文件图像分析系统概述

安德烈·提戈拉[1]

抽象

*本白皮书概述了文档图像分析系统及其组成
模块，这些模块使用的方法以及这些应用的用途。 之一
主要目标是提出一些最重要的技术和方法
文档图像分析域，以评估处理时的最佳方法
与真实世界的文件。 另一个主要目标是确保为那些开始的基础
建立这样复杂的软件系统并给出一个详尽的技术答案
问题：“如何为大量人员提供物理文件？”*

**关键词：文档图像分析，字符识别，OCR，图像数据
提取，图像导出**

1.引言

扫描物理页面并以数字格式存储它们是制作物理页面的一种手段
数据可用于数字世界。 它也解决了存储，纸张的问题
恶化，无障碍等等。 然而，它没有做的是让它更快
查明与某项工作相关的数据; 给像素结构和
将基础信息提取到计算机可理解的表示中
文档图像分析（DIA）系统的工作。 在谈论DIA系统时，第一个
想到的东西是光学字符识别[2] [16]，都是因为它
在美国邮政系统中使用相当长一段时间，因为它是用户的东西
甚至在与非文本文件进行交互时也会期待。 但是，OCR不在了
而不是大型应用程序的一小部分，并且代表狭义的内容
DIA真的是。

文档图像分析通常针对合成图像[1]，包含的图像
符号对象，例如书页，字母上的邮政地址，工程图纸，
乐谱，地图等。 但是，它也可以处理自然发生的模式
好，如指纹分析[16]。 另一类图像由表示
“自然”图像，如照片，卫星图像，X射线 – 可能的物体
用标准或非标准相机拍摄。 中国交通标志的图片可能是
这两个类别的一部分，取决于它的预期，但这不应该代表一个
其余讨论的问题。

2.处理步骤

首先应该指出的是，没有这样的东西作为一个独特的公认的分类
处理步骤图像从原始形式经历到计算机可理解
数据。 Nagy [1]提出了一个综合性的分类标准，并对程序进行了分类

1 工程师，Jinny Software Romania SRL 13 C Pictor Ion Negulici，罗马尼亚布加勒斯特，
andrei.tigora@jinnysoftware.com

第2页

基于他们处理的实体的粒度。 五个确定的级别，开始
从最低粒度开始，如下所示：

- 像素级别
- 原始级别
- 结构级别
- 文档级别
- 语料库级别

纳吉的分类标准还区分了输入图像的性质，即分离
那些内容大部分是文本的，大部分是由非文本图形组成的。
字符和图形主导图像之间的类似差异也被反映出来
在[2]中。 但是，对于某些输入图像，取决于所需的输出，处理步骤
来自这两个类别可能是必要的。 例如，电路图的扫描图像
将需要OCR处理和专用线路检测以及电路识别
零件。

本文中使用的分类松散地基于Nagy的分类[1]。 该
下一节将介绍每个级别的概述，重点介绍文本文档。

3.像素级处理

像素级处理处理图像到图像，试图变换给定的图像
进入更适合以下级别的版本。 属于这个的算法
分类处理降噪，二值化，字符分割，字符缩放和
矢量。

降噪

图像中的噪点有多种原因，例如输入图像质量下降，拍摄不完美
设备，以及不正当使用以前使用的设备，压缩和
传输错误。 在使用图像之前，光学异常必须是
补偿。

降噪旨在提高信噪比，而不仅仅是为了降低噪音
图像，但对于所有形式的信号。 在数字处理领域，它发挥着重要的作用
角色不仅在数字图像分析中，而且在医学[20]和天文学[23]中也是如此。 噪声
可以应用于双色调[16]，灰度[21] [22]，以及直接对色彩
图像[19] [20]。 大多数降噪机制都会对噪声做出假设
分布模式[18]。 由于其简单性，高斯分布被广泛假定
[19] [5]，但也考虑了泊松分布[20]以及非高斯分布
分布[5]。

现在用于降噪的方法非常多样化[5]以及更多
比十年前使用的那些复杂[2]。 经典的方法是
形态学方法，它依赖于侵蚀和膨胀变换的顺序
结合分割启发式[63]。 这些是非线性滤波器，并与其一起使用
线性滤波器，如基于高斯滤波器[17]的滤波器，属于空间域滤波

第3页

类别。 最近，特别关注变换域滤波;
其中，最好的结果似乎是由小波域滤波器产生的[21]。
虽然它们比自然图像的其他方法表现更好，但它们看起来并不像
成功地使用了合成文件，它们代表了文件图像分析的重点
系统。

