数学公式图像的结构理解与重现

史广顺, 肖 萃, 王庆人
(南开大学 机器智能研究所,天津 300071)

摘 要:数学公式图像识别与理解是文档图像处理领域的重要组成部分,目前尚无满足一般应用的处理方法.提出了一种鲁棒的数学公式结构理解方法,使用公式图像识别结果、语法规则和句法规则分析数学公式结构,对数学公式的类型进行了完整的划分,对识别结果的错误进行自动的检查和纠正,能够自动分析数学公式符号的优先级和计算顺序.既可以应用于数学公式图像的识别与格式转换,也可应用于对数学公式的检索和辅助编辑.基于 1 000个真实公式图像的实验结果证明了分析方法的有效性和稳定性.

关键词:数学公式识别;版面结构分析;语法结构分析;数学公式结构理解
中图分类号: TP391 **文献标识码**: A **文章编号**: 1673-4785(2008)05-0401-07

Reconstructing mathematical expressions from image data

SH I Guang-shun, X IAO Cui, WANG Q ing-ren
(Institute of Machine Intelligence, Nankai University, Tianjin 300071, China)

Abstract: Mathematical expressions appear in many kinds of scientific documents and technical reports. Understanding and reconstructing mathematical expressions has become an important problem in the domain of document image analysis. The authors developed a robust method for the analysis of structure in mathematical expressions. After images are processed, generating recognition results, this method analyzes the structure of mathematical expressions according to syntax rules and syntactic rules. Classification into different types of mathematical expressions is then made. Syntax errors in the recognition process are checked and corrected automatically. The preferential level and the computing sequences of arithmetical operation signs in mathematical expressions are also automatically analyzed. This method can be applied to the recognition of images containing mathematical expressions and transforming between formats, and is useful in retrieval and editing of mathematical expressions. About 1000 images of mathematical expressions from real documents were used for performance evaluation. The test results proved the stability and efficiency of this method.

Keywords: mathematical expression recognition; layout analysis; syntactic analysis; mathematical expression understanding

数学公式存在于各类文档之中,对其进行精确的识别和理解是文档图像处理领域的重要问题.由于数学公式的二维空间结构以及数学符号语义的多义性,使得对数学公式结构的描述与理解变得非常复杂和困难,所以对其的研究具有很高的科研价值和挑战性.近 20年来,研究者们提出了多种数学公式结构分析的处理方法.既包括利用版面信息[1-2]对数学公式进行结构拆解的方法,也包括使用语法规则[3-5]对数学公式结构进行理解和描述的方法. Tian[6]提出了利用基准线构建初始结构树,并利用语法和语义知识进行树转换的方法. U. Garain[7]采用词法分析与句法分析相结合的方法,来提高树结构的准确度.

在上述的各种方法中,单纯依靠版面信息无法消除数学符号的歧义性,不能理解数学公式的计算含义.近年来的一些新方法中,虽然加入了语义语法规则,可只是作为辅助信息,无法有效检查并纠正数

收稿日期: 2008-04-16.
基金项目:天津市自然科学基金资助项目 (05YFJMJC01500).
通信作者:史广顺. E-mail: gsshi@nankai.edu.cn.

学公式图像识别结果中的错误,鲁棒性不足.本文提出了一个句法规则驱动的、对数学公式进行结构分析的方法,将版面结构信息、符号语法规则、公式句法规则相互结合,对数学公式图像的识别结果进行分析.这种方法既可以实现对数学公式结构的重现,同时又可准确的理解数学公式所表达的计算含义,为数学公式的语义分析和高级应用提供帮助.

本文的研究重点在于如何建立一种介于图像和语义之间的描述结构,对数学公式的版面形式、符号内容、语法关系、句法结构进行完整的描述,实现不同类型信息之间的统一,从而为更加精确地理解数学公式结构奠定坚实的基础.

1 数学公式句法结构描述模型

1.1 数学公式的结构组成

文档图像处理的目标在于获取文档的表面结构或深层结构.对数学公式而言,其文档结构层次如图 1所示.

