**Computer Science and Application** 计算机科学与应用, 2015, 5, 218-224
Published Online June 2015 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2015.56028

Hans汉斯

# Research Status of Mathematical Formula Recognition

**Dongming Liu[1,2], Lian Chen[1], Ming Li[1,2,3], Ju Zhang[3]**

[1]Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu Sichuan
[2]University of Chinese Academy of Sciences, Beijing
[3]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing
Email: dacenzon@163.com, 248690205@qq.com, liming@cigit.ac.cn, zhangju@cigit.ac.cn

Received: Jun. 3rd, 2015; accepted: Jun. 22nd, 2015; published: Jun. 25th, 2015

Copyright © 2015 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

## Abstract

**In order to search and edit the documents which contain mathematical formulas, we must automatically recognize the expression. Mathematical formula recognition is an active research field and many approaches have been proposed over the years. Nowadays, there are several forms of input data format such as document images, strokes, vector images and so on. Different ways of inputs determine the methods to extract mathematical formulas and different ways of mathematical formula recognition. This article describes the currently researching work of mathematical formula recognition, discusses the four components problems in mathematical formula recognition: the detection of expression, symbol recognition, structural analysis, interpretation and so on, and points out the future research directions of mathematical expressions.**

## Keywords

**Mathematical Formula Recognition, Research Status, Document Images, Strokes, Vector Images**

# 数学公式识别研究现状

**刘东明[1,2]，陈　联[1]，李　明[1,2,3]，张　矩[3]**

[1]中国科学院成都计算机应用研究所，四川 成都
[2]中国科学院大学，北京
[3]中国科学院重庆绿色智能技术研究院，重庆

Email: dacenzon@163.com, 248690205@qq.com, liming@cigit.ac.cn, zhangju@cigit.ac.cn

收稿日期：2015年6月3日；录用日期：2015年6月22日；发布日期：2015年6月25日

## 摘　要

**文档的编辑和检索要求能够自动识别数学公式，数学公式识别是一个活跃的研究领域，经过多年的发展提出了许多解决方法。公式的输入数据格式有文档图像、笔划、矢量图形、特殊语言等几种形式，不同的输入方式决定数学公式的提取和和识别方式的不同。本文介绍了数学表达式识别邻域的研究现状，讨论了表达的检测、符号识别、结构分析、语义分析等四部分的问题，并提出未来数学表达式的研究方向和热点。**

## 关键词

**数学公式识别，研究现状，文档图像，笔划，矢量图形**

## 1. 引言

数学公式广泛存在于各类文献之中，是科技文档重要的组成部分，最近识别 PDF 文档中的公式的需求与日俱增，但是公式的识别远比文字段落的识别困难。文档编辑环境中也要求能够对用户输入的各种格式的数学符号进行识别。Web 上含有越来越多的数学公式文档，由于数学公式有自己独特的结构，使用传统的自然语言搜索系统不容易处理这些公式，数学信息检索是数学公式识别领域的一个重要的研究热点。在数学信息检索中，文档集合可以使用含数学符号的查询语句进行检索。检索系统需要识别查询语句以及集合中的所有文档，而且必须注明出数学表达式的位置以及识别结果解释，这就对数学公式的识别提示出了新的要求。伴随着基于光学扫描和笔写输入等硬件设备的发展，公式识别系统的软件实现方面成为关键问题。

数学符号提供具有挑战性的模式识别问题，包括分词歧义，符号识别的挑战和意义的模糊性。数学表达式的识别分为如下几个阶段，分别是预处理、公式检测、符号识别、符号间的空间关系确定、逻辑关系确定、意义构造等。本文介绍了在数学表达式识别的研究现状，讨论表达的检测、符号识别、结构分析、数学内容的解释等四部分的问题，并提出未来数学表达式的研究方向和热点。

## 2. 数学公式识别概述

数学公式识别主要由字符识别以及符号之间结构关系的分析两个阶段组成。现今数学公式识别的 4 个重要方向：表达式定位、符号抽取和识别、结构分析、语义分析。

### 2.1. 公式定位

数学公式的输入方式主要包括键盘\鼠标输入、语音输入、手写输入等方式。特殊语言法和图形界面输入方法，都不及手写数学公式自然、简便。相应的数学公式的数据格式有 4 种形式：文档图像、笔划、PDF 等矢量图形、Latex 等特殊标记语言。不同输入方式的数学公式的提取和识别方式不同。经过近 50 年的研究，用于检测独立表达方法是相当成熟，内嵌公式的检测仍然是一个挑战。