二值化

就像降噪一样，二值化[64]与多种类型的图像相关，不仅如此
用于文档图像。 通过二值化过程，图像的像素是
分配给两个类别：前景和背景。 这也许是最激进的
由于重大的信息损失，图像经历了变换; 然而，对于一个文本
图像，这种简化不应该使图像像字符一样失去任何价值
将成为前景和图像背景的其余部分。

最常用的算法是基于阈值的[24] [25] [26]; 阈值方法
计算一个全局或本地值（阈值）[64] [65]，它使用黑色之间的分隔符
和白色像素。 二值化和降噪可以在一个步骤中结合在一起
以更好地分类什么是前景和什么是背景。 [62] [63]一个不同的
方法用于抖动算法：半色调[27]和有序[28]。 这些
试图仅使用预定义重现给定边界内的颜色密度
一组颜色值。 最后一种二值化算法是误差分散技术
[29]; 他们的目标是尽量减少图像中的错误表示
近似和传播相邻像素之间的色差。

歪斜校正

倾斜检测而非纠偏本身是所使用的各种算法的主要焦点
进行偏斜校正。 图像歪斜是不正确处理文档的结果
数字化过程; 取决于设备和文档的性质，不正确
放置可能导致文档边界周围出现黑暗区域，
称为边际噪声[8]。 尽管纠错不关心删除
边际噪音，这两种增强通常是一个接一个地进行的。 一个
正确对齐的文档简化了字符和文本行检测算法[7]，
假定文本本身具有预期的有序分布。

偏斜估计的一些方法依赖于投影轮廓或Radon变换。 这些
算法在一定限度内旋转图像并评估获得的投影
型材; 变异最大的一个对应于倾斜角[5] [66]。 就像
投影轮廓，霍夫变换方法执行可能的角度值的扫描
试图确定图像歪斜; 他们依赖最多线的观察
像素将沿着平行于文本行的基线[4] [7] [31]的线遇到。
霍夫空间中的过滤可以提高结果[69]，但不幸的是这些都没有
方法在处理也包含图像的文档时产生良好结果，如
他们有不同的像素分布影响统计。

与以前的方法不同，最近邻方法首先确定字符块
然后通过对最近的那些进行分组。 这是在他们假设下完成的

第4页

必须是同一文本行的一部分。 一旦确定了邻居组，斜交角
通过分析相邻实体之间的变化来计算[5] [30] [67]。 这些
方法也有一个缺点：它们对噪声非常敏感，可能会引入噪声
扰乱邻里创造过程的人造实体。 然而，有

页面检测到的旋转镜像使用[63]并且甚至可以校正弯曲的方法

字符分割

字符分割涉及识别图像中的特定区域
对应于单个字符。 高质量地分割基于拉丁文字的印刷文本
图像几乎没有问题，因为很少 - 最多3个，但通常只有一个 - 连接
组成角色的组件。 这通常通过确定连接来解决
组件并运行最近邻居算法来为其分配各种变音符号
相应的中央实体。

但是，对于低质量的图像，字符倾向于合并或分裂为多个
实体，一些字体比其他字体更容易出现某种行为。 这种情况
对于其他脚本是常见的，例如阿拉伯语，字符链接或汉语
[56]，其中字符可能由几个不同的组件或脚本组成
人们知道如何解读，就像兰纳那样具有感人和重叠的字母。 [14]
因此，最初设计的一些机制解决了低质量问题
图像实际上用于这些特定脚本的字符分割。

分离合并的字符可以通过明确的字符识别完成[32] [33] [35]。
然而，这限制了对特定脚本的分割，使其不那么通用。
另一种解决方案依赖于垂直投影[33] [34]，其目的是确定
以最少的像素进行坐标并沿该坐标分割。 细化[33] [34]也可能
因为它可能导致合并点的“自然”消除，但它可以
同样过分地分割现有的角色。

将分离的实体合并为单个字符通常归结为评估
各种组件之间的距离。 这些组件可以用它们来表示
边界框或像素本身或通过Voronoi图，Delaunay三角测量
或3D网格。 [70] [71] [72]

4.本体级加工

原始级处理的目的是确定元素的性质
在预处理步骤之后变得可访问。 而文本文件则是重点
是对角色的最终认可，对于目标所认识的非文本文档
基本的几何实体。[13] 对于扫描地图，文本元素的扣除变为
更难，因为文本有各种不同的方向，并放置在可变的彩色背景上。
形态学操作与统计数据一起帮助识别有趣的
区域。 [15]

