2009年第十届国际文档分析与识别会议

通过Tab-Stop检测进行混合页面布局分析

雷·史密斯

Google Inc. 1600 Amphitheatre Parkway，Mountain View，CA 94043，USA。

theraysmith@gmail.com

抽象

一种新的混合页面布局分析算法是
提议的方法，该方法使用自下而上的方法来形成初始数据类型假设，并找到格式化页面时使用的制表位。检测到的制表位用于推断页面的列布局。然后以自上而下的方式应用列布局，以将结构和读取顺序强加到检测到的区域上。

完整的C ++源代码实现是
可作为Tesseract开源OCRengine的一部分获得，网址为http://code.google.com/p/tesseract-ocr。

1.简介

物理页面布局分析，第一步之一
OCR的功能是将图像分为文本和非文本区域，以及将多列文本分成多个列。本文不涉及逻辑布局分析，而是检测页眉，页脚，正文，编号列表和文章分段。

物理布局分析对于实现
OCR引擎可处理任意页面的图像，例如书籍，杂志，期刊，报纸，
信件和报告。物理布局分析的方法大致分为两类：

自下而上的方法都是最早的方法[1]甚至更多
最近发表的[2,3]方法。他们将图像的一小部分（像素，像素组或连接的组件）分类，并像类型一样聚集在一起以形成区域。自下而上的方法的主要优点是它们可以轻松处理任意形状的区域。关键缺点是它们努力考虑图像中的高级结构，例如列。这通常导致过度碎片化的区域。

自上而下的方法[4]以递归方式切割图像
沿着空白的垂直和水平方向，这些空白应该是列边界或段落边界。尽管自上而下的方法具有
优点是，它们从查看页面上最大的结构开始，因此无法处理许多杂志页面上出现的各种格式，例如非矩形区域和交叉列标题无缝地融合到下面的列中。

第三类方法[5-7]是基于对
图像中的空白。通过对间隙进行自底向上的分析，在列之间查找间隙，从而明确查找白色矩形，从而解决了递归自顶向下方法中的某些缺陷。这些算法大多仍然存在无法处理非矩形区域的问题。

2.通过制表符停止检测进行页面布局

布置页面时，由专业人员
发布系统或公共文字处理程序，页面区域由制表符限制。边距，列边缘，缩进和atable列都放置在固定的x位置，在这些位置上文本行的边缘或中心垂直对齐。制表位停止将表格与正文分开，并也
制了矩形的非列

元素，例如插图
图片和引号。

制表符在
图1的示例是带有用于页面缩进的附加制表符的列边界，这对于查找页面布局不是必需的。非矩形插图通常会偏离列边界。


The Gist of Translation

How long will it be before machines make the Web multilingual?

图。1。输入图像

从某种意义上说，白色矩形与制表符匹配，但是
背景矩形或背景图像可能会干扰白色矩形。也是白色的两端
矩形与制表位限制的区域的末端不匹配，因为白色矩形一直延伸到垂直空白中。

978-0-7695-3725-2 / 09 25.00美元©2009 IEEE 241

DOI 10.1109 / ICDAR.2009.257

所提出的算法类似于空白

矩形方法，它使用自下而上的方法来查找自上而下的结构，但是它没有查找列之间的空格，而是寻找标记其边缘的制表位，并通过自下而上和自上而下的方法的进一步组合，轻松应对非矩形区域。

主要阶段有：预处理，其中

自下而上的形态学和相关成分分析形成了本地数据类型的初始假设；自下而上的制表位检测；寻找列布局；最后应用列布局以创建有序的一组类型化区域。这些阶段将在第3-6节中详细介绍。

