文章编号:1000－582x(2000)06－0064－03

高校网上录取数据录入管理系统的开发

杨 波,叶俊勇,彭 健,汪同庆
(重庆大学 人工视觉研究室,重庆 400044)

391.4
G647.32 TP391

摘 要:高校网上录取数据录入管理系统是为适应网上录取工作而编制的,主要实现了数据采集以及电子档案的制作功能。系统在硬件上采用通用的扫描仪作为光电录入设备;在软件上实现了 OCR 和 OMR 共存情况下的通用识别,并且同时存储了图象数据。系统采用分布式处理,因此,系统与其他相似的系统相比,具有识别速度快、精度高、通用性好的特点。

高校 招生 网上录取数据录入管理系统
关键词:识别;分布式处理;OCR;OMR;检错技术
中图分类号:TP 391.4 **文献标识码**:A

网上录取就是由省、直辖市、自治区招生办公室将考生的纸介质档案转成电子档案(数据部分、图象部分等)后,利用 CERNET 教育科研网络,将考生的电子档案传送给相关招生院校,由招生院校进行审查录取[**]。

从恢复高考制度以来,我国二十二年的普通高校录取新生都是使用纸制档案,通过人工投放传递方式进行。教育部从 1998 年开始试行网上录取,经过两年的试验,有了相当大的进步。

然而,在网上录取的过程中,电子档案的制作是一项十分复杂而细致的工作,它是网上录取是否能取得成功的关键。传统的电子档案制作方式是采用扫描仪或数字相机等图象输入设备获取考生的图象信息,用手工方式输入考生的数据。由于考生信息多(包括报名信息、体检信息和志愿信息),采用传统的工作方式工作量大,而且数据的准确性难以得到保障。

由于模式识别技术,特别是文字识别技术(OCR)的日益成熟,并得到广泛的应用,将其引入到电子档案的制作是一种非常自然的想法。重庆市 2000 年普通高校招生就采用了这项技术来制作电子档案。

1 系统总体设计

系统的总体设计包括系统工作方式的选择,核心识别系统的选择。以下主要介绍核心识别系统的选择,至于工作方式的选择,笔者将放到系统物理设计中介绍。

与 OMR 技术相比,OCR 技术可以涵盖更多信息,特别是在某些数据量很大的场合,OCR 技术更显示出其突出的优势。但长期以来,OCR 的识别精度一直是困扰着广大 OCR 工作者的主要问题,并影响了 OCR 技术的推广应用。

通常,OCR 的识别精度(我们以下涉及的有关识别精度的指标均为系统的识别精度指标,而非单个识别器的指标)主要由三个指标控制,即正确识别率(Correct),误识别率(Error)和拒识率(Reject)。为表示方便,我们定义待识别字符总数(T_C)、正确识别字符总数(C_C)、拒识字符总数(R_C)、错误识别总数(E_C),我们很容易得到下面的关系:

$$T_C = C_C + R_C + E_C \tag{1}$$

各指标的定义如下:

正确识别率

$$C = C_C/T_C \times 100\% \tag{2}$$

误识别率

$$E = E_C/T_C \cdot 100\% \tag{3}$$

拒识率

$$R = R_C/T_C \times 100\% \tag{4}$$

显然,我们可以很容易得到

$$C + E + R = 100\% \tag{5}$$

根据戴汝为对系统可靠性[2](Reliability)的定义,

· 收稿日期:2000-04-19
作者简介:杨波(1973-),男,重庆人,重庆大学博士研究生。主要从事图象处理、文字识别系统领域研究与应用。
·· 重庆市大学中专招生委员会办公室.重庆市 2000 年普通高校招生考生手册.