第5页

字符识别

早期的字符识别工作是在拉丁文字上进行的，因为它是主要的
西方世界的脚本。 虽然还不完美，但在理想情况下，角色
识别率接近100%[36]。 所有方法在某种程度上都依赖于匹配a
候选人角色反对一组功能。 最基本的方法是使用参考
然后将该字符的位图与候选的缩放版本进行比较。 更多
可靠的参考集由线条笔划，笔划交叉，相对角度[1] [37]组成。 该
目的是发现一些可以唯一且明确地将某些特征分类的特征
候选对象恰好属于一个组。 参考设置的功能是
通过机器学习确定，通常是一个神经网络。

字符分类传统上以每个字符为基础进行确定
其特征最适合拟议的角色。 适应度函数通常是相对的
简单地说，就距离而言，计算候选人与参考模式的偏差
从参考的功能列表。 然而，多年来的经验表明
这种方法具有高度的模糊性，这只能通过使用更高的解决方案来解决
级别的信息，如语言模式。 通过考虑语言信息，
使用广义隐马尔可夫分类可以变得更加精确
模型或贝叶斯网络，从而表明更可能的字符序列
发生。

在某种程度上可以说，这个领域已经成熟了，与开发者和
研究人员同样承认目前OCR产品的性能[11] [36] [41]。

目前关于字符识别的研究大多针对阿拉伯语和中文
脚本以及手写文档[2]。 基于手写和阿拉伯文脚本的文本
在两种情况下，形成单词的字符彼此相关。 也，
而阿拉伯字符的形状因上下文而异，字符的形状为
手写文本往往会显示更大的变化，不仅在写文档之间
不同的人，但也在同一个文件。

对于基于中国的剧本，以及其他受其启发的剧本，主要难点在于
大量的可用字符，这使得它们不适合用于评估
用于拉丁文脚本的方法。 然而，有一些机制可以回应
以及该脚本强加的要求，如Tesseract [37]。 总体而言，大部分
为分析拉丁文字而开发的算法可以适用，但次要
修改/定制任何类型的脚本[39] [40]。

5.结构层面处理

对于这一级别的处理，目标是给诸如“实体”这样的实体组赋予“含义”
识别单词，重构文本行和解释表格。[10]

分词问题必须再次接受不同的处理
在使用的脚本上。 现代拉丁语剧本使用空格作为分隔单词的手段

第6页

并在一些特殊情况下使用其他符号，如连字符。 原始拉丁纪念碑
题词虽然，并没有分开的话; 这同样适用于中文。

对于基于拉丁文字的文本，分隔单词是评估两者之间距离的问题
连续的字符。 然而，对于嘈杂的文档图像，字符之间的距离
会受到意想不到的暗像素的影响，因此需要分割单词
某种语言模型信息。 这对于中文文本也适用
不使用任何形式的分隔符，并依靠读者的知识来决定分裂的位置
的话。

使用语言信息意味着分割成为OCR类型的处理。
最古老的方法可以追溯到20世纪60年代，当时N-grams被认为是错误的
校正[1]，后来扩展到分词[55]，作为高位的替代
使用完整词典的要求[57]。 N-gram和字典都可能是
从已有的知识库中获得[55]，或从“相关”分割产生
来源[57]。 实际的分割评估候选序列，选择那个

具有最高的概率，并且在长度模式不同的情况下略微偏向
更长的序列，因为它们不太常见。 语言信息可以在上面生成
从扫描的文档中找到，然后用于进一步的分割
文件或后期处理中的更正。 [74]

6.文档级处理

现代文档级处理集中在文档布局分析上，这是
将元素分组为逻辑顺序，并在一定程度上识别出优秀的
文本的元素，例如作者，标题等。对于布局分析，假定这是
诸如单词（甚至文本行），表格和图像等实体已被识别，
连同他们的包围盒。 有关内容的性质的信息
边框不是强制性的。 分组不同的元素是一个观察问题
在撰写文本时通常会遵守的距离惯例，寻找解释
空格，行[75]和任何表单分隔符[76]。 检测字体的方法
特征准确地提供有价值的信息[70]。 作者评估这些距离
相对于文本大小[77]，因为分隔两个文本列的空白列需要很大
足以明确分割具有相同垂直坐标的两条线[42]。 对于
自顶向下的方法，所谓的XY切割算法被采用，基于XY树，
通过两个轴中的一个递归地分割/剪切图像中的区域，最终结束
a [45]

如前所述，布局分析并不局限于分组元素; 它也是
执行文字标签。 这可以通过自下而上和自上而下来实现
方法。 标签可以完全依赖块页面分发[44]，或者是
增强了OCR信息[45]。