分析级别	分析层次	分析目标	重现目标
高	语义理解	实现机器自动计算	C/C++或Matlab代码
	句法重现	实现公式的再编辑	MathML文档
低	版式重现	实现公式的图像重现	Latex、MathML文档

图 1 数学公式文档结构层次

Fig 1 Overview of expressions structures

对印刷体数学公式图像的识别和理解包含 3个处理目标:1)数学公式版面结构重现,使其能够转换为 Latex或 MathML格式的电子文档;2)数学公式重新编辑,使其能够导入到类似于 MathType的编辑器中;3)数学公式重新计算,分析其计算含义并进行自动求解或计算.

由于数学公式的结构复杂性和符号多义性,以及图像质量低下和识别错误造成的信息失真.只利用版面结构信息无法深入理解数学公式的语义含义,而只利用语义规则无法有效克服各种前期处理的结果错误,只有将版面信息、字符内容、语法规则、句法规则相互结合,才能实现鲁棒准确的数学公式结构分析.

1.2 数学公式句法结构描述模型

本文使用四元组结构描述数学公式的句法结构,如图 2所示.

$$\text{syntactic structure} = (\text{layout information}, \text{symbol set}, \text{grammar rules}, \text{syntactic rules});$$

式中:layout information为版面结构,数学公式中所有符号和公式结构的版面位置信息.版面信息可用于判断作用范围、提取不同层次的子表达式,是重要的辅助信息;symbol set为符号集,数学公式中所有出现的操作符和操作数.根据符号内容可调用相应的语法规则,确定符号之间的组合关系,检查符号出现的合法性;grammar rules为语法规则,不同符号之间的语法约束与组合关系.它用于确定操作符的作用域和子表达式内容,同时检查发现识别结果中存在的错误,保证子表达式提取的完备性和正确性;syntactic rules为句法规则,子表达式的分解与约束关系.此类规则主要负责分析不同运算符之间的优先级顺序,消除数学公式符号的多义性,并可被快速解析转换为其他的数学公式描述形式.

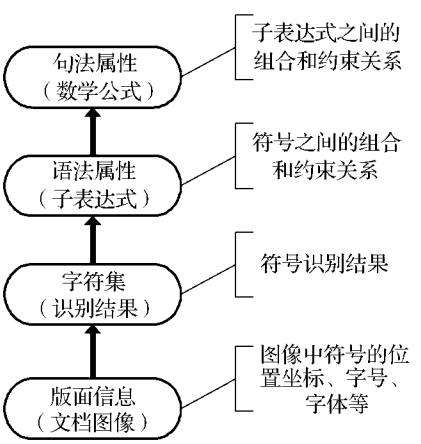


图 2 数学公式句法结构模型组成

Fig 2 Syntactic structure model of expressions

1.2.1 数学公式的版面结构

版面结构包括公式中字符的版面信息和公式结构版面信息两部分内容.

字符的版面信息包括字符的外界矩形位置坐标、字符中心线、字体、字号等信息.

公式结构版面信息包括公式的版面类型、基准线位置以及多条基准线间的相互关系等.数学公式中的符号根据其语法属性不同体现出不同的排版风格.根据公式符号的水平中心线 (HCL)排列情况,可对数学公式的类型进行如下划分.

1)基元表达式 (unit expression)

数学公式中的独立符号,不可再分,主要是指不具有运算功能的运算数;

2)普通表达式 (common expression)

绝大部分符号排列在相同的 HCL上,呈现一维版式结构;

3)角标表达式 (script expression)

角标是一种特殊的语法约束关系,角标符号的 HCL位于其描述符号的左上、左下、右上、右下 4个方向;

4)组表达式 (group expression)

一些特殊的运算符会与其他符号组合成 2D结构的版面形式,如根式、求和、积分、分式等.

5)矩阵表达式 (matrix expression)

由特殊定界符包含多行多列符号组成的表达式,如行列式、矩阵等;

6)堆叠表达式 (stack expression)

描述说明符号在数学公式中常以堆叠的形式出现,它们不是具有固定语法规则的组表达式,如帽子符号等;

图 3描述了不同版面类型的基础数学公式.对基础类型进行准确的划分和分析,有助于对公式整体结构的分解和重构.

