基于经验模式分解的汉字字体识别方法*

杨志华[1], 齐东旭[1,2], 杨力华[3+], 吴立军[1]

[1](中山大学 信息科学与技术学院,广东 广州 510275)

[2](澳门科技大学 资讯科技学院,澳门)

[3](中山大学 数学与计算科学学院,广东 广州 510275)

A Chinese Font Recognition Method Based on Empirical Mode Decomposition

YANG Zhi-Hua[1], QI Dong-Xu[1,2], YANG Li-Hua[3+], WU Li-Jun[1]

[1](School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China)

[2](Faculty of Information Technology, Macao University of Science and Technology, Macao, China)

[3](School of Mathematics and Computing Science, Sun Yat-sen University, Guangzhou 510275, China)

+ Corresponding author: Phn: +86-20-84115508, E-mail: mcsylh@zsu.edu.cn, http://www.zsu.edu.cn

Received 2004-09-07; Accepted 2005-03-11

Yang ZH, Qi DX, Yang LH, Wu LJ. A Chinese font recognition method based on empirical mode decomposition. ***Journal of Software*****, 2005,16(8):1438−1444.** DOI: 10.1360/jos161438

Abstract: This paper gives a novel approach to recognize Chinese fonts based on Empirical Mode Decomposition (EMD). By analyzing and comparing a great number of Chinese characters, 8 basic strokes are selected to characterize the structural attributes of Chinese fonts. Based on them, stroke feature sequences of each text block are calculated. Once decomposed by EMD, their first two intrinsic mode functions (IMFs), which are of the highest frequencies, are used to calculate the stroke energy of all the 8 basic strokes, forming the average of the energy of the two IMFs over the length of the sequence. To distinguish bold fonts from their regular fonts, average of the pixel's gray levels of the text is calculated and appended to the feature vector to form a 9 dimensional feature. Finally, the minimum distance classifier is used to recognize the fonts. Experiments show encouraging recognition rates.

Key words: font recognition; empirical mode decomposition (EMD); Hilbert-Huang transform (HHT)

摘 要: 提出了一种基于经验模式分解(empirical mode decomposition,简称EMD)的汉字字体识别方法.通过对大量汉字字体的研究比较,选取了能反映汉字字体基本特征的 8 种基本笔画.以这 8 种汉字笔画为模板,在汉

* Supported by the National Natural Science Foundation of China under Grant Nos.60133020, 60475042 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318000 (国家重点基础研究发展规划(973)); the Guangdong Provincial Natural Science Foundation of Guangdong Province of China under Grant No.036608 (广东省自然科学基金); the Foundation of Scientific and Technological Planning Project of Guangzhou of China under Grant No.2003J1-C0201 (广州市科技计划项目)

作者简介: 杨志华(1964－),男,湖南沅江人,博士生,讲师,主要研究领域信号分析,模式识别,信息安全;齐东旭(1940－), 男,教授,博士生导师,主要研究领域为数值分析,计算机图形学;杨力华(1962－),男,博士,教授,博士生导师,主要研究领域为小波与时频分析,模式识别,图像处理;吴立军(1965－),男,博士生,副教授,主要研究领域为人工智能,模型检测.

字文档图像块中随机地抽取笔画信息,形成笔画特征序列.通过对笔画特征序列作 EMD 分解,提取每个笔画特征序列的高频能量,并结合汉字文档图像块的平均灰度,形成字体识别的一个 9 维特征.

关键词: 字体识别;经验模式分解(EMD);Hilbert-Huang 变换

中图法分类号: TP391 文献标识码: A

字体识别就是判断文本图像中字体的类型,是计算机自动文档分析和处理中重要的研究内容之一.近 20 多年来,OCR(optical character recognition)获得了迅猛的发展.目前,印刷体字符识别技术已基本成熟,识别率已经达到了商业应用的要求.然而,现有的 OCR 系统主要面向于“识字”的层面,即提取版面的文字内容,而对文档结构信息的研究与实用化的要求还相距甚远.这对于版面结构的完整保存、重现和再编辑是一个核心的技术问题.另一方面,单体(mono-font)字符的 OCR 系统显然比多体(multi-font)和全体(omni-font)字符的 OCR 系统简单,并具有更高的识别率.因此,如果我们能够准确识别文档图像的字体信息,就能够将多体字符识别转换为单体字符识别,获得更好的识别效果.