其他重要问题是检测表格，图表，图片和文本区域的问题
一个文件图像。 算法应该能够从图像中切出每个元素
不包括其他物体的零件，即使这意味着以不规则的方式切割
[78] [71]。

第7页

检测表格区域的一种方法是识别组成的图形元素
该表格：线段和交点[58]。 这往往是相对有弹性的噪音，
但它不能应用于没有图形组件的表格。 其他解决方案
尝试分析某个地区内的单词分布并基于此做出决定
他们的立场相对于其他的话。

在[59]中，作者试图根据行线稀疏来确定表格，这是多大
特定行上的空格是。 但是，这假定表格单元格很少
内容，这可能并非总是如此。 基于对齐的方法被提出在
[60]首先识别窄文本列，然后将它们组合成表格。 一个
混合实现在[61]中描述; 这试图结合非表格
诸如标题和预告位置的信息，将空格单元格分隔为的空格/行
以及背景颜色的变化。

7. CORPUS LEVEL处理

一旦文件图像被完整的分析和信息增强，
互动，这些互动可能适用于扫描的应用程序特定数字版本
文件应该可供用户使用。 这些包括内容索引和搜索，
以及验证和文档更新。 在编目论文时，人类首先应对
名称，例如出版物的标题，所以期望机器会这样做
一样的东西。 不幸的是，真正需要自动索引的文档是

通常较旧，质量较差，导致OCR性能不佳[36]。因此，
已经开发了其他方法，例如基于文档布局的方法。该，
算法首先在一组图像上训练以学习现有的布局，然后进行
分类阶段，为每个新图像确定布局[71]，并且健身评分为
根据图像中的多边形与参考上的多边形的匹配程度来计算
[43] [46]。 这种布局识别方法也可以用作检索机制
可以证明是比OCR更强大的机制。

然而，信息提取机制可以依赖像素这样的低层次元素。
[47]的作者描述了一种使用像素分布的图像识别算法
测量来识别特定的图像。 稍高一点的方法分析
字符形状，更精确地说是字符的垂直段，以构建图像
功能集，然后用于识别一个特定的文件[9]。

8.绩效评估

评估特定算法的性能不应成为争议的问题，然而，
在文档图像分析领域它通常是。 第一个问题来自于
模块化，文档分析系统的分层结构。 在这方面，a
二值化算法不能简单地通过将其输出与地面进行比较来进行评估
真理形象。 图像清晰度指标是一个有用的起点[73]; 更可靠
指标将是正确识别人物和其他特征的较高者
级分析仪。 即便如此，一个先进的OCR分析仪实际上可能会隐藏一个平均值
性能较低，所以应该使用替代方法进行比较
模块。

第8页

选择参考图像也是一个重要组成部分。 虽然有一个一般的
一致认为使用大量的文件是可取的[1]，当社区分裂时，社区就会分裂
涉及到那些应该使用的文件。 这个问题存在于二值化中
达到高级处理算法的方式。 参考输出，即所谓的
“地面真相”是最大的问题。 例如，当对艺术图像进行二值化时，
照片是非常不可能有一个独特的令人满意的二进制图像，而对于
合成图像，二进制参考通常是唯一的 – 除非图像中的噪声是
由于某种原因更相关的实际内容。 尽管它的独特性，另一个
产生地面真实信息的问题出现了。 一方面，可能会开始
从实际出发并以各种方式扭曲它，以便处理的输入是
原始基础事实的扭曲版本。 替代方法是生成
来自通常由程序或平台处理的文档的基本事实。 如
预计，这两种方法都有其缺点; 第一个是相对较快并允许的
设计师可以很好地控制他们期望测试的扭曲类型，但扭曲
图像往往是“不自然的”。 因此，针对这些评估的算法
对于真实文档中从未遇到的情况，输入信息正在进行微调。 该
另一方面，后者的解决方案虽然提供了切合实际的输入，但没有快速手段
迅速产生地面真相 – 地面真相必须由手工产生
使得这种方法对于当前的测试需求是不可扩展的[38]。

9.分析系统方法

理想情况下，一个软件包应该能够从任何类型中提取信息
它提供的文件，但事实是，大多数时候高度专业化的软件是
用于解决特定的问题。 为了获得最佳结果，分析系统以及他们的
组件，往往具有有限的适用性。 这可能涉及有限的一系列
处理比可以执行的处理，对文档的质量施加限制，
有限的语言支持，这意味着只有使用特定语言编写的文档

