中 文 信 息 学 报
第 13 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 3

基于统计与神经元方法相结合的手写体相似字识别*

张德喜　马少平*　朱绍文　金奕江*

华中师范大学电子与计算机研究所　武汉　430079

*清华大学计算机科学与技术系　智能技术与系统国家重点实验室　北京　100084

摘要　本文提出了一种基于统计识别方法与人工神经元网络相结合的手写体相似汉字识别方法。该方法充分利用了统计识别方法和神经元网络识别方法的优点,不仅显著地提高了相似字的识别率,而且有效地提高了系统的整体性能。对相似字的识别率由 79.02 %提高到 84.32 %,提高了五个百分点,整体识别率提高了 1.3 个百分点。

关键词　神经元网络　汉字识别　相似字识别

一、问题的提出

近年来,汉字识别技术有了很大的发展,印刷体汉字识别、联机手写体汉字识别技术基本上已经成熟,在汉字输入中起着越来越重要的作用。脱机手写体汉字识别,虽然在方法上不断创新,识别率逐年提高,但由于其复杂性,距离人们普遍能接收的实用要求,还有一定的差距。其主要问题在于,每个人书写的汉字都有其各自的特点,差异性比较大,当识别率不高时,人们的研究重点是如何将这些由不同人书写的、差异性比较大的相同汉字统一起来,找出它们之间共同的部分,尽可能地消除这些差异,提高系统的总体识别率。如文献[1~3]采用整形变换和模糊特征的方法,较好地解决了手写体汉字中的各种变形问题,使得识别率有了较大的提高。这种方法的特点是通过非线性变换和特征的模糊化,使得系统对汉字小范围的变换不敏感,从而达到提高识别率的目的。其副作用是使得相似字更加难于区分了,虽然系统的整体识别率有了很大的提高,但相似字的识别率却反而降低了。这是因为相似字间的差别是很小的,系统在努力消除同一汉字不同写法间的差异性的同时,将相似字间的那些本来就很小的差别也消除得所剩无几了。人工神经元网络方法在小规模分类中显示出很强的分类能力[5,6],在手写体阿拉伯数字识别中也取得了较好的结果[7,8]。经小规模的试验,神经元网络方法对手写体相似字也具有比较满意的区分能力。但目前神经元网络方法存在的一个重要的问题是分类规

* 本文由国家自然科学基金和国家 863 高科技项目资助
本文于 1999 年 1 月 7 日收到

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

模问题,对于像数字识别这样的小规模分类问题,它显示出比较好的分类性能,随着规模的逐步加大,其分类能力则急剧下降,以至于不能使用。如何利用统计识别方法的大规模分类能力,缩小识别范围,在小范围内利用人工神经元网络对相似字进行识别,是提高手写体汉字识别水平的一个重要研究方向。

本文提出一种将传统的统计识别方法与人工神经元网络方法相结合的混合手写体汉字识别方法,充分发挥统计识别方法整体识别率高和人工神经元网络方法对相似字区分能力强的特点,使得相似字的识别率由原来的 79.02 %提高到 84.32 %,提高了约 5.3 %。

二、相似字与相似字集

何为相似字?并没有一个统一的定义,一般来说,相似字是指字形相近的汉字。但对于汉字识别来说,这样的定义并不全面。因为识别器是在特征级进行识别的,特征可以认为是对字形的采样,因此,在人看来,字形不太相似的字,对识别器来说也可能是相似的。所以是否是相似字与特征的选择、匹配方法等有很大的关联。为此,我们将相似字定义为客观相似字和主观相似字两种。客观相似字是指在字形上相近的汉字,而主观相似字则是指在人看来字形上不太相近的字,但由于特征的选取和识别方法的关系而容易混淆的字。主观相似字与识别器的关联比较大,对于某个识别器来说是相似的汉字,对于另外的识别器则可能不是相似的。而客观相似字则与识别器几乎没有关系,对任何识别器来说都是相似的。因此,对于汉字识别问题来说,两个汉字之间是否相似,主要取决于所选定的识别器。

