

图片中印刷体数学公式的自动识别

MathOCR 0.0.2 的设计与实现

陈颂光

1m02math@126.com

中山大学数学与计算科学学院 2011 级数学与应用数学

2015 年 5 月 7 日

摘要

未能识别数学公式妨碍了科技文献的电子化，故有必要研究数学公式识别技术。数学公式识别可分为符号识别和结构分析两个主要步骤。在符号识别中，除横线和圆点用经验规则判断外，大部分字形可以采用粗分类、细分类到模板匹配的流程，接着利用位置和识别结果合并分体符号的各个字形，对于特殊符号则可用动态生成模版方法匹配。在结构分析中，利用自底向上从局部到整体逐步进行合并的方法，已经支持角标、帽子、分式、根式、矩阵、多行表达式等主要的数学公式结构，并可继续扩充。实验表明此法可以对于高质量图片给出不错的识别效果。

〔关键词〕 数学公式识别；字符识别；结构分析

内容纲要 I

- 1 背景
- 2 图形预处理
- 3 数学符号识别
 - 字形分割
 - 字形识别
 - 字形合并
- 4 结构分析
 - 数据结构
 - 算法
- 5 总结

背景

重要性

- 现存科技文献中的大量数学公式保存于不便于再次利用的形式，导致了很多繁杂且容易出错的重复输入工作
- 为了整合和盘活数学公式资源，有必要建立一种有效机制把现存的数学公式转换为一种统一、便于重用的形式
- 这样不但可节省重复输入数学公式的繁琐工作，同时可为数学公式的搜索和相应的进一步处理提供可能的基础
- 当前的实际需要主要是处理较成熟的出版物，故优先考虑印刷体而非手写体
- 印刷体数学公式识别对于科技文献电子化以至科学技术传播有重要意义

背景

研究现状

- Anderson 在他于 1968 年的博士论文已经提出了数学公式识别问题 [1]
- 数学公式识别分为数学符号识别和结构分析两个主要步骤 [2]
- 数学符号识别可作为字符识别的特例，而后者获广泛研究且已实用化 [3]
- 20 世纪 70 年代的工作集中于建立完备的文法，20 世纪 80 年代的工作则集中于特定类型数学公式的识别 [1]
- 20 世纪 90 年代以来，公式识别的研究热度日益增加 [2]
- 虽然发表了不少论文，但可获取的数学公式识别系统依然是稀缺品

背景

难点

- 由于存在大量数学符号，而且需要区分字体，这使得分类数较大
- 数学公式中经常存在许多形状极为相似甚至相同的符号（例如分数线、减号、上划线、下划线形状相同），难以区分
- 数学公式是一种平面结构，符号间可能存在多种位置关系，并且局部误识容易导致全局错误
- 不同领域、不同作者有不同的符号体系，这限制了通用识别系统中语义信息的使用

图形预处理

目的和方法

- 由于原始图像自身的瑕疵和转换过程中产生的失真，待识别图像往往存在一些质量问题
- 这些质量问题包括但不限于墨点、纸张纹理、渗透、光照、纸张倾斜、纸张弯曲等原因造成的干扰
- 图形预处理的目的是使图像更能反映数学公式的原貌，为后续识别过程创造良好的条件
- 数字图像处理领域提供了现成的解决方法，可以直接选择简单且被推荐的方法加以实现

图形预处理

采用手段

- 为消除背景噪声，对图像进行二值化。本系统把图像先化为灰度图像再用以下方法之一化为二值图像：
 - Otsu 方法（一种全局阈值化方法）[4]
 - Sauvola 方法（一种局部阈值化方法）[5]
- 为消除椒盐噪声，可以应用低通滤波器。本系统提供了以下可选的滤波器：
 - 均值滤波器 [1]
 - 中值滤波器 [1]
 - kFill 滤波器 [6]
- 倾斜校正和卷曲校正更适合于对整个页面而非单个数学公式进行，故本系统暂不考虑这些问题

数学符号识别

目的与方法

- 数学符号识别的目的是确定数学公式中各符号是什么和在什么位置
- 可以借助常规字符识别的技巧，仅改造不适合数学符号识别之处
- 本系统以字形为基本识别单位
- 三个步骤分别是
 - ① 字形分割
把数学公式分割为比符号更细的单位
 - ② 字形识别
找出与各个字形接近的已知字形
 - ③ 字形合并
把字形合并为符号

数学符号识别

字形分割

- 注意到不同符号间互不连通，故作 8 连通域分割
- 连通域分割基于游程编码和图遍历（可改为采用带路径压缩和按秩合并的不相交集合算法）
- 为减少待匹配的单元数，把一些明显属于同一字符的连通域合并为字形
- 如果两个连通域物理矩形之交的面积大于面积较小者面积的若干分之一，并且两个连通域都未有被判为根号，则这两个连通域被认为属于同一字形
- 使用字形为识别单位适合于处理常有斜体字符和二维结构的数学公式

