广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(3):177~179
Vol. 23,No. 3 August 2007

基于结构布局的数学公式识别*
Recognition of Printed Mathematical Expressions Based on Structure Layout

黄 潇,李奋华
HUANG Xiao,LI Fen-hua

(运城学院计算机科学与技术系,山西运城 044000)
(Department of Computer Science and Technology, Yun Cheng University, Yuncheng, Shanxi, 044000, China)

摘要:介绍一种采用"自顶向下"和"自底向上"相结合的印刷体数学公式识别系统。该系统主要由局部结构分析和整体结构分析两个模块组成,能够较好地识别印刷体数学公式。该系统实际用于识别98个印刷体数学公式的正确识别率为0.867。

关键词:识别系统 数学公式 结构 字符 函数
中图法分类号:TP391.41 **文献标识码**:A **文章编号**:1002-7378(2007)03-0177-03

Abstract: This paper introduces a recognition system of printed mathematical expressions, which combines the method of "Top-Down" with the way of "Bottom-Up". The system, mainly consisting of local structural analysis and global analysis, is cabaple of recognizing printed mathematical expressions. The actual rate for accurately recognizing 98 printed mathematical expressions is 0.867.

Key words: recognition system, mathematical expression, structure, symbol, function

近年来,随着网络技术的飞速发展,信息交流和资源共享,特别是技术资源的共享日趋频繁,科技文献的电子化就显得尤其重要。科技文献电子化首先需要准确高效的识识系统来识别印刷文字。科技文献不仅包含普通文字、图像和图形,还包含大量的数学公式,准确高效的识别科技文献识别系统既要能识别文字、图像和图形,还要能识别数学公式。目前主流的OCR系统能够高效、准确地识别文档中的文字,但一般不具备数学公式的识别与重构功能,仍需要按照图片来处理公式,存储数据量大且无法编辑、修改。因此,研究数学公式识别,对于拓宽OCR系统的应用领域,具有重要意义。

国外于20世纪60年代后期开始数学公式识别的研究,进入90年代,这个领域的研究热度逐渐增加[1]。数学公式识别分为字符识别和结构分析识别两个阶段。字符识别首先采用连通域搜索算法对公式字符进行分割,提取包围结构字符,并根据分割字符的大小和连接矩阵决定的相对距离对多结构字符(如:i、j、=、>=等)进行合并;然后采用水平、垂直投影轮廓切割的方法对根号公式进行字符切割;最后利用模式匹配算法对切割结果进行识别[2]。使用句法识别公式最先是Anderson[3]采用纯粹"自顶向下"的分析方法以句法为标准分割数学公式进行识别。该方法由于分析策略的缘故对公式识别不是十分有效。后来Chang[4]提出用结构说明方案对公式结构进行分析,该方法主要采用操作码优先级和它的操作数范围,它考虑到效率,但是该方法冗长,难于理解和实现。其它数学公式识别方法[1]都是在假定字符识别完全正确的情况下进行,没有进行错误校正,整体效果并不理想。所以我们根据公式中字符的结构布局,采用"自顶向下"和"自底向上"相结合的结构分析方法识别数学公式。

1 结构分析方法的思想

本方法主要由两个模块组成:(1)整体结构分析

收稿日期:2006-10-27
修回日期:2007-06-12
作者简介:黄 潇(1978-),女,讲师,主要从事电子商务研究工作。
*运城学院院级项目(26)资助。

模块。该模块采用"自顶向下"的思想，使用算法将整个表达式分割成若干个子表达式，而后对每一个子表达式用同样的算法递归分割，直至不能再分为止。(2)局部结构分析模块。该模块采用"自底向上"的思想对子表达式中的特殊结构(根式、上/下标、矩阵)进行特定分析处理，最后达到需要的结果。设 csize(χ)，center(χ)，D(χ，y) 分别是字符 χ 的大小及中心，字符 χ，y 间的距离，则

$$\begin{cases}\text{top}(\chi)=\text{true}, \chi \text{ 正上方有字符；}\\ \text{btm}(\chi)=\text{true}, \chi \text{ 正下方有字符；}\\ \text{con}(\chi)=\text{true}, \chi \text{ 包含一个下标。}\end{cases}$$