定义 1:设 R 是给定的识别器,对于给定的汉字 y,其相似字集 S(y)定义为:

$$S(y) = \{x \mid P(I(R) = x \mid O(R) = y) > P_0\} \quad (1)$$

其中,I(R)表示识别器 R 的输入,O(R)表示识别器 R 的输出,P(·)表示概率,P_0 为给定的常数。当 S(y)只含有 y 一个元素时,我们称 y 不存在相似字集。

根据定义以及实验分析,相似字集具有以下性质:

1. 依赖性。在给定的 P_0 下,相似字集的数量与大小完全依赖于识别器的性能,识别率越高的识别器,其相似字集的数量越小,相似字集的平均元素数也越少;相反,识别率越低的识别器,其相似字集的数量越大,相似字集的平均元素数也越多。

2. 非对称性。设 x、y($x \neq y$)是两个不同的汉字,当 $x \in S(y)$ 时,$y \in S(x)$ 不一定成立,反之亦然。也就是说,当 x 是 y 的相似字时,y 不一定是 x 的相似字。

但这与人对相似字的认识完全不同,对人来说,如果 x 是 y 的相似字,则 y 也一定是 x 的相似字,即相似字间具有对称性。但对于识别器来说并不具有这样的性质。为什么会出现这种情况呢?设 A、B 是两个不同的汉字,其空间分布如图 1 所示,其中"+"表示 A 的分布情况,"O"表示 B 的分布情况。由于两类汉字在特征空间分布的不均匀,A 分布的比较稠密,而 B 分布的比较松散,当采用欧氏距离进行分类时,一部分 B 的样本会被识别为 A,而 A 的样本则不会识别为 B。根据相似字集的定义,当 B 的样本被识别为 A 的概率大于给定值时,B 就成为了 A 的相似字,但 A 不会成为 B 的相似字。

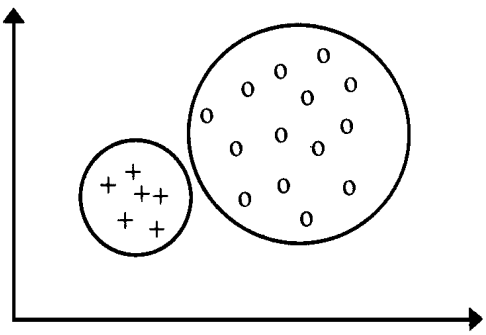


图 1 汉字的空间分布示意

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

即使不使用欧氏距离,而采用其他的策略进行分类,由于汉字在其特征空间的分布并不严格地遵循固定的分布,因此这种情况总是要出现的,这就导致了相似字集的非对称性。

图 2 给出了一个实际的例子。图中所表示的是某识别器对于汉字"廓"和"廊"这两类字的所有样本,到这两个汉字的标准样本间的距离。其中,横坐标是到"廓"的距离,而纵坐标表示的是到"廊"的距离。"廓"的样本用"●"表示,而"廊"的样本用"▲"表示。图中 45°实线表示的是两类字的分界线,位于该线左上方字被识别为"廓",而位于该线右下方的字被识别为"廊"。在图中可以很清楚地看出,一部分"廊"会被识别为"廓",但"廓"不会被识别为"廊"。也就是说,"廊"是"廓"的相似字,但"廓"不是"廊"的相似字。