**1) 文档图像公式定位**

在文档图像中检测出的表达式是页面分割问题的一部分，区分出页面区域包含的文本、表格、数学

公式以及图形图片。文档图片格式的数学公式字体规范、结构整齐，字符分割和识别都相对容易。独立公式独占一行与文本自然分开，而内嵌公式出现在文本行的中间。独立公式可以使用如高度、字符尺寸以及符号布局等信息与文本行区分出来。嵌入式表达式难以准确的检测出来，尤其是只含有比较少符号的表达式。一些方法利用光学字符识别系统，而其它方法仅使用几何特征定位表达式。

**2) 矢量图形公式定位**

从矢量图形文档如 PDF 文件中提取内容的处理与文档图像处理不同。以矢量图形为基础的文件格式如 PDF 不包含数学区域的明确划分信息。为了支持数学信息检索，需要提取符号然后自动在 PDF 文档中检测表达式的位置。对基于 PDF 等矢量格式表示的符号提取是容易的，可以使用符号位置和标签的编码信息区分歧义字符，基于矢量的表示需要一些进一步的处理以形成完整的符号。

**3) 笔式输入公式定位**

手写环境下的手势交互主要表现为通过笔书写一些广泛使用的符号或标识，来完成一定的意图表达，使用手势作为激活命令的一种特殊符号。事实上，笔划所记录的也就是手持笔在基于笔的应用，表达式经常用手势分割，通常情况下，手势给出了表达式部分或近似的范围。

## 2.2. 符号识别

数学公式识别可以划分为脱机识别和联机识别。脱机识别主要指识别扫描仪扫描得到的数学表达式图像或包含数学表达式的文档，包括印刷体和手写体 2 种情况；联机识别是指识别使用电子笔书写的表达式，主要是手写体数学表达式识别。对于印刷体公式，由于其结构形规则，一般的公式识别方法都比较适用，而手写公式的识别就要更多地考虑其拓扑结构，识别手写体公式的难度远远大于对文档扫描图像的识别。PDF 文档由于在信息储存和交换方面的广泛应用，如何对 PDF 文档中的数学公式进行识别成为很重要的研究方向。

数学符号不利于自动识别的特征：一些数学符号有多种含义，数学符号呈现出多种类型的歧义，运算符的含义有时隐含地用空间关系来表示；数学符号的字符集比较大，包括罗马和希腊字母、数字以及众多操作符；难以区分常见的噪声、逗号、点、和其他小的符号；字体、大小、粗体、斜体等使得数学符号的识别更为复杂的；字符的黏连；数学符号提供冗余度小，在许多情况下是不可能根据上下文来猜出符号的意思；在很多外观字符非常相似的情况下，为了消除歧义非常需要上下文信息，而有时可用信息比较少；在诸多学科领域产生的方言加上作者使用许多符号变量，完全结构化书写数学公式非常困难；数学符号由于其两维结构空间关系复杂，有六个空间关系；表达式的二维性质产生的复杂性，实际应用中从语料库寻找上下文识别表达式是不实际；在缺乏可用的训练数据，通过拼写检查不可能最终验证正确性；手写体公式具有较大的书写自由度，极大地增加了字符识别和公式结构分割的难度。

## 2.3. 结构分析

结构分析包括确定符号之间空间关系逻辑关系以及构造意义。健壮的系统能够将低字符识别率与高水平的结构分析相结合。数学表达式中的符号之间的关系在某些情况下是复杂的，判断一个数学表达式的空间结构是困难的，特别是手写的公式符号。即使能够正确的识别出数学公式符号，在某些情况下它们的空间关系也是非常模糊的。在结构分析阶段，符号之间的空间关系产生一个结构树，结构树提供足够的信息用于表达式的排版或基于表达式外观的搜索。很难通过简单的上下文的位置关系判断一些很复杂的空间操作符,需要借助于上下文的语义信息。由于大型公式矩阵以及方程组结构的复杂性使得对其进行的分析识别更加困难，例如要对上面的矩阵进行结构分析就相当复杂。

### 2.4. 语义分析

许多当今数学识别系统进行结构分析，但不能继续确定表达式的数学含义。这样的系统足够用于公式的排版，但不支持表达式与计算机代数系统的交互。为了支持表达式的评价，产生某种形式的表示逻辑关系和语义领域的运算符树是必要的。运算符树包含理解表达式含义所需的信息，可以用于数学表达式语义搜索。无论是通过手动或由计算机代数系统理解表达式，必须在运算符树中补充表达式中的变量及运算符的定义和值。使用符号结构信息和符号含义的相关信息创建一个运算符树，来表达的表达式的语义。由于数学符号的递归和嵌套结构，经常用上下文无关文法来定义合法的符号结构和运算符。确定的符号和结构的意义是困难的，特别是如果有限的上下文可用的时候。