3.预处理

预处理步骤的目的是识别线

分隔符，图像区域，并将其余连接的组件分离为可能的文本组件和较少数量的不确定类型。

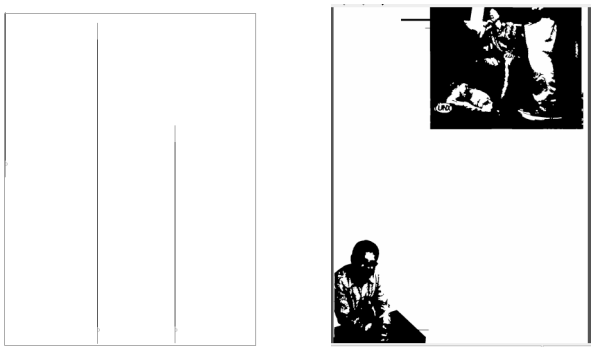

图2。（a）垂直线，（b）图像元素。

从图1的图像开始，形态

Leptonica [8]的处理检测到图2（a）所示的垂直线和图2（b）所示的图像蒙版。在将清洗后的图像传递到连接的分量分析之前，从输入图像中减去这些检测到的元素。

连接的组件（CC）通过

宽度，宽度w和高度h分为以下大小：h <7（在300ppi时）的CC很小。计算其余高度h75的第75个百分位数，h << h75 / 2很小； h> 2h75或w> 8h75大，其余中等。

由于小CC（噪声

或变音符号）和较大的非文本抄送（线条图，徽标或框架）可能会混淆文本线算法，但是较大的文本标题对于阅读顺序检测很重要。如果左或右邻居的笔触宽度相似，则在此阶段将大型CC视为文本。在“强调”字体上，笔触

垂直线上的宽度大于水平线上的宽度，因此笔划宽度是在两个方向上分别计算的。笔划宽度是根据CC二进制图像上距离函数的水平和垂直局部最大值来计算的。图3显示将CC过滤为中文本或大文本。

图3。过滤后的抄送

4.查找制表符位置作为线段

查找制表位线段的过程包括

几个主要子步骤：找到看起来像在文本区域边缘的候选制表位CC，然后将它们分组为制表位线，然后找到制表位线之间的连接，从而消除误报。

4.1。查找候选制表位组件

初始候选制表位CC可以通过以下方式找到

径向搜索从预处理中的每个已过滤CC开始。假设CC位于制表符停止位置，则搜索将查找对齐的邻居和应在排水沟中应有空间的邻居。每个CC独立处理，并根据其是否为候选左选项卡，右选项卡或两者都不标记而进行标记。图4（a）示出了候选制表位CC。

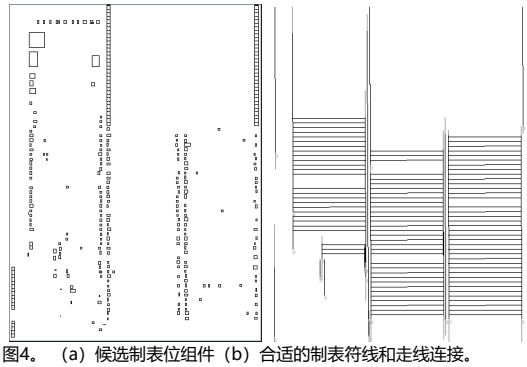

图4。（a）候选制表位组件（b）合适的制表符线和走线连接。

4.2。分组候选选项卡组件

候选标签CC分为几行，并且

如果组中有足够多的抄送，则保留它们。最小二乘平方中值算法用于将线拟合到组中每个CC的适当（左或右）边缘。找到所有制表符停止线段后，所有线都重新调整至页面均值方向，

978-0-7695-3725-2 / 09 25.00美元©2009 IEEE 242

DOI 10.1109 / ICDAR.2009.257

这样所有成员标签CC都会落到线段的一侧。

4.3。跟踪文本行以连接制表位

下一步通过跟踪文本来连接制表位

线从一个制表位到另一个。紧密相邻，垂直重叠的CC合格，但不能跳过较大的间隙。带有文本行将其连接在一起的制表位彼此关联，就像在文本列的相对两侧一样。图4（b）显示了制表符停止线和连接文本线。没有连接的制表位停止线将被丢弃。

