软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(2):234−245 [doi: 10.13328/j.cnki.jos.005032]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种面向问卷图像的版面分析算法*

段 露, 宋永红, 张元林

(西安交通大学 软件学院,陕西 西安 710049)

通信作者: 宋永红, E-mail: songyh@mail.xjtu.edu.cn

摘 要: 针对目前已有的问卷图像版面分析算法无法自动识别信息填写区域和无法处理无固定格式的问卷图像等问题,提出了一种连通区域和神经网络相结合的问卷图像版面分析算法.首先获得扫描得到的问卷图像的中心有效图形,接着提出并应用了一种针对问卷图像的快速倾斜矫正方法,对中心有效图像进行倾斜矫正;再利用水平投影进行行分割得到问卷行;然后提取每个问卷行的首个连通区域判断是否存在表格区域即表格问卷行,若存在表格问卷行,则对其进行表格区域分布分析和表格类型判断,得到可能的答案区域,否则直接对文本问卷行进行分析,得到可能的答案区域;最后利用神经网络判断筛选区域的类型,得到最终的答案填写区域.针对问卷图像的实验结果表明,该算法可以准确地识别各种问卷图像中的信息填写区域.

关键词: 行分割;连通区域;表格处理;神经网络

中图法分类号: TP391

中文引用格式: 段露,宋永红,张元林.一种面向问卷图像的版面分析算法.软件学报,2017,28(2):234−245. http://www.jos.org.cn/1000-9825/5032.htm

英文引用格式: Duan L, Song YH, Zhang YL. Layout analysis algorithm of questionnaire image. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):234−245 (in Chinese). http://www.jos.org.cn/1000-9825/5032.htm

Layout Analysis Algorithm of Questionnaire Image

DUAN Lu, SONG Yong-Hong, ZHANG Yuan-Lin

(School of Software, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: The recognition of the information area with common format in the non-fixed format questionnaire is the major problem in existing questionnaire layout recognition algorithm. To address those problems, a new approach for questionnaire layout analysis based on regional connectivity and neural networks is proposed. First, a center valid graphics is generated by preprocessing the scanned image firstly. Then, a rapid skew correction algorithm is applied for questionnaire images. Next, many questionnaire rows are obtained by using horizontal projection profile segmentation algorithms. After that, the first connected region for each row is extracted to estimate the existence of form region. Based on the analysis of general questionnaire row and table row, a large amount of possible answers region are generated. Finally, the neural network is used to determine the type of possible information areas. Experiments show that the proposed algorithm can automatically identify common questionnaire.

Key words: line split; connective region; table recognition; neural network

问卷是在社会研究中用来采集样本数据的工具,主要通过书面或者互联网[1]等形式来获得.在实际调查中,往往需要大量的样本,这给数据的回收和统计工作带来了极大的困难.尽管基于网页,Email 等电子化问卷在一定程度上简化了调查数据的回收和统计工作,但在实际工作中为了兼顾调查地区或对象信息化水平的差异,纸质问卷这种传统模式依然将长期存在.随着应用领域和应用对象的扩展,问卷的答案填写区域也慢慢从固定、

* 基金项目: 国家自然科学基金(61238303)
Foundation item: National Natural Science Foundation of China (61238303)
收稿时间: 2015-10-28; 修改时间: 2015-12-22; 采用时间: 2016-02-03

严格限制的格式逐渐向无固定格式的通用问卷发展.因此,近年来问卷文本图像的版面分析与识别技术已经成为图像分析应用领域比较活跃的话题.

现有的文本图像版面分析与识别定义技术有很多种,但大多数方法都是基于普通的文本图像[2],基于问卷文本图像的版面分析与识别算法还不是很多.特别是,目前已有的问卷文本图像识别方法都需要大量的人工交互且识别的内容版式比较单一,还远远达不到自动识别问卷文本图像的目的.邵中[3]提出了一种基于行块分割和复选框面积特征的复选框识别算法来识别问卷的答案区域,但是它仅能识别复选框类型答案,不能识别其他类型的答案区域,同时还需要利用人工交互来降低误识率.董世超[4]提出了一种基于 OCR 的问卷自动识别统计方法,它是完全依靠人工交互来获得信息填写区域.孙忠礼[5]提出了一种基于连通区域标记方法的答案填写区域识别算法,但是它也只能识别复选框类型的答案,并且需要大量的交互.此外,有学者提出利用表单的配置文件识别表格并从表格中提取数据,根据空白表单建立,且是固定的布局,这种方法需要对文档结构具有先验知识,应用领域有限,不适用于任意格式的文档.

本文提出了一种连通区域与神经网络相结合的问卷图像版面自动分析算法,主要贡献如下:根据问卷图像最左列为题号列,且处于同一直线的特点提出了一种快速的倾斜矫正方法;通过分析问卷行,尤其是分析不同类型的表格行,得到所有可能的答案区域,再利用神经网络筛选出最后的答案区域,即完成问卷图像的自动识别.本文方法不需要任何人工交互和问卷版面的先验知识,可以识别出复杂的问卷版面.

