一种基于颜色统计聚类的医学图像检索技术[J].计算机工程与应用，2004，40(5)：40-41.

[3] Chen Yi-xin，Wang J Z，Krovetz R.Content based image retrieval by clustering [C]//MIR'03，November 7，2003，Berkeley，California，USA.

[4] Yin Xiao-xin，Li Ming-jing.Semantic image clustering using relevant feedback.IEEE，2003.

[5] 莫宏伟，金鸿章.人工免疫系统：一个新兴的交叉学科[J].计算机工程与科学，2004，26(5)：70-73.

[6] 刘韬，王耀才，王致杰.一种基于人工免疫系统的聚类算法[J].计算机工程与应用，2004，40(19)：182-184.