然而,字体识别远没有引起人们应有的重视.关于字体识别的研究还相对较少.已有的研究主要包括:(a) 利用字体的局部特征,如衬线、字体粗细等进行字体识别[1];(b) 基于局部或全局版面特征进行字体识别[2−4];(c) 基于纹理分析的字体识别[5,6].这些方法分别利用了文字图像的不同特征,获得了较为满意的识别效果.由于汉字字符的复杂性,使得汉字字体的识别比英文字体更加困难.近年来,我国学者在这方面已经开展了富有成效的研究,并取得了令人瞩目的成果.文献[5]基于纹理分析的观点利用 Gabor 滤波器提取字体特征,得到了较好的识别效果.但是特征维数较高,计算比较复杂.文献[7]利用小波变换提出了单字符汉字字体识别方法,对单个字符的字体平均识别率达到 97.35%.然而,这种方法的特征维数高达 256 维,有很高的计算复杂度.

本文基于经验模式分解(empirical mode decomposition,简称 EMD)[8]提出了一种新的汉字字体识别方法.实验结果表明,此方法具有高识别率和低特征维数等优点.

本文第 1 节引入字体特征的提取方法.第 2 节给出基于 EMD 的字体识别方法.第 3 节给出字体识别系统的构成框架.第 4 节是实验结果及分析.第 5 节是总结.

1 字体特征提取

1.1 预处理

由于不同的文字图像其文字的大小、字间距、行间距往往不相同,而且,有些原始的文字块还可能包含许多空格(例如在大多数段落的结尾处).这些因素都给字体识别带来麻烦.为了消除这些非本质因素的影响,在进行字体识别之前进行预处理是有必要的.预处理的主要目的是:(a) 将文字块中的文字规范到预先指定的大小;(b) 将文字块中的行间距和字间距规范到预先指定的大小;(c) 填补文字块中可能的空格.由于我们的字体识别方法是与内容无关的,所以,用来填补空格的文字可以直接从原文字块中的任何非空格部分抽取.

由于预处理方法不是本文的研究重点,因此本文直接采用了文献[5]中的预处理方法.关于预处理方法的细节,请参阅文献[5].

1.2 笔画特征序列的形成

通过对大量不同字体和字形的汉字的观察和比较,我们发现一些基本笔画在不同字体中是不相同的.图 1 列出了 6 种常见的不同字体中的几种基本笔画,从左至右分别是宋体、楷体、黑体、仿宋、隶书和幼圆.这些基本笔画在不同字体中所表现的特征告诉我们,如果我们能找到它们在一幅汉字文档图像中的统计特性,就有可能以此来区分不同的汉字字体.为此,我们选择了如图 2 所示的 8 种基本笔画.其中,每种笔画的线宽均为 1,且除了笔画(g)以外,每种笔画都包含 8 个像素点.对于笔画(g),我们在实际处理中将转角点的像素做乘 2 的加权处理.因此,在这种意义下,图 2 中的每种笔画均包含 8 个像素.为方便描述,称这种基本笔画为 8×1 像素笔画.

这样,图 2 中的每种基本笔画实际上就是具有一定几何结构的 8 个点.将该笔画随机地拼贴到文字块的某

个局部,计算该笔画上8个点所在位置的文字块上像素的灰度值的均值(即8个灰度值的平均值).显然,计算所得的值反映了这个位置的文本与所使用的基本笔画的相似程度.由于上述位置的随机性,我们需要进行大量这样的计算才能获取整个文本块与所使用笔画的相似性.将上述计算进行 N 次,我们便得到一个长度为 N 的序列,称为该笔画的笔画特征序列.对如图 2 所示的每种笔画均进行上述计算,便能得到反映文本块结构特征的 8 个笔画特征序列.

