

清华大学学术专著

Computer Writer Identification and Verification
Theory and Method

计算机笔迹鉴别与验证的 理论和方法

丁晓青 李昕 等著



清华大学出版社

清华大学学术专著

Computer Writer Identification and Verification
Theory and Method

计算机笔迹鉴别与验证的 理论和方法

丁晓青 李昕 等著



清华大学出版社
北京

前 言

文字是人类记录和传承信息的载体。人们借助印刷或书写的文字文本记载历史和传递信息。但是,无论是手写或印刷的文本图像中,除了包含有人们普遍关心的文本内容的信息外,还包含有表示文字风格的信息。前者毋庸置疑的是人们关注的主要对象;对于后者,从古至今,也一直受到人们的关心。中国书法的研究古已有之;自古就有“字如其人,人如其字”的说法;“签字画押”作为书写者的身份签证也由来已久,得到人们越来越多的关注和重视。

随着现代计算机信息技术的发展和研究者的不懈努力,利用光学字符识别技术,经过文档的分析、理解和识别,可以自动从文本图像中直接获得文本图像包含和表达的内容信息。文档内容的自动识别对文档自动化信息处理发挥着极其重要的作用。与此同时,也推动着文本图像风格信息的计算机自动分析和识别的研究。因为风格信息往往是依附于表示内容的笔画结构之上,虽然它的识别和理解往往比以笔画结构为主的内容信息的识别和理解要困难,自动对文本图像风格信息的识别和理解,自动对笔迹风格信息的辨识,可以达到对书写者身份的识别,近年来也取得了显著的进步。

作者所在的清华大学电子工程系智能图文信息处理研究室,长期从事文字识别、图文分析理解、人脸检测识别、视频监控等文档、图像及视频的智能信息处理等研究工作,迄今已逾 20 年。其中,除了对于文档内容的分析和文字识别的研究取得全面开创性成果以外,对于人脸识别、文档风格的识别理解也进行了全面和深入地研究。

文档风格的识别,特别是笔迹鉴别(writer recognition)是生物特征识别的一个重要分支,其研究具有重要的理论意义和实用价值。在世界范围,利用签字作为书写者的身份签证由来已久,笔迹鉴别在许多领域都有重要应用,如安全、金融、司法等。由于书写风格往往呈现为在字符结构的微弱变化,使书写风格的识别理解具有更大的特殊性和困难程度。

复杂困难的文档图像风格识别问题按照书写方式不同,可以分为联机书写笔迹鉴别和脱机书写笔迹鉴别两大类;按照识别的方法不同,基本上分为文本相关和文本无关两大类;按照笔迹鉴别的对象不同,又可以分为基于少量单个字符的和基于篇章的两种笔迹鉴别方法。我们的研究创新成果涵盖了

文档风格识别领域这些主要方面内容：包括基于脱机手写笔迹和联机签名鉴别的身份识别，基于单个字符的文本相关和文本无关的笔迹识别，以及高性能的基于篇章文本无关笔迹鉴别等内容。本书在介绍已有文献中笔迹识别的各种方法的基础上，将着重介绍作者研究团队在笔迹识别领域理论和方法的研究成果，最后对笔迹识别领域的研究进行总结和展望。

本书内容主要包括如下几个方面：

由于文档内容信息和风格信息是相互紧密交织于文字/文本图像中，因此从文本图像中提取风格信息是文档风格识别的关键。无论是对单字笔迹或对篇章笔迹鉴别，都将对笔迹风格鉴别特征的提取重点加以分析和研究。

单个字符文档风格识别的研究从印刷文本字体识别开始。我们提出了未知单个字符的文本无关字体识别算法。利用线性鉴别方法，对每种字体风格的文本无关字符图像获取字体风格鉴别信息，建立字符字体风格对内容变化的线性模型，成功解决了未知单个字符的印刷汉字字体识别困难问题。

手写单字符笔迹的笔迹识别的困难是如何从单字笔迹提取书写者稳定的笔迹特征。显然，从相同字符内容的笔迹图像中提取书写者笔迹信息，是一种有益的途径。我们提出的基于汉字单字笔迹的文本相关笔迹鉴别方法，基于对特定字符汉字笔迹的统计分析，仅利用少数几个相同字符，就可以取得很高的书写者身份识别正确率，实现了完整的笔迹鉴别验证实用系统，并在现实事件中得以成功地应用。

