脱机手写字符识别的 DP 算法的设计*

刘 璟 白 刚

(南开大学计算机与系统科学系,天津 300071)

摘 要 把手写体汉字识别视为一种允许空间偏移的弹性模式识别匹配问题,提出了一种用于这类二值图象的两层动态规划算法,初步实验表明该算法具有十分满意的识别效果。

关键词 近似串匹配,脱机手写体汉字识别,空间偏移匹配,动态规划

1 引 言

1.1 脱机手写体字符识别

手写体字符识别是一个难度很大的研究课题,由于手写文本、信封、票据表格和签名等的计算机自动阅读具有十分诱人的应用背景,因此吸引了许多研究者的关注。目前在国内外,多字体印刷体字符的识别系统已达到相当高的识别率,不少 OCR 系统已成为成熟的软件产品。然而,手写体字符的识别,特别是脱机手写体字符阅读系统的开发却遇到了很大的困难。多数手写字符识别算法属于启发式的算法,根据字符的结构特征进行分类和识别。但由于手写体字符的随意性,很难找到共同的规律,因此对手写字符限制过严或识别率低。本文介绍的方法,针对手写体汉字识别问题,把每个汉字视为二维图象,不考虑各个不同汉字的结构特征,把同一汉字的不同写法视作某种"弹性"形变的结果,因此把它称为空间偏移匹配(Space-warp Match)或弹性匹配算法。

1.2 近似串匹配

近似串匹配(Approximate String Match)又称子串搜索。允许至多有 K 个编辑错误的子串搜索以及含通配符 * 的串匹配算法的研究近几年十分活跃,其中许多有效的 ASM 算法是动态规划(Dynamic Programming)算法[1~3]。

一般的近似串匹配问题是:在长度为 m 的文本 Text[0…$m-1$]中搜索长度为 n($n<m$)的样本串 *Pattern*[0…$n-1$]的近似出现。所谓近似出现,即指至多有 k 个编辑错误的匹配出现。问题的核心是要计算文本 Text 的所有子串与样本 Pattern 的编辑距离。

两个字符串之间距离的计算不是一个简单问题。因为不同的编辑错误组合可能造成相同的结果,串 a 与 b 的编辑距离被定义为把 a 错印成 b 的所有的错误序列中最小错误数。因此,这种距离的计算实际上是一个组合优化问题,近似串匹配算法的设计,主要是要找到一个有效的距离计算方法。其中动态规划方法的引用十分成功。

1.3 二维图象的弹性匹配

把输入字形与模板字形的比较归结为二维图象的空间偏移匹配,非常适合于手写汉字的本质特征。手写汉字与图象中的零件、地形的轮廓识别不同,它允许弹性形变。换句话说,两个手写汉字图象的相似程度(或距离),应在对两者做了各种不同的弹性形变的条件下,以其中最小的差别来度量。由此可以看出手写体汉字识别问题与近似串匹配问题的某种内

* 国家自然科学基金和天津市自然科学基金资助.
收稿日期:1996-09-15;收到修改稿日期:1997-02-25

在联系。

本文所指出的算法是把手写体汉字的二维图象视为二维串(字符串的串),把识别问题归结为类似的近似串匹配问题,采用动态规划思想设计出的一种识别算法,实验表明该算法获得了相当满意的结果。

2 基本概念和问题描述

2.1 位向量串

定义1:串(string)$a=a_1a_2\cdots a_m$称为一个(m长的)位向量串。其分量$a_i\in\sum=\sigma^n, i=1,2,\cdots,m$,其中$a_i=a_{i1}a_{i2}\cdots a_{in}, a_{ij}\in\sigma=\{0,1\}$对于更一般的情形,可令$\sigma=\{0,1,\cdots s-1\}$。

2.2 基本变换

类似于一般符号串,两个向量串a和b的相似程度(或距离)是通过计算从a经过一系列的基本变换成为b的最小变换代价而得到的,因此我们首先给出向量串的基本变换定义。

定义2:用于一个位向量串的基本变换有3种:

(1) 代换:以n长位向量a_i代替a的分量a_i,

$$a=a_1a_2\cdots a_{i-1}a_ia_{i+1}\cdots a_m\rightarrow a=a_1a_2\cdots a_{i-1}a_ia_{i+1}\cdots a_m$$

(2) 插入:把n长位向量a_i'插到a的分量a_i的右侧,

$$a=a_1a_2\cdots a_ia_{i+1}\cdots a_m\rightarrow a=a_1a_2\cdots a_ia_i'a_{i+1}\cdots a_m$$

(3) 删除:从串a中删去它的分量a_i,

$$a=a_1a_2\cdots a_{i-1}a_ia_{i+1}\cdots a_m\rightarrow a=a_1a_2\cdots a_{i-1}a_{i+1}\cdots a_m$$

