分类号：TP391　　单位代码：10422

密　级：　　学　　号：0543-0470318

山东大学

硕士学位论文

论文题目:基于手写体数字识别的信息录入与处理系统

AN INFORMATION INPUT AND PROCESSING SYSTEM BASED ON HANDWRITING NUMERAL RECOGNITION TECHNOLOGY

作 者 姓 名　吕　蓉

专　　业　计算机技术

指导教师姓名　朱大铭　教 授

专业技术职务　刘亚军　高级工程师

2007 年 4 月 5 日

原创性声明和关于论文使用授权的说明

原 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

论文作者签名：_____ 日 期：_____

关于学位论文使用授权的声明

本人完全了解山东大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权山东大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

(保密论文在解密后应遵守此规定)

论文作者签名：_____导师签名：_____日 期：_____

目　　录

TABLE OF CONTENTS

摘　　要

手写体数字识别是信息录入的关键步骤，广泛应用于公安、税务、交通、金融等行业的实践活动中。虽然识别方法多种多样，但是目前技术尚不能使识别率达到 100%。为了能够将手写体数字识别真正应用到实际工作中，本文除了在手写体识别算法的识别率提高上下功夫之外，还设计了一个基于手写体数字识别的信息录入与处理系统。

手写体数字识别的主要难点在于手写体数字字形小，特征信息量少，不同的人群书写习惯不同造成数字的形态千变万化；在某些应用中对于单字识别来说，手写体数字的正确识别要比其他字符严格得多。在对手写体数字识别技术做了充分比较后，本文选择了 BP 神经网络算法作为识别算法。BP 网络实质上实现了一个从输入到输出的映射，理论上它具有实现任何复杂的非线性映射的能力，适合于求解内部机制复杂的问题。

BP 神经网络的识别效果还依赖于训练神经网络样本集合的质量。为了获得识别率高、误识率低的分类神经网络，本文采集了不同人群中的 5 万余个有代表性的手写体数字图像作为训练、测试样本。有了这些样本以后，我们对样本图像作了二值化、去噪声、纠偏、细化、定位分割、尺寸归一化等一系列处理，经过这些处理后的单个字符图像被离散成神经网络的输入样本。将样本划分为训练集和测试集，对神经网络进行训练，所得到的神经网络能够达到较满意识别效果。其中，我在处理数字样本的时候发现，某些样本是对神经网络有害的‘坏’样本，在研究中我尝试将这些坏样本‘剔除’以后发现神经网络的识别率和拒识率有所提高、误识率明显下降。

将手写体数字识别技术投入实际应用中是我们的最终目标，为此本文设计了一个基于手写体数字识别的信息录入与处理系统，此系统可以混合处理照片、文字（仅保存图像）、OMR 以及手写体数字，本文仅重点研究及介绍其中的手写体数字识别模块。系统使用扫描仪将信息卡内容扫描到计算机中以后，程序对手写体数字图像进行二值化、去噪声等预处理，然后形成神经网络分类模型的输入，并由神经网络进行识别。识别后的结果会显示在计算机终端上，并由操作人员对拒识字符进行处理。考虑到识别算法可能有误识别的情况，本文设计了一种批量校

对的方法来处理可能发生的误识。

经过测试，BP 神经网络手写体数字识别算法识别率可达到 96.8%以上、拒识率小于 2.7%、误识率小于 0.5%；在实际系统应用中，批量手写体数字校对方式可以成倍提高误识校正的速度,熟练操作者的处理速度可以达到 100 字符/秒甚至更高，经过人工干预后，最终误识率小于万分之一。

本文主要贡献如下：

(1)实现了 BP 神经网络的训练和识别算法，并在实际工作中得到应用。

(2)发现了‘坏’样本对 BP 网络识别效果的影响，给出了‘坏’样本的判定方法。

(3)设计实现了识别结果的批量校对软件，提高了误识字符校对的效率。

在以后的工作中，我还会在快速神经网络算法、神经网络集成和图像预处理等技术上多下功夫，争取为神经网络和手写体数字识别的应用和推广做出贡献。

关键词：手写体数字识别；BP 神经网络；去噪声；批量校对；样本筛选

ABSTRACT

Handwriting numeral recognition is a key step in information input. It is widely used in various domains, such as public security, revenue, traffic, finance. Recently none of the handwriting numeral recognition technology could reach 100% recognition rate though there were various technologies. For the purpose of using this technology in our daily work, in this paper, we not only do a lot of work to improve the recognition rate, but also design an information input and processing system based on handwriting numeral recognition technology.