但是，将这种文本相关笔迹鉴别方法应用到大规模笔迹检索系统中时，由于现实中不能保证未知查询笔迹样本与已知参考笔迹样本间具有足够数量的相同字符，使实际应用受到很大影响。

和人类对于笔迹验证能力相比，为克服文本相关方法对字符相同内容严格要求的局限，我们提出一种基于不同单字笔迹样本的笔迹验证方法，即所谓的半文本无关的单字笔迹鉴别算法，它是利用手写汉字识别获得被鉴别字符内容的条件下，从不同内容的字符图像中提取笔迹风格信息，顺利地解决了文本无关单字笔迹鉴别问题。这种半文本无关的单字笔迹鉴别方法，克服了文本相关笔迹鉴别要求训练与参考字符相同的限制，极大地开拓了基于单字的汉字笔迹鉴别的应用范围。综上所述，我们提出和有效解决了基于单个字符的文本无关笔迹鉴别方法，并在实际中得以成功地推广应用。

对于基于篇章的文本无关笔迹鉴别算法，以往的学者已进行了长期的研究，提出了很多有效的文本无关笔迹识别方法，但是在大规模汉字笔迹库的鉴别实验结果并不理想。本书研究的重点是如何提取更好的笔迹特征，进一步提高文本无关方法的识别性能，满足实际的笔迹鉴别系统对更高鉴别性能和更广泛适应性的要求。例如，很需要一种不仅能适用于英文等西方文字、

汉字等东方文字，也能适应藏文、维吾尔文、阿拉伯文等不同文种的高性能文本无关笔迹鉴别算法和系统。我们提出了适用于不同文种笔迹识别一致性的高性能的文本无关笔迹识别方法，并在此基础上，还首次对笔迹文本的篇幅、笔迹文本内容变化等对文本无关方法性能的影响也进行了深入和细致的分析。

本书在对作者提出的方法从理论上给予深入的分析研究的同时，给出了定量的实验验证结果及其分析讨论，并提出和构建了大规模实际的笔迹查询系统，供读者参考。本书介绍的这些研究成果已经应用于相关安全领域，并取得了很好的实际成果，验证了算法的有效性。特别需要提出的是在本书交付印刷之时，即在 2011 年 9 月 28—31 日召开的第 11 届国际顶级、权威的文本分析和识别国际会议 ICDAR2011 (11th International Conference on Document Analysis and Recognition) 举办的 ICDAR2011 文档笔迹鉴别的国际评测中，我们提出的多文种的文本无关笔迹鉴别算法的识别性能，即使是在对我们从未训练过的法、德、希腊文笔迹鉴别评测中，均超过其他参赛组的成绩，获得第一和竞赛奖励。这是对我们所提算法性能优越的一个佐证。

全书共分为 8 章，除第 1 章引言外，其余部分的章节安排如下：

第 2 章研讨了如何从内容和风格信息紧密交织的字符笔迹原始图像中，将书写者的笔迹风格信息分离出来，即解决笔迹鉴别特征提取这一笔迹识别的关键问题。

第 3 章提出并介绍了基于单个字符笔迹的文本相关笔迹验证和笔迹鉴别方法，包括最能体现不同书写风格的单字笔迹的原始特征的提取、最佳笔迹鉴别特征的选择，以及分类器的设计等。

第 4 章在引入基于单个汉字的文本无关的印刷字体识别算法后，介绍基于单字笔迹的半文本无关笔迹验证方法，即提出一种在已知字符文本内容的条件下来解决文本无关的笔迹鉴别的问题。引入了代表字符内容信息的内容模板的“规一化”，以降低笔迹中字符内容结构差异的影响，这也是一种在一定字符集内容变化范围内利用线性鉴别分析提取笔迹风格特征的半文本无关笔迹验证方法。

第 5 章探讨如何解决能从篇章笔迹中提取反映书写者书写风格的笔迹特征的关键问题，提出一种将局部特征的全局篇章统计概率分布作为书写者的篇章的笔迹特征的统一提取方法。