可能会被处理，或者是特定的文档布局，用于归档其格式的文档
一遍又一遍地重复[12] [48]。 然而，即使输入材料受限于
对内容的观点来看，所需的处理可能会非常复杂，因为处理得很高
各种其他参数，如纸张颜色，非标准文本布局和广泛的范围
手写风格是[53]的情况。

文档图像分析系统可能在范围上受限于诸如提取等
文件名[4]，或仅处理表格，图形和图像[3]。 这是
简化依赖于只有某些信息值得使用的假设
文件索引。

文献[6]中描述的系统完全避开了字符识别，提取“单词”
图像“，然后通过逐字匹配用户输入的关键字
匹配。

有些系统根本无法弥补文件的低质量
处理，所以设计人员开发了依赖于用户反馈的系统。 如[36]所述，
没有人为干预，结果往往会显着降低，因此人为干预
对于良好的结果是强制性的。 在[7] [12]中提出的系统使用最初的人类输入
创建之后用于提取类似构造的信息的模板

第9页

文档，而不需要其他用户干预。 更常见的是，允许
人类用户检查分析的建议结果并相应地进行修正
[12] [51]。

其他系统允许用户完全控制自己的能力[49]，这可能是
每当批处理模式处理不会产生预期结果时都很有用。 更多
激进的做法是将整个文件分析留给最终用户[50]
一个大型的科学家社区的优势，这些科学家对这些文档感兴趣并且能够
合作加强它们。 一定程度的自动化处理可以被整合
尽管如此，已经在[51]中尝试过了。 但是，为了消除努力
应该使用单一共享的文件模型[52]。 人群采购
也用于[54]，以便检查自动化处理的结果并加以改进
最终文件的质量。

10.结论

由于输入材料种类繁多，所有问题都没有通用的解决方案。

但是，为特定文档输入类别和特定文档找到最佳解决方案
（预）处理阶段是可能的。 为了充分利用“最佳”解决方案
合适的“处理环境/阶段一个智能文档图像分析系统
这将允许选择算法/方法来解决特定问题
以及最适合该任务的参数设置。

11.参考文献

1.G.Nagy，“20年的PAMI中的文档图像分析”， *IEEE Transactions on Pattern 分析和机器智能* ，卷。 第22号，第1号，2000年1月。
2. Rr Kasturi，L. O'gorman，V. Govindaraju，“Document image analysis：A primer”， *Sadhana* ，Vol。 第27号，第1号，2002年2月。
3. X. Lu，S. Kataria，WJ Brouwer，JZ Wang，P. Mitra，C. Lee Giles，“Automated 分析智能文件搜索文件中的图像“， *国际期刊 文件分析与识别* ，卷。 12.第2号，2009年6月。
4. L. Likforman-Sulem，P. Vaillant，AB de laJacopière，“Automatic name extraction from 退化的文件图像“， *模式分析和应用* ，卷。 9号，2-3号，10月

2006年。
5. T. Saba，G. Sulong，A. Rehman，“文档图像分析：问题，方法比较和剩下的问题“， *Artificial Intelligent Review* ，Vol。 35，No. 2，2011年2月。
6. CB Jeong，SH Kim，“用于关键词识别的文档图像预处理系统”，*第七届数字图书馆国际会议论文集：国际合作和交叉受精* ，2004年12月。
7. C. Antonio Peanho，H. Stagni，FSC da Silva，“语义信息抽取复杂文件的图像“， *Applied Intelligence* ，Vol。 2012年12月号第37号，第4号。
8. Z. Hu，X. Lin，H. Yan，“基于多密度特征的文档图像检索”，*中国电子与电气工程前沿* ，Vol。 2，2007年第2期。
9.陈，，W. Huang，SY Sung，Z. Yu，Y. Xu，“从文档图像中检索文本基于Word形状分析“， *Applied Intelligence* ，Vol。 18，第3号，2003年5月至6月。