1 问题的描述

作为采集信息的载体,基于不同的功能需求和个性化因素的影响,问卷的设计形式千变万化,在组成结构上呈现出由各种问卷基本元素(题型、复选框、表格、单选框等)构成的排列组合.题目类型的多样性和版面设计形式的多变性,构成了自动识别问卷文本图像的两大难题.基于信息统计分析的考虑,问卷文本图像版面分析的主要任务是从复杂的版面中自动识别出所有可能的答案填写区域.但是,如何自动地在复杂的版面中识别出正确的信息填写区域?要达到自动识别版面进而自动识别答案区域的目的,我们面临的首要问题就是如何适应通用问卷的多样性.尽管目前存在许多针对问卷的版面分析方法,但它们要么需要大量的人工交互,要么仅针对某种特定的类型,这些好像都有自动识别和通用性相差甚远.针对通用问卷样式多样、不定的问题,本文提出了基于连通区域和神经网络相结合的版面分析方法,相应的框图如图 1 所示.

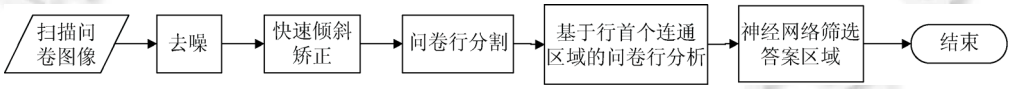


Fig.1 Algorithmic flowchart of this article

图 1 本文算法框图

2 问卷图像的预处理

由扫描仪等图像采集设备获得的图像总有噪声和倾斜等现象,这些因素会给问卷版面分析及后续的识别工作带来一定的影响,需要对图像进行预处理.对图像的噪声[6],本文采用中值滤波方法来去除扫描图像中含有的椒盐噪声.本节主要介绍一种快速的倾斜矫正方法及其有效性.

2.1 问卷图像的快速倾斜矫正

基于问卷图像最左列的像素处于同一条直线上以及图像略微的倾斜并不影响后续分析这一特点,本文提出了一种快速、有效的方法来估计图像的倾斜角.通过搜索问卷图像的最左边界像素点集合 E,再拟合 E 得到 $y=ax+b$,就可以得到倾斜角 θ.校正过程如图 2 所示,倾斜校正基本的流程图及步骤如下.

(1) 提取去噪后图像的中心有效图形 X[7].

(2) 得到每个像素行的首个像素点 $T(a)$.$T(a)$为第 a 行的第 1 个像素点.取 x_a 值最小的点 $T_{\min}$(图像最左像素

点)作为 E 的第 1 个元素, $E=\{(x_{t_{\min}}, y_{t_{\min}})\}$.

(3) 如果 $y_{t_{\min}} > H/2$,则图像向左倾斜,取图像 X 上半部分 x_a 值最小的点 $T^t_{\min}$, 由 $T_{\min}$ 到 $T^t_{\min}$ 从下向上反向搜索图像最左边满足条件的点加入 E;反之,图像向右倾斜,取图像 X 下半部分 x_a 值最小的点 $T^b_{\min}$, 由 $T_{\min}$ 到 $T^b_{\min}$ 自上向下正向搜索图像最左边满足条件的点加入 E.其中,H 为 X 的高.

(4) 将得到的所有满足条件的点用最小二乘法拟合成直线 $y=ax+b$,从而得到倾斜角 θ.

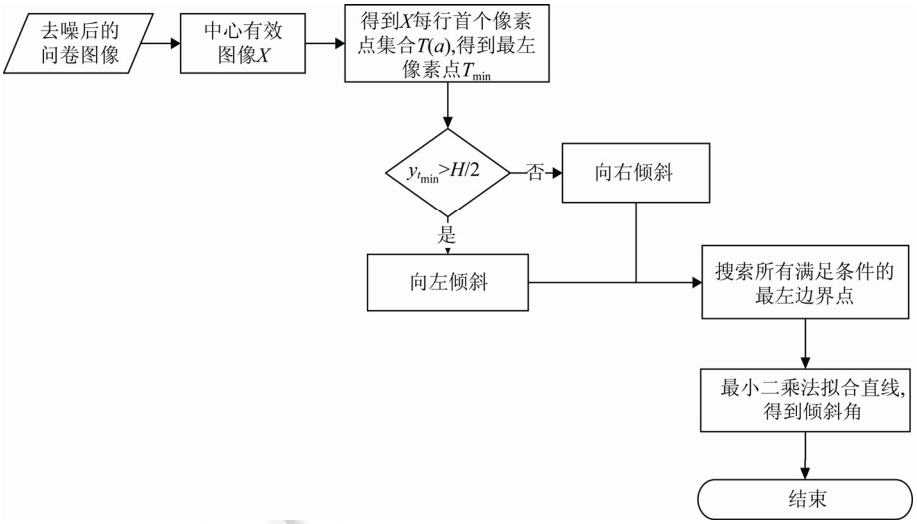
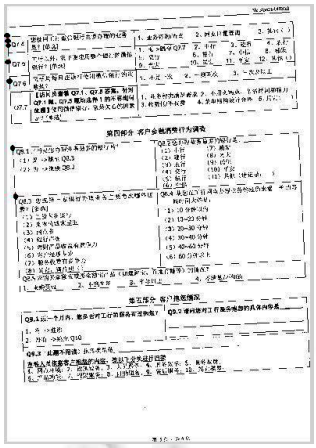


