

基于连通区域的传真图像版面分割与分类算法

郭晓宇, 平西建, 周 林

(信息工程大学 信息工程学院, 郑州 450002)

摘 要: 针对传真图像, 为了提高版面分割与分类准确率, 提高处理速度, 以连通区域为处理元素, 通过合理设定阈值, 将水平和垂直相邻连通区域进行合并, 快速准确地分割图像。并且将版面分割过程与分类过程相结合, 根据连通区域的大小建立矩阵, 提取能够表征区域信息的八维特征, 然后使用 BP 神经网络将版面区域分为文字区域和非文字区域两类。实验中得到版面分割准确率为 89.2%, 版面分类准确率为 94.22%。实验结果证明, 该算法能够快速准确地对传真图像进行分割和分类, 具有较强的实用价值。

关键词: 传真图像; 版面分析; 连通区域; 版面分割; 文字区域

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)08-3198-03

doi:10.3969/j.issn.1001-3695.2012.08.107

Fax image layout analysis based on connect components

GUO Xiao-yu, PING Xi-jian, ZHOU Lin

(Information Engineering Institute, Information Engineering University, Zhengzhou 450002, China)

Abstract: In order to improve accuracy of layout segmentation and classification of fax images, segment images by merging horizontally/vertically adjacent connect components blocks by setting threshold properly. And combined image classify process with image segment process, obtained matrix based on connect components. Then, abstracted the feature vector of eight dimensions for representing block information. And classified document images as two types with BP artificial neural networks. In experiments, the accuracy of layout segmentation and classification was 89.2% and 94.22%. Experimental results show that the method is effective and feasible in document images segmentation.

Key words: fax image; layout analysis; connect component; page segmentation; text region

0 引言

随着文本图像数量的日益增多, 对文本图像进行自动处理是计算机应用领域的一项重要课题。文本图像的种类繁多, 版面结构日趋复杂, 不仅包含字号不一的文字区域, 往往还包括图像、表格、图形等元素, 排版形式也多种多样, 不仅有矩形版面, 也有非矩形版面。要对文本图像进行自动化处理, 版面分析是一个重要的环节。它利用计算机对文本图像进行处理, 确定文本图像的物理结构, 将图像划分为文本、图像、图形、表格等不同属性的区域, 以满足文本图像字符识别(OCR)、表格识别、图标识别等不同应用的需要。版面分析一直是文本图像处理领域的研究热点^[1,2], 作为文本图像的预处理过程, 版面分析的结果直接影响到后续处理的准确性。

传真图像是文本图像中重要的一类, 利用传真图像中包含的电话号码、电子邮件地址等对传真图像进行检索和分类得到了广泛的关注。这一工作首先需要对传真图像进行版面分割以及文字区域和非文字区域的分类。版面分割及分类的准确性直接影响到后续的提取工作, 特别是对文字区域的分类错误有可能会影响信息提取的失败。另外考虑到实用性, 对处理速度也有一定要求。本文针对这一目的提出了传真图像的版面分割及分类算法。

1 基于连通区域合并的版面分割

传统的版面分割的方法通常可以分为三类, 即自顶向下法、自底向上法和混合法。其中自顶向下法的优点是速度快, 但需要根据版面结构的先验知识进行分割, 对不同版面适应性较差, 其代表性算法是基于 X-Y 分割法^[3]、游程平滑法(RLSA)^[4]等。自底向上法的优点是不需要对版面有先验知识, 对不同的版面适应性较好, 但需要的存储量及计算量较大, 耗时较多, 其代表性算法是近邻线密度法^[5]、连通分量分析法^[6]等。混合法是将自顶向下和自底向上两种算法组合运用, 以某一类算法为主。

在自顶向下方方法中, RLSA 算法得到广泛的应用, 许多版面分析的算法都是在 RLSA 算法的基础上, 通过设定合理的阈值对其进行扩展和改进的^[7]。它的计算步骤为: 分别从水平和垂直方向扫描图像, 如果连续白色像素点个数小于给定的阈值, 则把这些白色像素点变换为黑色像素点, 得到两幅图像; 将两幅图像进行逻辑与操作, 当某个像素在两幅图像中都为黑像素时, 合并结果为黑像素, 否则合并结果就为白像素; 最后对合并得到的图像在垂直方向上将连续长度小于给定阈值的白像素变换为黑像素, 填补内部的空白, 分割出版面的不同类型区域。

收稿日期: 2011-12-07; 修回日期: 2012-01-15

作者简介: 郭晓宇(1984-), 男, 河南新郑人, 硕士研究生, 主要研究方向为图像信息处理、文本图像分析与处理(xiaoyu.guo@qq.com); 平西建(1953-), 男, 教授, 博导, 主要研究方向为图像分析与处理、模式识别、信息隐藏; 周林(1983-), 男, 博士研究生, 主要研究方向为图像信息处理、模式识别。