文字最常出现的宽度

记录连接制表位的线以用于指示列布局。

4.4。清理凸舌止挡端

最后一步尝试

通过允许端点在最后一个成员CC（其边缘用于该tab线）与该行的第一个非成员CC之间移动，使连接的标签线在相同的y坐标处结束

相交。图5显示了最终的标签线段。

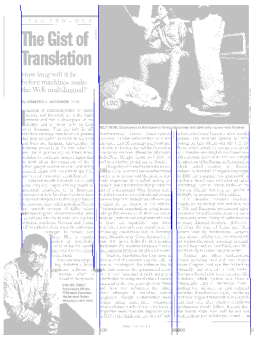

图5。已清洗制表位

施工后

制表位，CC会重新分类为“文本”或

使用与上面用于查找制表位之间的连接相同的文本行跟踪算法的“未知”。如果一组宽度很大的CC形成文本行，则将它们分类为文本。通过形态预处理从图像蒙版中创建与正文CCd大小相同的人造imageCC。

5.查找列布局

下一步是找到以下内容的列布局

这一页。所有其他步骤都使用现在创建的“列分区”（CP）对象。

从左到右，从上到下扫描CC

在底部，将类似分类（文本，图像或未知）CC的运行收集到CP中，但要遵守以下约束：没有CP可以越过制表符停止线。图6显示了此过程的结果。来自单个水平扫描的CP的集合存储在ColumnPartition Set（CPset）中。

每个CPset都有可能

在该垂直方向上将页面分为多个列

位置。寻找

因此，列布局是寻找

最佳“解释”（参见下文）页面上所有CPset的最佳CPset集，但首先需要进行一些定义：

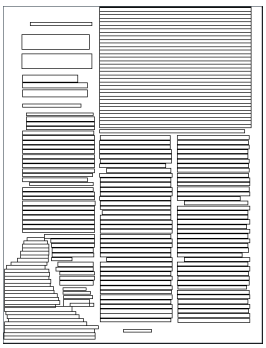

图6。列分区（CP）

一个好的CP

触碰其两个垂直边缘上的制表符线

边界框或其宽度接近经常出现的宽度。（请参阅4.3。）

CPset的覆盖范围是所有

它包含的良好CP。

如果A具有更大的CPset A，则CPset A优于CPset B

覆盖率，或相等的覆盖率，但更多的CP，或等好的CP，但更多的CP。

CPset A解释集合B，除非一个或多个

以下是正确的：1。 B的一个CP的边缘位于A的所有CP的外部。这是不允许的，因为它表明B具有比A.2更多的文本。 B的CP之一的边缘落在A的不同CP中，并且B CP的宽度是公共的。这意味着A已拆分了具有相同宽度的列3。 B的一个CP的右边缘与下一个B CP的左边缘在同一ACP中，并且B CP的宽度大致相同。看起来A与B的列数不同。相同宽度的条件允许A用引出线解释B。4。 B的两个CP的两个边缘都落在A的同一CP中。这意味着A合并了B的两列。

请注意，

B的一个CP允许落入A的两个CP，只要宽度不常见即可。这允许

合并B中列的标题，由A解释。

栏目清单

候选者是从页面上的CPset集合中选出的，顺序最佳，并且A被重复的B规则所消除。在此过程中，将忽略所列。有图像CP。

最初的候选人制成后，他们是

通过在不同的CPSet中使用CP的边缘，增加了新的CP并扩展了现有的CP，从而进行了改进，而扩展不会引起CP的重叠。

978-0-7695-3725-2 / 09 25.00美元©2009 IEEE 243

DOI 10.1109 / ICDAR.2009.257

然后，迭代过程会标记最长的段

列候选对象之一解释的连续（允许非常小的区域故障）页面y坐标的坐标。图7示出了该处理的结果。

6.寻找地区

找到这些列后，为CP赋予一个类型

根据它们跨越多少列。具有单列的CP正在流动，触及多于一列但不跨任一列的外边缘的分区为拉出式，而完全跨过一列以上的分区为行进。

6.1。创建CP流

每个CP选择其最佳匹配的上下限

伙伴，即垂直最接近的CP，重叠在一起。由于每个CP向其选择的伙伴注册自己，因此每个CP可能具有零个或多个注册的上，下伙伴。

注册合作伙伴列表的大小被迫

依次使用以下规则将上下限分别设为零或一：1。类型。如果存在多种类型，则文本只能保留其自己的（精确）类型，而图像可以保留任何其他图像类型; 2。传递伙伴的快捷方式已损坏。如果A具有2个伙伴B和C，并且B在相同方向上具有C作为伙伴，则删除C作为A的伙伴，留下一条干净的链A-B-C。同样，如果A有一个伙伴B，并且B在同一个方向上有一个伙伴A，则中断循环3。（仅文本）如果A仍然有2个伙伴B，C，请追随Band C的伙伴，看看哪个拥有最长的链。从A删除拥有最短链的伙伴，然后将最短链的类型转换为拉出。 4。（仅图像）选择水平重叠最大的伙伴CP。