Fig.2 Flow chart of tilt correction

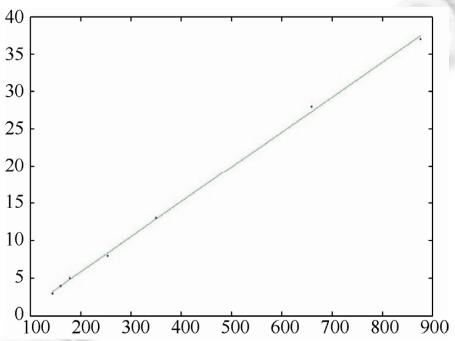
图 2 倾斜矫正流程图

常见的倾斜校正方法可以分为 5 种:基于交叉相关性的方法[8]、基于投影的方法[9]、基于 Fourier 变换的方法[10]、K-最近邻簇方法[11]和基于 Hough 变换的方法[12].但这些方法都需要较多的计算.基于交叉相关性的方法需要维护一个相关矩阵,内存开销较大;基于投影的方法需要在每个角度对图像做投影变化,以时间为代价来获得较高的准确率,计算量很大;基于 Fourier 变换的方法需要从空间域向时域变化,计算复杂性较高;K-最近邻簇方法需要提取大量的连通区域来获取中心点;基于 Hough 变换的方法容易受噪声干扰,时间和空间的代价都很高.为了加快倾斜矫正的速度,我们提出了一种针对问卷图像的快速倾斜矫正方法,该方法主要利用问卷图像最左列为题号列的特点,每次仅搜索图像的边缘像素即可,只需涉及 H 个像素点和最多 H 次比较操作,无论是时间复杂度和空间复杂性都远远低于上述已有的倾斜矫正方法,通过实验,本文方法在 1700×2800 的问卷图像上处理平均时间为 0.135s,具体的实验过程如图 3 所示,图 3(a)为原始的倾斜图像;图(b)为找到的符合条件的点,图中以黑圆点来表示;图(c)为拟合得到的直线;图(d)为最后的倾斜校正结果.

(a) 倾斜图像



(b) 黑圆点为符合条件的点



(c) 拟合直线

(d) 倾斜校正结果

Fig.3 Sample graphs of tilt correction

图 3 倾斜校正过程示例图

3 问卷图像的行分析

根据问卷图像不同行间有明显间隔且问卷图像仅由文本行和表格构成的特点,本文采用投影轮廓分割算法对图像进行水平投影分析,将图像分为不同的行.在得到问卷行之后,首先判断问卷行的类型,再对不同类型的问卷行分别进行分析.本节将主要详细介绍如何对文本行和表格行进行分析,得到所有可能的答案填写区域.

3.1 问卷图像行的类型判定

针对提取出来的每个问卷行,取首个连通区域进行分析,将结果记为 i.

$$O(i)=\begin{cases}1, & h_i > h_{\text{ave}}\\ 0, & \text{other}\end{cases} \tag{1}$$

如果 $h_i>h_{\text{ave}}$ 则该问卷行为表格行;否则,则是文本问卷行. $h_{\text{ave}}=\frac{1}{n}\sum_{i=1}^{n}h_i$, 其中,$h_i$ 为提取得到的第 i 个问卷行的高度,总共有 n 行.

3.2 文本问卷行分析

在确定问卷行为普通的文本问卷行后,本文采用顺序搜索法,从左往右搜索所有的连通区域,得到所有可能

的答案区域,如图 4 所示,具体步骤如下.

(1) 得到文本行的所有 M 个连通区域集合.

(2) 得到选项间的最小间距 d.从左往右按顺序取两个相邻连通区域的间距:

$$d_i = l_{i+1} - r_i,\ i = 1,2,3,...,M-1 \tag{2}$$

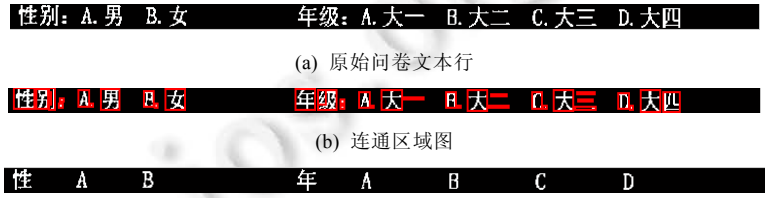
其中,l_i,r_i 是第 i 个连通区域外接矩形的左、右边界.在对所有 d_i 从大到小排序后,取前 k 个值求平均,得到最小间距 d,本文中 k 取值为 8,因为通过统计得知,问卷行中可能存在的选项间隔数不可能超过 8.

$$d = \frac{1}{k}\sum_{i=1}^{k} d_i \tag{3}$$

(3) 从第 2 个连通区域开始从左往右搜索连通区域集合,得到所有潜在的信息填写区域.

$$f(i) = \begin{cases} 1, & d_{i-1} > d \text{且} SUM(r_{i-1}, l_i) = 0 \\ 0, & \text{其他} \end{cases} \tag{4}$$

$SUM(x,y)$表明文本行在横坐标 x 和 y 之间像素点的总数,$f(i)$=1 表明第 i 个连通区域为潜在的答案区域.同时,该文本行首个连通区域也被认为是潜在的答案区域.特别地,如果存在某个连通区域位于问卷行下半部分且宽高比相对较大,则该连通区域为下划线,即答案填写区域.