不难证明,对任意2个σ^n上的位向量串a和b,都存在着一个基本变换序列,把a变换为b。

2.3 距离

定义3:两个σ^n上的m长位向量串a和b之间的距离$\mathrm{dist}(a,b)$的值定义为:

(1) 当$m=1,n=1$时,$\mathrm{dist}(a,b)=|a-b|$;

(2) 在一般情况下,

$\mathrm{dist}(a,b)=\min_{\pi}\{\mathrm{cost}(\pi)\mid\pi$:把$a$变换为$b$的所有基本变换序列集$\}$;

其中$\mathrm{cost}(\pi)$为π中所有基本变换的代价之和。

3种基本变换的代价分别按下列各式计算:

代换:在串a中以分量a_i'代替a_i的代价为$\mathrm{dist}(a_i,a_i')$;

插入:在串a中插入分量a_i'代价为$\alpha+\beta\times\mathrm{dist}(a_i,a_i')$;

删除:从串a中删去分量a_i的代价为$\alpha+\beta\times\mathrm{dist}(a_{i-1},a_i)$,其中$\alpha,\beta$为2个常数。

这个定义是递归的,例如,σ^n上的m长串a与b之间的距离计算归结为σ上的n长串之间的距离计算,而后者归结为σ上的元c,d差的绝对值$|c-d|$的计算。

2.4 手写体汉字识别

把每个手写体汉字表示为$m\times n$维二值图象,进而视为$\sigma^n=\{0,1\}^n$上的m长位向量串,若干对应于同一汉字的字形串组成位向量串组,若干不同汉字的字形串组组成手写体汉字的样本库。

对任意一个手写体汉字的识别,就是把该手写体汉字的图象数据表示为σ^n上的m长位向量串,然后计算输入串与样本库中字形串之间的距离,根据计算结果确定一个或几个汉字识别结果。

手写体汉字识别,特别是脱机识别十分依赖于手写字体的规范性,有时同一个汉字的不同手写体相差极大,很难获得较高的识别率,为了使手写体汉字识字系统达到可以接受的识别率,往往利用多种识别算法表决来提高识别率。另外,字符图象的预处理,样本库的合理分类组织,利用有关的语义知识结合上下文进行修正后处理等都是有效的手段,本文给出的新识字算法将推进这一系统实用化的进程。

3 空间偏移匹配算法的设计及初步分析

3.1 二维图象的空间偏移匹配

把两个二维的二值图象在允许某种"弹性"形变的条件下进行近似匹配的问题归结为上节定义的两个σ^n上的m长向量串的距离计算问题。

已知σ^n上的m长向量串a,b,计算串a,b之间的距离$\mathrm{dist}(a,b)$。

3.2 平凡算法

设从串a变换为串b由基本变换序列:$\pi:t_1t_2\cdots t_k$完成。

当采用把所有可能变换的π的代价计算出来后求最小值的办法,其计算复杂度为指数阶,更何况我们的距离定义是递归的,对于串a的分量进行基本变换的代价又是2个σ上的n长串的距离。如此,不

难看出,采用平凡算法计算 a,b 距离方法的复杂度为 $O(3^{m+n})$ 阶,当 m,n 稍大时,就是实际不可计算的,所以我们考虑用动态规划(DP)算法。

3.3 动态规划算法 Distance

为了快速而又准确的计算串 a,b 之间的距离,我们设计了一种两层的 DP 算法:

```
int distance(int *x,int *y,int p,int q){
    int  dist[m][n],d1,d2,d3,*a,*b;
    if( p==1 )&&( q==1 )
    return( (*x=*y) ? 0 : 10 );
    dist[0][0]- distance(x,y,q,1);
    dist[0][1]=distance(y,y+q,q,1)*β+α+
              dist[0][0];
    dist[1][0]=distance(x+q,x,q,1)*β+α+
              dist[0][0];
    for(int i=2;i<p;i++)
      dist[0][i]= maxint;
    for(i=2;i<p;i++)
      dist[i][0]=maxint;
    for(i=1;i<p;i++)
      for(int j=1;j<p;j++){
          d1 = dist[i-1][j-1]+distance(x+
               q*i,y+q*j,q,1);
          d2 =dist[i-1][j]+α+β*distance
               (x+q*i,x+q*(i-1),q,1);
          d3 =dist[i][j-1]+α+β*distancc
               (y+q*j,y+q*(j-1),q,1);
      dist[i][j]= min (d1,d2,d3);
        }
  return(dist[p-1][p-1] );
```

