

概率统计应用于手写模式识别系统

熊晓龙¹, 钟满田²

(1. 武汉工程大学理学院, 湖北 武汉 430074; 2. 罗定职业技术学院, 广东 罗定 527200)

摘要:介绍了模式识别及手写模式识别系统相关知识,进而叙述了手写模式识别系统的结构设计、实现和实验,通过实验得出该手写模式识别系统在理论上和实际上是可行的结论。

关键词:概率统计; 手写; 模式识别

中图分类号:O21

文献标识码:A

doi:10.3969/j.issn.1674-2869.2010.01.031

0 引言

模式识别(Pattern Recognition)是人类的一项基本智能,在日常生活中,人们经常在进行“模式识别”。随着20世纪40年代计算机的出现以及50年代人工智能的兴起,人们希望能用计算机来代替或扩展人类的部分脑力劳动。(计算机)模式识别在20世纪60年代初迅速发展并成为一门新学科。

模式识别是指对表征事物或现象的各种形式的(数值的、文字的和逻辑关系的)信息进行处理和分析,以对事物或现象进行描述、辨认、分类和解释的过程,是信息科学和人工智能的重要组成部分^[1]。

在计算机安全领域中,一般识别方式为:信息(密码、PIN或其他用户个人信息)、实物(钥匙、智能卡、IC卡等)。硬件设备容易丢失、被偷窃和复制;冗长的口令容易忘记,“数字手写”技术在难以分配和管理数量庞大的密钥的同时,也难以避免泄密的可能。在这些识别方法中,每个人唯一的生物特征无疑是最安全的识别工具。在基于生物特征的识别方法中,指纹、视网膜识别等需要复杂的信息获取设备和技术,个人不便于携带,成本高,一般只用于特殊场合。而基于人的行为特征的手写模式识别技术具有方便、实用、低成本并符合人的生活习惯等特点,受到了社会广泛的关注。

1 手写模式识别系统介绍

模式识别系统具有巨大的实用价值,而要想对用户的身份加以识别,则需要对用户的手写字

进行识别。

1.1 手写模式识别算法

计算机笔迹识别主要分为在线和离线两类。离线笔迹识别的对象是写在纸上的字符,通过扫描仪和摄像机转化为计算机能处理的信号;而在线的笔迹识别则通过专用的数字板或数字仪实时地采集书写信号,它不仅可以采集到笔迹序列并转化成图像,而且可以记录书写的压力、速度、停顿时间等信息,可为笔迹识别提供更丰富的信息。

1.2 手写模式识别系统特点

在设计基于生物特征的安全系统时,最主要的研究方向就是如何对所获取生物特征进行特征提取和模式识别,以及如何将生物特征验证系统集成到自己的应用程序安全控制中^[1]。同时还需要考虑到其他一些因素:易用性、突发事件、精确度、成本、用户的接受率、所需安全级别、系统长时间运行的稳定性。

2 模式识别系统结构设计

模式识别系统一定要能方便、高效地嵌入到某个应用系统中才能发挥它的作用。为了能提供一种更灵活、方便的方式与各种OA系统相接合,这里将该手写字识别部分以OCX控件和DLL文件的方式提供,可根据需要定制自己的界面以及安全策略。识别服务器可通过ODBC接口方便地与各种流行的数据库服务器进行连接。

手写模式识别系统一般的登录过程:新用户首先需要在放松的状态下通过写字板在控件上自然书写5~10个样本手写字,然后系统会通过这些样本来提取用户的行为特征,并形成

唯一字符串作为模板保存在任意存储介质中。当他(或其它人)想以他的身份登录时,就用从这次测试手写字中提取的行为特征与相应模板做比较,如果本次手写字是真实的,那么这个比较的过程就会得到一个较高的相似度,否则所得到的相似度会较低。