(a) 原始问卷文本行

(b) 连通区域图

(c) 可能的答案区域

Fig.4 Process of extracting the answers of text line

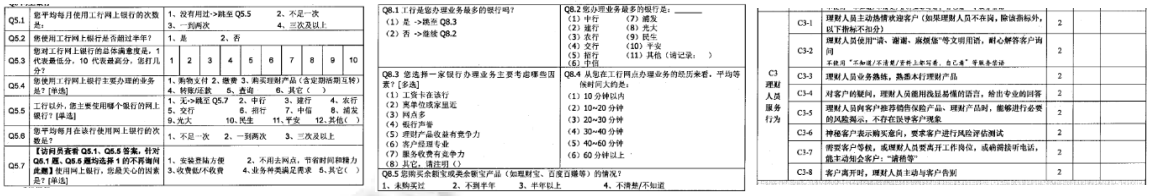
图 4 文本问卷行提取可能答案区域过程

3.3 表格问卷行分析

表格的识别[13]一直是文档图像版面分析的重要领域之一,也是其中的难点.而与普通文档图像中的表格不同,问卷图像中的表格主要存在如图 5 所示的 3 种形式(最后一列单元格为答案选项,单元格为题目,为多区域填写型),此时识别重点在于找到答案填写区域.由于表格中可能存在分隔行(将表格分为不同区域的行)的情况,在分析表格时如果直接对表格进行列分析会产生错误的表格列,所以先只将表格进行行分割,分为表格行.同时,考虑到表格线分为实线和虚线两种,虚表格线的存在会使扫描图像的表格线出现很多断裂的情况,导致断裂表格线被误认为是噪声,考虑到这种情形,本文在后续阶段使用噪声图像来消除这种影响,具体流程图如图 6 所示.

针对表格问卷行的分析和识别算法如下.

1) 对去除点状噪声后的图像继续进行去噪,去除表格图像中面积较小的连通区域.



(a) 最后一列为答案选项 (b) 表格单元为问卷题目 (c) 多区域选择填写型

Fig.5 Major types of tables

图 5 表格的主要类型

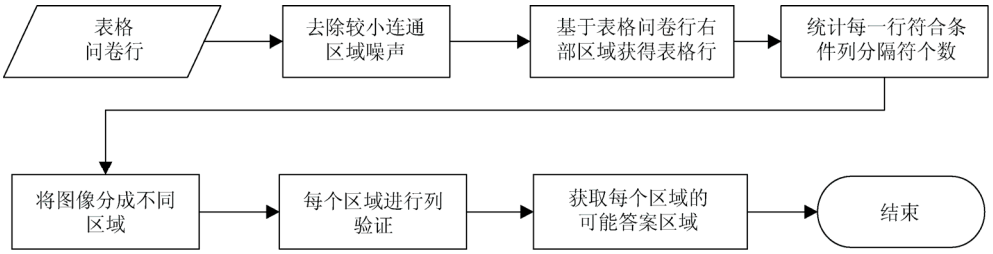


Fig.6 Flow chart of the analysis of table line

图 6 表格问卷行分析的流程图

2) 采用投影法得到表格行.考虑到表格图像左边可能出现部分行缺失的情况,如图 5(c)所示,取图像右边的区域进行水平投影分析,得到表格行,本文取 90%的区域即可达到较好的效果.

3) 统计每一表格行中满足条件的列分隔符个数 num_j,得到 num_j 集合 Num,m 为表格行个数.要满足的条件如下(满足任一条件即可).

(1) 连通区域外接矩形的宽度 $w_{j,p}\leqslant W_{th}$,$h_{j,p}\geqslant h_j-1$,其中,h_j 为当前表格行的高度;$w_{j,p}$ 和 $h_{j,p}$ 分别为第 j 行,第 p 个连通区域的宽度和高度.

(2) 考虑到像素断裂的情况,引入一个水平间隔 d^h,即在$(-d^h,d^h)$的区域且宽度 $w_{j,p}\leqslant W_{th}$ 的连通区域聚合为同一个连通区域,如果新的连通区域的高度 $h'\geqslant h_j/2$ 且参与聚合的连通区域个数大于 2.

4) 表格区域分布分析:对 Num 排序从而得到潜在的分割行,将表格区域分为 t 个区域.t 应该满足的条件为

$$num_{t+1}>\frac{num_1+num_m}{2},\ num_t<\frac{num_1+num_m}{2} \tag{5}$$

(1) 取前 t 个最小行进行分析.

(2) 分隔行验证:在仅含噪声的图像中按次序分别对 t 个潜在分隔行进行验证.验证时将 3)中得到的 num_t 代表的列画在噪声图对应行中,通过垂直投影分析得到投影积分图 Sum,为了得到更有效的极值,取 $sum_y=sum_y-h_t$ 且$-h_t\cdot 0.7$ 为阈值从而得到极大值集合 N_t.

(3) 如果$|N_t|>num_t$,则认为该行为分隔行,否则不是分隔行.