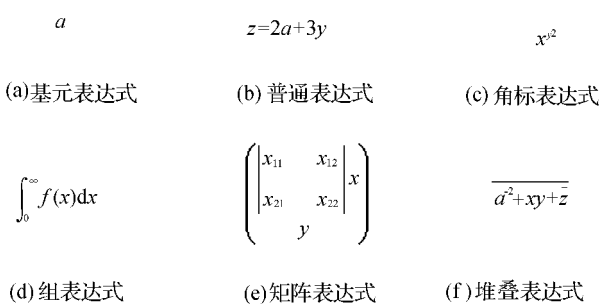


图 3　数学公式版面类型示例

Fig 3　Visual samples of layout structures

1.2.2　数学公式符号集的组成和类别

数学公式的符号可分为操作符和操作数 2类.操作符:包括运算符、函数名及某些特殊符号,在数学公式中表示对一个或多个操作数的某种操作关系,或某种特殊数学规律;操作数:是指由数字、英文字母、希腊字母等代数符号构成,在数学公式中表示数量、变量等含义.根据符号的类型可选用对应的语法规则进行深入的子表达式分析.

本文研究工作针对正体/斜体英文字字母、数字、标点、希腊字母、数学符号、三角函数共计 220个字符,覆盖了科技文献中所有数学公式的常用字符.

1.2.3　数学公式的语法规则

语法规则规定了数学公式中字符的语法属性,以及不同符号间的语法约束与组合关系.对数学公式的分析过程中,语法规则具有如下作用:

字符语法属性分类.语法规则可以通过对字符间空间关系的判断,确定字符的惟一语法属性.图 4描述了对"+"进行语法判断的过程.

组合"子表达式"(定义见 2.3节).语法规则可以通过作用域信息,将具有组合规则的运算符与其附属字符合并,成为一个子表达式.

语法规则验证.结合识别结果和版面信息,语法规则还可以用来纠正识别错误、消除多语义字符的语法歧义、实现对树结构的校验和修改.(详见 2.5节)

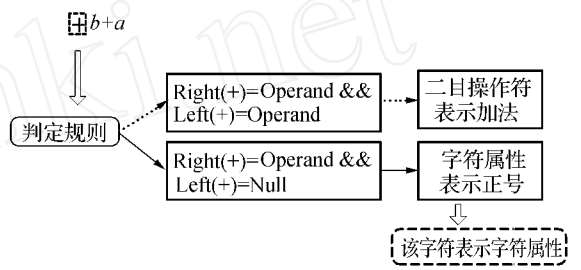


图 4　利用语法规则确定字符属性

Fig 4　Check the grammar attributes by grammar rules

1.2.4　数学公式的句法规则

不同类型表达式的组合形成了多重层次的数学公式结构,同层次操作符之间的优先级关系决定了数学公式的计算顺序.

句法规则描述每个操作符的子表达式形成规则,每一个操作符都有一个固定的树型结构模板,其中子结点的个数和属性均根据语法规则预先填充.图 5描述了根据句法规则中不同的优先级关系生成的不同子表达式结构.

句法规则同时负责判断操作符之间的优先级,采用"相对优先级"的形式设计了包含所有操作符的矩阵结构,任意 2个操作符均可通过查找矩阵以确定哪个操作符的优先级更高.

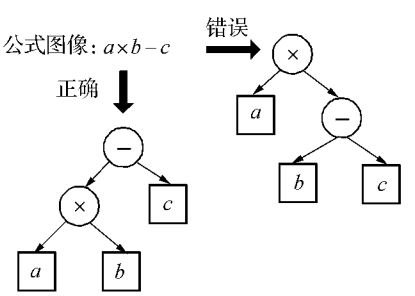


图 5　优先级比较示例

Fig 5　Sample of priority comparison

1.3　结构分析底层知识库

以上提到的四元组是结构分析的直接依据.通过前期大量的统计工作,对数学公式符号的操作属

性、语法属性、组合关系、操作符优先级、目类型等进行了统计分类与描述. 同时,定义了底层知识库,用于以上各类规则和信息的存储. 知识库的内容如图6所示.