得到：

$$Re = C/(1 - R) \cdot 100\% \quad (6)$$

由式(5)，很明显，在保证正确识别率不变的前提下，降低拒识率，误识别率则会相应提高；而要得到低的误识别率，那么拒识率会相应上升！显然，一个可靠性为 100% 的系统是无法实现的，那么如何选择一个合理的识别系统就需要根据应用系统的要求而定。

在这个系统中，考虑到 OMR 的识别精度通常要比 OCR 的识别精度好，于是在设计阶段就充分考虑了采用 OMR 技术发挥其识别精度好的特点。而在有些用 OMR 不合理的场合，就采用 OCR 技术。例如：在设计志愿卡的时候，考虑到考生的调配、定向等项目的取值范围只有是和否(即 0 和 1)的情况，就采有 OMR，以保证考生的信息能够准确进入计算机；而象院校代码、专业代码等取值范围大的项目，用 OMR 也能实现，但在表格中占用的区域太大，所以采用 OCR 的方式。同时，为保证 OCR 的识别精度，笔者采用了通讯技术中常用的检错技术[3]，即把一张表格当作一个数据包，把表格的每一行当作一个数据片(Data Pitch)，在每一行的末尾加入检错码以检查数据片的正确性。这样，在不降低拒识率的情况下降低了误识别率，系统的识别精度得到提高。

2　系统物理设计

为了得到一个好的系统，应当充分考虑系统的工作效率。

传统的串行工作方式虽然简单，但其工作效率却较低，由于这个系统需要在较短的时间内(五天)完成大量数据的录入(约 60 000 张表格，超过 6 000 000 个字符)，那么考虑分布式处理是很有必要的。

整个系统的物理结构见图 1。

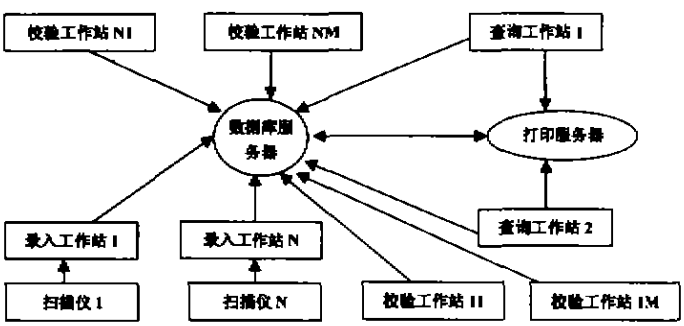

图 1　系统构成

由图 1 可以看到，把表格的录入与校验分开进行，采用一台录入工作站对应几台校验工作站的方式进行数据处理。

采用这种方式，数据的录入就只与扫描仪的扫描速度有关，而与识别系统的人工干预量无关了；即如果校验工作站的处理速度比录入工作站的处理速度慢，则只需要增加校验工作站的数目就可以了。在实际情况中，发现数据的校验工作要比录入工作复杂得多，这是因为录入工作站只是采集了图象数据，并做相应的变化，很少进行人工干预，而校验工作站进行了包括字符识别，人工排错等工作，有不少需要进行人工干预(特别是填写错误需要改正)，因此其处理速度相对较慢。为此，每台录入工作站录入的数据用两台校验工作站来校验，以分担校验的工作量。

从测试阶段及试运行阶段的工作来看，采用这样的工作方式是合理的，避免了数据阻塞现象的出现。

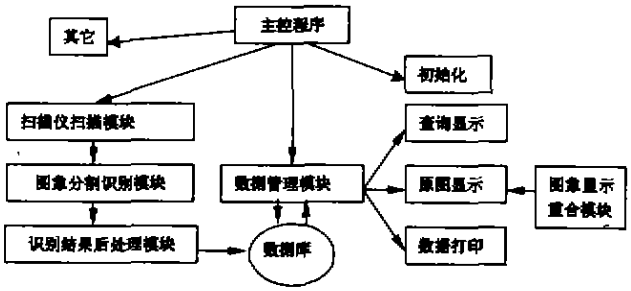

图 2　软件系统构成

3　软件系统设计

软件系统设计主要包括录入端软件系统和校验端软件系统的设计，软件系统的框图可由图 2 表示。

由软件系统构成框图可以看到，系统要实现的功能是比较复杂的，不仅有硬件的处理，图象处理，还有数据库的操作，它不仅是一个实用的录入系统，同时也是一个简单的 MIS 系统。