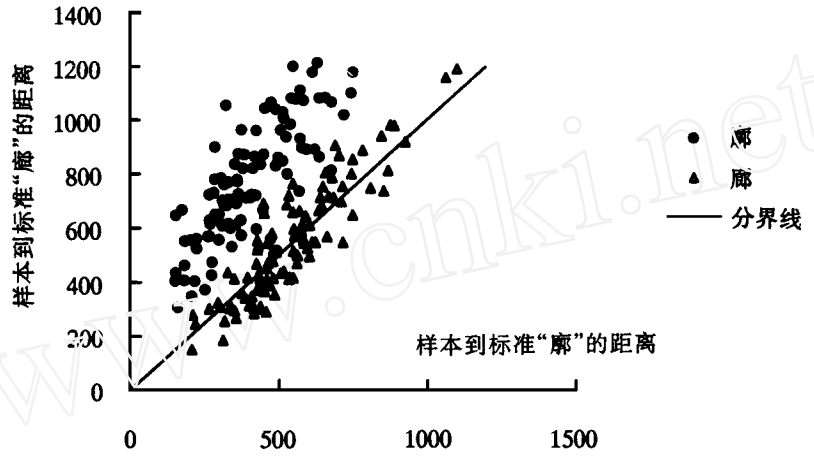


图 2　相似字非对称性的一个实例

3. 可采纳性。由于建立相似字集的目的是为了对相似的字用神经元网络进行区分,如果相似字集的规模比较大的话,神经元网络的识别能力将急剧下降,达不到对相似字高精度识别的目的。经大量实验表明,相似字集的规模一般比较小,大多数的相似字集所含有的元素数为 2～3 个,最多的为 6 个,完全适合于用神经元网络进行分类。因此这样的相似字集是可采纳的。

4. 可操作性。由于相似字集是基于识别器的输出定义的,因此很容易根据识别器的输出结果确定待识汉字所属的相似字集,具有很好的可操作性。

当给定了识别器 R 及具有一定规模的测试集后,由相似字集的定义,很容易求出一个汉字的相似字集。

如何评价一个相似字集的相似性呢?我们给出以下相似度的定义。

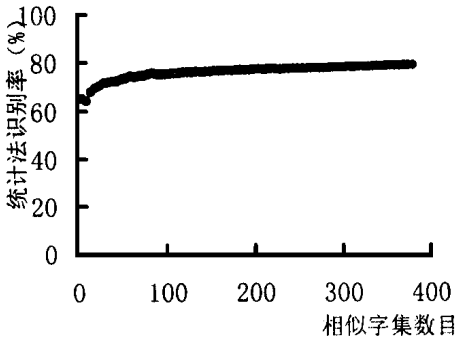


图 3　相似度与网络识别率的关系

定义 2:一个相似字集 S(y) 的相似度 γ(y) 定义为:

$$\gamma(y) = \min\left\{1, \frac{\max\limits_{x \in S(y), x \neq y} (p(I(R) = x \mid O(R) = y))}{p(I(R) = y \mid O(R) = y)}\right\} \tag{2}$$

其中,各符号的含义见定义 1。

由定义可以看出,γ(y) ∈[0,1],γ(y)越接近于 1,表明 S(y)的相似性程度越大,反之,越接近于 0,则相似性越小。图 3 给出了相似度与识别率的关系,图中横坐标为按相似度从大到小排序的相似字集,纵坐标为相应的识别率。由图中可以看出,随着相似度的降低,识别率基

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

本上是逐渐上升的,可见相似度的定义是合理的。

相似度最大的 20 个相似集为:

1 竞竟
2 菜莱荬某粟
3 廓廊
4 鸣呜
5 曰日
6 日曰
7 睛晴暗
8 间问同
9 竟竞
10 玉王壬
11 好奸抒扦
12 乌乌
13 究宪完
14 诚诫城减
15 人入
16 采来呆果栗苯
17 晴睛暗
18 呜鸣坞
19 凤凤凡讽
20 骋聘

这与我们人的认识,基本上也是一致的。

三、相似字的识别

神经元网络在小规模识别问题中显示出很好的性能,这一点已经得到人们的广泛认可,经实验,即使是对于相似字,当网络规模不大时,也能有很好的区分能力。而相似字集的规模大部分在 2～6 个字之间,很适合于用神经元网络求解。为此,我们将每个相似字集建立一个神经元网络,然后组成一个神经元网络群,具体的实现方法见第 4 节。