## 3. 基于文档图像的公式识别研究

多年来对扫描的文档数学公式的识别，已经提出了许多方法。上下标关系数学公式中出现频繁又难于解决的特殊结构，容易与其它关系混淆。

Kumar 等[1]研究印刷体数学表达式的识别和多组件字符部分的标记.提出的方法由符号产生、结构分析、编码生成三个阶段组成。符号形成过程，形成多组件字符，利用空间关系处理依赖于上下文的符号标识，处理矩阵等多笔划数学公式以及枚举函数。提出了一个基于规则的方法，在规则的形成过程中采用基于空间关系的启发式规则，依照输入数据的不同采用不同规则。采用类似专家系统的规则同数据隔离方式。对单行和多行公式采用统一的处理方式。Yoo 等[2]针对图像字符分割提出了一种改进的递归投影轮廓切割法，采用深度优先搜索算法对分割字符进行整理，再采用双链法进行第二次整理。Amit Pillay [3]提出了一种使用多个结构识别算法的新方法，使用图形变换网，它是以函数为基础的网络系统，每个系统以图形作为输入，使用函数并产生一个图作为输出。图形变换网用于组合多个结构解析器和参数使用基于梯度的学习。

郭育生等[4]提出了一种基于多候选方法的数学公式识别系统。符号分割阶段采用 3 次动态规划方法对数学公式图像进行多候选公式符号切分。在结构分析阶段，采用层次结构方法产生符号间的结构关系，建立了符号的空间关系模型，然后数学公式的识别结果表示成 Latex 格式和 MathType 格式。肖建于[5]等提出了一种基于字符凸壳和模糊识别的字符空间关系判别方法。首先，对数学公式符号进行分类，然后对每一类采用字符凸壳对上下标关系进行判别，最后运用模糊理论对数学公式中符号的空间区域关系进行划分。

## 4. 基于文档图像的公式识别研究

由于笔式界面本身的非精确性，以往的模式识别技术并不能够解决笔式界面下的全部问题，还需要多种人机交互技术综合来解决。

Lei Hu [6]提出了递归基线提取算法用于符号结构分析，并将手写笔划作为输入数据。基线提取在改进的 LL(1)分析器中用于词法分析，当解析器要求沿当前基线输入最左边的或左到右的下一个符号时返回一组候选符号。候选符号被用来产生解析树的森林，返回排名最高的解析结果。隐马尔可夫模型(HMM)用于符号分类，和符号之间的水平邻接使用两个概率二次分类器，一个用于上行符号，另一个用于中心以及下行符号。MacLean 等[7]提出一个系统，它可以捕获所有对输入的可识别的解释，并将它们组织在解析森林中。如果排名第一的解释是不正确的，用户可以要求交替，并选择他们想要的识别结果。树提取步骤采用一种新颖的概率树评分策略，其中贝叶斯网络是基于输入的结构构成，各关节变量赋值对应于不同的语法分析树，然后为了降低概率产生解析树。

Le 等[8]使用上下文无关文法表示数学公式，并使用科克-雅戈尔-卡西米(CYK)算法来分析联机手写

数学公式的二维结构，并选择在符号分割，识别和结构分析过程中产生的最好的解释。通过使用两个 SVM 模型，从没有任何启发式决策的训练模式中学习结构关系。采用笔划顺序以减少解析算法的复杂性。Simistira 等[9]符号识别基于笔的方向特征的弹性模板匹配距离，结构分析是基于提取数学表达式的基线，然后符号分成上方和下方的基线水平。然后符号依次使用六个空间关系和各自的二维结构被处理。Álvaro 等[10]使用二维随机上下文无关文法和隐马尔可夫模型来识别联机手写数学表达式，隐马尔可夫模型被用来识别数学符号，随机上下文无关文法用于对这些符号之间的关系进行建模。该模型能够捕获许多的在训练过程中出现联机手写数学表达式的变异现象。分析过程仅仅使用随机的信息便可做出决定以及避免启发式决定。Awala 等[11]的系统，允许直接从完整的表达式学习数学符号和空间关系。提出了新的结合语法和结构信息的上下文模型，这些模型被用来找到二维分割方案中的分割/识别假设的最可能的组合，模型是基于关于所述符号的布局结构信息。MacLean 等[12]提出了一种使用关系语法和模糊集解析二维输入的新方法。根据公式的二维结构，设计了一种快速增量解析算法。检查模糊集合同手写输入之间的相似性以及隶属度。应用并改进现有矩形划分和共享解析森林等的技术，并引进如关系类和互换性等新的思想。提出一种在识别错误或模糊的情况下允许用户浏览解析结果并选择正确的解释的修正机制，然后将这样的修改纳入后续的增量识别结果。Hirata 等[13]基于表达式匹配提出了一种新的方法。在符号输入过程中使用手工标注模型表达匹配符号模型，将匹配过程归结为一个图的匹配问题，考虑表达式的局部和全局特征计算出两两匹配代价。