第10页

10. J.–Y. Ramel，S. Busson，ML Demonet，“AGORA：交互式文档图像BVH项目的分析工具“， *第二届国际文档图像会议2006年图书馆分析2006年* 4月27 – 28日。
11. A. Antonacopoulos，D. Karatzas，“第二次世界大战个人档案图像分析”记录“， *第一届国际文献图像分析研讨会论文集图书馆* ，2004年。
12. J. He，AC Downton，“用于数字图书馆的用户辅助归档文档图像分析”建设“， *第七届国际文件分析和会议论文集确认* ，2003年8月3日至6日。
13. EE Regentova，S. Latifi，D. Chen，K. Taghva，D. Yao，“通过处理分析文档JBIG编码图像“， *国际文件分析和识别杂志* ，Vol。 7，第4号，2005年9月。
14. S. Pravesjit，A. Thammano，“历史兰纳手抄本的分割”，*智能系统（IS），2012年第6届IEEE国际会议* ，2012年9月6 – 8日。
15. S. Biswas，AK Das，“从扫描的土地图图像中提取文本”， *2012年国际版信息学，电子与视觉会议（ICIEV）* ，2012年5月18 – 19日。
16. LO Gorman，R. Kasturi，“文档图像分析”， *Los Alamitos：IEEE CS Press* ，1995年。
17. J. Lee，R.–H. Park，S.Chang，“使用多尺度双边分解进行噪声降低数字彩色图像“， *信号，图像和视频处理* ，2012年8月。
18. H.–Y. Lim，D.–S. 康，“使用定向修改的图像高效降噪西格玛过滤器“， *超级计算杂志* ，2012年12月。
19. J. Harikiran，R. Usha Rani，“用于高斯噪声的彩色图像恢复方法去除“， *计算机信息和通信技术通信和Information Science* ，Vol。 101，2010，pp 554–560。
20. S. Okawa，Y. Endo，Y. Hoshi，Y. Yamada，“减少测量时间的泊松噪声 – 解决时间域漫反射光学层析成像数据“， *医疗与生物工程与计算* ，50（1），2012，pp 69–78。
21. ADE Stefano，PR White，WB Collis，“图像阈值选择方案的选择”Noiselet on Wavelet“，2004，第225–233页。
22. M. Nachtegael，S. Schulte，D. Weken，D. Van Der，V. De Witte，EE Kerre，“Do Fuzzy技术为图像降噪提供附加价值？“ *高级概念智能视觉系统* ，2005年，第658–665页。
23. S. Harmeling，B. Scholkopf，HC Burger，“从天文图像中消除噪音使用像素特定噪声模型“， *IEEE国际计算会议摄影* ，2011，第1–8页。
24.N.Otsu，“灰度直方图的阈值选择方法”， *IEEE Transactions系统，人与控制论* ，1979。
25. M. Sezgin，B. Sankur，“调查图像阈值技术和定量性能评估“， *电子成像杂志* ，2004年。
26. J. Zhang，J. Hu，“基于二维Otsu方法和直方图分析的图像分割”，*国际计算机科学和软件会议论文集*

工程，*IEEE计算机协会，华盛顿特区，美国*，2008年。
27. AR Ulichney，“频域中的半色调表征”，*影像科学和科技第47届年会*，1994年。
28. AR Ulichney，“用于抖动阵列生成的空白和聚类方法”， *SPIE*，1913年，1993年。
29. RW Floyd，L. Steinberg，“自适应算法的空间灰度”， *Proceedings of 2004信息显示学会*，Vol。17，pp.75–77,1976。

第11页

30. Y.Lu，CLTan，“基于最近邻链的偏差估计方法文件图像“， *模式识别字母*，卷。24，pp.2315–2323,2003。
31. YM Alginahi，“关于阿拉伯字符分割的调查”，*国际期刊文件分析和承认*，2012年。
32. J. Wang，J. Jean，“用神经网络和最短路径分割合并字符”路径“， *Proceedings of the 1993 ACM / SIGAPP symposium on Applied computing*，769–1993。
09.7.62
33. YM Alginahi，“关于阿拉伯字符分割的调查”，*国际期刊文件分析和承认*，2012年。
34. A. Zahour，B. Taconet，L. Likforman–Sulem，W. Boussellaa，“Overlapping and multi–通过块覆盖分析来触摸文本行分割“， *模式分析和应用程序*，2008。
35. A. Nomura，K. Michishita，S. Uchida，M. Suzuki，“检测和分割在数学表达中接触人物“， *第七届国际会议文件分析与识别*，第126–130页，2003。
36.霍利，R.“它有多好？分析和提高大规模OCR精度历史报纸数字化计划“，*D–Lib Magazine*，Vol。15，第3期，第1–13页，2009年。
37.R.Smith，“Tesseract OCR Engine概述”， *第九届国际会议关于文件分析和识别*，第2卷，第629–633页，2007。
38. J. Callan，P. Kantor，D. Grossman，“信息检索和OCR：来自转换内容抓住意义“， *ACM SIGIR Forum*，Vol。2002年第36卷第2期，第58–61页。
39. F. Hedayati，J. Chong，K. Keutzer，“用西藏木刻印刷品的认识”广义隐马尔可夫和核化修正二次距离函数“，*关于噪声的多语言OCR和分析2011年联合研讨会会议论文集非结构化文本数据*，2011年。
40. MK Jindal，M. Kumar，M.，RK Sharma，“脱机手写Gurmukhi字符识别：不同特征 – 分类器组合研究“， *论文集文件分析和识别页面研讨会*，第94–99页，2012。
41. H.Déjean，J. Meunier，“根据目录建立文件”，*ACM关于文档工程的讨论会*，2005年2月9日。
42. PE Mitchell，H. Yan，“采用连接线的报纸文件分析”细分“， *Pan–Sydney区域视觉信息研讨会论文集处理*，卷。11，pp.77–81,2001。
43. L. Golebiowski，“Automated Layout Recognition”， *第一届ACM硬拷贝研讨会文件处理*，第41–45页，2004年。
44.罗森菲尔德，R.费尔德曼，Y.奥曼，“从视觉布局提取结构文件“， *第十一届国际信息和通信会议论文集知识管理*，第203–210页，2002年。
45. A. Takasu，K. Aihara，“从随机扫描文件中提取信息”页面布局分析“， *2008年ACM应用计算研讨会论文集*，第447页，2008年。
46. L. Lecerf，D. Maupertuis，“基于布局的文档检索和可扩展索引”排名“， *2010年ACM应用计算研讨会论文集*，第28–32页，2010。
47. Z. Hu，X. Lin，H. Yan，“基于多密度特征的文档图像检索”，*中国电子与电气工程前沿*，Vol。2，2007年第2期。
48. ABS Almeida，RD Lins，GDF Pereira e Silva，Thanatos：“自动检索