知识库	符号信息	符号图像 符号内容
	语法规则	符号的操作属性 符号的语法属性(即符号的类别) 符号语法属性的判定规则 符号具有的组合关系
	句法规则	符号的目类型 子表达式的组合关系和判定规则 操作符的优先级别

图 6 底层知识库构成

Fig 6 Structure of mathematical knowledge database

知识库是为语法结构分析处理过程而建立的,为结构分析提供重要的信息. 同时,这种知识统一存储与管理的方式,大大的增强了系统的可扩充性和灵活性.

2 算法与关键技术

基于数学公式结构描述规则库,可采用"自顶向下"的处理流程对数学公式的结构进行迭代式的分析. 首先通过版面信息找到公式的核心骨干层次,然后利用语法和句法规则将该层次转换为一棵能反映公式正确计算顺序和结构的句法树. 当该层次全部分析完成,再从公式中找到次级核心骨干层次,对句法树进行扩充. 不断重复这一过程,直到公式结构分析全部完成.

2.1 数据结构设计与核心处理算法

本文采用树型结构描述数学公式,每一个操作符的树型结构都是与其对应的句法规则的一个实例.

本文方法的处理流程描述如下:

算法 1:

初始状态:处理对象为公式中所有符号. 创建空的根结点.

1)进行版面分析,提取第 1层次的所有字符;

2)应用语法规则,确定核心操作符集;

3)应用句法规则,判断操作符的优先级,按优先级将核心操作符的子表达式结构填充到结构树中;

4)选择公式中次高级别的骨干层次作为下一个处理对象,跳至 1),循环重复,直至结构分析完成.

采用以上算法,数学公式图像的识别结果可以被组织成遵循计算顺序的树型结构.

2.2 版面分析技术的应用

算法 1的 1)需要通过版面分析以确定属于第 1层次的操作符. 可利用的版面信息包括:操作符的 HCL、符号的大小、表达式图像外接矩形的水平中心坐标等.

首先,将所有字符按照 HCL的值进行聚类,得到公式中所有骨干线信息;然后,挑选出具有最高优先级的骨干线作为当前分析层次对象.

图 7描述了提取公式 $y=\int_0^d \sin x\mathrm{d}x-\frac{\ln x}{2a}$ 最高级别骨干层次的处理过程.

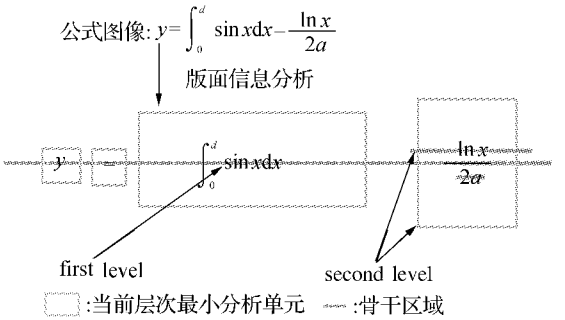


图 7 版面结构分析

Fig 7 Layout structure analysis

2.3 子表达式提取

定义子表达式:由一个或多个当前层次运算符与其附属字符组合而成,例如图 6公式中的 $\int_0^d \sin x\mathrm{d}x$ 和 $\frac{\ln x}{2a}$,分别是以积分号和分数线为核心操作符的子表达式. 组合后的子表达式在当前的运算层次的属性为操作数. 因此,在句法结构分析之前,根据语法规则提取子表达式,可以有效筛选出公式核心操作符,避免附属操作符对公式计算顺序的影响,保证公式结构的准确性. 根据数学公式的阅读顺序(从左至右),对所有当前层次的操作符应用语法规则,执行以下算法:

算法 2:

输入:核心骨干层次字符集(FLOS),包含公式经版面信息分析提取出的所有处于第 1层次的字符.