1.1 连通区域合并过程

RLSA 算法是在像素级别操作,需要对图像进行多次扫描,计算量较大。一些自底向上的算法以连通区域为单位,通过合并连通区域来进行版面分割^[8,9]。受这些算法思想的启发,通过借鉴 RLSA 算法处理过程,本文以连通区域为单位,通过先合并水平相邻连通区域,再合并垂直相邻连通区域的方法进行版面分割。算法的具体步骤为:

a) 采用八邻域定义的连通区域,为图像中的每个连通区域建立一个矩形框,使用这个框架的左上角坐标 (x_s, y_s) 及右下角坐标 (x_e, y_e) 表示这个矩形框的位置,计算每个连通区域的长度与宽度,并将长度或宽度过小的连通区域去除以消除噪声的影响。以图 1(a)为例,建立的连通区域矩形框如(b)所示,为方便观察,矩形框内部采用灰色色彩进行填充。

b) 计算三维序列 $H(d_h, b_1, b_2)$, b_1, b_2 表示两个水平方向相邻的连通区域矩形框, d_h 是水平方向矩形框 b_1, b_2 之间的距离。不同于 K 近邻法^[9]采用的连通区域中心距离,这里采用的是连通区域矩形框边缘之间的距离,它的定义如图 2 所示。

c) 按照距离 d_h 由小到大的顺序对三维序列 H 进行排序。

d) 当序列 H 的第 1 个元素距离参数 d 小于给定的阈值 th_v 时,合并两个矩形框为一个矩形框,并将 H 首个元素从列表中移除。图 3(d) 给出了算法步骤(d)的运行结果,从图中可以看到,在同一行的多个矩形框合并为一个。

e) 计算三维序列 $V(d_v, b_1, b_2)$, b_1, b_2 表示垂直方向相邻的连通区域矩形框, d_v 是垂直方向矩形框 b_1, b_2 之间的距离。

f) 按照距离 d_v 由小到大的顺序对三维序列 V 进行排序。

g) 当 V 的第 1 个元素距离参数小于给定的阈值 th_h 时,合并两个矩形框为一个矩形框,并将 V 首个元素从列表中移除,图 3(b) 给出了算法最终分割结果。

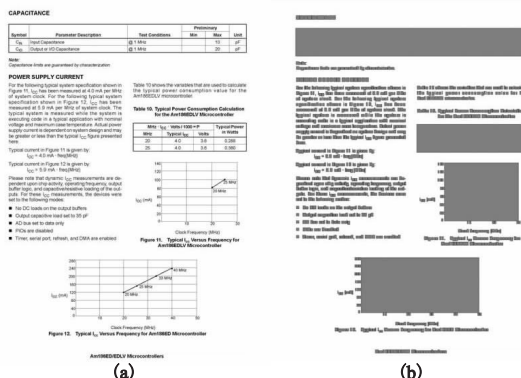


图1 原始图像与建立连通区域矩形框后图像

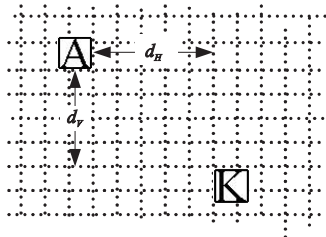


图2 连通区域边缘距离

算法步骤 b) ~ d) 类似于 RLSA 算法中的水平扫描过程,步骤 e) ~ g) 类似于垂直扫描过程,通过先后对水平方向和垂直方向连通区域的合并,使相同属性的连通区域合并到一个大的矩形框内,达到版面分割的目的。相对于文献[8]算法需要计算所有连通区域之间的距离,并建立最小生成树,而本算法仅需计算相邻连通区域之间的距离,需要的计算量更少。水平方向矩形框合并的阈值 th_v 选择为水平矩形框之间距离的平

均值,在算法步骤 b) 之后计算。在垂直方向矩形框合并中,因为各文本行之间距离是一致的,且文本行数量较多,通过在步骤 e) 之后,统计垂直方向相邻的矩形框之间的距离分布情况,可以确定文本行之间的距离 d ,阈值 th_h 选择为 $5/4d$ 。步骤 a) 中对连通区域宽度与长度的计算是为了后续版面区域分类的需要。