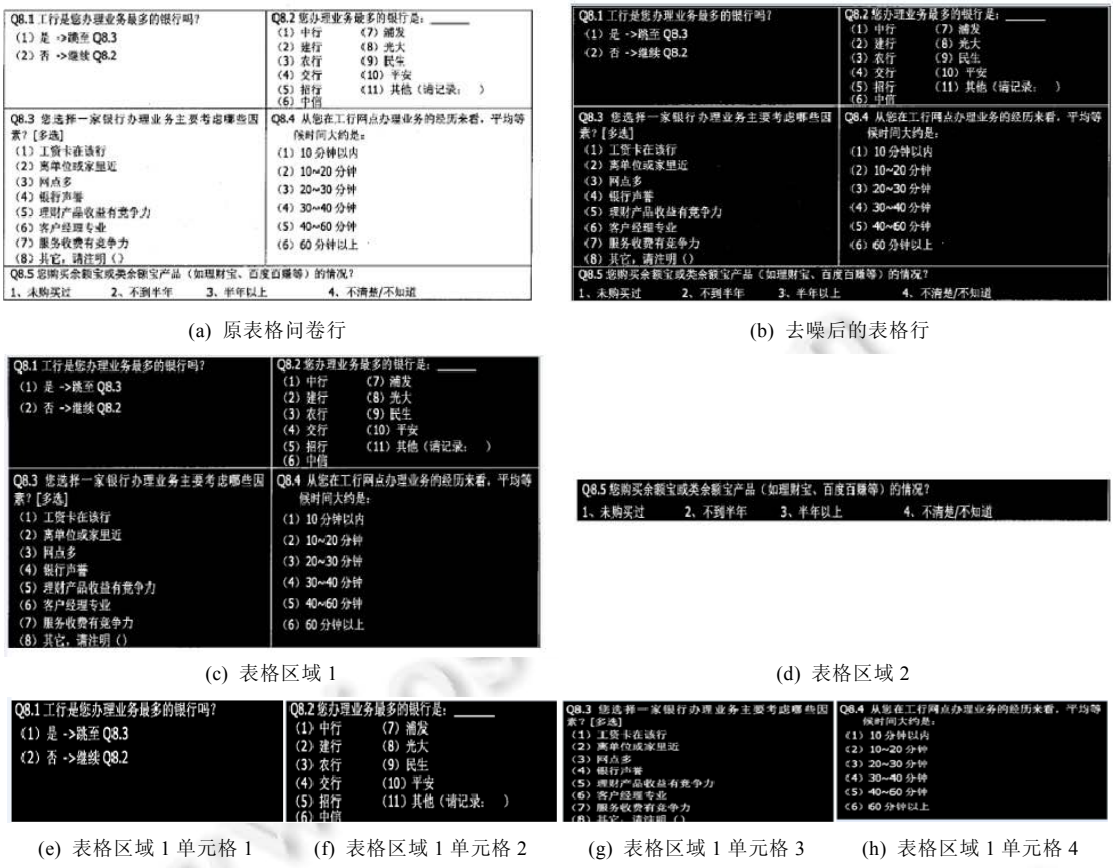
5) 在得到列分隔符后,将图像分为不同的区域,对每一个区域分别得到具有最多和最少列分隔符的行,先对具有最多列分割符和最少列分隔符的行,然后对有最多列分割符的行进行 4)中第(2)步的分隔行验证,再将得到的结果 N_t 加入具有最少列分割符的行进行校验,删除不存在的列,得到的 N_t 为该区域的列位置.

6) 区域表格类型判定:对每个区域进行答案填写区域提取,取每个区域的中间行进行分析.

(1) 如果区域的列数 $N_t\leqslant 2$,则分析所有列中的连通区域个数.如果连通区域个数大于 th,则认为表格单元的内容为问卷题目,将其看成普通文本问卷题目进行处理.本算法中 th 取值为 5,因为对于一个问卷题目而言,最少的题目也应该有 5 个以上的连通区域.

(2) 如果区域的列数 $N_t>2$,则先分析最后一列的表格单元中连通区域个数,记为 c.如果 $c>th$,则最后一列的单元格为问卷题目,将其看成普通文本问卷题目进行处理;反之,如果 $c\leqslant th$,则应该是普通填写的表格单元格,此时应继续向前搜索,直到搜索到某一列连通区域个数 c 发生明显变化时,认为该列的后一列到最后一列为答案区域.尤其是当最后一列的连通区域个数为 0 时,搜索前面所有连通区域个数为 0 的单元格,均为答案区域.

对表格问卷行的分析结果如图 7 所示.



(a) 原表格问卷行 (b) 去噪后的表格行

(c) 表格区域 1 (d) 表格区域 2

(e) 表格区域 1 单元格 1 (f) 表格区域 1 单元格 2 (g) 表格区域 1 单元格 3 (h) 表格区域 1 单元格 4

Fig.7 Analysis of table line

图 7 表格问卷行分析

4 神经网络分析

在得到所有的潜在答案区域后,我们训练了一个神经网络[14]来排除非答案区域.

本文采用一个 2 层的,含有 375 个输入、22 个输出的人工神经网络进行识别.在训练和测试开始之前首先把所有的样本初始化为 25×15 大小,共 375 维,作为本网络的输入,识别的结果总共有 22 类:A,B,C,D,E,F,G,Q,1,2,3,4,5,6,7,8,9,□,○,(,),汉字.识别的逻辑如下.

(1) 如果某行的首个连通区域是汉字或“Q”,后续该行的潜在答案区域中某两个连续的连通区域为“(”和“)”,则这两个连通区域和它们之间的区域为答案填写区域.

(2) 如果某行的首个连通区域是数字,后续该行的潜在答案区域中某两个连续连通区域为“(”和“)”,则这两个连通区域和它们之间的区域为答案填写区域.如果某个潜在答案区域也为数字,则该区域也是答案区域.

(3) 如果某行的首个连通区域是字母或者复选框或圈,后续该行的潜在答案区域中某个潜在答案区域也为同一类型,则该区域也是答案区域.

5 实验结果与分析

我们将本文提出的融合连通区域和神经网络的面向问卷图像的版面识别算法在类别齐全的问卷数据集上进行了实验.

5.1 数据集

本文实验采用的数据集是通过线上和线下搜集的包含 7 种类别,共 184 张图片的测试集.每张图片是由 A4 大小的问卷通过扫描仪在 200dpi 的环境下得到的单色位图,图像大小为 1701×2800.通过调研发现,问卷图像的基本组成单元如图 8 所示,主要有以下 3 种:一行 1 个选项的选择题、一行多个选项的选择题、表格式的题目.本文未将较开放的问答式题目考虑在内,因为其识别比较简单,同时不便于进行自动统计.

