# 印刷体科技文档识别技术实践研究

*MathOCR 0.0.3 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 17 日

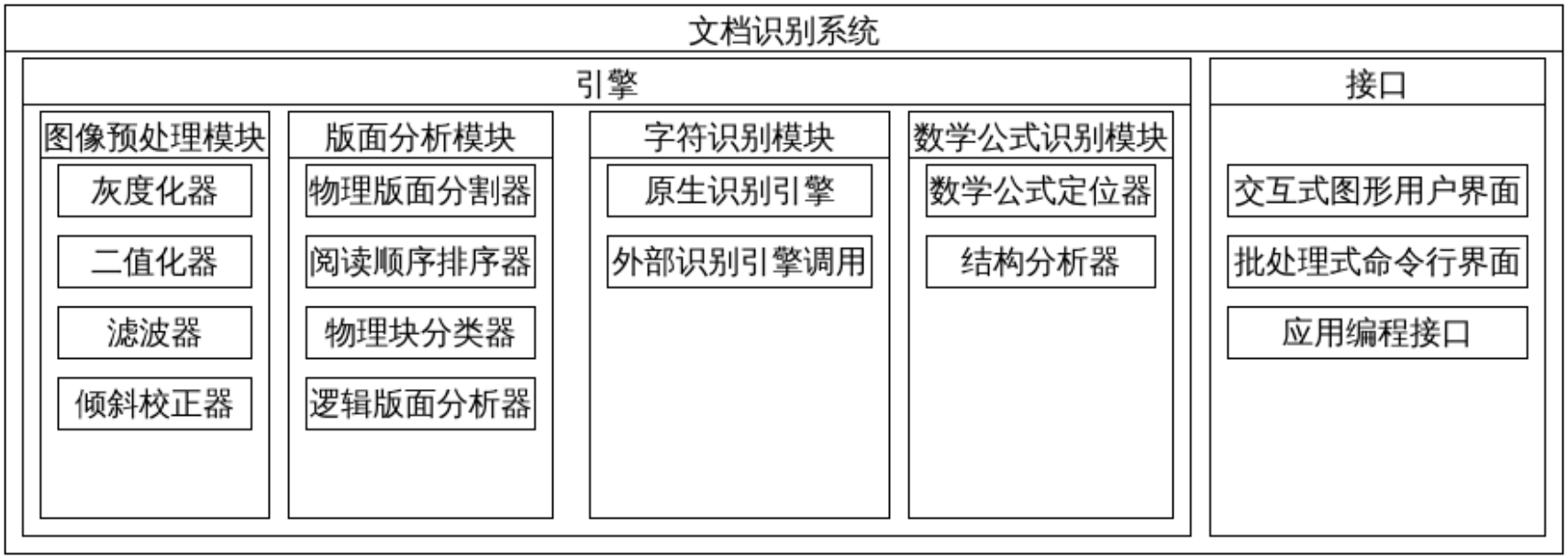
# 引言

## 背景

- 纸质版或扫描版的出版物不便于被发现和重用
- 有必要建立一种有效机制把现存的文献转换为便于编辑和处理的形式
- 现有的文档识别系统大多不支持数学公式识别，逻辑版面分析能力也不足够
- 开发科技文档识别系统对科学技术传播有现实意义

# 引言

## 研究范围


文档识别系统
引擎
图像预处理模块
灰度化器
二值化器
滤波器
倾斜校正器
版面分析模块
物理版面分割器
阅读顺序排序器
物理块分类器
逻辑版面分析器
字符识别模块
原生识别引擎
外部识别引擎调用
数学公式识别模块
数学公式定位器
结构分析器
接口
交互式图形用户界面
批处理式命令行界面
应用编程接口

# 图形预处理

## 目的与方法

本阶段使用数字图像处理的技巧以求尽可能恢复文档的原貌

- 交互式地确定页框的机制被用于去除边缘噪声
- 二值化被用于去除背景噪声，其中 Sauvola 方法为首选，Otsu 方法为备选
- 滤波器被用于去除椒盐噪声，可选的滤波器有均值滤波器、中值滤波器、kFill 滤波器、二值化后处理等
- 交互式地修改连通域集合的机制被用于去除不规则噪声
- 投影方法为默认的倾斜检测方法，也支持另外 6 种方法

# 图形预处理

## 实例

### 灰度化后图像

The Initiative Application of Tool Like SCIgen

Nameless Robot , Bye Another Earth and Hello World

**Abstract**

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. Since almost all academic paper of these days are meaningless, papers generated by SCIgen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be saved for that.

**1 Introduction**

SCIgen uses a hand-written context-free grammar to form all elements of the papers. The aim is to maximize amusement, rather than coherence. One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to SCI 2005! Our plan is:

- fill up all journals with auto-generated articles;
- Break all myth on scholar.

**2 Methodology**

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the whole academic industry. Sooner or later(if not currently), all journal and conference papers will be completely meaningless. To uncover the sacred veil of scholars, we should tell to truth, no dream is good dream. Or even better, no one will trust education anymore, which bring significant benefit to our society.