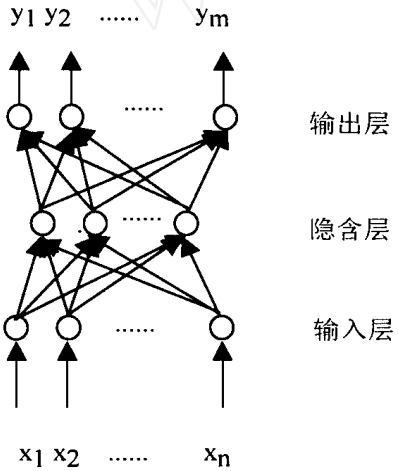


图 4 神经元网络示意图

图 4 给出了神经元网络的结构示意图。

神经元网络采用一个输入层、一个输出层和一个隐含层的三层结构,其中输入为汉字的模糊方向线素特征,输入层节点个数 256 个,输出为相应的相似字集的元素,根据相似字集的不同,输出层的节点个数为 2～6 个不等,隐含层的个数根据输出层节点个数的不同而异,由实验决定,一般选取在 15～20 之间。神经元网络的激励函数选择为 S 函数,即:

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (3)$$

神经元网络的训练采用标准的 BP 算法。

四、系统构成

系统的总体构成框图如图 5 所示。

其中,识别器是一个基于统计识别方法构成的识别器,它采用模糊特征方法[1][2]对待识汉字进行分类,并得到含有 c 个汉字的候选集。分配器根据识别器的误识模型[4]对识别结果进行评价,对于可信度大于给定值的识别结果直接送出,对于其他结果,则判断其可能属于的

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

相似集,根据相似集选定 1 到 m(1 < m < n, n 为神经元网络群中神经元网络的个数)个神经元网络,被选定的神经元网络对待识字进行相似字识别。综合器对识别结果进行模糊综合评判[2],给出可能性最大的作为最终的识别结果输出。

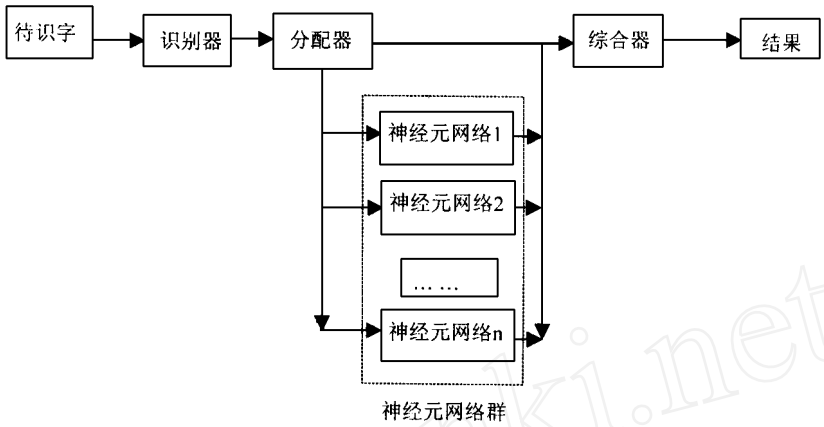


图 5 系统构成框图

五、实验结果

为了验证本文所述方法的有效性,我们用实际收集的手写体汉字样本对系统进行了测试。

测试环境:中文 Windows 95,Pentium/ 120 计算机

测试对象:GB2312 - 80 一级汉字

训练样本:90 ×3755 = 337950

测试样本:20 ×3755 = 75100

测试结果如下:

根据相似字集的定义,我们求出了 384 个相似字集,并分别建立了神经元网络,各神经元网络训练集和测试集的识别率如图 6 所示,其中纵坐标为累积识别率,横坐标为按相似度排列的相似字集。从图中可以看出,相似度大的相似字集其识别率仍然较低,这是因为很多相似字在手写情况下确实很难区分,如"土"和"士"、"未"和"末"等,如果没有上下文的帮助,人也很难确认。但是从第 80 个相似字集开始,累积识别率基本稳定在 90 %以上,效果还是比较理想的,说明了神经元网络确实具有较强的分类能力。测试集与训练集的识别率相差还比较大,相似度越大,这方面的差别越明显。这一方面说明对于相似字来说,神经元网络的泛化能力还不够,同时也说明训练样本的数量还不足,如果增加训练样本的数量,二者的差别会进一步缩小。