为了能够实现分布式处理，选择 MICROSOFT 公司的 SQL SERVER 数据库平台。同时选用 C++ BUILDER 作为开发工具完成大部分系统功能。

此外，需要特别指出的是，在数据校验的过程中，采用了拒识(人工干预)、检错码、逻辑关系检查等措施以降低系统的误识别率，并得到较好的效果。

4　系统性能评价

4.1　识别精度评价

在测试及试运行阶段，处理某学校的报名表(共

181 个考生,共 18 109 个字符),拒识字符总数 789 个,校验码发现误识别 7 个字符,经过逻辑检查未发现误识别字符,所有最终识别结果经人工核对无误识别字符。

根据前面对各项指标的定义,可以得到以下的结果:

识别器的拒识率

$$R = 789/18\ 109 \times 100\% = 4.36\%$$

识别器的正确识别率

$$C = (18\ 109 - 789 - 7)/18\ 109 \times 100\% = 95.60\%$$

识别器的误识别率

$$E = 7/18\ 109 \times 100\% = 0.04\%$$

识别器的可靠性

$$Re = 95.60\%/(1 - 4.36\%) = 99.96\%$$

系统的拒识率

$$R = (789 + 7)/18\ 109 \times 100\% = 4.40\%$$

系统的正确识别率

$$C = 95.60\%$$

系统的误识别率

$$E = 0$$

系统的可靠性

$$Re = 95.60\%/(1 - 4.40\%) = 100\%$$

上述识别器的指标是未经人工干预的结果,而系统的指标是经过人工干预后的结果。系统的误识别率为 0 和可靠性为 100%,是因为测试样本的数据量较小,不足以准确发现系统的误识别情况.

这表明系统的拒识率为 5% 以下,正确识别率为 95% 以上,无识别出错的字符。很明显,系统在拒识率为 5% 时,可以得到相当低的误识别率.从而为那些对识别结果要求非常准确的场合提供了一种可行的方法。

4.2 处理效率评价

由于采用了分布式处理,使系统的处理速度得到了很大的改善,特别是采用了多机校验的方式,使数据的处理速度仅与扫描仪的扫描速度有关! 在运行期间,采用两台校验机对应一台录入机的方式,基本上使校验的处理速度能够跟上扫描速度。而采用的扫描仪能够达到 15 张/分钟的处理速度,在实际运行中,用两台扫描仪仅用了 10 分钟就完成了所有的数据处理工作。因此系统的工作效率是相当高的。

5 结语

本系统的开发由于受到时间的限制,在一些方面还欠考虑,比如识别系统的服务器化、数字和汉字共存的表格识别处理,分布式处理的优化以及识别器的选择优化等,都有待进一步完善。

参考文献:

[1] 戴汝为. 汉字识别的系统与集成[M]. 杭州:浙江科学技术出版社,1998.

[2] 肖国镇. 编码理论[M]. 北京:国防工业出版社,1993

Development of Management System for Network Matriculation

YANG Bo, YE Jun-yong, PENG Jian, WANG Tong-qing

(Laboratory of Artificial Vision, Chongqing University, Chongqing 400044, China)

Abstract: The management system is designed for network matriculation, which achieves data collection and electro-files making. As for aspect of the hardware, a general scanner is applied. Meanwhile, as for aspect of the software, the general recognition is completed under the condition of OCR and OMR existed in the same system and the image is also saved. We applied distributed processing in the system, which possesses the characters of faster recognizing speed, higher accuracy and better generality compared to the similar system.

Key words: recognition; distributed processing; OCR; OMR; check-up errors

(责任编辑 张小强)