2 结构分析方法

2.1 结构分析预处理

为了便于结构分析，需要提取函数名与子表达式字符串，需要对字符的大小和中心进行归一化。如图1所示，定义归一化尺寸(NSize)为一个符号经过扩展后的尺寸，即包括 x,y,z 三部分之和；归一化中心(NCenter)为一个符号经过扩展后的外接矩形的中心。NSize 和 NCenter 可以通过符号的外接矩形根据 x,y,z 的比值计算得到，比如，基于32篇数学文档统计出这个比值平均为 28∶51∶21。给定一对符号，h_1 和 h_2 分别表示两个符号的归一化尺寸，c_1 和 c_2 分别表示两个符号的归一化中心在垂直方向上的坐标值。令 $H=\frac{h_1}{h_2}\times 1000, D=\frac{c_1-c_2}{h_1}\times .1000$ 则符号对之间的位置关系可以用归一化坐标平面上的点 (H,D) 来表示[5]。

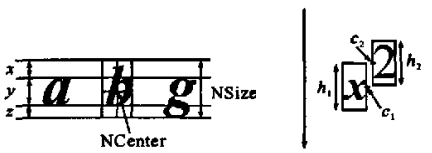

图1 使用的符号定义

对字符归一化后，采用最长字符串匹配的方法提取函数型字符(三角函数，abs()等)，并把它看作一个整体符号。同时，还要将满足条件下列条件的连续字符合并为一个字符：(1)具有相同大小和中心；(2)两相邻字符间的空格小于它们大小。

2.2 局部结构分析

局部结构分析主要是对根号表达式、上/下标和矩阵进行处理，将处理结果交给整体结构分析模块进行分析，然后该模块将分析结果通过相应的链接连接到结构树的目标字符上。

2.2.1 根号表达式

根据根号运算符的水平线及其高度确定其内的子表达式，利用整体结构分析模块对该子表达式进行分析，将分析结果通过包含链接连接到结构树的"√"运算符上。

2.2.2 上/下标表达式

下上标的处理类似，仅以上标的处理进行介绍。上标处理过程分两步：(1)搜索子表达式的第1个字符；(2)子表达式范围的确定。

2.2.2.1 搜索子表达式的第1个字符

设 χ，S 分别表示目标符、极限符集合；op(χ)是目标符类型；A，C，D 分别表示图2(a)、(c)、(d)的阴影搜索区域；Search(w) 是在 w 区域的搜索函数。z_i 表示子表达式的第 i 个字符；搜索算法如下：

```
if (op(χ)∈S) then break;
else {search (A);
if (top(χ)=true and btm(χ)=true) then break;
else if (top(χ)=true and btm(χ)=false)
    then search(C);
    else if (top(χ)=btm(χ)=false and con(χ)=
    true)
    then search (D);}
```

同时，z_1 应满足条件：①z_1 在搜索区内。② csize(z_1) <=csize(χ)。③ D(χ，z_1) <D(χ，z_i)，$i=2,3\cdots$。④z_1 符合上下文。

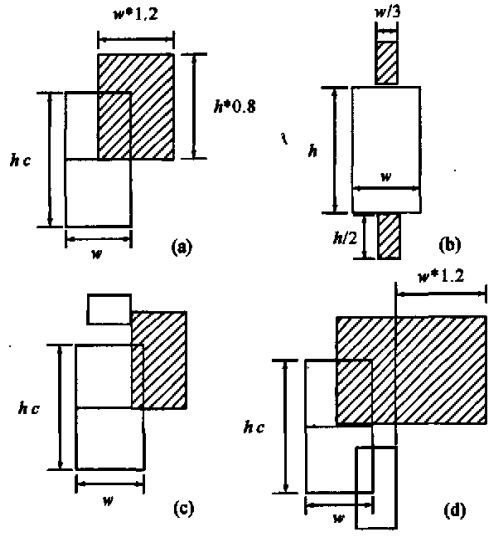

图2 目标符 χ 的搜索区域

c，w，h 分别是目标符 χ 的中心、宽度、高度。

2.2.2.2 子表达式范围的确定

一般子表达式中字符须满足：①D(z_i，z_{i+1}) <=0.6 * csize(z_1)，$i=2,3\cdots$。②csize(z_i) <=csize(χ)，$i=1,2,3\cdots$。子表达式范围可以通过结构