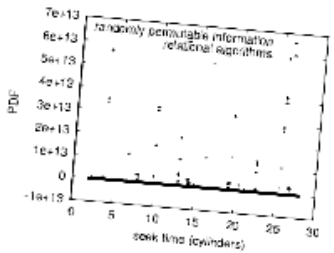



Figure 1: A figure generated by SCIgen

### 二值化后图像

The Initiative Application of Tool Like SCIgen

Nameless Robot , Bye Another Earth and Hello World

**Abstract**

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. Since almost all academic paper of these days are meaningless, papers generated by SCIgen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be saved for that.

**1 Introduction**

SCIgen uses a hand-written context-free grammar to form all elements of the papers. The aim is to maximize amusement, rather than coherence. One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to SCI 2005! Our plan is:

- fill up all journals with auto-generated articles;
- Break all myth on scholar.

**2 Methodology**

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the whole academic industry. Sooner or later(if not currently), all journal and conference papers will be completely meaningless. To uncover the sacred veil of scholars, we should tell to truth, no dream is good dream. Or even better, no one will trust education anymore, which bring significant benefit to our society.

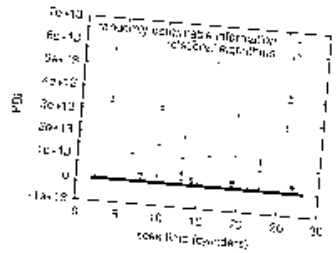



Figure 1: A figure generated by SCIgen

# 图形预处理

## 实例

### 倾斜校正后图像

The Initiative Application of Tool Like SCIgen

Namcless Robot , Bye Another Earth and Hello World

**Abstract**

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. Since almost all academic paper of these days are meaningless, papers generated by SCIgen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be saved for that.

**1 Introduction**

SCIgen uses a hand-written context free grammar to form all elements of the papers. The aim is to maximize amusement, rather than coherence. One useful purpose for such a program is to auto generate submissions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to SCI 2005! Our plan is:

- fill up all journals with auto-generated articles;
- Break all myth on scholar.

**2 Methodology**

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the whole academic industry. Sooner or later(if not currently), all journal and conference papers will be completely meaningless. To uncover the sacred veil of scholars, we should tell to truth, no dream is good dream. Or even better, no one will trust education anymore, which bring significant benefit to our society.

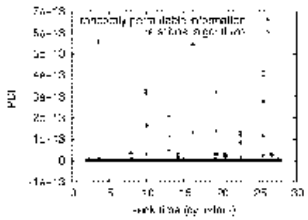
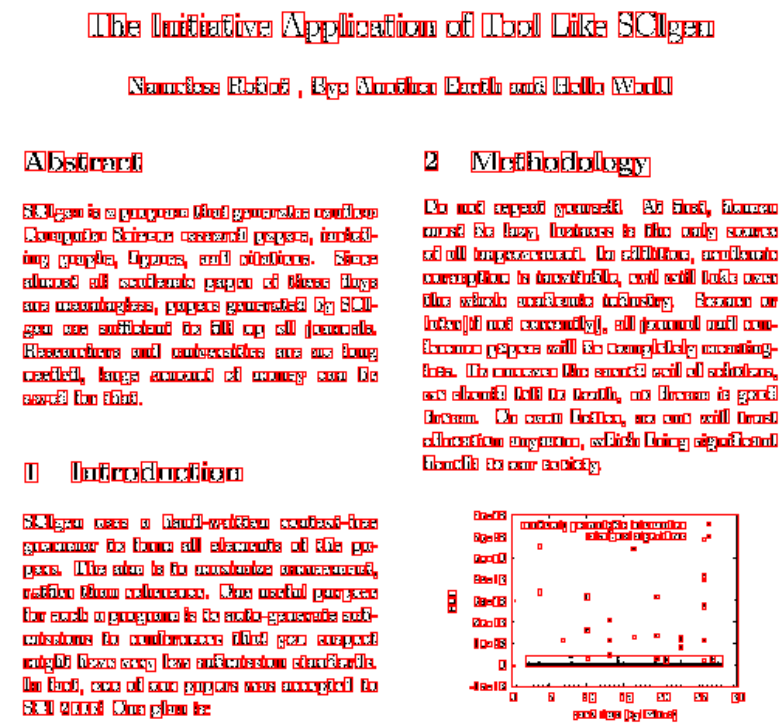


Figure 1: A figure generated by SCIgen

### 连通域分割结果



The Initiative Application of Tool Like SCIgen

Namcless Robot , Bye Another Earth and Hello World

Abstract

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. Since almost all academic paper of these days are meaningless, papers generated by SCIgen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be saved for that.

1 Introduction

SCIgen uses a hand-written context-free grammar to form all elements of the papers. The aim is to maximize amusement, rather than coherence. One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to SCI 2005! Our plan is:

- fill up all journals with auto-generated articles;
- Break all myth on scholar.