Yang Hu 等[14]提出了一个新的框架来分析手写数学表达式的布局和语义信息。符号分割阶段，笔划被分解，然后使用动态规划来寻找对应于最佳分割方式的路径，并降低了笔划的搜索复杂性。符号识别阶段，空间的几何形状和方向元件的特性，使用通过最大期望算法学习的高斯混合模型进行分类。语义关系分析阶段，一个三叉树被用于存储通过计算操作优先级来排序的符号。Frank 等[15]侧重于数学符号识别的问题，提取完整的手写数学公式的符号假设，训练的人工神经网络对正确的假设和拒绝错误假设进行符号分类。提出了一个新的基于上下文的符号描述的形状：模糊形状上下文。

卢晓卫等[16]提出了一种基于分块树的结构分析方法，使用分块处理方式，根据数学表达式内部结构特征将数学表达式分解为若干子模块，然后采用树型结构对每个子模块内部字符之间的结构关系进行表示，最终形成整个表达式的树型表示。定义了一系列的再现了公式的结构特征的字符结构属性作为结构分析的结果。陈临强等[17]对手写输入符号进行有效的预处理，获取并简化字符笔画区域码序列以及字符之间的位置关系，然后应用模糊矩阵的区域码序列匹配运算进行字符匹配，在匹配过程中提供字符的自学习能力，最后通过基于数学公式规则的语法分析，实现了数学公式的识别。

## 5. 基于矢量图形的公式识别研究

从 PDF 文档中识别数学表达式是文档分析一个新的重要领域，它与图像文件中提取数学表达式大不相同。现在关于 PDF 格式有很多研究，但是很少方法利用 PDF 格式的特点来提高公式识别的准确性、可靠性和速度。有些系统可以直接从文档中提取数学公式使用的字符信息，这样可以避免传统的 OCR 方法引入的模糊，还可以利用现有的公式识别技术来获得正确的结果和高性能。

Baker 等[18]应用了两个阶段的方法解析公式，第一阶段在 Anderson 原先的线性方法基础上，为了克服它的不足进行扩展和改进，将 2 维数学公式转换成线性表示。然后使用标准 Yacc-style LALR 分析器，将结果表达式转换成一个抽象语法树。最后对这棵树进行遍历产生 Latex 结果输出。

田学东等[19]提出了一种直接从 PDF 文件提取数学组件的方法。与传统的基于图像的方法相比，该方法充分利用了 PDF 文档中如字体大小、基线、字符包围盒等的内部信息，提取的数学符号及其几何信息。该方法能满足公式结构分析处理、重建和检索的需要，并且具有更高的效率。林晓燕等[20]提出了一

种结合基于规则和基于学习的方法来检测 PDF 文档中的独立与内嵌数学公式的新方法。此外，数学公式的各种特征，包括几何布局、字符和上下文内容，用于识别多种不同类型公式。

## 6. 结束语

数学公式的识别研究到今天，经过研究者们的努力，识别技术日渐成熟，今后的工作包括：提高分割算法，既用于检测内嵌公式，以及手写表达式符号的分割。通过使用更可靠的符号结构模型来提高结构分析的可靠性。提高矩阵处理的可靠性和灵活性。这可能通过结合的现有结构分析方法来产生整体效果的解析器的开发来实现。对数学公式识别几个阶段进行整合，整合可以充分利用上下文信息有效利用减少识别错误。现在已经有多种识别系统集成方法，尤其在语法和动态处理方面，系统的集成有很大的研究前景。开发可以识别和理解数学符号方言的识别系统。提高向用户显示识别结果的便利性和有效性。研究公式的逻辑结构和公式的含义，使得公式能够自动化输入；增加的多种输入方式的使用，如允许用户既可以语音输入也可以手写输入。提高数学识别系统的性能预测。数学公式的识别和检索整合是一个重要趋势，这就要求具有便利的表达式查询界面，以及有效利用嘈杂的识别结果的检索算法；像 C 编译器检测源文件的语法错误一样，我们希望在数学公式 OCR 输出中有一种发现错误验证机制。

## 基金项目

科学技术部国家科技支撑计划“教育云服务关键技术攻关”项目资助(2013BAH72B01)。

## 参考文献 (References)

[1] Pavan Kumar, P., Agarwal, A. and Bhagvati, C. (2011) A rule-based approach to form mathematical symbols in printed mathematical expressions. *Lecture Notes in Computer Science*, **7080**, 181-192.