Fig.1 Some basic strokes of six kinds of Chinese fonts

图 1 6 种不同字体中的基本笔画

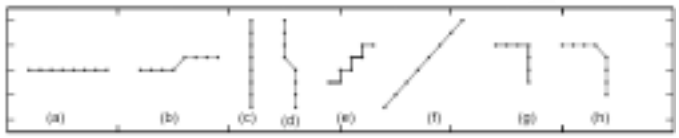


Fig.2 Templates of eight selected basic strokes

图 2 8 种基本笔画示意图

1.3 经验模式分解(EMD)

1998 年,Huang 等人提出了一种具有自适应时频分辨能力的信号分析方法[8],称为 Hilbert-Huang 变换(HHT).其核心包括两部分:经验模式分解和 Hilbert 谱分析.该方法首先通过 EMD 提取复杂信号在每一个时间局部的振荡模式,将复杂信号分解为有限个固有模式函数(intrinsic mode function,简称 IMF)之和;然后对每个 IMF 作 Hilbert 变换,进而计算每个 IMF 的瞬时频率和振幅(能量);最后将所有 IMF 的时间、瞬时频率和振幅的关系形成一个时间-频率-振幅的三维表示,即 Hilbert 谱.Hilbert 谱不仅具有很好的时频局部性,而且基于 Hilbert 谱的信号分析还具有很好的物理意义.近年来,关于 HHT 理论和应用的文献不断涌现[9,10].应用领域已经遍及地震、海洋、医学、语音及图像分析与处理[11,12].

有关 EMD 的详细介绍请参阅文献[8].本文中,由于我们提取的笔画特征序列是非平稳随机序列,因此,用 EMD 分解能够很好地提取它们在各个局部的频率结构,使得最终提取的特征信息能够很好地反映不同字体的特征.

1.4 提取字体特征

不同字体包含的各种基本笔画是不相同的.例如,宋体、黑体字体的汉字比楷体、仿宋字体的汉字包含有较多的笔画(a),这意味着宋体、黑体字体的笔画(a)特征序列比楷体、仿宋字体的笔画(a)特征序列具有较大的高频振幅.因此,宋体、黑体字体的笔画(a)特征序列经 EMD 分解后提取的高频能量将大于楷体、仿宋字体的笔画(a)特征序列的高频能量.

图 3 是两幅大小为 128×128 的汉字图像,左边为黑体,右边为楷体.在它们的下方分别给出了取 N=200 时的笔画(a)特征序列和它们经 EMD 分解以后得到的 IMFs 及其余量.为叙述方便,我们将黑体和楷体文本图像块的笔画(a)特征序列记为 x_a^H 和 x_a^K ,记 x_a^H 和 x_a^K 的第 i 个 IMF 为 $imf_a^H(i)$ 和 $imf_a^K(i)$.在下面的实验中,我们所使用的文本块均为 256 级灰度图像,按习惯,0 表示完全黑像素,255 表示完全白像素.这样,当用基本笔画模板按上述均值方法计算相似度时,值为 0 表示相似度最大,值为 255 表示相似度最小.由于在黑体汉字中有较多的笔画(a),因此,我们看到在 x_a^H 中几处出现 x_a^H =0 或接近于 0.这意味着笔画(a)模板与黑体图像块中的某些笔画(a)达到了完全耦合或几乎完全耦合.而在 x_a^K 中,几乎没有 x_a^H =0 或接近于 0 的点,这说明在楷体汉字中的笔画(a)很少.通过对 x_a^H 和 x_a^K 作 EMD 分解可以看到 $imf_a^H(1)$ 和 $imf_a^H(2)$ 比 $imf_a^H(1)$ 和 $imf_a^H(2)$ 具有更大的振荡幅度.这表明,x_a^H 将比 x_a^K 具有更高的高频能量.