2 Methodology

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the whole academic industry. Sooner or later(if not currently), all journal and conference papers will be completely meaningless. To uncover the sacred veil of scholars, we should tell to truth, no dream is good dream. Or even better, no one will trust education anymore, which bring significant benefit to our society.

Figure 1: A figure generated by SCIgen

# 图形预处理

## 实验结果

利用 2013 年文档分析与识别国际会议（ICDAR）的文档图像倾斜检测竞赛（DISEC）基准数据集中的 1550 个二值文档图像对七种倾斜检测算法进行性能评估的结果：

| | 成功返回率 | 平均误差（弧度） | 均方误差（弧度） | 误差中位数（弧度） | 运行时间中位数（毫秒） |
|---|---|---|---|---|---|
| 参照方法 | **100.00%** | 0.1298 | 0.1504 | 0.1286 | **0** |
| 分片填涂方法 | 93.48% | 0.0621 | 0.1947 | 0.0061 | 50 |
| 分片覆盖方法 | 99.10% | 0.0154 | **0.0299** | 0.0073 | 300 |
| 投影方法 | 99.35% | **0.0068** | 0.0335 | **0.0033** | 197 |
| 交错数方法 | 99.35% | 0.0101 | 0.0458 | 0.0035 | 198 |
| Hough 变换方法 | 99.35% | 0.1448 | 0.1827 | 0.0452 | 99 |
| 行间相关方法 | 93.81% | 0.2115 | 0.3811 | 0.0346 | 1615 |
| 最近邻聚类方法 | 96.71% | 0.0883 | 0.1207 | 0.0690 | 67 |

# 版面分析

## 目的与方法

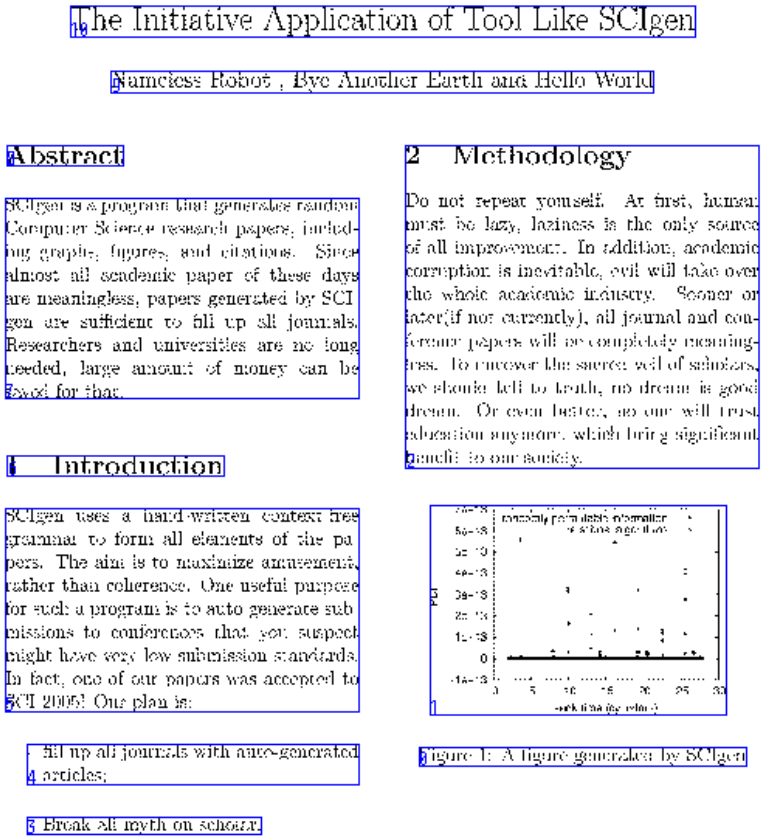
本阶段逐步把文档分解为各种逻辑块并决定它们的先后顺序

1. 物理版面分析把文本和非文本区域分开，并把不同栏中的文本也分开，同时不切开行
   - 基于递归投影切分的方法被选作物理版面分割方法
   - 基于连通域高度分布的方法被用于区分文本块与非文本块
   - 经本文修正、基于拓扑排序的方法被用于阅读顺序排序
2. 逻辑版面分析把文本块进一步分解为更细的逻辑单元如题名、作者、标题、段落和列表项目
   - 横向投影被用于提取行
   - 行对齐方式被用于生成逻辑块
   - 通过行对齐判断和对识别结果作正则表达式匹配来区分逻辑块类型