The difficulties in handwriting numeral recognition research are: 1) handwriting numeral figure is very small so that it contains little information, 2) different people has different habit in writing so that the numeral figure is daedal. What's more, for some application, the request of exact handwriting numeral recognize is even stricter. In this paper, BP network arithmetic was selected as handwriting numeral recognition arithmetic after fully comparison in various recognition technologies. In fact, BP network maps input to output. In theory, it has the ability to implement any nonlinear mapping no matter how complex it is. Further more, it is good at handle problem which was very complex on its inner mechanism.

The effect of handwriting numeral recognition is also depend on the quality of pattern set. In this paper, to get a high recognition rate, low mis-recognition rate classify neural network, about 50000 representative handwriting numeral image was collected from different people. After the collection, a list of pretreatment operation such as two-valued, removes noise, thinning, locate and cut, size normalization, was performed to build a input pattern for neural network. All numeral patterns are split as train

set and test set, then train the neural network repeatedly, finally, we get a good classify neural network. It's discovered in this paper when training that some patterns do 'harm' to neural network. We try to take away the 'bad' patterns and train and test again, then the identify rate and reject rate increase, while misidentify rate decrease.

It's our final goal that to use handwriting numeral recognition technology in actual application. To reach the goal we designed a information input and processing system based on handwriting numeral recognition technology. After the content of information card was scanned and saved in computer system using scanner, program The result of recognition would display on the screen of computer, and the numeral character which couldn't be recognized would proofread by operators. Because of the existence of misidentify, a batch proofread method was designed to deal with those characters which were misidentified in this paper. The result of the test which using BP neural network as handwriting numeral recognition arithmetic indicate that the identify rate is over 96.8%, the reject rate is below 2.7%, and the misidentify rate is below 0.5%. In the actual application, the batch proofread method could multiple the rates at handwriting numeral misidentify proofread. A skilled operator could process 100 characters one second or even more.

The test in this paper indicates that: the recognition rate of BP neural network classifier is at less 96.8 percent, refuse rate is less than 2.7 percent; mis-recognition rate is less than 0.5 percent. When work in practical application, batch proofread could speed up handwriting numeral processing, and reduce the harm caused by misidentify efficiently. A skill operator could check more than 100 characters per second, the misidentify rate will less than 1/10000 after manual intervention.

In this paper, first, an algorithm for numeral recognition and training

using BP neural network is implemented and used in practical application; second, it's discovered that 'bad' patterns affect the effect of handwriting numeral recognition, and a method to find out the 'bad' patterns given by author; third, a recognize result batch proofread software is designed, which increase the efficiency of proofread.

In the future, I will make great efforts on quick neural network arithmetic and image pretreatment. I will do my possible on neural network and handwriting numeral recognition's application and promotion.

Keywords: handwriting numeral recognition; BP neural network; remove noise; batch proofread; pattern screening

第 1 章 绪　论

1.1 研究背景

在教育考试领域，每年都有各种各样的招生、报名、考试阅卷等工作，这些工作信息量大、种类繁多、时效性强，同时又要求必须有近乎100%的准确度，人工进行这项工作不仅工作量大，耗时耗力，可靠性和安全性也难于保障。因此对可靠的字符识别技术的需求已迫在眉睫，因为它是机器智能化的瓶颈。手写体数字识别系统即为适应此要求而研制开发的,它具有广泛的社会应用前景。

进入90年代后，国内各地陆续开始使用OMR（光学标记阅读）产品进行大规模的数据录入，收到了良好的效果。OMR技术是对信息卡中的有效信息点（涂点）的灰度值进行筛选和识别，并对其进行数据形式转换，从而得到识别结果字符串。