数学符号识别

常规字形识别

- 每个字形按被经验规则判为圆点、横线还是其它来分配初始候选集
- 依次应用一系列匹配器来筛除一些不合适的候选
- 分别利用低阶矩、投影、孔洞数、高宽比、穿线数和网格特征 [1, 7] 这几个较直观、容易计算的特征构造了匹配器
- 为了完全利用信息，使用基于 Hausdorff 距离 [1] 的模版匹配作为验证和多候选排序的主要依据

数学符号识别

特殊字形识别

- 数学公式中有些符号不能仅由字体模板通过缩放得到，还涉及局部延长 [8]。这些符号包括定界符（如 “|”、“(”、“)”、“[”、“]”、“{”、“}”、“[”、“]”、“[”、“]”、“<”、“>”）、根号（如 $\sqrt{\quad}$ ）、箭头（如 “ \rightarrow ”）和水平括号（如 “ \frown ”、“ \smile ”）
- 为了识别这些符号，使用基于缩放不变特征的方法是不合适的
- 人手设计可以惟一标识各个特殊字形的结构特征会带来的繁琐工作和与常规字形识别方法的不协调性
- 本系统使用动态生成模板的方法，即根据待识别字形的特点（主要是高宽比）生成各个特殊字形的模板，然后让待识别字形与各个特殊字形模板进行模板匹配
- 为了节省开销，只有在待识别字形按常规方式得不到理想的识别或高宽比悬殊时才启用特殊符号识别

数学符号识别

字形合并

- 在进行字形识别后，需要利用位置和识别结果合并分体符号的各个字形
- 对属分体符号的一部分的候选字形，检查该分体符号其它字形应在的位置是否有字形：
 - 如全部出现，则计算候选符号的像素矩阵和各个字形合并所得像素矩阵间的 Hausdorff 距离以代替候选的距离，再删去其它对应字形中相应于该分体字的候选
 - 如不全出现，删去此候选
- 可以生成多候选，以便把一些困难的选择留待结构分析阶段处理

数学符号识别

性能评估

- 利用的 AMSFonts 字体中的符号作为训练集和测试集
- 训练字体大小为 40
- 二值化方法选为 Sauvola 方法，粗分类方法为高宽比，距离判别器使用网格特征和投影
- 利用一个自动化测试程序生成完美的孤立符号图片供识别并统计第一候选给出正确识别结果的比例
- 对于字母，由于字体常影响含义，必须连字体也正确识别才算正确识别；而对于非字母的符号，并不区分字体

数学符号识别

性能评估

字体	大小			
	40	30	20	10
CMB10	123/125	106/125	97/125	34/125
CMBSY10	126/126	120/126	104/126	55/126
CMEX10	95/95	82/95	77/95	52/95
CMMI10	110/110	98/110	77/110	3/110
CMMIB10	111/111	99/111	69/111	25/111
CMR10	126/128	99/128	62/128	5/128
CMSY10	125/126	115/126	106/126	29/126
MSAM10	114/114	105/114	93/114	24/114
MSBM10	85/85	80/85	68/85	3/85
RSF10	26/26	25/26	18/26	0/26
总计	1041/1046	929/1046	771/1046	230/1046

结构分析

目的与方法

- 结构分析的目的在于符号识别的基础上根据符号间位置关系重组出数学公式的结构
- 本系统采用自底向上的合并策略
- 结构分析的输入不一定依赖于上节的符号识别技术，例如可来自解析 PDF 文件

结构分析

数据结构

- 盒子这个数据结构被用于表示子公式
- 每个盒子记录排版代码、基准点位置、逻辑高度和逻辑宽度、上升和左偏移、参照大小、基线确定程度
- 不同的子公式结构可以用不同的盒子类型来表示，盒子类型可以有符号型、同行型、分式型、根式型、帽子型、大型操作符型、多行型，并可继续扩充
- 合并算法在盒子的集合上进行

结构分析 I

算法

① 初始化

对于每个符号，创建一个符号型盒子

② 合并

① 合并基线确定且一致的盒子

如有两相异盒子基线一致、水平距离较小且它们共同最小外接矩形不与其它盒子相交，则两盒子被认为是同行相邻两部分而合并，按各自基准点横坐标决定左右，然后回到合并步继续

② 合并基线确定的上下标

如有两相异盒子基线不一致但差异在一定范围内、水平距离较小且它们共同最小外接矩形不与多于一个其它盒子相交，则一盒子被认为是另一盒子的上标或下标而合并，按各自基准点横坐标决定左右，按各自参照大小和基线决定上下，然后回到合并步继续