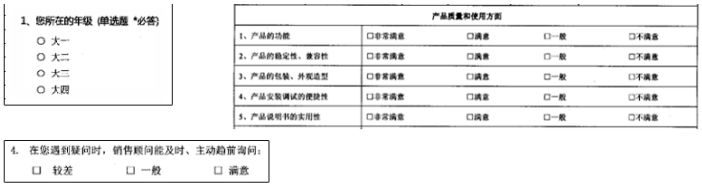


Fig.8 Basic components of the questionnaire

图 8 问卷基本组成类型

基于实验数据完备性的考虑,在选择数据集时,本文将以上 3 种基本组成类型组合成不同的问卷类型,同时考虑到表格类型又分为 3 个大类,所以本文的数据集包括以下 7 种类型:一行一个选项型、一行多个选项型、表格嵌入型、表格单元为问题型、表格单元为选项型、表格行为题目型、混合表格型.

5.2 实验环境

为了验证本文算法的可行性和准确率,在 i5 3.10GHz,4G 内存的 PC 机,Visual Studio 2010 环境下分别对测试集中的 184 幅图像进行实验.

5.3 分析与结果

为了有效地评价本文方法,本文采用识别的查全率和查准率来评价识别结果.

5.3.1 神经网络参数实验

本文采用一个含有 375 个输入、22 个输出的两层神经网络进行符号的识别.为了确定神经网络的学习率和迭代次数,我们做了一系列的实验,实验结果见表 1,神经网络的训练集中共有 22 类,每类 650 张图片,测试集中共有 22 类,每类约 100 张图片.通过图 9 所示的结果可知,在学习率为 0.005,迭代次数为 30 时效果最佳,识别率为 99.94%.

Table 1 Results of 7 types of questionnaire

表 1 7 个类别的实验结果

图像类别	查全率(%)	查准率(%)
一行一个选项型	100	100
一行多个选项型	99.4	100
表格嵌入型	100	100
表格单元问题型	100	100
表格单元选项型	100	100
表格行为题目型	99.65	100
混合表格型	98.78	100
平均	99.57	100

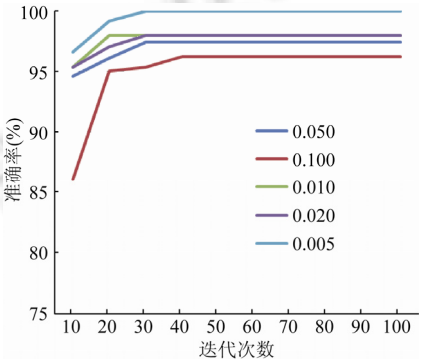


Fig.9 Experiment of NN's parameter

图 9 神经网络参数实验

5.3.2 典型问卷图像分析

为了测试本算法的性能,首先将本算法在一行一个答案型、一行多个答案型、表格嵌入型这 3 类比较典型的问卷上进行实验,实验效果如下.

如图 10 所示,所有问卷文本图像中可能的答案区域都被算法直接用黑框框出来,可以看出,算法能够很好地识别所有的基本类型.

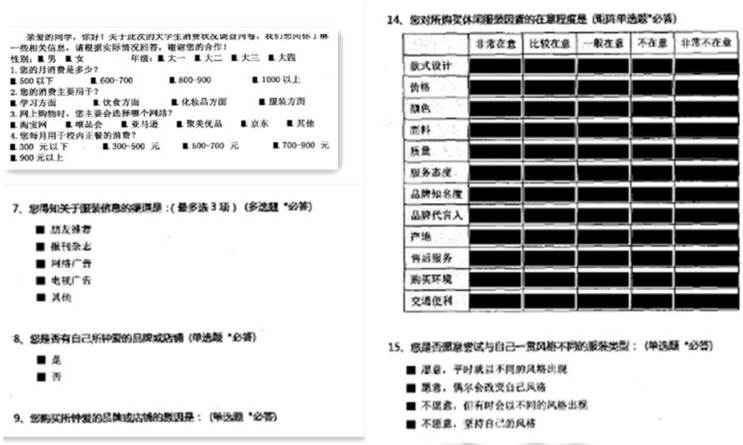


Fig.10 Experimental effect chart

图 10 实验效果图

5.3.3 7 个类别实验结果综合分析

在确定神经网络的结构以后,我们在数据集上验证本文的算法,图 11 所示为部分实验结果,左边是原图,右边用黑色矩形框出的是答案区域,从上到下依次是 7 个类别的结果示例.实验结果表明,本文的算法是有效的.在 7 个类别上分别测试得到的实验结果见表 1.在 1701×2800 的问卷图像上,本文算法的平均处理时间为 4.89s,基本符合实时性的需求.