取大小为 128×128 的黑体、宋体、仿宋和楷体字体汉字图像块各 25 幅,当 N=1000 时,用笔画(a)提取每幅汉字图像块的笔画(a)特征序列.对每一个笔画(a)特征序列作 EMD 分解,保留其前面的两个 IMF,并计算这两个 IMF 对长度 N 的平均能量,以此作为该文本块中笔画(a)的特征能量,我们称其为该文本块的笔画(a)能量.对每种字体的 25 幅图分别计算其笔画(a)能量,并将它们用折线连接起来.图 4 给出了黑体、宋体、仿宋和楷体字体汉

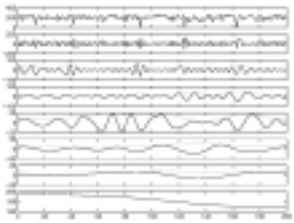
字图像块各 25 幅图的笔画(a)能量.不难发现,黑体和宋体的笔画(a)能量明显高于仿宋和楷体的笔画(a)能量,而楷体和仿宋体的笔画(a)能量则基本上没有差别.

A Chinese text block of HeiTi

黑体汉字图块

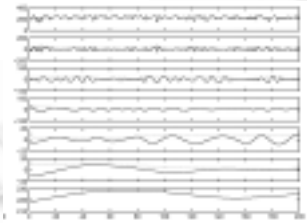
A Chinese text block of KaiTi

楷体汉字图块



The stroke(a) feature series of 200 samples corresponding to the Chinese text blocks of HeiTi and its EMD decomposition

长度为 200 的黑体汉字图块的笔画(a)特征序列和它的 EMD 分解



The stroke(a) feature series of 200 samples corresponding to the Chinese text blocks of KaiTi and its EMD decomposition.

长度为 200 的楷体汉字图块的笔画(a)特征序列和它的 EMD 分解

Fig.3

图 3

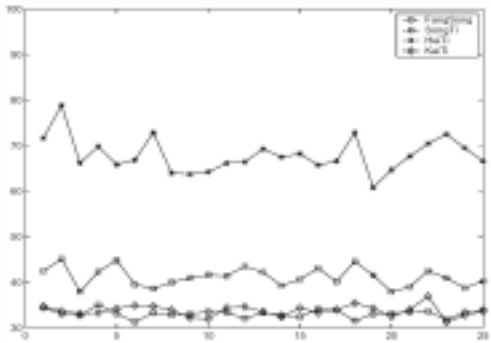


Fig.4 The distribution of the stroke(a) energy extracted from 100 text blocks of HeiTi, SongTi, FangSong and KaiTi

图 4 黑体、宋体、仿宋和楷体字体的笔画(a)能量分布

类似地,其他几种笔画的能量特征也反映了其他不同字体的结构特征.8 种笔画能量组成一个 8 维的特征向量.除此之外,汉字图像块的总灰度平均可以用于区分粗体和非粗体字型,且易于获取.因此,我们在特征向量中加入该特征,最后形成一个 9 维的特征向量.

综上,将本文提出的特征提取算法归纳如下.

算法. 设 I 是经预处理后的文字图像块,设定特征序列长度为 $N\approx 10\max(W,H)$,其中,W 和 H 分别是图像块的宽度和高度.

第 1 步.令 $i=1$;

第 2 步.按第 1.2 节所述方法提取第 i 种笔画的笔画特征序列 x_i;

第 3 步.对 x_i 作 EMD 分解,得到前两个 IMF,即 IMF1 和 IMF2;

第 4 步.利用 IMF1 和 IMF2,按照第 1.4 节所述,计算该文本图像块的笔画(i)能量;

第 5 步.令 $i=i+1$,若 $i\leq 8$,转第 2 步;否则,转第 6 步;

第 6 步.求 I 的均值,作为特征向量的第 9 个特征,形成 9 维的特征向量.

2 字体识别

在取得字体的特征向量之后,下面的任务是设计分类器进行字体分类.分类器的设计是模式识别中的关键问题.本文的研究重点不是分类器设计.所以,为了更好地与文献[7]中的结果进行比较,我们采用与文献[7]中相同的最小距离分类器.

设 f_j 是待识别字体的第 j 个笔画能量,f_j^k 是第 k 种字体的第 j 个笔画能量的均值,δ_j^k 是第 k 种字体的第 j 个笔画能量的标准差.当满足下面的条件时,待识别字体被识别为第 k 种字体:

$$k = \arg\min_{1\le k\le K}\left\{\sum_{j=1}^{9}\frac{(f_j - f_j^k)^2}{(\delta_j^k)^2}\right\} \tag{1}$$

其中,K 为总的字体类别数.