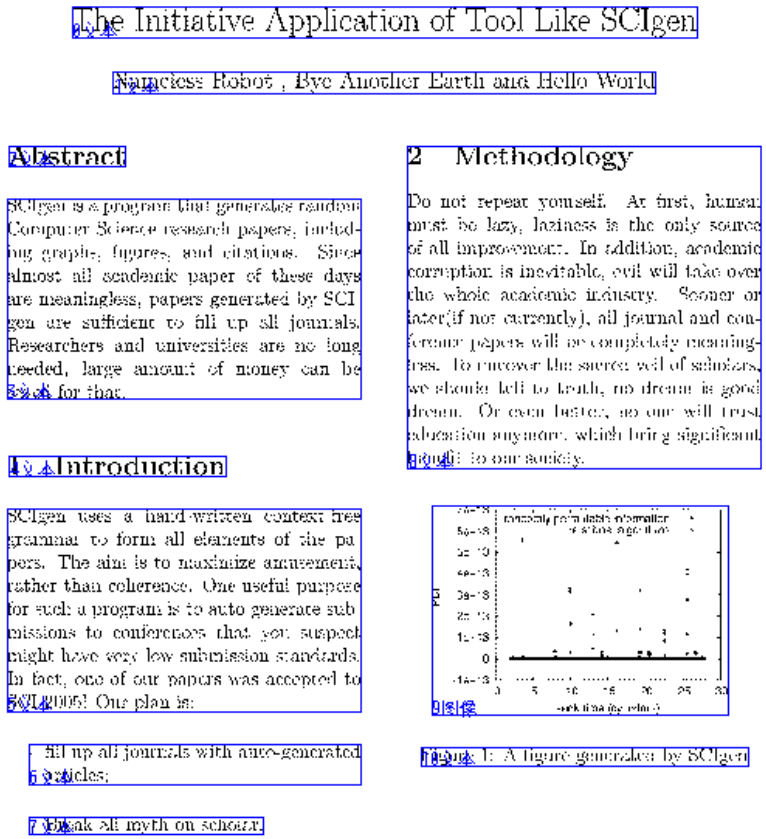
# 版面分析

## 实例

### 物理版面分割结果



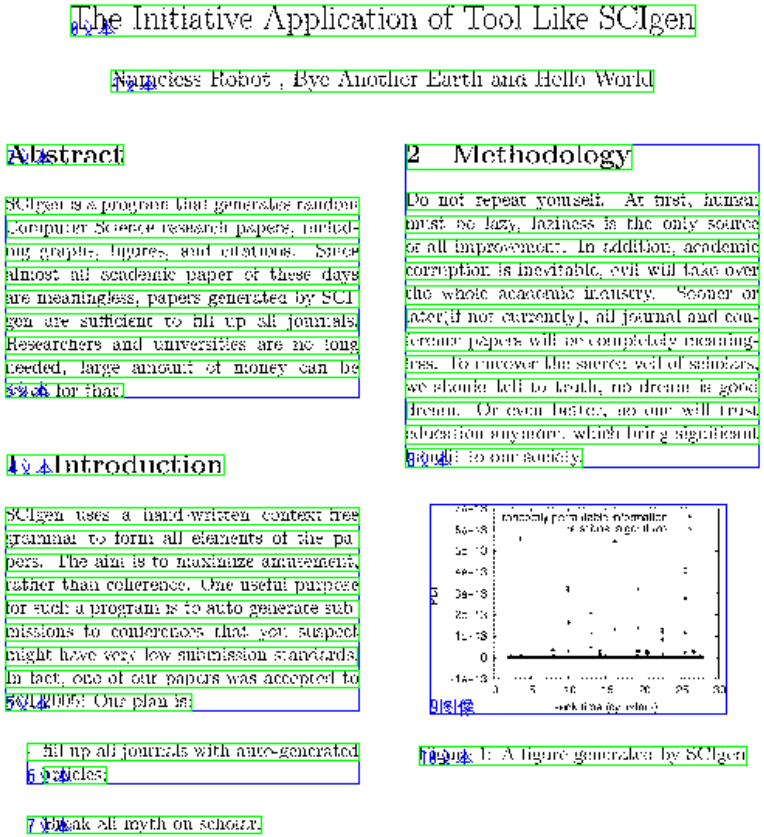

### 块分类及阅读顺序排序结果



# 版面分析

## 实例

### 文本行提取结果



The Initiative Application of Tool Like SCIgen

Narcicss Robot , Bye Another Earth and Hello World

Abstract

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. Since almost all academic paper of these days are meaningless, papers generated by SCIgen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be for that.

2 Methodology

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the whole academic industry. Sooner or later(if not currently), all journal and conference papers will be completely meaningless. To uncover the sacred veil of scholars, we should tell to truth, no dream is good dream. Or even better, no one will trust education anymore, which bring significant to our society.