1)依据字符左边界 X 坐标值,从左至右排列所有字符;

2)从 FLOS中提取第 1个操作符,定义为 A;

3)在知识库中查询 A的语法规则;

4)若 A有多个语法属性,则结合版面信息判定,得到 A的惟一语法属性;

5)若 A具有组合规则,则根据作用域,将所有附属于 A的字符合并,形成以 A为核心运算符的子表达式;否则,跳至 7);

6)将被组合的操作符从 FLOS中删除,子表达式作为新的操作数;

7)在 FLOS中提取下一个操作符,跳至 3).

经过以上处理步骤,在 FLOS中留下的操作符就是第 1层次的核心操作符,可根据它们进一步建立数学公式的句法结构框架.

2.4　句法结构树的生成

得到核心操作符集 (Core-FLOS)之后,首先,利用知识库中的句法规则,对第 1层次核心操作符进行优先级比较、公式结构拆分、子表达式拆解等句法结构分析,得到当前层次的句法树;然后,提取公式的下一层次字符,重复算法 1,直到生成一棵完整的公式结构树.

图 8完整描述了图 7中公式的结构分析过程,处理结果被保存在结构树之中.

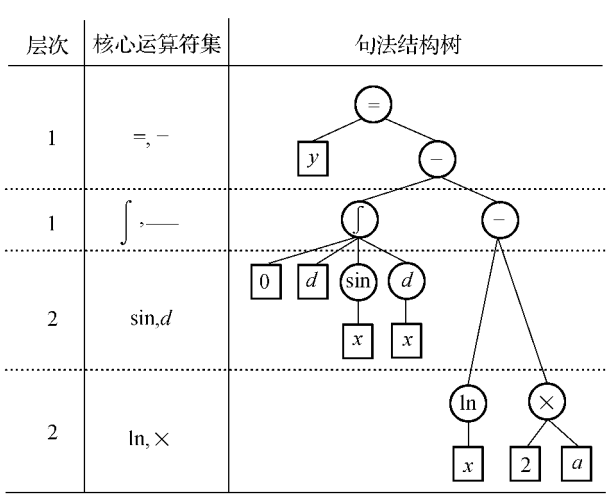


图 8　句法结构分析过程示例

Fig 8　Sample of syntactic analysis

2.5　结构重现与格式转换

在获得如图 8所示的数学公式结构树之后,可以非常容易的将其转换为 Latex格式或 MathML格式.

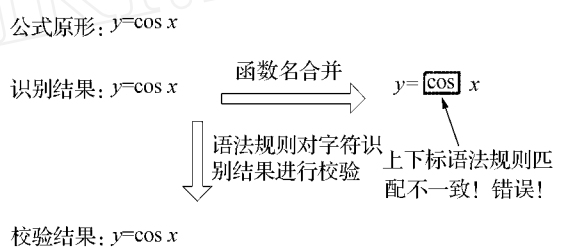
本文方法不仅可以实现版面结构的重现,同时可以快速的将数学公式结构导入到各种编辑器之中,结构树中蕴涵的语法信息和句法信息均可通过特殊的编译器将其转换为可直接运行的计算代码.将在后续的论文中描述这种方法.

3　语法规则对结果的校验

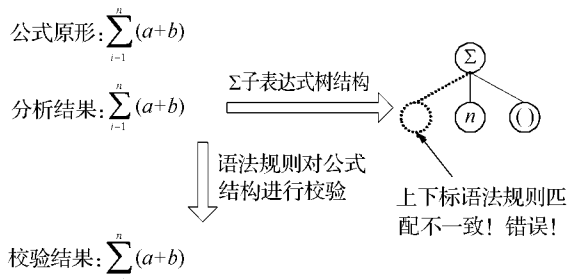
3.1　识别结果校验

在数学公式的符号识别结果中可能包含识别错误,利用语法规则中的约束条件,可以发现并消除部分识别错误,本文设计了以下几类容错处理:

1)字母和数字的拼写检查:‘0’和‘o’,‘1’和‘l’非常容易混淆.但是如果出现‘c0s’,则根据函数名拼写规则应将其改为‘cos’;在数字串中如果出现字母,也可以将其改为形式相似的数字.图 9(a).