3 字体识别系统

根据以上分析,一个完整的字体识别系统如图 5 所示.

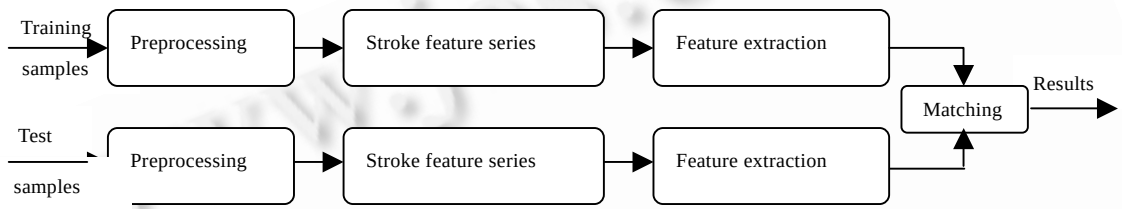


Fig.5 The skeleton of a font recognition system

图 5 字体识别系统示意图

4 实验结果及分析

为了验证本文方法的有效性,我们选取了 6 种常用的汉字字体,即宋体、楷体、黑体、仿宋、隶书和幼圆.每种字体又分 4 种字形,即标准体、粗体、斜体和粗斜体,总共 24 种字体(形).训练和测试样本分为两类,即计算机生成的文档图像和经扫描仪获得的文档图像.计算机生成的文档图像由 Photoshop 7.0 生成,图像分辨率为 72pixels/inch,灰度模式.扫描图像由 HP scanjet 3670 扫描仪扫描获得,扫描分辨率为 100dpi,灰度模式.对每种字体,我们生成计算机图像和扫描图像各 50 幅,大小为 128×128 像素.每种字体 100 个样本,总共 2 400 个样本.在每种字体的 100 个样本中,40 个作为训练样本(计算机图像和扫描图像各 20 个),其余 60 个作为测试样本.

对上面的训练和测试样本,表 1 给出了测试结果.从表 1 的统计结果看,平均识别率达到了 96.4%.需要说明的是,我们的实验中有一半的样本来自于扫描样本.这些扫描样本都是由实际文本经普通扫描仪获得.能够得到平均 96.4%的高识别率,这本身就说明了我们的方法具有很好的抗干扰能力.具有这种良好的抗干扰能力的主要原因是我们在提取笔画特征序列时,特征序列的每一个值是基本笔画模板中 8 个像素点的均值,因此,如果实际文本图像块中仅存在某些孤立的干扰点,作 1/8 平均以后,对特征序列的影响很小.然而,由于实际文本纸质发黄而引起的扫描图像整体发暗,或因扫描设置不当引起的图像发暗或发白,将会对识别结果产生较大的影响,对于这种情况,我们需要对图像块进行预处理.

表 2 给出了各种字体之间的混淆率.其中,仿宋误识为楷体的混淆率最高,为 4.3%.其主要原因是这两种字体的多种笔画存在很大的相似性.

文献[7]用基于全局纹理分析的方法对以上常见的 6 种字体和 4 种字形的汉字字体进行识别,得到了平均 98.5%的识别率,比我们的识别率略高.但是,由于我们无法得到文献[7]的字体数据,所以,这一比较并不是完全客观的.而且,文献[7]的训练和测试样本都是计算机生成的样本,而在我们的实验中所用到的样本既有计算机生成的样本,也包括扫描样本.需要指出的是,文献[7]使用了 16 维特征,而我们的方法只用到 9 维特征.