1 Introduction

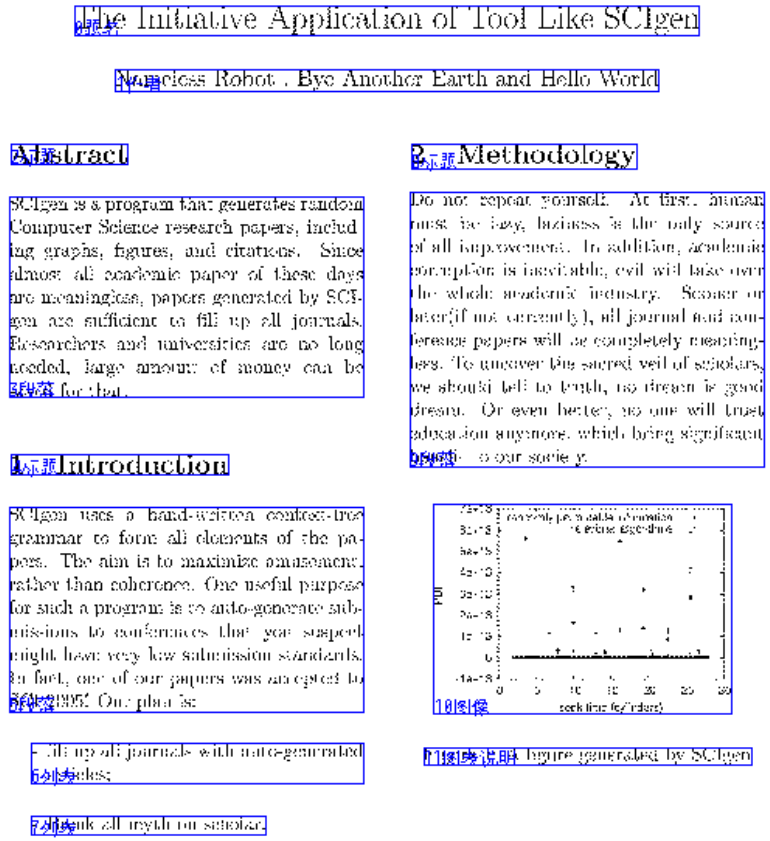
SCIgen uses a hand-written context-free grammar to form all elements of the papers. The aim is to maximize amusement, rather than coherence. One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to 2005! Our plan is:

- fill up all journals with auto-generated articles;

7 Rank all myth on schoolr.

Figure 1: A figure generated by SCIgen

### 逻辑版面分析结果



The Initiative Application of Tool Like SCIgen

Narcicss Robot , Bye Another Earth and Hello World

Abstract

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. Since almost all academic paper of these days are meaningless, papers generated by SCIgen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be for that.

2 Methodology

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the whole academic industry. Sooner or later(if not currently), all journal and conference papers will be completely meaningless. To uncover the sacred veil of scholars, we should tell to truth, no dream is good dream. Or even better, no one will trust education anymore, which bring significant to our society.

1 Introduction

SCIgen uses a hand-written context-free grammar to form all elements of the papers. The aim is to maximize amusement, rather than coherence. One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to 2005! Our plan is:

- fill up all journals with auto-generated articles;

7 Rank all myth on scholar.

Figure generated by SCIgen

# 版面分析

## 实例

### 输出效果（HTML）

**The Initiative Application of Tool Like SCIgen**

*Nameless Robot , Bye Another Earth and Hello World*

**Abstract**

SCIgen is a program that generates random Computer Science research papers, includ-ing graphs, figures, and Citations. Since almost all academic paper of these meaningless, papers generated by SCI-gen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be save

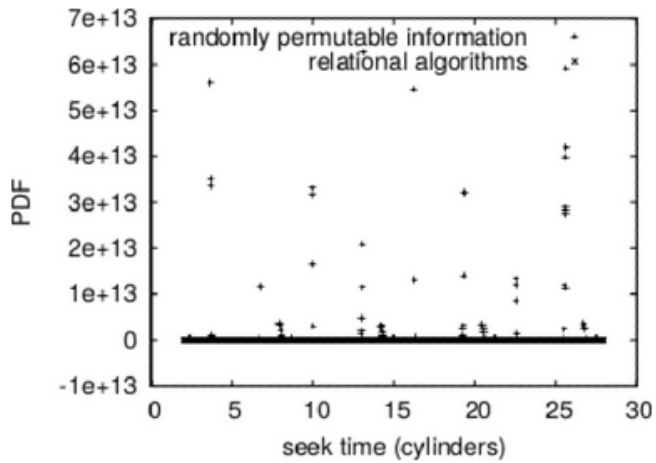
**1 Introduction**

SCIgen uses a hand-written context-free grammar to form all elements of the pa-pers. The aim is to maximize amusement, rather than coherence. One useful purp program is to auto-generate sub-missions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to S is:

- -fill up all journals with auto-generated articles;
- -Break all myth on scholar.

**2 Methodology**

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over academic industry. Sooner or later(if not currently), all journal and con-ference papers will be completely meaning-less. To uncover the sacred veil of scholars, we s no dream is good dream. Or even better, no one will trust education anymore, which bring significant benefit to our society.




: A figure generated by SCIgen

### 输出效果（LaTeX，第一页）

The Initiative Application of Tool Like SCIgen

Nameless Robot , Bye Another Earth and Hello World

May 10, 2015

**Abstract**

SCIgen is a program that generates random Computer Science research pa-pers, includ- ing graphs, figures, and Citations. Since almost all academic paper of these days are meaningless, papers generated by SCI- gen are sufficient to fill up all journals. Researchers and universities are no long needed, large amount of money can be saved for that.