现在所有CP都具有0或1

伙伴。即使这样，（重新）运行

上面的规则1。这会将文本的所有链纯化为单一类型，并将文本链与图像链分开。通过将链中的所有CP设置为链中最通用的类□□型，可以纯化图像链。图8显示了最终键入的CP，其中流动文本为蓝色，标题文本为青色，标题图像为洋红色，拉出图像为橙色。

图8。已输入分隔链。

文本CP链进一步分为以下几组

统一的行距，使文本块。现在每个CP链代表一个候选区域，但是必须对这些区域进行排序。

6.2。阅读顺序确定

回想一下，图像和文本分区的类型为

三种可能性之一：流动，拉出和标题。此外，页面分为一致的列布局的各个部分。有了这些信息，合理的阅读顺序就会脱离一些简单的规则：1。流动的块在列中紧跟着y位置2。拉出块在它们接触的实列之间的虚列中紧跟y位置3。标题跨越多个列，并在所跨越的列中或其之间的上方的任何内容之后。该标题之后的所有内容都位于该标题下的同一列中。4。列布局的更改就像标题一样，任何更改的列（或它们之间）中的任何内容都将在新列中的任何内容之前发生。未更改的列不受列更改的影响.5。在标题之间，列的内容从左到右排序。

6.3。找到每个区域的多边形边界

为了简化实施，该地区

多边形是等规的：即边在水平和平行于平均制表符线之间交替

（大约垂直。）

多边形的边缘是

选择最小化

满足所有CP都包含在其区域多边形内且没有来自其他区域的CP的约束

相交。图9显示了为图1的输入图像创建的最终块。

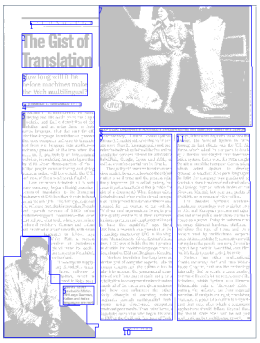

图9。最终块

7.测试与结果

本文描述的算法在以下位置实现

C ++，并且源代码作为Tesseract开源OCR系统的一部分提供[9,10]。它在3.4 GHz Pentium 4上以大约1秒的时间在非典型8MPixel图像上运行。

978-0-7695-3725-2 / 09 25.00美元©2009 IEEE 244

DOI 10.1109 / ICDAR.2009.257

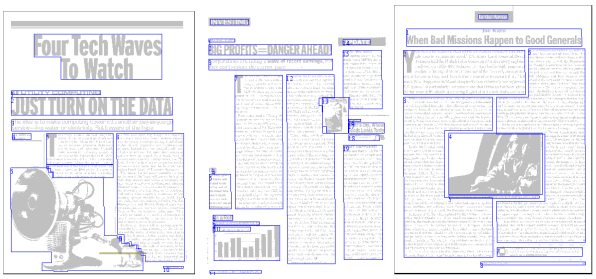

图10。 ICDAR2007上的部分结果。

正确测试页面布局分析是困难的

问题[11]，对于复杂的杂志页面而言，很少有公开可用的真相。 UNLV测试集[12]仅测量文本区域并计算错误，除非在所有正文文本后放置图形标题。

ICDAR页面布局分析比赛

提供了更好的整体精度度量，该算法的结果出现在2009年的竞争中[13]。一些图形结果如图2所示。表1给出了10个数据，并与ICDAR 2007竞赛中的参赛者进行了数值比较。表1中的结果仅基于2007年测试集计算，作者感谢ApostolosAntonocopoulos所提供的这些结果。有关测试方法的详细信息，请参见参考文献[11]和[13]。