第 6 章主要介绍适用于多文种笔迹的基于特殊网格微结构特征的文本无关笔迹鉴别方法；提出了一种新的篇章笔迹风格特性——网格微结构特征，以及使用加权距离度量的最近邻分类器，取得了最好的鉴别性能。分析了文本篇幅和文本内容对笔迹鉴别性能的影响。

第 7 章利用本书提出的基于网格微结构特征的文本无关方法和基于单字符笔迹的文本相关方法,实现了基于大规模数据库的笔迹检索系统。其中,文本无关方法可以适用于汉字、英文、藏文、维吾尔文等各种文种的笔迹。

第 8 章讨论了联机手写签名的计算机自动认证的关键问题,提出了描述联机手写签名的动态结构统计模型 DSSM,并在 DSSM 模型基础上,提出了一种和基于动态时间规整 DTW 匹配相融合的联机签名认证算法,更进一步利用了笔迹的均值特性和高阶的变化统计特性,实现了具有更高性能的联机签名认证算法和系统。

附录 D 为由国际模式识别学会 (IAPR) 召开的文档分析和识别国际会议 ICDAR2011 举办的 ICDAR2011 笔迹鉴别竞赛的评测报告,报告对来自 7 个参赛机构共 8 种方法,使用数据集的完整图像和剪裁图像两种方案进行了评测。竞赛第一名均是来自中国北京清华大学电子工程系我们的方法。

本书总结了笔者和众多博士研究生在笔迹识别这一领域的研究工作所取得的学术成果,其中包括王贤良博士在单字符汉字笔迹鉴别的研究成果 (第 2、3、7 章),陈力博士、吴涛在关于单个汉字文本无关字体识别 (第 4 章) 的研究成果,李昕博士关于基于单个字符的半文本无关笔迹鉴别和新网格微结构文本无关笔迹鉴别的研究成果 (第 4、5、6、7 章),以及陈彦博士关于联机签名验证理论方法的研究成果 (第 8 章) 等。李昕博士对本书最后的成文承担了大量细致的工作。

本书是对文本图像风格的笔迹鉴别和认证这一重要问题的理论和方法的研讨和总结,由于在此领域的研究工作比较分散,且尚未见到有类似著作出版,本书提供了对于笔迹鉴别这一重要文档风格识别问题的较全面和深入的理论分析和研讨,提出了解决各种复杂情况下有效的笔迹鉴别算法和系统。本书提供给读者一个有益的参考,以期普及对于笔迹识别的认知和理解,进而推广其应用。本书也能对同行研究者、对文字识别、图像识别、模式识别、机器学习、智能信息处理相关领域的研究者和爱好者,具有一定的参考意义。

由于作者水平所限,书中难免存在纰漏及不当之处,敬请读者不吝指教!

丁晓青

2011 年 12 月于北京清华大学

目 录

第 1 章	引言	1
1.1	笔迹识别的背景与意义	1
1.1.1	生物特征识别之笔迹识别	1
1.1.2	计算机自动笔迹识别	3
1.1.3	笔迹验证和笔迹鉴别	5
1.2	笔迹识别方法文献综述	7
1.2.1	文本相关笔迹识别方法	7
1.2.2	文本无关笔迹识别方法	9
1.2.3	现有方法总结	12
1.3	关于本书	13
1.4	本书的内容和结构安排	15
第 2 章	单字笔迹风格鉴别特征分析	19
2.1	引论	19
2.2	主要文本相关方法比较	20
2.3	单字笔迹图像中的内容和风格	22
2.4	单字笔迹风格特征提取	22
2.5	利用线性鉴别分析提取风格	24
第 3 章	基于单个字符的计算机文本相关笔迹鉴别	27
3.1	单个字符笔迹原始特征的提取	27
3.1.1	文检专家和已有文献使用的特征	27
3.1.2	笔迹边缘特征的提取	30
3.1.3	对比实验及结果	37
3.2	基于单个特征字的笔迹鉴别算法	40
3.2.1	主分量分析	41
3.2.2	线性鉴别分析-最大鉴别分析	42
3.2.3	正则化线性鉴别分析	43
3.2.4	特征分布整形	44