**1 Introduction**

SCIgen uses a hand-written context-free grammar to form all elements of the pa- pers. The aim is to maximize amusement, rather than coherence. One useful purpose for such a program is to auto-generate sub- missions to conferences that you suspect might have very low submission standards. In fact, one of our papers was accepted to SCI 2005! Our plan is:

- fill up all journals with auto-generated articles;
- Break all myth on scholar.

**2 Methodology**

Do not repeat yourself. At first, human must be lazy, laziness is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the Whole academic industry. Sooner or later(if not currently), all journal and con- ference papers will be completely meaning- less. To uncover the sacred veil of scholars, we should tell to truth, no dream is good dream. Or even better, no one will trust education anymore, which bring significant benefit to our society.

1

# 版面分析

## 实验结果

- 以 MediaTeam Oulu 文档数据库中类别为文章、手册、数学、程序或地理下的 274 个页文档作为测试集
- 本文系统的递归投影切分算法对 164 页文档 (占总数的约 $59.9\%$) 给出了可接受的物理版面分割
- 本文系统文本块的识别准确率达到 $98.7\%$，文本块的召回率达到 $99.3\%$
- 本文系统的阅读顺序排序算法对 265 页文档 (占总数的约 $96.7\%$) 给出了合理的阅读顺序
- 实验中发现文本块中文本行总数为 $15749$，基于投影的行提取未能正确把两行分开的情况只出现了 $129$ 次，而把一行切开的情况也只出现了 $97$ 次

# 字符识别

## 目的与方法

本阶段需要分割出各个字符并确定它们分别是什么

- 以字形而非字符作为基本识别单位
- 字形的识别使用从初始候选集开始用多个匹配器依次进行筛选的策略
- 基于 Hausdorff 距离的模板匹配被用于多候选排序的依据
- 可局部伸缩的特殊字形通过动态生成模板的方法匹配

# 字符识别

## 实验结果

在一个大小为 1131 的字符集上不同匹配器组合的字符识别准确率（区分字母的字体）

| 匹配器组合 | 待识别字体大小 | | | | |
|---|---|---|---|---|---|
| | 10 | 20 | 30 | 40 | 50 |
| 穿线数 | 17.06 | 42.53 | 54.64 | 99.20 | 61.27 |
| 网格 | 20.69 | 75.42 | 85.94 | 99.20 | 93.72 |
| 矩 | 13.97 | 76.75 | 86.91 | **99.29** | 94.25 |
| 投影 | 16.89 | **78.78** | **88.15** | 99.20 | 93.55 |
| 网格 + 矩 + 投影 | 18.30 | 77.72 | 87.44 | 99.20 | **94.96** |
| 高宽比 + 矩 + 投影 | 16.18 | 77.90 | **88.15** | 99.20 | 94.61 |
| 高宽比 + 网格 + 矩 | 19.36 | 76.22 | 86.74 | 99.20 | 93.99 |
| 高宽比 + 网格 + 投影 | **21.04** | 77.90 | 87.62 | 99.20 | 94.25 |
| **高宽比 + 网格 + 矩 + 投影** | 20.51 | 78.34 | 87.62 | 99.20 | **94.96** |
| 孔洞数 + 高宽比 + 网格 + 矩 + 投影 | 17.15 | 64.46 | 80.11 | 99.20 | 89.83 |
| 穿线数 + 高宽比 + 网格 + 矩 + 投影 | 16.27 | 42.53 | 54.47 | 99.20 | 61.63 |

# 数学公式识别

## 目的与方法

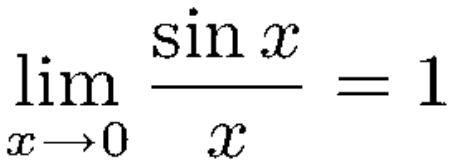
本阶段需要把数学公式找出来并分析其结构

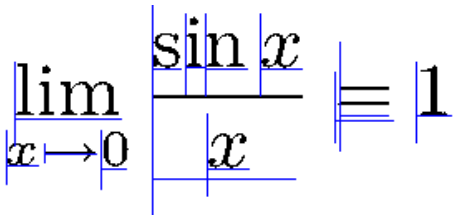
- 版式特征被用于快速定位编号公式
- 其它公式利用二维结构检测和符号类型定位
- 利用其它字符的信息来修正一些字符的识别结果
- 基于符号邻接图的方法被用于数学公式结构分析
- 已支持上下标、帽子、分式、根式和矩阵等多种数学公式类型

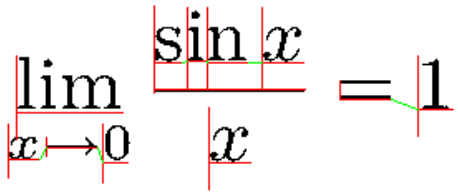
# 数学公式识别

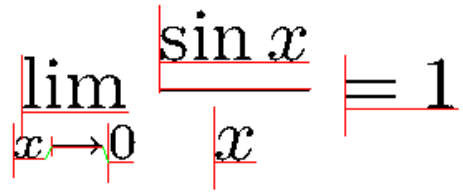
## 实例

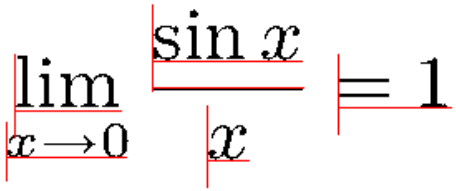
$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$$