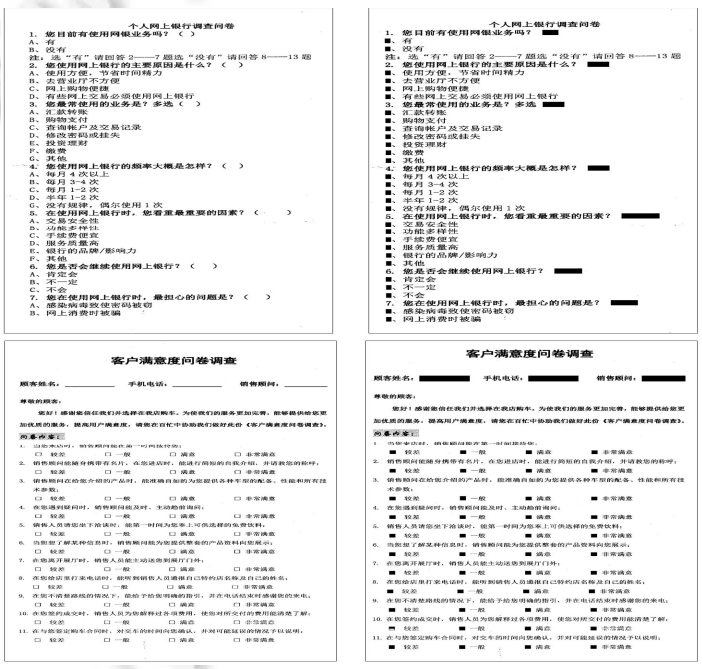


Fig.11 Results of seven categories of questionnaires

图 11 7 个类别的问卷结果

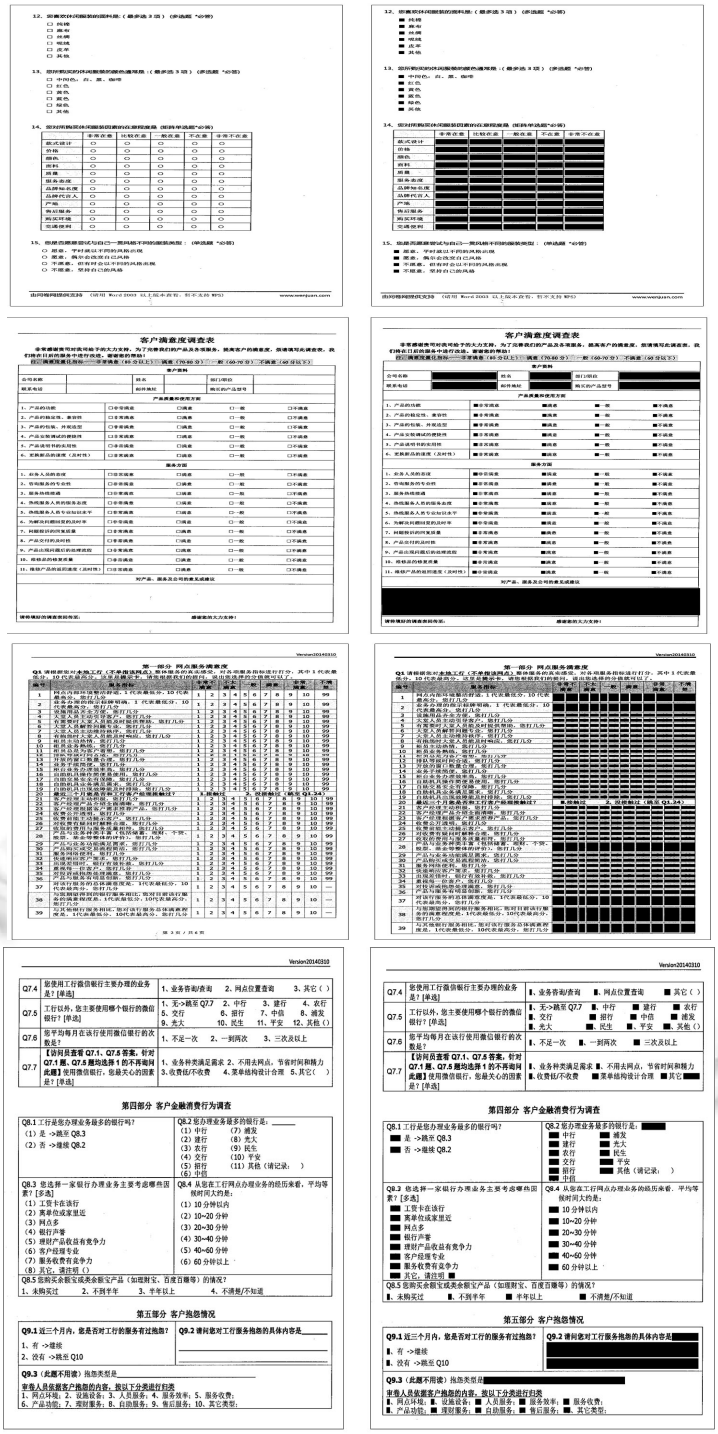


Fig.11 Results of seven categories of questionnaires (Continued)

图 11 7 个类别的问卷结果(续)

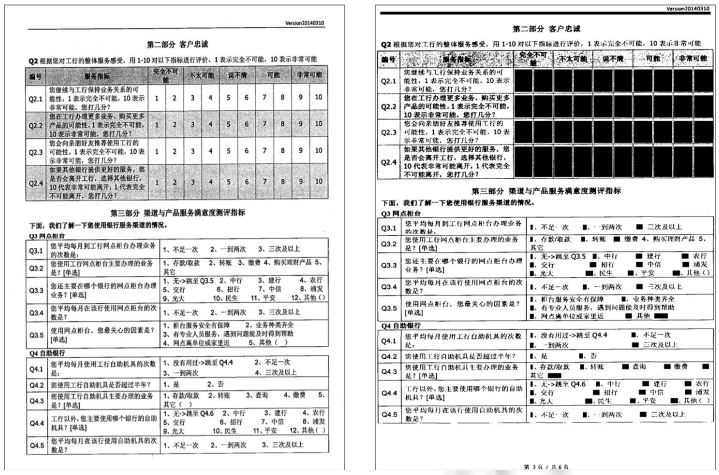


Fig.11 Results of seven categories of questionnaires (Continued)

图 11 7 个类别的问卷结果(续)

从表 1 中可以看出,本文方法在 7 个类别上均得到了很好的识别效果.就查准率而言,本文提出的算法可以达到 100%,同时平均查全率也达到了 99.57%,基本上可以满足实用性的需求,这也为后续的问卷信息统计工作奠定了良好的基础.而这也同时说明了本文的算法基本可以完全适应目前的问卷图像的版面,即不受问卷版面格式的约束.在混合型表格问卷,即第 7 类和第 2 类上查全率没有达到 100%,主要是由于半色调阴影图像造成的漏检,但总的来说,本文算法具有很强的实用性.