3.2.5	单个特征字笔迹鉴别的分类器设计	46
3.3	基于单个字符的笔迹鉴别实验结果	47
3.3.1	总体流程	47
3.3.2	参数确定	48
3.3.3	方法比较	49
3.3.4	字符训练样本数的影响	52
3.3.5	大样本集性能	53
3.4	融合多个特征字进行笔迹鉴别	55
第 4 章	基于单个字符的计算机文本无关笔迹鉴别	63
4.1	引论	63
4.2	基于单字的文本无关汉字字体识别方法	63
4.2.1	印刷体字体识别介绍	63
4.2.2	笔画分布特征	67
4.2.3	小波特征	69
4.2.4	特征选择	80
4.2.5	分类器设计	82
4.3	基于单字笔迹的半文本无关笔迹验证方法	93
4.3.1	半文本无关	94
4.3.2	不同字符笔迹风格特征的传递性	95
4.3.3	半文本无关基本方法流程	97
4.3.4	基本方法的笔迹验证实验	100
4.3.5	对半文本无关方法的改进	103
4.3.6	改进方法的笔迹验证实验	110
4.3.7	半文本无关方法总结	113
第 5 章	篇章笔迹风格鉴别特征分析	115
5.1	引论	115
5.2	本书研究的几种文种笔迹	115
5.3	主要文本无关方法比较	117
5.4	篇章笔迹风格特征提取	120
第 6 章	基于篇章笔迹的文本无关笔迹鉴别方法	123
6.1	引论	123
6.2	网格微结构特征	124

6.2.1	笔迹样本预处理和图像边缘检测	124
6.2.2	特征提取方法及基本的网格微结构特征	125
6.2.3	基本版网格微结构特征示例	128
6.2.4	改进的网格微结构特征	132
6.2.5	网格微结构特征的特点和对笔迹风格的描述性能	135
6.3	分类器设计	139
6.3.1	简单距离度量方法	140
6.3.2	加权距离度量方法	141
6.3.3	分类器选择实验	142
6.4	笔迹鉴别实验	144
6.4.1	网格窗口大小	144
6.4.2	不同文种笔迹	145
6.4.3	与其他方法的比较	147
6.4.4	鉴别性能受文本篇幅及文本内容的影响	149
6.5	文本无关方法总结	151
第 7 章	实用的笔迹检索系统	153
7.1	引论	153
7.2	基于大规模数据库的笔迹检索系统	153
7.3	笔迹检索系统性能	157
第 8 章	联机手写签名的计算机自动认证	161
8.1	引论	161
8.1.1	手写签名认证研究的背景与意义	161
8.1.2	签名校验研究方法综述	164
8.2	联机手写签名的预处理及匹配对齐	169
8.2.1	签名的信号描述及预处理	169
8.2.2	签名序列的匹配对齐	172
8.3	基于空间信号描述的签名认证	178
8.3.1	统计特征的抽取	179
8.3.2	统计分类器的设计与结果分析	181
8.4	基于时间序列信号处理的联机签名认证	186
8.4.1	时间序列信号的差别度量	187
8.4.2	分类算法设计与结果分析	188
8.5	基于签名动态结构统计模型的签名认证	192