[2] Yoo, Y.-H. and Kim, J.-H. (2013) Mathematical formula recognition based on modified recursive projection profile cutting and labeling with double linked list. *Advances in Intelligent Systems and Computing*, **208**, 983-992.

[3] Amit, P. (2014) Intelligent combination of structural analysis algorithms: application to mathematical expression recognition. Rochester Institute of Technology. http://scholarworks.rit.edu/theses/7874

[4] 郭育生, 黄磊, 刘昌平 (2007) 基于多候选的数学公式识别系统. *计算机研究与发展*, **44**, 1144.

[5] 肖建于, 王潜平, 洪留荣 (2008) 基于凸壳和模糊识别的数学公式识别. *计算机应用与软件*, **29**, 208.

[6] Hu, L. (2012) Baseline extraction-driven Parsing of handwritten mathematical expressions. 21*st International Conference on Pattern Recognition* (*ICPR*), 11-15 November 2012, 326-330.

[7] MacLean, S. and Labahn, G. (2014) A Bayesian model for recognizing handwritten mathematical expressions. *Thu*, 18 September 2014 14:45:24 GMT.

[8] Le, A.D., Van Phan, T. and Nakagawa, M. (2014) A system for recognizing online handwritten mathematical expressions and improvement of structure analysis. 11*th IAPR International Workshop on Document Analysis Systems* (*DAS*), 7-10 April 2014, 51-55.

[9] Simistira, F., Papavassiliou, V., Katsouros, V. and Carayannis, G. (2012) A system for recognition of on-line handwritten mathematical expressions. *ICFHR* 12 *Proceedings of the* 2012 *International Conference on Frontiers in Handwriting Recognition*, 193-198.

[10] Álvaro, F., Sánchez, J.-A. and Benedí, J.-M. (2014) Recognition of on-line handwritten mathematical expressions using 2D stochastic context-free grammars and hidden Markov models. *Pattern Recognition Letters*, **35**, 58-67.

[11] Awala, A.-M., Mouchèreb, H. and Viard-Gaudinb, C. (2014) A global learning approach for an online handwritten mathematical expression recognition system. *Pattern Recognition Letters*, **35**, 68-77.

[12] MacLean, S. and Labahn, G. (2013) A new approach for recognizing handwritten mathematics using relational grammars and fuzzy sets. *International Journal on Document Analysis and Recognition* (*IJDAR*), **16**, 139-163.

[13] Nina, S., Hirata, T. and Honda, W.Y. (2011) Automatic labeling of handwritten mathematical symbols via expression matching. Graph-based representations in pattern recognition. *Lecture Notes in Computer Science*, **6658**, 295-304.

[14] Hu, Y., Peng, L.R. and Tang, Y.J. (2014) On-line handwritten mathematical expression recognition method based on statistical and semantic analysis. 11*th IAPR International Workshop on Document Analysis Systems* (*DAS*), 7-10 April

2014, 171-175.

[15] Julca-Aguilar, F., Hirata, N., Viard-Gaudin, C., Mouchère, H. and Medjkoune, S. (2014) Mathematical symbol hypothesis recognition with rejection option. 14*th International Conference on Frontiers in Handwriting Recognition*, Crete, 500-504.

[16] 卢晓卫, 林嘉宇 (2010) 一种基于分块树的手写数学公式结构分析算法. *计算机工程与科学*, **23**, 69.

[17] 陈临强, 李云霞, 沈俊 (2009) 联机手写数学公式识别系统的设计和实现. *杭州电子科技大学学报*: *自然科学版*, **29**, 36.

[18] Baker, J.B., Sexton, A.P. and Sorge, V. (2009) A linear grammar approach to mathematical formula recognition from PDF. *Lecture Notes in Computer Science*, **5625**, 201-216.

[19] Yu, B.T., Tian, X.D. and Luo, W.J. (2014) Extracting Mathematical components directly from PDF documents for mathematical expression recognition and retrieval. *Lecture Notes in Computer Science*, **8795**, 170-179.

[20] Lin, X.Y., Gao, L.C., Tang, Z., Lin, X.F. and Hu, X. (2011) Mathematical formula identification in PDF documents. *International Conference on Document Analysis and Recognition* (*ICDAR*), 1419-1423.