巴西死亡证书的信息“， *2011年研讨会论文集*
历史文件成像和处理”，第146–153页，2011年。

第12页

49. RD Lins，GDF Pereira e Silva，ADA Formiga，“加强处理平台
历史文献“， *2011年历史文献研讨会论文集*
成像和处理，卷。0，pp.169–176。
P. P. Tranouez，S. Nicolas，V. Dovgalecs，A. Burnett，L. Heutte，Y. Liang，R. Guest，R.，
“DocExplore：克服文化和物理障碍获取古代文件”，
2012年ACM关于文档工程的讨论会议录，第205–208页，2012。
51. AH Toselli，E. Vidal，A. Juan，“交互式布局分析和转录系统”
为历史手写文件分类和主题描述“， *Proceedings of*
第10届ACM文件工程研讨会，第219–222页，2010。
52.R.Sanderson，B.Albritton，R.Schwemmer，H.Van De Sompel，“SharedCanvas：A
中世纪手稿布局传播的协作模型“， *Proceedings of the*
第11届国际ACM／IEEE数字图书馆联合会议，第175–184页，
2011。
53. E. Matthaiou，E. Kavallieratou，“患者历史信息提取系统
文件“， *第27届ACM年度应用计算研讨会论文集*，
787，2012。
54.T.Ishihara，T.Intoko，D.Sato，A.Tzadok，H.Takagi，“Transforming Japanese Archives into
无障碍数字图书分类和主题描述“， *第12届论文集*
ACM／IEEE–CS数字图书馆联合会议，2012年第91–100页。
V. Zhikov，H. Takamura，“一种有效的无监督词分割算法
与分支熵和MDL“， *2010年会议实证会议记录*
自然语言处理中的方法，第832–842页，2010。
56. A. Chen，“使用最小语言知识的中文分词”，
对中国语言处理第二SIGHAN研讨会论文集，第148–
151，2003。
P. P. Simon，S. Hsieh，L. Prevot，C.–R. 黄，“重新思考中文分词：
Tokenization，Character Classification，or Word Break Identification“， *Proceedings of*
ACL互动海报和示范会第45届年会，
69–72，2007。
58. LA Pereira Neves，JM de Carvalho，F. Bortolozzi，“用表格形式提取
Artefact Removal“， *2007 ACM Symposium on Applied Computing会议论文集*，
622–626,2007。
59. Y. Liu，P. Mitra，CL Giles，“通过稀疏识别数字文档中的表格边界”
在线检测“， *第17届ACM信息和知识会议论文集*
采矿，第1311页，2008年。
60. F. Shafait，R. Smith，“异构文档中的表格检测”， *第八届会议记录*
IAPR国际文件分析系统研讨会，第65–72页，2010年。
61. G. Harit，P. Art，“使用标题和预告模式的文件图像中的表格检测”，
第八届印度计算机视觉，图形和图像*会议论文集*
处理，2012年。
62. CA Boiangiu，AI Dvornic。“现代和现代的双色图像转换方法”
Classic Documents“， *WSEAS Transactions on Computers*，第7期，第7卷，第1081页 –
1090，2008年7月。
63. CA Boiangiu，AI Dvornic，DC Cananau，“二值化数字化项目使用
Hybrid Foreground–Reconstruction“， *Proceedings of the 2009 IEEE 5th International*
智能计算机通信与处理会议，克卢日纳波卡，8月
27–29，pp.141–144。
64. CA Boiangiu，A. Olteanu，AV Stefanescu，D. Rosner，N. Tapus，M. Andreica，“本地
Thresholding Algorithm Based on Variable Window Size Statistics“， *Proceedings CSCS–*
18日，第18届国际控制系统和计算机科学会议，5月
2011年24月27日，罗马尼亚布加勒斯特，第2卷，647–652。