8.5.1 签名动态结构统计模型的建立 196

8.5.2 基于签名动态结构统计模型的联机签名认证 199

8.5.3 多种方差假设下模型参数的训练 201

8.5.4 测试结果与结论 205

8.6 基于 DSSM 模型与 DTW 匹配集成的联机签名认证..... 209

8.7 联机手写签名认证系统 214

8.7.1 完整的联机手写签名认证系统及应用 214

8.7.2 联机手写签名认证的其他问题 216

8.8 联机手写签名认证的总结与展望 218

参考文献 221

附录 A 本书所用脱机笔迹样本库..... 241

A.1 汉字篇章笔迹库 241

A.2 英文篇章笔迹库 242

A.3 藏文和维吾尔文篇章笔迹库 242

A.4 汉字单字笔迹库 243

A.5 印刷字体样本集 244

附录 B PCA 和 LDA 251

B.1 主分量分析 PCA 介绍 251

B.2 线性鉴别分析 LDA 介绍 253

B.3 Box-Cox 变换 255

附录 C 联机手写签名数据库..... 259

附录 D ICDAR 2011 笔迹鉴别竞赛 261

D.1 引言 261

D.2 参赛者和参赛方法 263

D.3 性能评价 264

D.4 评测结果 265

D.5 结论 271

致谢 271

参考文献 271

第1章 引言

1.1 笔迹识别的背景与意义

1.1.1 生物特征识别之笔迹识别

在当今的信息化社会里，如何确定一个人的身份，毫无疑问是一个非常重要的问题。身份识别就是通过各种技术或者非技术手段，证明当事人的身份，以向有关方获得行使当事人某种权利的许可。传统的身份认证方法主要通过身份标识物品（如钥匙、证件、银行卡等）和身份标识知识（如用户名和密码）来证明身份。这种方法简单，但有一个非常大的缺点，就是不大方便，如钥匙容易丢失，密码难于记忆。随着计算机技术的发展，各种利用人的生物特征（biometric feature）来进行身份识别的方法得到了长足的发展。

图 1.1 给出了几种常见的生物特征。生物特征是我们人体所固有的各种生理特征或者行为特征的总称。生理特征大多是先天性的，不随外在条件和主观意愿发生改变，比如指纹、掌纹、虹膜、人脸等；或特有的，如行为特



图 1.1 多种不同的生物特征

征（笔迹、声音、步态等）则是人们在长期生活中养成的行为习惯，很难改变。人的生物特征具有唯一性、长期不变性，以及不会被遗忘和丢失、不易伪造或被盗、随身“携带”以及随时随地都可使用等优点。所谓生物特征识别（biometrics）技术就是，通过计算机与光学、声学、生物传感器和生物统计学原理等高科技手段密切结合，利用人体固有的生理特性和行为特征来进行个人身份的鉴定。生物识别技术比传统的身份识别方法更安全、保密和方便。

笔迹识别（writer recognition）是生物特征识别的一个重要分支。中国古代就有“字如其人，人如其字”的说法，“签字画押”作为书写者的身份签证也由来已久。笔迹是人类运用肢体书写在一定物体上的一系列文字符号^[1]。对于每个书写者而言，其笔迹总体上具有相对稳定性，而笔迹的局部变化则是每个书写者笔迹的固有特性；而对于不同的书写者而言，其笔迹的差别则比较大^[2]。笔迹特征体现在多个方面。比如书写动作习惯特征，有笔画的运笔特征、笔顺特征、连笔方式等；再比如文字布局，包括字结构、行结构、段落结构、整体纹理等；当然还有语言特征，指用语习惯、词汇搭配等。在这里，语言特征不是我们考虑的重点。从生物特征识别的角度来看，笔迹是一种稳定的行为特征，笔迹的获取具有非侵犯性（或非触性），笔迹识别是易为人所接受、非常有应用前景的身份识别方式。

笔迹识别在社会生活中具有广泛的应用，如重要协议合同的签署，银行等金融部门相关单据的签名，公安司法部门的刑事调查，民事法庭文检证据的判定，等等。图 1.2 给出了一张带签名的银行单据和一份犯罪嫌疑人笔迹的示例。此外，随着计算机技术和网络技术的发展和普及，笔迹识别技术的