表1. ICDAR 2007上的结果。

方法噪声Sep文本图像整体					
PRImA图标					
2007年-Besus 86.8%76.9%37.4%42.5%35.9%					
2007-TH1 68.0%79.7%76.1%46.2%67.6%					
2007-TH2 67.6%79.6%72.9%48.4%65.7%					
Tesseract 65.6%74.1%72.1%55.3%68.4%					
F测量					
2007年-Besus 62.9%76.2%95.8%57.2%90.2%					
2007-TH1 79.2%80.7%91.9%72.1%88.2%					
2007年第二季度79.2%80.6%92.3%72.4%88.6%					
Tesseract 79.2%70.9%93.3%82.0%91.3%					
召回					
2007年-Besus 65.7%71.7%94.9%67.0%88.2%					
2007-TH1 65.6%79.5%96.9%66.4%89.8%					
2007-TH2 65.6%79.5%97.2%66.9%90.2%					
Tesseract 65.6%81.4%97.9%76.5%93.8%					
精确					
2007年-Besus 60.4%81.3%96.7%50.0%92.2%					
2007-TH1 100.0%81.9%87.4%79.0%86.7%					
2007-TH2 100.0%81.7%87.9%79.0%87.0%					
Tesseract 100.0%62.8%89.0%88.3%88.9%					

10.结论和进一步工作

制表位是一种有趣且有用的替代方法

到白色矩形以查找a的列结构

978-0-7695-3725-2 / 09 25.00美元©2009 IEEE 245

DOI 10.1109 / ICDAR.2009.257

页。结合专栏的自上而下的概念

使用自底向上分类方法的结构使页面布局分析能够轻松处理现代杂志页面的复杂非矩形布局，而不会遗忘单独使用自底向上方法时经常发生的“大图”。

所描述的算法没有表检测或

分析，但制表位对这两个功能特别有用，因此将来会添加表格分析功能。

11.参考

[1] F. Wahl，K。Wong，R。Casey，“混合文本/图像文档中的块分割和文本提取”，计算机图形学和图像处理，1982年第20期，第375-390页。[2] M. Chen，丁小琴，“基于HMM的统一布局分析框架和算法”，SCI CHINA Ser F，46（6），2003年12月，pp401-408。[3] SP Chowdhury，S.Mandal，AK Das，B。Chanda，“从文档中分割文本和图形

图像，“第9届国际文档分析和识别大会，IEEE，巴西库里提巴，2007年9月，pp619-623。[4] G. Nagy，SC塞思，“光学扫描文档的分层表示”，第7版。国际会议，模式识别，加拿大蒙特利尔，1984，pp347-349。[5] HS Baird，SE Jones，SJ Fortune，“图片

[1] T. Pavlidis，J。Zhou，“基于页面的分割和形状分割的分割”，Proc。第十届国际模式识别会议，IEEE大西洋城，新泽西，1990，pp820-825。[6]

分类”，CVGIP：图形模型和图像

Processing，54（6），1992年11月，pp484-496。[7] TM值。 Breuel，“用于布局分析的两种几何算法”，Proc.Natl.Acad.Sci.USA。第五国际关于文档分析系统V的研讨会，Springer-Verlag

2002，pp188-199。[8] Leptonica图像处理和分析库。

http://www.leptonica.com。[9]史密斯（R. Smith）。 “ Tesseract OCR引擎概述。” Proc 9th Int。 Conf。关于文档分析和识别，IEEE，巴西库里蒂巴，2007年9月，pp629-633。[10] Tesseract开源OCR引擎。

http://code.google.com/p/tesseract-ocr。[11] A. Antonacopoulos，B。Gatos，D。Bridson，

Proc 9th Int。Conf。 “

ICDAR2007页面细分竞赛”。关于文档分析和识别，IEEE，

巴西库里提巴，2007年9月，第1279-1283页。[12] UNLV ISRI OCR测试工具包和数据库

http://www.isri.unlv.edu/ISRI/OCRtk.[13] A. Antonacopoulos等。等 “ ICDAR2009页面

细分竞赛”，Proc10thInt。 Conf。上

文档分析和识别，IEEE，西班牙巴塞罗那，2009年7月