6 结 语

本文提出的面向调查问卷的版面识别算法将连通区域分析与神经网络结合在一起,同时充分运用了问卷的特有特性,不需要任何人工交互工作,真正达到了自动识别的目的.并且本文的测试集中包含了多种可能含有的问卷类型,而不是仅仅适用于特定的版面,因而本方法可以普及到通用的问卷识别中,具有通用性.

同时,本文提出的算法具有很强的可拓展性.当出现神经网络可识别的 22 种类型以外的类型时,仅需要在神经网络中加入新的类型重新训练即可.

通过与已有方法相比较,本文提出的方法具有明显的优势,因为不需要任何人工交互工作,无需定义固定的版面.但是,该方法在表格中出现较多的阴影区域,即半色调图像时,识别会出现错误.解决这个问题将成为我们今后工作的重点.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是向西安交通大学软件学院的宋永红老师、张元林副教授领导的实验组的老师和同学表示感谢.

References:

[1] Lan Y. Questionnaire survey and visual analysis system on the net [MS. Thesis]. Jilin: Jilin University, 2005 (in Chinese with English abstract).

[2] Liu J, Tang YY, Suen CY. Chinese document layout analysis based on adaptive split-and-merge and qualitative spatial reasoning. Pattern Recognition, 1997,30(8):1265−1278. [doi: 10.1016/S0031-3203(96)00165-3]

[3] Shao Z. Research and software design of automatic statistic method based on image processing [MS. Thesis]. Shenyang: Shenyang University of Technology, 2011 (in Chinese with English abstract).

[4] Dong SC. Development and design of questionnaire automated statistical analysis system based on OCR [MS. Thesis]. Shenyang: Shenyang University of Technology, 2012 (in Chinese with English abstract).

[5] Sun ZL. Design and implement of survey layout definer oriented survey automatic identification system [MS. Thesis]. Shenyang: Shenyang University of Technology, 2014 (in Chinese with English abstract).

[6] Fan KC, Wang YK, Lay TR. Marginal noise removal of document images. Pattern Recognition, 2002,35(11):2593−2611. [doi: 10.1016/S0031-3203(01)00205-9]

[7] Padfield D. Masked FFT registration. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2010. 2918−2925. [doi: 10.1109/CVPR.2010.5540032]

[8] Chaudhuri A, Chaudhuri S, Robust detection of skew in document images. IEEE Trans. on Image Processing, 1997,6(2):344−349. [doi: 10.1109/83.551708]

[9] Kwag HK, Kim SH, Jeong SH, Lee GS. Efficient skew estimation and correction algorithm for document images. Image and Vision Computing, 2002,20(1):25−35. [doi: 10.1016/S0262-8856(01)00071-3]

[10] Peake GS, Tan TN. A general algorithm for document skew angle estimation. In: Proc. of the Int'l Conf. on Image Processing, Vol.2. IEEE, 1997. 230−233. [doi: 10.1109/ICIP.1997.638728]

[11] Lu Y, Tan CL. Improved nearest neighbor based approach to accurate document skew estimation. In: Proc. of the 7th Int'l Conf. on Document Analysis and Recognition. IEEE, 2003. 503−507. [doi: 10.1109/ICDAR.2003.1227716]

[12] Srihari SN, Govindaraju V. Analysis of textual images using the Hough transform. Machine Vision and Applications, 1989,2(3): 141−153. [doi: 10.1007/BF01212455]

[13] Watanabe T, Luo Q, Sugie N. Layout recognition of multi-kinds of table-form documents. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1995,17(4):432−445. [doi: 10.1109/34.385976]

[14] Rowley H, Baluja S, Kanade T. Neural network-based face detection. Pattern Analysis and Machine Intelligence, 1998,20(1):203−208 [doi: 10.1109/34.655647]

附中文参考文献:

[1] 蓝鹰.基于.NET 的网上问卷调查及其可视化分析系统[硕士学位论文].吉林:吉林大学,2005.

[3] 邵中.基于图像处理的自动统计方法研究与软件设计[硕士学位论文].沈阳:沈阳工业大学,2011.

[4] 董世超.基于 OCR 的调查问卷自动识别统计分析系统的开发与设计[硕士学位论文].沈阳:沈阳工业大学,2012.

[5] 孙忠礼.面向问卷自动识别系统的版面定义器的设计与实现[硕士学位论文].沈阳:沈阳工业大学,2014.

段露(1993－),女,湖北荆州人,硕士生,主要研究领域为图像处理,模式识别.

张元林(1968－),男,副教授,主要研究领域为机器学习,计算视觉,文本图像处理.

宋永红(1967－),女,高级工程师,主要研究领域为图像处理,智能系统,软件工程新技术.