预处理中提取的字符串范围和与目标字符具有相同基准线的字符位置确定。

2.2.3 矩阵

设 search(a,b,T) 是一对同类型包围结构界定符 a,b 的查找函数,如果 a,b 的大小和中心都相等,T = true;Hproj(w,n,z) 是给定区域 w 内的水平投影函数,并返回行数 n,如果所有行高度相等,则 z = true;Vproj(w,m) 是给定区域 w 内的垂直投影函数,返回列数 m,并确定 w 内各元素所在的位置。算法如下:

```
Search (a,b,T);
if (T=true)
then {Hproj (w, n, z);
    if (n>=2 and z=true)
    then { w 是矩阵区域;Vproj (w, m);}
    else break;}
```

2.3 整体结构分析

先用最佳目标符和空格对公式进行整体分割,然后该模块在被分割的子表达式区域被递归调用。按下列优先级对给定公式分割:(1)用字符垂直分割;(2)用字符水平分割;(3)用空格水平分割。

设 w ,χ,segment(w ,χ)分别是给定的表达式区域、目标字符、分割函数。

2.3.1 用字符进行垂直分割

设 χ,y 分别是 w 内的最长字符、除目标符外的任意字符。分割算法如下。

```
if (∃ y|center(y)=center(χ))
   then break;
else  segment(w,χ);
```

调用整体结构分析模块分析上、下两个子区域,将结果用上/下链接连接到目标符上。

2.3.2 用字符进行水平分割

设 χ 是 w 内的最高字符。分割算法如下。

```
    if (top(χ)=true and btm(χ)=true)
    then break;
    else segment(w,χ);
```

调用整体结构分析模块分析左、右两个子区域,将结果用水平链接连接到目标符上。

2.3.3 用字符间空格进行水平分割

如果公式经上两步处理仍未分割成功,则用最佳空格将它分成左、右两个子区域。

(1)最佳空格估算。设 n,w,h,s 分别表示空格数目、给定区域的宽度与高度、被切分字符串大小;e_i,f_i,w_i 分别表示第 i 个空格的两个惩罚值及其宽度。用最佳适合值 d 和 e_i,f_i 来估算最佳空格。最佳空格满足:① 空格越宽越好,因此,$e_i = (w - w_i)$;② 空格应适合分割大的字符串及尽可能少切分字符串,使子表达式结构尽可能少被破坏,因此,$f_i = (h - s)$,最佳空格应使 $d = \min\{d_i = e_i + f_i\}$。

(2)用字符间空格进行水平分割。设 Find (w, χ, T) 是查找最佳空格函数,当空格数目非 0 时,T =true,并返回最佳空格 χ。分割算法如下:

```
Find (w,χ,T);
if (T=true) then segment(w,χ);
    else break;
```

调用整体结构分析模块分析左、右两个子区域,分析结果用水平链接连接到目标符上。

3 结束语

使用本文方法实际识别 98 个印刷体数学公式,能正确识别 85 个,正确识别率为 0.867,效果不算理想。其原因主要表现在两个方面:(1)对一些相似字符识别不清,如{o,O,0},{S,s,5}。(2)对一些空间运算符识别错误,如:"σ^y"错识为"δy"。这第二种错识原因则主要是由结构分析方法造成的,因此,该结构分析方法还不够完善,有待进一步改进。比如,一些很复杂的空间操作符,通过公式字符的结构布局很难判断,需要在结构分析策略中利用操作符优先级和上下文语义信息加以解决;应建立一些相应的语法规则,进行错误校正,提高该方法的结构分析能力。

参考文献:

[1] CHAN K F, YEUNG D Y. Mathematical expression recognition: a survey[J]. Int J On Document Analysis and Recognition, 2000, 13(1): 1-11.

[2] OKAMOTO M. Recognition of mathematical expressions by using the layout structures of symbols [J]. Proc ICDAR'91, 1991, 3(1): 242-250.

[3] ANDERSON R. Two-dimensional mathematical notion [M]//FU K S. Syntactic Pattern Recognition Applications. New York: Spring-Verlag, 1977: 147-177.

[4] CHANG S K. A method for the structural analysis of two-dimensional mathematical expressions [J]. Information Sciences, 1999, 2(3): 43-52.

[5] YUKO ETO, MASAKAZU SUZUKI. Mathematical formula recognition using virtual link network[J]. Int J On Document Analysis and Recognition, 2001, 11(2): 1-3.

(责任编辑:邓大玉)