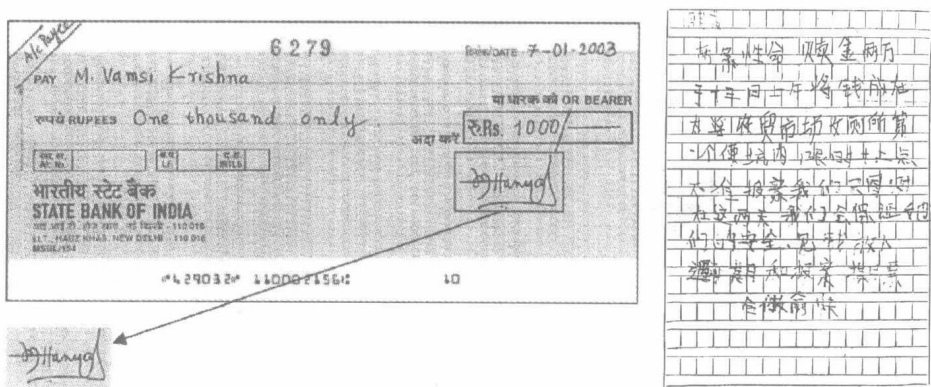


图 1.2 单据签名和勒索信笔迹示例

应用领域更为宽广，突破了原有的应用范畴，比如，计算机登录、信用卡签字、电子商务等。高准确率笔迹鉴别技术甚至可以使我们的生活质量得到提高。

笔迹是对人进行书写这一动态过程的记录，它的形成既受个体生物因素的影响，也受个体所处的社会环境、个性心理特征、情感过程的影响与制约。有些人书写的笔迹变化很大，即使连续的笔迹都有很大的差别。另外，笔迹可以被模仿。有些经过专门训练的人，模仿别人的笔迹足以达到以假乱真的地步。如果个人的签名或笔迹被别人伪造模仿，有可能引起严重的社会后果或巨大的经济损失。

现在，真正具有法律效力可在法庭上使用的笔迹鉴定是由笔迹专家来完成的。鉴定的准确性主要依靠笔迹专家的知识 and 经验判断来保证^[3]。人的书写习惯是在长期练习和书写实践活动中逐渐形成的，是由语言、文字作为主要刺激物在大脑皮层建立起来的一种巩固的动力定型。笔迹专家正是根据笔迹体现的习惯体系的异同来鉴定笔迹书写者的。书写动作的局部细节特征是书写动作习惯的直接反映。笔迹专家往往从样本中选择反映书写者书写技能和书写习惯的局部特征（例如运笔特征、笔画交叉与连接搭配特征、字的结构特征、笔顺特征以及特殊字等），再比较检材笔迹和样本笔迹两者之间的相同特征与不同特征，然后对相同特征与不同特征的价值进行科学分析来确定两者的符合点与差异点及其性质，最后根据差异点和符合点的评断做出检材笔迹与样品笔迹是同一人书写或不是同一人书写的确切结论。笔迹鉴别专业性强、鉴定程序复杂，笔迹专家的工作非常繁重，而且很难避免主观人为因素的影响。在十余年前香港发生的龚如心“媳翁”争产案中，为鉴定一份遗嘱的真实性，诉讼方委任的三位“国际级”文检专家和另一方聘请的国内三大权威笔迹专家得出的是完全相反的鉴定结论。如果用计算机来进行笔迹识别，将笔迹鉴定自动化，就会使鉴别结果更加客观。当然，用计算机来描述笔迹的特征、自动地进行笔迹识别还是个非常困难的问题。

1.1.2 计算机自动笔迹识别

计算机自动笔迹识别可分为联机笔迹识别（on-line）和脱机（off-line）笔迹识别两种^[4]。联机笔迹识别用手写板在线采集书写者书写时的各种信息，不仅包括笔迹图像，还包括书写时的速度、加速度、压力甚至握笔力度、笔杆倾斜角度等信息^[5,6]。联机笔迹识别虽然对输入方式有限制，但因为其获得的信息更多，识别性能也相对更高。脱机笔迹识别只能对通过扫描仪和数码

相机获得的笔迹文本的图像进行处理。脱机笔迹识别获取笔迹虽然自由，但因获得信息有限，识别在技术上更加困难。

人们书写字符的根本目的是传递信息。手写笔迹图像中包含两种信息，一种是语义信息（semantic information）另一种是奇异信息（singular information）^[4]。语义信息表示的是文本内容方面的信息，可称为文本内容信息；而奇异信息则反映书写风格的变化，表示书写者的身份，可称为笔迹风格信息。手写光学字符识别（optical character recognition, OCR）^[7]和笔迹识别是两种均以手写笔迹为处理对象的不同技术。前者的目的是提取文本内容信息，后者则为了提取笔迹风格信息。简单地说，前者是要知道手写文本中含有哪些字符，后者则是要判断谁手写了笔迹文本。在手写文本中，由字符决定的文本内容信息占主要地位，而笔迹风格信息则依附于文本内容信息上的微小变化，很难将两者截然分开。因此，字符识别必须尽量减小不同书写者书写风格的影响，往往通过学习不同人书写的同一字符来提取只与该字符文字类别有关的特征。而笔迹识别必须减小文本内容和字符结构变化的干扰，从同一人书写的不同或相同字符中提取反映书写风格的特征。