(a)利用语法规则纠正字符识别错误



(b)语法规则对子表达式树结构的校验

图 9　语法规则的验证功能

Fig 9　Grammar ambiguity eliminate

2)约束条件的错误修正.对定界符‘(’和‘)’而言,识别结果可能是‘[’或‘]’,根据符号的组合规则与约束条件,可以利用上下文信息进行可信度验证,从而修正其识别错误.

3.2　语法消歧

在数学公式中存在很多有歧义的运算符,消除歧义是分析数学公式结构的重要技术.本文设计的语法规则中包含一项约束条件,其用途就是消除运算符的语法歧义.

约束条件包含以下几项内容:

1)运算数的个数与位置.如 1.2.3节中图 4所示,以“+”为例,当其左右两侧均出现运算数或其他子表达式时,说明它是四则运算符.当只有其右侧

存在操作数和子表达式时,说明它是符号运算符.

2)与其他运算符的组合关系.以“(”为例,当其右侧出现“]”且两者之间存在“,”时,说明它是一个区间描述符“(,]”;当不存在“,”时,它应该被理解为一个定界运算符.

通过定义一系列的约束条件,可以准确地分析每个符号的语法属性,并根据约束条件和组合关系提取子表达式.

3.3 句法树结构校验

子表达式树反映了对应公式的句法结构.由于识别及版面结构分析的误差,有可能使树结构变形,从而无法正确表达公式的信息.应用语法规则,可以通过必要条件是否存在,来判断拆分的正确性,对错误情况依据语法规则进行修正,得到准确的子表达式结构.图 9(b)描述了对子表达式结构树的校验过程.

4 实 验

为验证本文研究工作的有效性,共选择使用了 500个公式样本作为训练样本.从 IEEE Transactions及其他学术期刊中扫描制作了 500页文档图像,并将这些图像中包含的 7 610个数学公式作为评测样本.在评测样本中,包含 90 913个公式符号,覆盖了本文使用字符集中的所有符号.由于尚无有效的数学公式结构分析自动评测工具,因此采用人工观察的方法对 1 000个评测样本进行了性能评价.结构分析的评测参数设计如下:

1)子表达式内容完整性.子表达式(一级核心运算符)是否能够正确提取,语法约束元素是否能够正确提取;

2)句法结构正确性.公式的计算顺序是否描述正确,子表达式的句法结构是否正确有效;

3)系统容错性.对于图像噪音造成的误识结果是否能够删除,对于符号误识结果是否能够修正.

采用 J. Mitra[8]提出的方法,以句法层次数目作为评价公式复杂性的标准.并按照版面结构复杂度对测试公式进行分类,进行测试.测试结果如表 1所示.测试结果表明,本文介绍的句法结构分析方法适用于多种类型的公式.特别的,对于占数学公式主体的中等复杂度公式,结构分析准确率得到了非常大的提升,平均准确率达到 96.8%,高于目前同类系统平均水平(85%～92%).而对于复杂度较高的公式,结构分析的准确率也得到了相应的提高,平均准确率为 81.7%,达到预期的目标.

除此之外,为了验证系统的容错性,对 200个数学公式图像的识别结果进行了人工修改,得到带有噪音和误差的评测样张集,进行系统容错性测试.测试结果如表 2所示.综合 2类测试结果,应用本文提出的结构分析方法,可以有效提高结构分析的正确性和稳定性,更好的满足各类应用的需求.