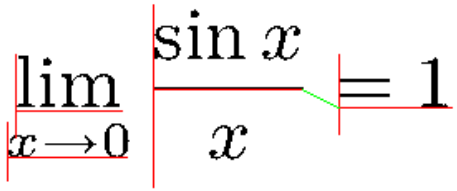


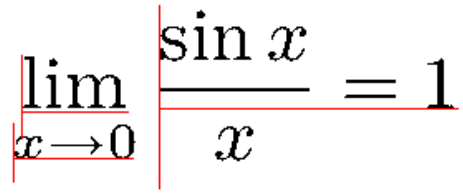


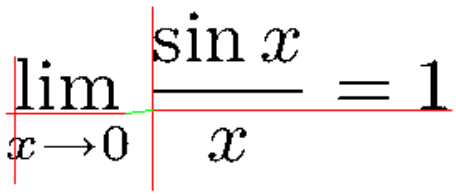


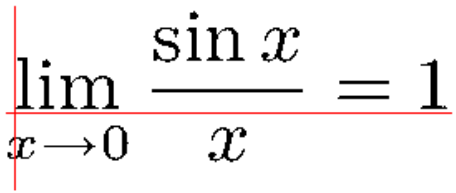












```
$$\lim_{\boldsymbol{x} \rightarrow 0}
\frac{\sin x}{x} = 1$$
```

# 数学公式识别

## 实验结果

使用 Infty 项目的 InftyCDB-2 数据库中的 4400 个公式作为测试集进行测试得到的结构分析结果

| 类型 | 样本个数 | 结构分析准确个数 | 准确率 |
|---|---|---|---|
| 简单表达式 | 2692 | 2096 | 77.9% |
| 含根号表达式 | 64 | 47 | 73.4% |
| 含分式表达式 | 676 | 391 | 57.8% |
| 含帽子表达式 | 769 | 450 | 58.5% |
| 含大型操作符表达式 | 714 | 310 | 43.4% |
| 总计 | 4400 | 3060 | 69.5% |

此外，值得注意的是，实验中对全部 $4400$ 个数学公式进行结构分析仅用时 $3041$ 毫秒，平均每秒处理了约 $1447$ 个数学公式

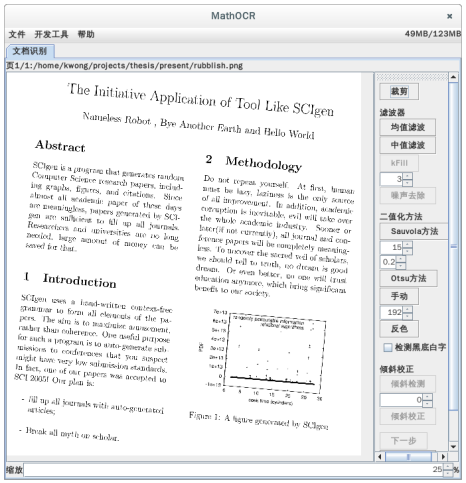
# 总结

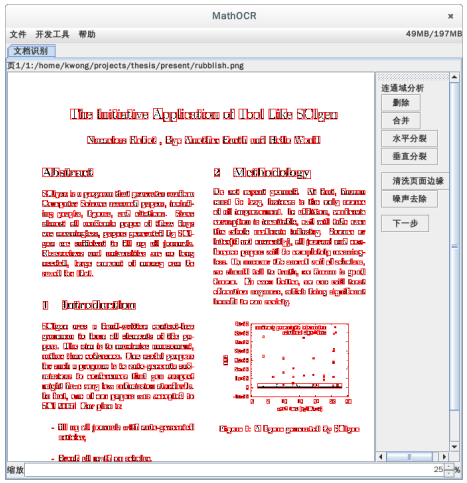
## 主要成果

- 形成了基本完整的科技文档识别技术链
- 实现了一个科技文档识别系统 MathOCR 0.0.3，支持数学公式识别和逻辑版面分析
- 以 MathOCR 0.0.2 作为项目成果的大学生创新训练计划项目《图片中印刷体数学公式的自动识别》获评优秀
- MathOCR 0.0.2 在 8 个月内取得超过 400 次下载，来自至少 20 个国家