进行笔迹识别研究，首先要收集笔迹样本。Srihari^[8]结合英文笔迹的情况，对如何合理收集不同人群的笔迹样本作了研究。下一步，如何从样本中提取有效的笔迹风格特征并加以比较就成为整个研究的核心。Klement^[9]指出，一个好的特征必须满足三个条件：① 书写者相关性（writer specificity）——也就是说特征必须要充分反映每个书写者的书写特性；② 完整性（completeness）——特征必须能区分所有的书写者；③ 环境不变性（environment invariance）——特征不应该依赖于特定的外部条件，如书写的笔、纸、扫描分辨率等。实际上，要想找到完全满足上述三个条件的特征是很困难的。对笔迹风格特征的分析研究，将有助于提高从笔迹样本中所提取特征的有效性，提高笔迹识别的正确率和鲁棒性。这是进行笔迹识别算法研究的基础。

笔迹风格有全局和局部两个层次。图 1.3 是不同书写者的多行英文笔迹的图像示例，而 1.4 是不同书写者所书写的英文字母和单词的对照。可以看出，不同书写者的笔迹风格差异既表现在整段整篇的笔迹文本中，也同时体现在某一个字符或单词上。如何选择处理对象并把表现笔迹风格信息的特征从中提取出来，是笔迹识别的关键。按所处理的文本材料类型来分，笔迹识别的方法可以分为基于篇章笔迹的方法和基于单字笔迹的方法。篇章笔迹图像往往被作为一个整体，从中提取一致的风格特征。而单字笔迹含有的笔画少、结构信息少，往往要利用多个单字笔迹图像来提高笔迹识别的性能。

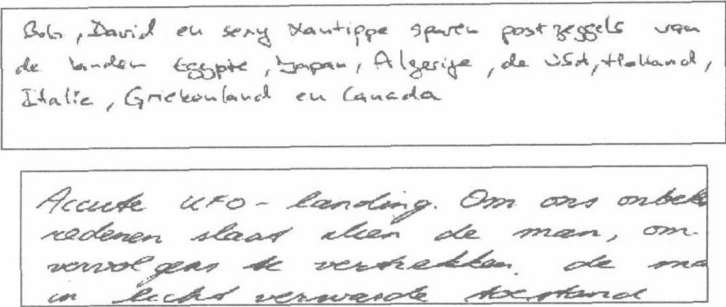


图 1.3 不同书写者书写的多行英文文本对照



图 1.4 不同书写者书写的英文字母和英文单词对照

1.1.3 笔迹验证和笔迹鉴别

笔迹识别中有两类问题，分别为笔迹验证（writer verification）和笔迹鉴别（writer identification）。笔迹验证，是通过比较两份笔迹样本，判断它们是同一人书写还是来自不同的书写者，这是一个一对一的两类分类问题；笔迹鉴别，又称为笔迹检索，是根据大量已知书写者的参考笔迹样本库，在这些书写者中找出查询笔迹样本最可能的一个或若干个候选书写者，这是一个一对多的多类分类问题。笔迹验证和笔迹鉴别有不同的性能评价指标。

笔迹验证的性能指标有错误接受率（false acceptance rate, FAR）、错误拒绝率（false rejection rate, FRR）和等错误率（equal error rate, EER）^[10]。笔迹

验证中，一般将两份笔迹样本间的风格相似度与一个决策阈值比较，当相似度大于该阈值时，确定两份笔迹样本是同一人书写，判定为接受；而当相似度小于该阈值时，认为两份笔迹样本来自不同的书写者，判定为拒绝。所谓错误接受率，是表示非同一书写者书写的笔迹样本对被错误判断为同一书写者书写的数目占有非同一书写者书写的样本对数目的百分比。所谓错误拒绝率，则表示确为同一书写者书写的笔迹样本对被错误判断为不是同一书写者书写的数目占有同一书写者书写的样本对总数目的百分比。FAR、FRR与决策阈值的关系如图 1.5 所示。可以看出，FAR 和 FRR 都是决策阈值的单调函数。而 FAR 和 FRR 关于决策阈值的曲线有一个交点，在该点处 FAR 等于 FRR。这个交点就是等错误率点，对应的错误率就称为等错误率。在评估验证系统的性能时，一般都采用 EER 作为性能指标。另外，表示真接受率（true acceptance rate, TAR）与误接受率 FAR 关系的 ROC（receiver operating characteristic）曲线^[11]也是一种评估标准^[10]。图 1.6 是一个验证系统 ROC 曲线的示意图，它同样可以表达 EER^[12]。