3.4 算法的时间复杂度

从上节的描述中不难看出,运行该算法的时间代价是参数 m,n 的函数,主程序的主要部分有 m^2 次循环,考虑递归调用,以计算 σ 中 2 个 $a,b\in\sigma$ 的差 abs(a,b)为基本运算,大约共需 $9m^2n^2$ 次,因此可计为 $O(m^2n^2$ 阶),比平凡算法要快得多。

4 手写体字符的实验结果

我们的实验不是直接面向应用系统的开发,因为那将涉及许多因素。我们的目的仅仅是评估一下我们给出的算法的性能。

在由 360(36 个汉字,每字有 10 种不同手写字型)个手写体汉字的二值图象组成的样本中,以库中的每一字形作为输入的待识字符,分别与其余 359 个样本比较,对同一汉字的每组样本计算平均距离,获得相当好的识别效果:

识别正确者 281 个,占 78%;正确结果位于前 3 名者 327 个占 91%;正确结果位于前五名者 345 个,占 96%。

5 结　论

从初步的实验结果来看,我们给出了一个强有力的识别算法,由于其思路与一般的启发式结构算法不同,因此有着较好的互补作用。

另一方面,较高识别率的代价是计算复杂度仍然较高。不过,进一步的设计完全可能大幅度地缩短计算时间。

对于大规模的实用系统,由于动态规划算法自身的特性,极适于并行计算[4~6]。一种利用 DP 算法进行手写体字符识别的并行计算 ASIC(专用集成电路)芯片肯定有美好的前景。

参 考 文 献

1 Sara Baase. Computer Algorithms: Introduction to Design and Analysis. California: Addison－Wesley, second edition, Menlo Park 1988.

2 Cormen T H, Leiserson C E, Rivest R L. Introducyion to Algorithms. , MIT Press, Cambridge, Massachusetts, 1990.

3 Bunke H, Buhler U. Applications of Approximate String Matching to 2D Shape Recognition. Pattern Recognition, 1993, 26 (12): 1797～1812.

4 Cheng H D, Fu K S. VLSI Architecture for Dynamic Time-warp Recognition of Handwritten Symbols. IEEE ASSR, 1986, 34 (3): 603～613.

5 刘璟,卢桂章,韩维桓. 一种用于曲线检测的动态规划算法及相应的 VLSI 阵列结构. 计算机学报,1990,13 (2): 101～106.

6 陈国良. 并行算法的设计与分析. 北京:高等教育出版社,1994.

刘璟，南开大学计算机与系统科学系教授，主要研究：算法的设计与分析，VLSI 算法与结构，ASIC 设计，并行计算技术及应用，图象处理与模式识别。国内外发表论文 20 余篇。中国计算机学会理论计算机科学分会理事，天津市电子学会常务理事，天津市计算机基础教学指导委员会副主任等。

A DP Algorithm for Off-line Recognition of Handwritten Symbols

Liu Jing, Bai Gang

(Department of Computer and System Science NanKai University, Tianjin 300071)

Abstract In this paper we regard recognition of handwritten Chinese symbols a space-warp pattern matching problom, and design a two-level dynamic programming algorithm for it. The tentative results state clearly it's satisfactory

Keywords Approximate string match, Off-line recognition of handwritten Chinese symbols, Space-warp matching, Dynamic programming

VTEL 美国视讯公司——南京邮电学院多媒体视讯技术培训中心成立

（江苏南京 1997 年 12 月 31 日讯）为了更好地服务于 VTEL 美国视讯公司广大的中国用户，使他们了解世界最新视讯技术、加强技术水平，VTEL 美国视讯公司与中国会议电视先驱及图象处理学术权威之——南京邮电学院签署了协议，共同合作成立 VTEL 美国视讯公司——南京邮电学院多媒体视讯技术培训中心。

VTEL 美国视讯公司—南京邮电学院多媒体视讯技术培训中心，在南京邮电学院校园内挂牌成立。该中心将面向会议电视设备的使用者、加强他们对会议电视、多媒体通信原理及其应用、方案设计及发展方向方面的深入了解；提高他们对会议电视设备操作及对常见故障进行检修等方面的技能。同时，该中心还是会议电视领域国内首家中外合作的技术性培训中心。

中心将由南京邮电学经验丰富的教授、讲师执教，以灵活生动的授课形式向学员提供包括会议电视、多媒体通信基础原理及最新技术应用、发展方向等课程。另外，还会通过安装实习及交互式答疑等形式，进一步提高学员对会议电视设备操作和常见故障检修等方面的技能。

此协议是双方本着互相促进、互相学习的原则签定的，目的在于进一步推动中国蓬勃发展的教学科技与现代通信，为不断壮大的 VTEL 中国用户提供一个集理论与实践有机结合的技术学习中心。