Table 1 Recognition rate of fonts (%)

表 1 字体识别率 (%)

	SongTi	KaiTi	HeiTi	FangSong	LiShu	YouYuan	Average
Regular	98	100	90.2	100	100	100	98
Bold	90.6	88.7	100	92.1	100	93.5	94.2
Italic	100	100	95.6	91.3	100	100	97.8
Bold italic	100	92.8	100	89.8	94.5	95.9	95.5
Average	97.2	95.4	96.5	93.3	98.6	97.4	96.4

Table 2 Font confusion rate (%)

表 2 字体混淆率 (%)

	SongTi	KaiTi	HeiTi	FangSong	LiShu	YouYuan
SongTi	97.2	0	1.3	1.5	0	0
KaiTi	0	95.4	0	3.8	0.8	0
HeiTi	3.4	0	96.5	0	0	0.1
FangSong	2.4	4.3	0	93.3	0	0
LiShu	0	0	1.4	0	98.6	0
YouYuan	0.4	0	2.2	0	0	97.4

5 结 论

本文提出了一种基于 EMD 分解的汉字字体识别方法.该方法具有如下优点:

- 特征维数低.
- 在训练和测试中只需要较少的样本个数.尽管在实验中,我们使用了大量样本进行训练和测试.但是,由于笔画特征序列是用笔画模板在汉字图像块中随机抽取的,因此事实上,同一汉字图像块可以多次使用,而不会对测试结果产生较大的影响.
- 训练样本和测试样本的大小和形状可以不相同.
- 较高的识别率.

由于本文的方法是采用汉字笔画来抽取笔画特征序列,因此,除非选取适合于西文字体特征的基本笔画,否则,该方法不适合西文字体的识别.另外,本文方法的识别率还直接依赖于图像分割质量的好坏,当一个图像块中包含有多种字体时,由于它包含了多种字体的统计特征,因而很容易造成误识,这是本文方法的一些缺陷.然而,EMD 作为一种新的信号分析方法,被成功地用于提取汉字文本的笔画特征,是有意义的探索.它为拓展包括 EMD 在内的 Hilbert-Huang 变换理论的应用给出了一个新的尝试.

References:

[1] Khoubyari S, Hull JJ. Font and function word identification in document recognition. Computer Vision and Image Understanding, 1996,63(1):66−74.

[2] Shi H, Pavlidis T. Font recognition and contextual processing for more accurate text recognition. In: Proc. of the ICDAR'97. ULm: IEEE Computer Society Press, 1997. 39−44.

[3] Zramdini A, Ingold R. Optical font recognition using typographical features. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,220(8):877−882.

[4] Jung MC, Shin YC, Srihari SN. Multifont classification using typographical attributes. In: Proc. of the ICDAR'99. Bangalore: IEEE Computer Socety Press, 1999. 353−356.

[5] Zhu Y, Tan TN, Wang YH. Font recognition based on global texture analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001,23(10):1192−1200.

[6] Zeng L, Tang YY, Chen TH. Multi-Scale wavelet texture-based script identification method. Chinese Journal of Computers, 2000,23(7):699−704 (in Chinese with English abstract).

[7] Chen L, Ding XQ. Font recognition of single Chinese character based on wavelet feature. Acta Electronica Sinica, 2004,32(2):177−180 (in Chinese with English abstract).

[8] Huang NE, Shen Z, Long SR. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. of the Royal Society of London, 1998,A(454):903−995.

[9] Flandrin P, Rilling G, Goncalves P. Empirical mode decomposition as a filter bank. IEEE Signal Processing Letters, 2004, 11(2):112−114.

[10] Deng YJ, Wang W, Qian CC, Dai DJ. Boundary-Processing technique in EMD method and Hilbert transform. Chinese Science Bulletin, 2001,46(3):954−961 (in Chinese with English abstract).

[11] Yang ZH, Huang D, Yang LH. A novel pitch period detection algorithm based on Hilbert-Huang transform. LNCS 3338, 2004. 586−593.

[12] Yang ZH, Qi DX, Yang LH. Signal period analysis based on Hilbert-Huang transform and its application to texture analysis. In: Proc. of the 3rd Int'l Conf. on Image and Graphics. Hong Kong: IEEE Computer Society Press, 2004. 430−433.

附中文参考文献:

[6] 曾理,唐远炎,陈廷槐.基于多尺度小波纹理分析的文字种类自动识别.计算机学报,2000,23(7):699−704.

[7] 陈力,丁晓青.基于小波特征的单字符汉字字体识别.电子学报,2004,32(2):177−180.

[10] 邓拥军,王伟,钱成春.EMD 方法及 Hilbert 变换中边界点问题的处理.科学通报,2001,46(3):257−263.