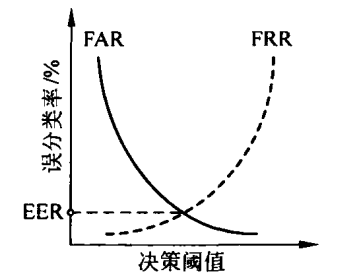


图 1.5 FAR、FRR、EER 与决策阈值的关系

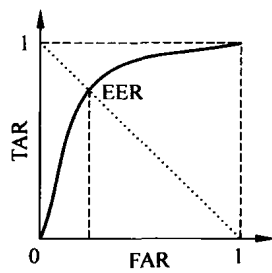


图 1.6 ROC 曲线

对于笔迹鉴别这个多类问题而言，通常就直接用鉴别正确率来评估其性能。鉴别正确率就是指其书写者身份被正确鉴别的笔迹样本数目占参加鉴别的总样本数目的百分比。在笔迹鉴别中，除给出最可能的首选候选书写者之外，往往会给出包含前若干个候选书写者的候选名单。这时，就需要使用前 n 选正确率。前 n 选正确率就是指其真实书写者出现在前 n 选名单中的笔迹样本数目占总样本数目的百分比。鉴别正确率实际就是首选正确率。

因为笔迹的类内变化很大、训练样本往往较少，直接进行多类分类判别比较困难。Cha 等^[13]提出了一种将多类的笔迹鉴别问题两类化的方法。它先提取笔迹的特征，然后计算同一书写者特征间的距离（within-writer distance）和不同书写者特征间的距离（between-writer distance）。对未知书写者的查询

样本和每一个已知书写者的参考样本，都计算彼此特征间的距离，通过判断该距离是同一书写者特征间的距离还是不同书写者特征间的距离，从而判断查询样本是否与该参考样本来自同一个书写者。当与所有参考笔迹进行比较后，进一步确定查询样本的书写者是否是其中一位已知书写者。因此，笔迹鉴别和笔迹验证的核心问题是一致的，即如何提取能表达书写者笔迹风格差异的特征并计算特征间的距离。

1.2 笔迹识别方法文献综述

最早关于计算机笔迹识别的研究出现在 20 世纪 60 年代^[4]。1967 年，Kozinets 等人发表了用计算机进行笔迹识别的研究报告^[14]。他们采用数字“2”和字母“K”作为特征字，从字符的骨架中提取若干个代表点，采用最小距离分类器，分别得到 73% 和 75% 的识别正确率。此后，对笔迹识别的研究逐渐增多。早期关于笔迹识别方面的研究可以参见两篇综述文章^[4, 15]。随着图像处理、模式识别和计算机视觉各领域的发展，笔迹识别技术也取得了一定的成果。根据对笔迹样本文本内容的要求，现有的笔迹识别方法主要分为文本相关（text-sensitive or text-dependent）与文本无关（text-insensitive or text-independent）两类^[4]。文本相关方法要求笔迹样本的文本内容必须一致，而文本无关方法则不考虑笔迹文本中含有什么样的内容。文本相关方法多是基于单字或单词的，并主要用于对正确率要求很高的笔迹验证问题。一般来说，文本无关方法多是基于整个篇章的，且主要用于文本内容难以限制的笔迹鉴别问题。下面简要介绍一下这两类方法中的一些主要方法。

1.2.1 文本相关笔迹识别方法

文本相关的方法是通过比较相同文本内容的笔迹，以去掉文本内容的影响，达到突出笔迹风格差异的目的。签名识别是文本相关笔迹识别的一种特殊情况。与文本无关的笔迹识别使用一定篇幅的笔迹文本不同，文本相关的笔迹识别通常采用相同的单个字符（称为特征字）作为训练和测试笔迹。提取的笔迹特征依赖于具体的特征字，也就是依赖于文本内容。

Zuo^[16]应用类似人脸识别中本征脸（eigenface）^[17, 18]，直接将归一化后的字符原始图像的各个像素灰度值排成的高维列向量，采用主分量分析（principal component analysis, PCA）^[19]降维，再通过计算欧氏距离进行分类。