3 特征提取

在该系统中用来进行手写模式识别的特征都是经过广泛研究和应用的,包括书写总时间、总长度、速度零点次数等等。而在将参考模板和待测手写字进行比较时,笔者采用了另外一种算法:那就是使用了用来计算概率密度函数的高斯函数来为这些特征进行建模。也就是说通过密度函数,使用一个人的所有样本来对各个特性进行参数估计。在估计的过程中可以得到各个特征的均值和方差,并且使用它们作为手写字的唯一标识。在识别阶段,将待测特征手写字的各个特征与参考特征估计值进行匹配,并将每个特征上的匹配概率值进行叠加从而得到总的概率值。然后就可以通过这个概率估计值对待测手写字的真伪进行判断^[2,3]。

手写笔驱动程序可将笔每次移动的点的 X—Y 坐标报告给控件,即可进行特征提取。

3.1 计算手写字特征

要识别不同的手写字,必须从计算手写字特征开始。因此,使用如下的手写字特征作为手写字识别的依据^[4-6]:

- ①在 X 方向上的速度零点个数。
 - ②在 Y 方向上的速度零点个数(绝对零点是不可测的,所以采用与峰值相比低于某个小百分比的速度值代替零点)。
 - ③在 X 方向上加速度零点个数。
 - ④在 Y 方向上加速度零点个数(这些零点个数表明由加速到减速或者由减速到加速的变化)。
 - ⑤手写字所用的总时间。
 - ⑥手写笔在手写板上经过的总路程,即所有点之间的欧几里得距离值的和:
- $$D_{XY} = \sum_{i=0}^{n-1} \{(X_i - X_{i+1})^2 + (Y_i - Y_{i+1})^2\}^{\frac{1}{2}} \quad (1)$$
- ⑦总抬笔次数。
 - ⑧抬笔总时间(即从开始书写到最后一次书写间的抬笔时间和)。
 - ⑨压力高于某个阈值 T_{\max} 的次数(由手写笔硬件特性而定)。
 - ⑩压力低于某个阈值 T_{\min} 的次数(由手写笔

硬件特性而定)。

速度是时间的函数,可以用下面的公式计算:

$$V_{X_n} = \frac{X_{n+1} - X_n}{t_{n+1} - t_n} \quad (2)$$

$$V_{Y_n} = \frac{Y_{n+1} - Y_n}{t_{n+1} - t_n} \quad (3)$$

V_{X_n}, V_{Y_n} 分别表示在 X 方向和 Y 方向上在 t_n 时间点上的速度。

加速度是速度的导数,可以得出在 t_n 时间点上 X 方向和 Y 方向上的加速度可以表示为:

$$A_{X_n} = \frac{V_{n+1} - V_n}{t_{n+1} - t_n} \quad (4)$$

$$A_{Y_n} = \frac{V_{n+1} - V_n}{t_{n+1} - t_n} \quad (5)$$

3.2 特征提取

在向系统内注册一个新用户的过程中,需要用户在某段时间内输入 20 到 30 个手写字(越多越好),但这里根据应用系统的需要,要求用户一次性全部输入。然后从这些样本手写字中提取所有特征。如果用户输入了 20 个样本手写字,那么就得到了 20 个包含 10 个特征元素的向量。在笔者设计和实现系统时发现用户的手写字与其心理状态有很大关系,在不同的环境下、时间段内以及其他因素影响下手写字都有较大差异。因此,像总时间、总长度、抬笔时间等特征即使对于同一用户来说,虽然在其不同手写字中是不同的,但都围绕着可唯一表示一个人生物特性的平均值和方差上下波动因此考虑可以使用某种分布来描述这些特征。据发现用高斯密度函数来为这些特征建模是最合适的。平均值描述了数据的总体趋势,方差描述了各个分组在特征空间内的分布情况。 N 组手写字样本可以给出某一特征的 N 个值,并试图将高斯密度函数应用到这些值上,从而得到可参数化定义分布的均值估计和方差估计^[2]。均值的无偏估计可这样得到:

$$\mu_k = \sum_{i=1}^N \frac{X_i}{N} \quad (6)$$

式子(6)中的 X_i 表示手写字的特征值。

计算手写字方差的无偏估计:

$$V_{\sigma_k} = \sum_{i=1}^{N-1} \frac{(X_i - \mu_k)^2}{N-1} \quad (7)$$

式子(7)中的 k 表示手写字特征的个数。

举例说明这个计算过程。假设有某用户的 5 个样本手写字,只考虑其中的两个特征—手写字总时间和抬笔次数如表 1。

表 1 手写字的数字特征

Table 1 The digital characteristic of hand write

手写字样本	停留时间/s	次数
手写字 1	5.234	5
手写字 2	5.623	6
手写字 3	5.545	4
手写字 4	5.772	5
手写字 5	5.692	5

手写字总时间的均值是 5.546 s, 方差是 0.051 477, 抬笔次数的均值是 5, 方差是 0.666. 所以对于这 4 个样本手写字, 参考模板为 (5.546, 0.051 477)、(5, 0.666).

因此, 对于每个用户, 都可以用 (μ_k, V_{σ_k}) 来描述某一特征对于采集 10 个特征的系统, 可以得到 $(\mu_1, V_{\sigma_1}), (\mu_2, V_{\sigma_2}), \dots, (\mu_{10}, V_{\sigma_{10}})$.

3.3 仿真实验

当系统需要对一个用户进行身份验证时, 用户首先需要在控件上进行手写字. 从用户的手写字中可以提取上面所提到的特征, 并建立一个特征向量. 然后将测试手写字特征值带入高斯密度函数:

$$P(X_i) = \frac{\exp\left[-\frac{(X_i - \mu_i)^2}{2V_{\sigma_i}}\right]}{\sqrt{2\pi V_{\sigma_i}}} \quad (8)$$

这里 $P(X_i)$ 是由特征 X_i 产生的值.

由 (8) 式可以看出: X_i 越接近于均值, 则所得的 $P(X_i)$ 值越高. 将上式应用到所有特征上, 可以通过 (9) 式得到一个概率值 PS.

$$PS = \sum_{i=1}^{10} P(X_i) \quad (9)$$

在得到每个特征的均值和方差的估计值后, 还需要对阈值进行估计, 可将每个样本手写字分别作为测试手写字, 将其特征带入式 (8) 中, 得到包含所有特征估计值的 PS. 在得到所有样本手写字的 PS 后, 取这些 PS 的平均值作为阈值.

如果待测手写字的 PS 等于或者高于阈值, 那么就认为识别通过. 否则就识别失败. 因此, 可以通过调整阈值的高低 (PS 平均值的百分比) 来调整系统的性能.

再举一个具体的例子. 假设在上面的例子中参考模板的均值和方差分别为 (4.546, 0.051 477) 和 (5, 0.666). 现有一个伪造手写字, 总手写字时间为 5.2 s, 抬笔 6 次; 真实手写字用时 4.6 s, 抬笔 5 次. 通过式 (8) 和 $P(X_i)$, 可得伪造手写字的两个 P 值为 0.02 758 和 0.23 089, 总 PS 值为 0.258 477; 真实手写字的 P 值分别为 1.7 092 和 0.488 6, 总 PS 值为 2.197 8. 虽然这里没有真正

设定阈值, 但可以看出两者 PS 值的差别很大. 真实手写字的 PS 值应该超过阈值.

由于条件所限, 故只在小范围内进行了测试 (10 个人, 每个人 20 个样本手写字). 在没有笔迹模仿专家的情况下进行基础测试, 系统能识别出大约 85% 的模仿者. 另外试验用的手写板分辨率较低, 如果采用高级的手写板, 进一步提高输入的分辨率, 则识别率可以进一步提高.