表 1 结构分析测试结果

Table 1 The results of syntactic analysis

公式复杂度	样张数量	子表达式完整	结构正确	平均准确率
1	150	150	150	1.000
2～4	400	399	398	0.996
5	276	269	253	0.946
6	89	83	79	0.910
7～12	85	75	64	0.817
Total	1 000	976	944	0.960

表 2 容错性测试结果

Table 2 The results of fault-tolerant

错误类型	样张数量	纠正数量	平均纠正率
识别错误	100	82	0.820
结构错误	100	85	0.850
Total	200	167	0.835

5 结 论

本文的研究工作,将句法结构、语法规则、版面分析相互结合,更加完整的理解数学公式.与目前同类方法相比,主要有以下方面的优点:

1)定义了完整的数学公式句法结构模型,明确定义了语法规则和句法规则,可以更加准确地对数学公式进行描述和分类;

2)首次提出底层知识库概念,系统地总结、归纳并分类存储了大量先验数学属性和规则,为公式结构分析提供直接依据;

3)采用 HCL聚类的骨干线提取方法,通过多参数确定骨干区域,提高了骨干层次划分的稳定性;

4)采用句法规则驱动的结构分析方法,从结构分析的最初,加入验证信息,保证了公式结构分析的正确性.同时该结构分析过程又可准确的理解数学公式所表达的计算含义,为语义计算等高级应用提供基础信息;

5)使用语法规则进行结果验证,极大地增强了数学公式理解系统消除符号歧义、校验错误的能力,提高了分析的准确率与稳定性.

在今后的工作中,还要进一步研究如何分析并提取数学公式的计算含义,同时对数学公式结构分析方法的评价也将是重要的问题.

参考文献:

[1]CHEN Y, SHIMIZU T, OKADA M. Fundamental study on structural understanding of mathematical expressions[C]// Proceedings of IEEE SMC '99 Conference. Tokyo, Japan, 1999: 153-158.

[2]靳简明. 数学公式图像处理研究[D]. 天津: 南开大学, 2003.
JIN Jianming. Research on typeset mathematical expression-image processing[D]. Tianjin: Nankai University, 2003.

[3]GUO Yusheng, HUANG Lei, LIU Changping. A new approach for understanding of structure of printed mathematical expression[C]// Proceedings of the 6th International Conference on Machine Learning and Cybernetics. HongKong, China, 2007: 19-22.

[4]LAVIROTTE S, POTTIER L. Mathematical formula recognition using graph grammar[J]. Document Recognition V, 1998, 3305: 44-52.

[5]田学东, 王文姣. 基于综合纠错的印刷体数学公式识别后处理[J]. 计算机工程与设计, 2007, 28(20): 5039-5041.
TIAN Xuedong, WANG Wenjiao. Post-processing for printed formula based on synthetic error correction[J]. Computer Engineering and Design, 2007, 28(20): 5039-5041.

[6]田学东, 李　娜, 徐丽娟. 印刷体数学公式结构分析方法的研究[J]. 计算机工程, 2006, 32(23): 202-204.
TIAN Xuedong, LI Na, XU Lijuan. Research on structural analysis of mathematical expressions in printed documents[J]. Computer Engineering, 2006, 32(23): 202-204.

[7]GARAIN U, CHAUDHURI B B. A syntactic approach for processing mathematical expressions in printed documents[C]// Proceedings of 15th International Conference on Pattern Recognition. Barcelona, Spain, 2000: 523-526.

[8]MITRA J, GARAIN U, CHAUDHURI B B. Automatic understanding of structures in printed mathematical expressions[C]//Proceedings of Seventh International Conference on Document Analysis and Recognition. Edinburgh, Scotland, 2003: 540-544.

作者简介:

史广顺,男,1978年生,副教授,硕士生导师,先后负责省部级科研项目 4项,参与省部级与国家级科研项目 10余项.主要研究方向为模式识别与机器智能、数字图像处理、自然语言理解、软件开发技术.

肖　萃,女,1984年生,硕士研究生,主要研究方向为文档图像处理、模式识别.

王庆人,男,1965年生,教授,博士生导师. 1989年创造性的实现基于熵分类的 OCR引擎,并于 1992～1994连续 3年获得美国 UNLV全球析与 OCR评比冠军.先后承担国家级、省部级项目 20余项,主要研究方向为模式识别与机器智能.发表 IEEE期刊论文 10余篇.