第13页

65. CA Boiangiu，A. Olteanu，AV Stefanescu，D. Rosner，AI Egner，“Local Thresholding 图像二值化使用变量窗口标准偏差响应“， *Annals of DAAAM for 2010，第21届国际DAAAM研讨会论文集*，20–23 2010年10月，克罗地亚扎达尔，第133–134页。
66. B. Raducanu，CA Boiangiu，A. Olteanu，A.Stefănescu，F. Pop，I. Bucur，“倾斜检测使用Radon变换“， *Proceedings CSCS–18，第18届国际会议控制系统与计算机科学“*，2011年5月24 – 27日，罗马尼亚布加勒斯特，卷2，Pp。653–657。
67. D. Rosner，CA Boiangiu，A.Stefănescu，N.Tăpus，A. Olteanu，“Text Line Processing for 图像文件中的高置信度歪斜检测“， *ICCP 2010 Proceedings*，Cluj–Napoca, Romania, August 26–28 2010, pp. 129–132.
68. CA Boiangiu, D. Rosner, A. Olteanu, AV Stefanescu, ADB Moldoveanu, “Confidence Measure for Skew Detection in Photographed Documents”, *Annals of DAAAM for 2010, Proceedings of the 21st International DAAAM Symposium* , 20–23 October 2010, Zadar, Croatia, pp. 129–130.
69. CA Boiangiu, B. Raducanu, “Effects of Data Filtering Techniques in Line Detection”, *Annals of DAAAM for 2008, Proceedings of the 19th International DAAAM Symposium* , pp. 0125—0126.
70. CA Boiangiu, AC Spataru, AI Dvornic, DC Cananau, “Automatic Text Clustering and Classification Based on Font Geometrical Characteristics”, *Proceedings of the 9th WSEAS国际自动化与信息会议，WSEAS出版社*，第468页 – 473，2008年6月24 – 26日，罗马尼亚布加勒斯特。
71. CA Boiangiu，DC Cananau，B. Raducanu，I. Bucur，“分层聚类方法针对文档布局理解和分析“， *国际期刊应用科学中的数学模型和方法*，第1卷，2008年第2卷，Pp。413–423 422。
72. CA Boiangiu，B. Raducanu，“用于图像页面的3D网格简化技术 Clusters Detection“， *WSEAS Transactions on Information Science，Applications*，Issue 第5卷，第1200–1209页，2008年7月。
73. CA Boiangiu，AV Stefanescu，D. Rosner，A. Olteanu，A. Morar，“自动倾斜大规模数字化项目中的边缘目标验证“， *Proceedings of the 21st 21st 国际DAAAM研讨会*，2010年10月20 – 23日，第131–132页。
74. CA Boiangiu，DC Cananau，S.Petrescu，A.Moldoveanu，“基于OCR后处理关于字符模式匹配“， *第20届DAAAM世界研讨会*，奥地利141–144页 维也纳中心（ACV），2009年11月25 – 28日。
75. CA Boiangiu，B. Raducanu，“自动内容的线路检测技术” 转换系统“， *WSEAS信息科学交易*，应用，第7期， 第5卷，第1200–1209页，2008年7月。
76. CA Boiangiu，DC Cananau，AC Spataru，“检测任意形式的分隔符基于Filtered Delaunay Triangulation“， *第9届WSEAS国际会议论文集自动化与信息会议*，WSEAS出版社，第442–445页，布加勒斯特，罗马尼亚，2008年6月24日至26日。
77. CA Boiangiu，DC Cananau，AI Dvornic，“基于白色空间探测技术关于邻里距离测量“， *2008年DAAAM年报，会议记录第19届国际DAAAM研讨会*，第131–132页。
78. CA Boiangiu，M. Zaharescu，I. Bucur，“为图像建立非重叠多边形 Document Layout Analysis Results“， *The Journal of Journal ISOM*，Vol。6第2号/ 2012年12月，第428–436页。