4 结 语

该模式识别系统使用 VB6.0 软件进行开发, 通过调用 Windows GUI 函数已成功应用到企事业单位 (部门) 的文档管理系统中. 它主要起两方面作用: 一是身份识别, 二是报表手写字笔迹打印. 身份识别提高了系统的安全性; 报表手写字笔迹打印一可使得人们对文件的可信度一目了然, 提高高层领导审批文件的效率, 二可以增强手写字的不可抵赖性, 必要时可以由笔迹分析专家对实际手写字进行分析.

参考文献:

- [1] 侯媛彬, 杜京义, 汪梅. 神经网络 [M]. 西安: 西安电子科技大学出版社, 2004: 25-30.
- [2] 陈希孺. 概率论与数理统计 [M]. 北京: 中国科学技术出版社, 1993: 120-124.
- [3] Ito Kuen Kiat, Heiko Schroder, Graham Leedham. Codebooks for signature verification and handwriting recognition [C] // Seventh Australian and New Zealand Intelligent Information Systems Conference. New York: Springer, 2001: 221-224.
- [4] Edson J R Justino, Flavio Bortolozzi, Robert Sabourin. Off-line Signature Verification Using HMM for Random, Simple and Skilled Forgeries [C] // International Conference on Document Analysis and Recognition. New York: Springer, 2001: 1031-1034.
- [5] Robert Sabourin, Jean-Pierre Drouhard. Off-Line Signature Verification Using Directional PDF and Neural Networks [J]. Pattern Recognition, 2000, 29: 415-424.
- [6] Dariusz Z Lejtmann, Susan E George. On-line handwritten signature verification using wavelets and back-propagation neural networks [C] // International Conference on Document Analysis and Recognition. New York: Astin Bulletin, 2001: 992-996.

Research of application of the probability statistics in hand writing by pattern recognition system

XIONG Xiao-long¹, ZHONG Man-tian²

(1. School of Science, Wuhan Institute of Technology, Wuhan 430074, China;

2. School of Luoding Polytechnic, Luoding 527200, China)

Abstract: The author of this article first introduced the pattern recognition and the writing by hand pattern recognition system related knowledge, then narrated the writing by hand pattern recognition system's structural design, realization and testing. This writing by hand pattern recognition system through the experiment is the feasible conclusion theoretically and practically.

Key words: probability statistics; writing by hand; pattern recognition

本文编辑: 龚晓宁



(上接第 89 页)

[6] 张泽强, 沈上越. 二氧化钛 累托石净水材料的制备及其净水性能评价[J]. 化工环保, 2005, 25(1): 59 - 61.

[7] 曲晶, 韩丽. 累托石粘土改性试验研究及其进展[J]. 矿产保护与利用, 2004(1): 27 - 30.

[8] 唐新村, 黄伯云, 贺跃辉, 等. 低热固相反应的反应机理研究: 熔融熔机理和冷溶熔机理的模型. 无机化学学报, 2004, 20(7): 795 - 800.

[9] 陈静, 龙光斗, 艾春玲, 等. 累托石插层改性纳米材料的制备及应用[J]. 应用化工, 2005, 34(12): 785 - 787.

Preparation and XRD characterization of Cu₂O/ REC nanocomposite

LI Nian¹, PENG Meng², LI Wan-yi¹, HAI Ting-ting¹, LI Wei¹, CHEN Jin-yi¹

(1. School of Environment and Civil Engineering, Wuhan Institute of Technology, Wuhan 430074, China;

2. School of Chemical Engineering and Pharmacy, Wuhan Institute of Technology, Wuhan 430074, China)

Abstract: Rectorie (REC) is a representative layered clay mineral which has important effects on regulating and controlling the synthesis of semiconductor photocatalytic nano materials. In this paper, the Cu₂O/ REC can be sonochemically prepared and synthesized with CuCl, NaOH and various REC via low-temperature solid state method. The experimental conditions, effect factors, and the mechanism of the preparation of Cu₂O and Cu₂O/ REC nanocomposite were studied by X-Ray diffractions.

Key words: Cu₂O; rectories; nanocomposite; low temperature solid state method

本文编辑: 龚晓宁