随着OMR产品的普及，OMR产品的弊端也日益暴露出来。首先，OMR信息卡的包含的信息量较少，信息密度低。其次，填涂OMR信息卡的效率也较低且容易出错，无法在全社会范围内普及应用。

面对 OMR 产品的诸多弊端，我们迫切需要一种更加易于用户使用、效率更高的信息处理系统，由此我们开发了这套自动识别自由手写体数字的信息录入与处理系统。本系统针对于招生报名、考试阅卷、网上录取等领域的实际需要进行研发，目的是实现信息快速采集。它可对已录入的各类信息卡上的手写数字（邮编、汉字区位码、电话号码等）、照片、条码、OMR（光学标记识别）标记等多种信息进行分类识别、存储、校对、检索、定制输出等等。系统完成后，不仅可以应用于上述领域，也可用于统计、交通、档案管理、金融、邮政、图书馆等行业，完成各种信息表格、报表的图像数据采集与识别及数据录入工作，可见成熟的手写体数字识别技术将会大大地加快现代的信息化工程进展，具有广阔的应用前景。

1.2 国内外研究现状

本课题的研究重点和难点在脱机手写体数字识别方法上，手写体数字识别在学科上属于模式识别和人工智能的范畴。在过去的数十年中，研究者们提出了许

许多多地识别方法，按使用特征的不同，目前手写体数字识别方法的可以分为两类：基于结构特征的方法和基于统计特征的方法[1]。统计特征通常包括密度的测量、矩、特征区域等等；结构特征通常包括圈、半圈、交叉点、端点、节点、弧、突起、凹陷、笔画以及横纵两方向上的交叉次数等等。一般来说，两类特征各有优势。例如使用统计特征的分类器易于训练，而且对于使用统计特征的分类器，在给定的训练集上能得到相对较高的识别率；而结构特征的主要优点之一是能描述字符的结构，在识别过程中能有效地结合几何和结构的知识，因此能得到可靠性较高的识别结果。

多年的研究实践表明，对于完全没有限制的手写数字，由于不同手写习惯的数字体千差万别、风格迥异，且有的书写很不规范，目前还没有一种简单的方案能达到理想的识别率和识别精度。因此，我们所做的工作是努力使这项技术向着更为成熟、复杂、综合的方向发展。另一方面，研究工作者正努力把新的知识运用到预处理，特征提取，分类等技术当中，如：神经网络、数学形态学等。在手写数字识别的研究中，神经网络技术和多种方法的综合是值得重视的方向。

1.3 本文的工作

为了获得一个有效（识别率高、误识率低）的神经网络分类模型，本文采集了不同人群中的大量有代表性的手写体数字图像，设计并使用了带有一定格式的采样卡以便加快手写体数字字符的处理效率。有了这些样本以后，我们对这些图像作了二值化、去噪声、纠偏、细化、定位分割、尺寸归一化等一系列处理，经过这些处理后的单个字符图像被离散成神经网络的输入样本。我们将这些样本划分为训练集和测试集，对神经网络进行了充分的训练，取得了比较满意的神经网络分类模型。其中，我在处理数字样本的时候发现，某些样本是对神经网络有害的‘坏’样本，在研究中我尝试将这些坏样本‘剔除’以后发现神经网络的识别率和拒识率有所提高、误识率明显下降。

如何获得一个好的神经网络分类模型，并通过细化、去噪声等一些技术的应用得到更高的识别率，将手写体数字识别技术投入到实际应用中是我们的最终目标，为此本文设计了一个基于手写体数字识别的信息录入与处理系统，此系统可

以混合处理照片、文字（仅保存图像）、OMR 以及手写体数字，本文仅重点研究及介绍其中的手写体数字识别模块。要求系统能够进行大规模、高速度、高效率的数据表格图像信息处理，实现各种信息表格、报表的数据录入工作，在实际应用中对信息卡中手写体数字识别结果的人工干预量小于等于百分之三（按字符计算）；最终识别误差小于等于万分之一。