③ 合并基线不确定的行内盒子

与第 (a)、(b) 步类似

④ 合并特殊符号

结构分析 II

算法

- ① 帽子合并
对于上划线、下划线和向量箭头，如所管辖区恰有一个别的盒子，则把有关盒子合并；对于上花括号和下花括号，如果主管辖区恰有一个其它盒子，副管辖区至多有一个其它盒子，则把有关盒子合并，然后回到合并步继续
- ② 分式合并
若分子部分和分母部分分别恰有一个别的盒子，则把有关盒子合并，然后回到合并步继续
- ③ 根号合并
若根号内部恰有一个盒子而次数位置至多有一个盒子，则把有关盒子合并，然后回到合并步继续
- ④ 大型操作符合并
若下标域恰有一个盒子而上标域至多有一个盒子，则把有关盒子合并，然后回到合并步继续

结构分析 III

算法

- ⑤ 合并行
如有两相异盒子垂直距离较小且它们共同最小外接矩形不与其它盒子相交，则两盒子被认为是相邻两行而合并，按各自基准线决定上下，然后回到合并步继续
- ③ 如只剩下一个盒子，算法结束并得到整个数学公式的识别结果；否则，得到数学公式各个组成部分的识别结果

结构分析

性能评估

对由两本数学公式排版教程中数学公式转换得到的图片进行识别，用人手判断识别结果是否准确

类型	样本个数	结构分析准确个数	准确率
简单表达式	39	30	76.9%
含根号表达式	13	11	84.6%
含分式表达式	23	19	82.6%
含帽子表达式	22	15	68.2%
含大型操作符表达式	24	16	66.7%
含矩阵表达式	20	14	70.0%
多行表达式	35	30	85.7%
总计	147	113	76.9%

结构分析

性能评估

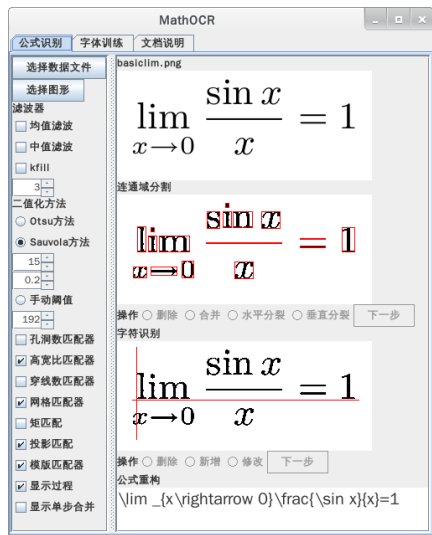
导致错误的原因包括：

- 上下标关系误判
- 存在符号识别错误
- 存在不支持的数学公式结构

总结

主要成果

最重要的工作在于用 Java 语言实现了一个演示性质的数学公式识别系统 MathOCR 0.0.2。



总结

创新点

- 使用字形为基本识别单位
- 使用动态生成模板方法把特殊字形识别问题化为常规字形识别问题
- 采用了对基线、逻辑高度、逻辑宽度的精确估计
- 避免作人为的假定，未有显式地使用任何统计模型或形式文法

总结

不足

- 准确程度有待提高
- 代码的容错性、自解释性和正交性并未达到工业界水平
- 沒有严谨的理论支撐，有点 ad hoc

总结

后续工作

可以继续改进的方向包括：

- 更细致的预处理
- 加入对粘连和断裂字形的处理
- 重新设计符号识别算法
- 扩充符号数据库
- 对交换图的识别
- 对识别结果的自动验证
- 对识别结果的自动修正





总结

应用前景

数学公式识别技术可能可用于以下方面：

- 科技文档识别
- 数学公式检索
- 论文中公式的重合率检测
- 知识管理
- 化学公式识别

主要参考文献 I

-  王科俊, 冯伟兴.
中文印刷体文档识别技术 [M].
北京: 科学出版社, 2010.
-  CHAN K F, YEUNG D Y.
Mathematical expression recognition: a survey[R].
Hong Kong: HKUST, 1999.
-  MOHAMED C, NAWWAF K, LIU C L, et al.
Character recognition systems[M].
Hoboken: Wiley-Interscience, 2007.
-  BURGER W, BURGE M J.
数字图像处理: Java 语言算法描述 [M].
北京: 清华大学出版社, 2010.

主要参考文献 II

-  SAUVOLA J, PIETIKÄINEN M.
Adaptive document image binarization[J].
Pattern Recognition, 2000, **33**(2):225–236.
-  O'GORMAN L.
Image and Document Processing Techniques for the RightPages
Electronic Library System[A].
Proceedings of the International Conference on Pattern
Recognition[C], Los Alimitos: IEEE, 1992:260–263.
-  TRIER D, TAXT T, JAIN A K.
Feature extraction methods for character recognition - A survey[J].
Pattern Recognition, 1996, **29**(4):641–662.

主要参考文献 III



KNUTH D E.

The T_EXbook[M].

Reading: Addison-Wesley, 1986.

相关资源

本课件（未删节版）、总结报告全文、程序（包括二进制包和源代码）可以在<http://mathocr.sourceforge.net/>获得。



谢谢大家！