图 7 给出的结果是对于相似字而言,系统集成在一起后的累积识别率和单独使用统计方法时的累积识别率的对比图,图 8 给出的是相应的二者识别率间的差别图。从图中可以看出,采用本文的方法后,相似字的累积识别率平均提高了 5 个百分点以上。但该图给出的识别率比图 6 所示的识别率要低,这是因为图 6 给出的是网络单独运行时的识别率,不存在选网的问题,而图 7 给出的是系统集成后实际运行的结果,存在选网问题,即根据统计识别方法的识别结果,选择一个网络运行,但这种选择不一定完全正确,可能出现错误,从而导致误识。图 9 给出了网络的正确选择率,相似度越大的网络,正确选择率越高,随着相似度的降低,网络正确选择率逐渐减少,最终基本稳定在 96 %左右。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

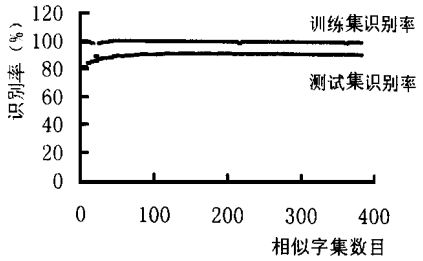


图 6　神经网络训练集和测试集的识别率

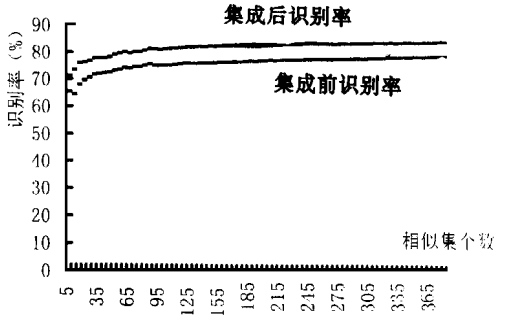


图 7　统计法与系统集成法识别率的比较

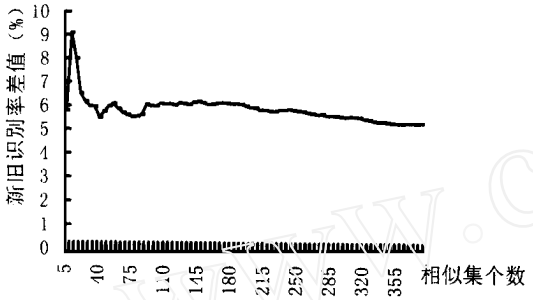


图 8　统计法与系统集成法识别率的差值

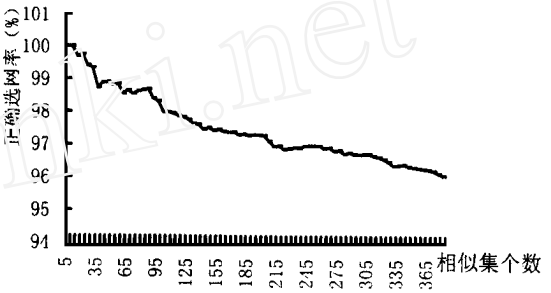


图 9　相似度与正确选网率的关系

以上结果中的识别率都是针对相似字集中的汉字来说的,如果将考察的汉字集扩大到整个字符集,即 3755 个汉字后,结果会如何呢？图 10 给出了这一结果。图中横坐标同样是相似字集的个数,但纵坐标为集成后的系统对全体字符集的识别率的增加数。随着相似字集网络的增加,系统的识别率基本上呈直线上升,当相似字集网络增加到 384 个时,系统的识别率增加 1.3 个百分点左右。这一结果是比较令人满意的。

为了评价一个系统集成了相似字识别后所产生的效果,我们提出了相对校正率这一概念。相对校正率定义如下:

$$\text{相对校正率} = \frac{\text{集成后识别率} - \text{原系统识别率}}{\text{原系统误识率}} \times 100\ \%$$