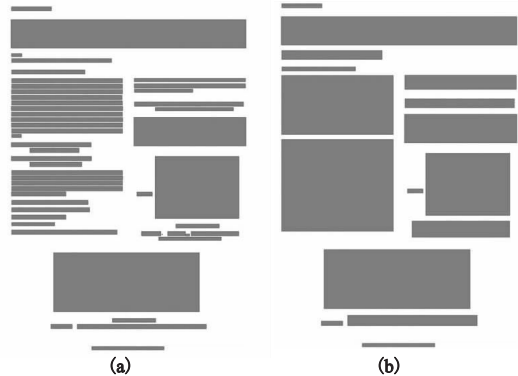


图3 水平方向与垂直方向矩形框合并后图像

1.2 版面分割实验

为了检验算法能否将图像不同属性的版块进行正确分割,本文选取 120 幅传真图像进行实验。图像主要包含英文、中文等语言种类,多为文字、图形、图像和表格等混合排版。在 120 幅图像中,正确分割数目为 107 幅,正确率达到 89.2%。图像分割错误主要包括以下三种情况:

- 标题栏文字之间间距较大以及因为排版原因造成的文字行之间间距过大,从而导致过度分割;
- 不同版面区域之间间距较小,甚至小于文字行之间间距,从而造成的分割不足问题;
- 部分传真图像噪声较多,质量较差导致的分割错误。

2 基于连通区域矩阵的版面分类

2.1 分类特征提取

在版面分割步骤 a) 中得到了每个连通区域的长度与宽度,根据连通区域的长度与宽度信息建立一个矩阵,这个矩阵的横向坐标表示连通区域宽度,纵向坐标表示连通区域高度的矩阵,矩阵的第 m 行第 n 列的数值表示宽度为 m 、长度为 n 的连通区域的个数。文字区域中所包含的文字,虽然字形、笔画不尽相同,但是其长度、宽度具有一定的一致性,而非文字区域不具有这样的特点。因此在连通区域宽度与高度信息生成的矩阵中,文字与非文字区域的数值分布具有不同的特点。

文字区域的连通区域矩阵数值分布较为集中,主要集中在有限的一些点上,而非文字区域数值分布比较广泛。这是因为文字区域连通区域的主体,即文字的大小是有规律的,而非文字区域则毫无规律可言。使用熵和非零比来描述这种差异,分别用 E_T 和 D_T 来表示。假设矩阵大小为 $N \times N$,矩阵元素为 R_{ij} ,进行归一化后为 p_{ij} ,集合 N 表示非零元素的个数。计算公式为

$$E_T = - \sum (p_{ij} \log p_{ij}) \quad (1)$$

$$D_T = \frac{N \{ p_{ij} | p_{ij} > 0 \}}{N^2} \quad (2)$$

文字区域中文字的高度一般是固定的或者是分布在有限的几个值上,因此在连通区域矩阵上的表现为数值分布在一行或者有限的几行。而非文字区域的连通区域高度分布没有规律可循,其连通区域矩阵的数值往往在各行均有分布。对矩阵的水平方向求和得到列向量 p_j ,再计算其熵和非零比,分别用 E_H 和

D_H 来表示,集合 N 表示非零元素的个数,计算公式如下:

$$E_H = -\sum (p_j \log p_j) \quad (3)$$

$$D_H = \frac{N \{p_j | p_j > 0\}}{N} \quad (4)$$

文字区域由于文字的宽度与高度是规律的,因此在矩阵的几个点上具有较大的值,而非文字区域连通区域矩阵的值分布比较散乱。计算两者的平均角度会发现存在较明显差异,平均角度 A 的计算公式如下:

$$A = \sum (p_{ij} \arctan j/i) \quad (5)$$

文字区域与非文字区域距离原点的距离也存在明显的差异。文字区域字符大小有限,其连通区域矩阵的值距离原点较近,而非文字区域往往存在尺寸较大的连通区域,因此其连通区域矩阵距离原点距离较远的值比较多。以距离原点平均距离来衡量这种差异,计算公式如下:

$$D = \sum (p_{ij} \sqrt{i^2 + j^2}) \quad (6)$$

文字区域连通域矩形框内一般为单个的字符或者是字符的组成部分,其黑色像素所占比例比较接近。而非文字区域连通区域大小差异较大,其黑色像素所占比例会有比较大的差异。因此计算每个连通区域内黑色像素所占比例,再计算其均值及方差来衡量两者的差异。

2.2 BP 神经网络的建立

BP 网络是一种按误差反向传播算法训练的多层前馈网络,是在模式识别中使用最广泛的神经网络模型之一,具有较强的分类能力。BP 神经网络分为三层,即输入层、隐含层和输出层,本文中从提取到的八维特征值作为输入层来建立网络模型,经多次实验确定隐含层个数为 18,网络的输出层个数为 2。采用 Sigmoid 激发函数,并运用改进的 BP 算法——Levenberg Marquardt 方法。