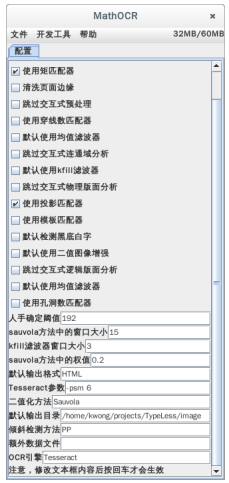
# 总结

## 主要成果


MathOCR
文件 开发工具 帮助
49MB/123MB
文档识别
The Initiative Application of Tool Like SCIgen
Nameless Robot , Bye Another Earth and Hello World
Abstract
1 Introduction
2 Methodology
裁剪
滤波器
均值滤波
中值滤波
kFill
噪声去除
二值化方法
Sauvola方法
Otsu方法
手动
反色
检测黑底白字
倾斜校正
倾斜检测
倾斜校正
下一步


MathOCR
文件 开发工具 帮助
49MB/197MB
文档识别
The Initiative Application of Tool Like SCIgen
Nameless Robot , Bye Another Earth and Hello World
Abstract
1 Introduction
2 Methodology
连通域分析
删除
合并
水平分裂
垂直分裂
清洗页面边缘
噪声去除
下一步

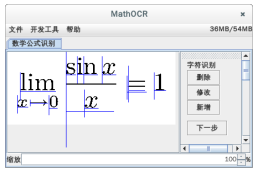

MathOCR
文件 开发工具 帮助
51MB/467MB
文档识别
The Initiative Application of Tool Like SCIgen
Nameless Robot , Bye Another Earth and Hello World
Abstract
1 Introduction
2 Methodology
版面分析
删除
合并
水平分裂
垂直分裂
区域分类
修正区域分类
文本
阅读顺序排序
手动重排阅读顺序
提取文本行
分裂行
与后一行合并
下一步


MathOCR
文件 开发工具 帮助
32MB/60MB
配置
使用矩形配器
清洗页面边缘
跳过交互式预处理
使用穿线数匹配器
默认使用均值滤波器
跳过交互式连通域分析
默认使用kFill滤波器
跳过交互式物理版面分析
使用投影匹配器
使用模板匹配器
默认检测黑底白字
默认使用二值图像增强
跳过交互式逻辑版面分析
默认使用均值滤波器
使用孔洞数匹配器
人手确定阈值 192
sauvola方法中的窗口大小 15
kFill滤波器窗口大小 3
sauvola方法中的权值 0.2
默认输出格式 HTML
Tesseract参数 psm 6
二值化方法 Sauvola
默认输出目录 /home/kwong/projects/TypeLess/image
倾斜检测方法 PP
额外数据文件
OCR引擎 Tesseract
注意，修改文本框内容后按回车才会生效

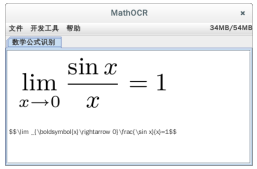

MathOCR
文件 开发工具 帮助
51MB/200MB
文档识别
LaTeX


MathOCR
文件 开发工具 帮助
19MB/52MB
字体训练
选择字体
字体大小
滤波器
均值滤波
中值滤波
kFill
噪声去除
二值化方法
Otsu方法
Sauvola方法
手动
显示数据
生成测试用图片


MathOCR
文件 开发工具 帮助
36MB/58MB
文档说明
源代码
关于本程序
通用公共许可证
开放字体许可证
帮助


MathOCR
文件 开发工具 帮助
36MB/54MB
数学公式识别
字符识别
删除
修改
新增
下一步


MathOCR
文件 开发工具 帮助
35MB/54MB
数学公式识别
字符识别
显示单步合并
下一步


MathOCR
文件 开发工具 帮助
34MB/54MB
数学公式识别

# 总结

## 可能的后续工作

- 非线性几何畸变纠正和更有效的自动去噪 (*)
- 基于案例的版面分析 (*)
- 识别表格、算法、代码、注释、页码引用、交换图、分子结构式等对象 (*)
- 加入对粘连和断裂字形的处理 (*)
- 利用已识别字符的信息改进分类器 (*)
- 基于结构的字符识别算法 (**)
- 结构分析规则的自动生成 (**)
- 基于语言模型的识别结果校正 (**)
- 手写文献的自动识别 (***)

# 总结

## 应用前景

科技文档识别技术的应用包括：

- 图书馆电子化
- 事实库自动生成系统

其中，数学公式识别技术的应用包括：

- 以公式作为关键词检索
- 检测论文中公式抄袭情况

# 相关资源

本课件、毕业论文全文、程序（包括二进制包及源代码）和训练数据可以在http://mathocr.sf.net/免费获得

# 致谢


- 感谢我的本科毕业论文导师黎培兴老师，正是他的鼓励使这个程序从构想变为现实
- 感谢全体答辩组老师，你们浪费时间来听我这种缺乏数学性的报告着实令人感动
- MathOCR 自带的数据文件为美国数学学会 amsfonts 的派生品，而用来读入 PNM 文件的代码取自 JAI 库，特此致谢
- 感谢所有下载过 MathOCR 的人，正是他们让我看到了我工作的价值，并让我形成工作的动力
- 感谢所有计算机黑客，他们搭建了我工作的基础，而且让我形成了现实的理想主义精神