图 11 给出了本集成系统的相对校正率曲线,基本是一个水平的直线,平均值约为 25 %左右。其物理含义为,原识别系统误识的汉字中,约有 25 %的汉字可得到校正,被正确识别。

六、结束语

本文将统计识别方法与神经元网络识别方法结合起来,利用前者解决大规模分类问题,用后者解决相似字的识别问题,充分发挥了各自的特点,较为成功地解决了两种不同识别方法的集成问题。集成后的系统对相似字的识别率提高了约 5 个百分点,对全体字符集的识别率提高了约 1.3 个百分点,集成后系统的校正率达到了 25 %的较为满意的水平。在相似字集达到一定数量后,如何提高系统的网络正确选择率是今后进一步研究的问题。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

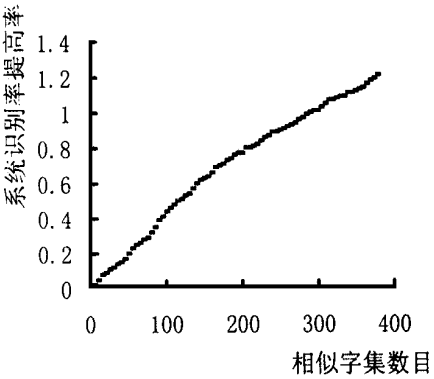


图 10　集成系统识别率与相似字集的关系

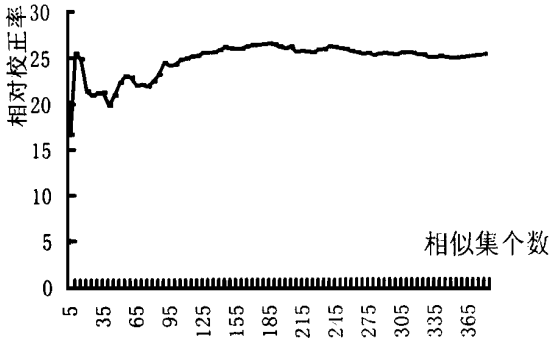


图 11　集成系统的相对校正率

参 考 文 献

[1] 马少平,夏莹,朱小燕. 基于模糊方向线素特征的手写体汉字识别. 清华大学学报,1997,37(3):42 - 45

[2] 马少平. 脱机手写汉字识别研究[博士学位论文]. 北京:清华大学,1995

[3] 马少平,夏莹,朱小燕. 基于非线性规格化的手写体汉字识别. 软件学报,1996,863 专刊:200 - 205

[4] 马少平,夏莹,朱小燕,姜哲. 汉字识别系统的误识模型. 清华大学学报,1998,38(S1):108 - 111

[5] 金连文,徐秉铮. 基于多级神经网络结构的手写体汉字识别. 通信学报,1997,18(5):21 - 27

[6] 王国胤,施鸿宝. 汉字识别的并行神经网络方法. 模式识别与人工智能,1996,9(1):96 - 101

[7] 洪沁,何振亚. 手写体数字的神经网络识别方法. 模式识别与人工智能,1994,7(1):66 - 71

[8] 谢光毅,钟义信. 神经网络用于手写体数字识别. 模式识别与人工智能,1994,7(4):334 - 337

Handwritten Similar Chinese Characters Recognition Based on Combining Statistics with Neural Networks Method

Zhang Dexi　Ma Shaoping *　Zhu Shaowen　Jin Yijiang *

Institute of Electronics and Computer　Central China Normal University　Wuhan 430079

*Department of Computer Science and Technology　Tsinghua University

State Key Laboratory of Intelligent Technology and System　Beijing　100084

Abstract　This paper presents a method to recognize handwritten similar Chinese characters based on combining statistics model recognition method with artificial neural networks (ANN). This method takes advantage of the peculiarities of above two methods efficiently. It not only increases the recognition rate to similar Chinese characters, but also improves the performance of the system. The recognition rate to similar Chinese characters can be improved from 79.02 % to 84.32 %, about 5 percentage points improving; the recognition rate to system can be improved about 1.3 percentage points.

Keywords　neural networks　Chinese characters recognition　similar Chinese characters recognition

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net