分类的过程如图 4 所示,首先提取训练样本的八维特征对 BP 网络进行训练,待分类器性能达到要求后结束训练,然后进行测试样本集的分类。

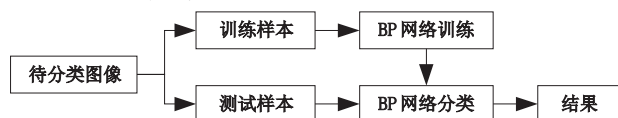


图4 版面分类流程

2.3 版面分类实验

为了验证分类算法的性能,本文对版面分割得到的 1 340 幅子区域图像样本进行人工分类,其中包括文字区域样本 852 幅,非文字区域样本 488 幅,非文字区域样本包括图像、图形和表格等类型。从中各随机选择 200 幅作为训练集,其余样本作为测试集。分类结果用准确率 P 和召回率 R 衡量,准确率和召回率的定义如式(7)(8)所示。表 1 给出了算法的实验结果。

准确率 P :

$$P = \frac{A \text{ 中分类正确的图像数量}}{\text{分类为 } A \text{ 的图像数量}} \quad (7)$$

召回率 R :

$$R = \frac{A \text{ 中分类正确的图像数量}}{\text{样本中 } A \text{ 的实际数量}} \quad (8)$$

其中: A 表示类别中的某一类。

文献[10]中的方法提取出区域的几何、形状规则化、边缘、笔划和空间相关性五类特征用于版面分类。本文对图像样本提取这些特征并进行分类,得到其平均分类准确率为 86%,而本文算法平均准确率达到 94.22%,特别是对文字区域的分类准确率与召回率较高,这对从文字区域提取信息的应用而言非常重要。另外传统算法的版面分割与分类是两个互相独立

的过程,版面分割完成后,版面分类用到的参数需要重新计算,计算量较大,耗时较多。本文分类算法主要利用版面分割过程中获取的信息,从而减少了计算量。此外,分类算法无须提取图像灰度信息特征,更适用于处理传真等二值图像。

对分类错误的版面区域进行分析发现,错误的主要原因是部分文字区域包含字符数量较少,以及部分图形区域具有文字区域的一些特点,从而导致分类错误。如果提取针对这两种情况进行分类的特征,分类准确率有进一步提高的空间。

表 1 版面分类算法实验结果

区域类型	样本数量	准确率/%	召回率/%
文字区域	652	95.89	96.78
非文字区域	288	92.55	90.62
总计	940	94.22	93.7

3 结束语

本文通过对传真图像版面区域的连通区域进行分析,针对传真图像用户信息(传真、电话号码、电子邮箱地址)提取这一实际应用,提出了一种合并水平相邻和垂直相邻连通区域的版面分割方法,以及通过连通区域尺度信息建立连通区域矩阵,提取八维特征向量来描述文字区域与非文字区域之间的差异,并通过 BP 神经网络进行分类的版面分类算法。算法的优点为将版面分割与分类过程相结合,分类过程充分使用分割过程中获取到的信息,减小了计算量,提高了运算速度;另外文字区域分类准确率较高,对从文字区域提取信息的应用有较大价值。实验结果证实算法具有较强的实用性。作为传真图像用户信息提取系统的组成部分,本文算法在实际中也取得了较好的效果。

参考文献:

- [1] 李化,崔荣一.基于游程熵的图文分割算法[J].延边大学学报,2010,36(2):165-169.
- [2] 党兴,龚声蓉,刘全.基于 SVM 区域构造的复杂中文版面分析[J].计算机工程,2010,36(12):200-203.
- [3] NAGY G, SETH S, VISWANATHAN M. A prototype document image analysis system for technical journals[J]. IEEE Computer, 1992,25(7):10-12.
- [4] WONG K Y, CASEY R G, WAHL F M. Document analysis system[J]. IBM Journal of Research and Development,1982,26(6):647-656.
- [5] FAN K C, WANG L S. Classification of document blocks using density feature and connectivity histogram[J]. Pattern Recognition Letters,1995,16(9):955-962.
- [6] KRUATRACHUE B, SUTHAPHAN P. A fast and efficient method for document segmentation for OCR[C]//Proc of IEEE Region 10 International Conference on Electrical and Electronic Technology. 2001:352-354.
- [7] FERILLI S, BIBA M, ESPOSITO F, et al. A distance-based technique for non-Manhattan layout analysis[C]//Proc of the 10th International Conference on Document Analysis and Recognition. Washington DC: IEEE Computer Society,2009:231-235.
- [8] SIMON A, PRET J C, JOHNSON A P. A fast algorithm for bottom-up document layout analysis[J]. Pattern Analysis and Machine Intelligence,1997,19(3):273-277.
- [9] O' GORMAN L. The document spectrum for page layout analysis[J]. Pattern Analysis and Machine Intelligence,1993,15(11):1162-1173.
- [10] 蒋人杰,戚飞虎,徐立,等.基于连通分量特征的文本检测与分割[J].中国图象图形学报,2006,11(11):1653-1656.