手持终端自动阅卷的表格数据定位算法的研究

夏 禾，冯军焕
（西南交通大学 信息科学技术学院，四川 成都 610031）

摘 要：针对手机等手持终端的客观题自动阅卷系统，提出了一种基于区域生长的表格定位算法，该算法对表格外框的确定有较好的适应性，并且对表格外噪声数据有较强的抗干扰能力。测试表明，该算法提高了表格定位精度和阅卷准确率。

关键词：表格识别；模板匹配；OpenCV；Android
中图分类号：TP391.41 **文献标识码**：A **文章编号**：1674-7720(2014)06-0033-04

Study of a form localization algorithm for automatic marking for handheld terminals

Xia He, Feng Junhuan
(School of Information and Science Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: This thesis presents a form localization algorithm based on region growing for objective topic automatic marking system for handheld terminal such as mobile phone. The algorithm has good adaptability for determining form frame and anti-interference ability to the outside noise of form. Tests show that this algorithm improves positioning accuracy and the accuracy for marking.

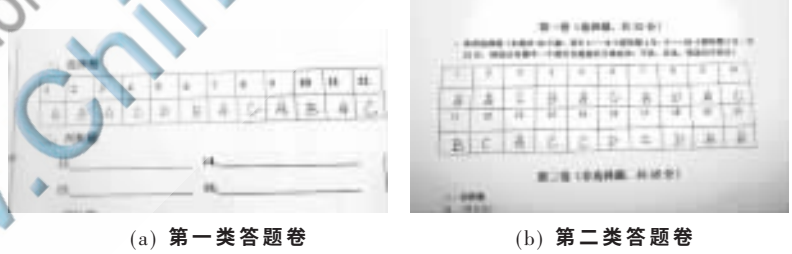
Key words: form recognition; pattern matching; OpenCV; Android

传统的人工阅卷方式往往花费大量的人力物力和时间，效率低下，却不能得到一个很好的效果。因此，采用一种高效准确的阅卷方式显得尤为重要。参考文献[1]提出了一种基于图像识别的阅卷系统，这种阅卷系统需要对现有的常用答题卷格式进行改造，并且需要用扫描仪对学生的答题卷进行扫描，这势必会增加考试阅卷的复杂度。近年来，带有照相机功能的智能手机等移动终端快速增长且广泛使用，为高效地自动智能阅卷提供了很好的解决方案。

本文利用带有照相机功能的智能手机设计实现自动阅卷，教师只需将标准答案输入 Android 智能手机中，然后对学生答题卷进行拍照，如图 1 所示，通过手机阅卷软件便可以轻松得到学生客观题的考试成绩。这种阅卷方式可以降低人工评卷误差，减少教师阅卷工作量。

1 智能手机自动阅卷系统设计

各类考试的试卷中一般都会包括客观题部分，大多数考试答题卷的客观题部分都是由表格设计构成的。本文重点介绍系统在 Android 平台下识别表格图像，实现智能手机自动阅卷功能，具体流程如图 2 所示。



(a) 第一类答题卷 (b) 第二类答题卷

图 1 学生答题卷

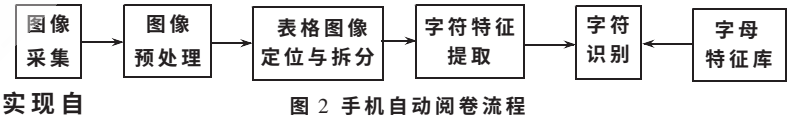


图 2 手机自动阅卷流程

1.1 图像采集

与扫描仪获取图像资源相比，采用智能手机中的照相机获取图片的方式具有方便、快捷、应用一体化等特点。通过 Android 操作系统自带的照相机进行图像采集，并将采集到的图像存储到 SD 卡中等待后台应用程序进行处理。

1.2 图像预处理

照相机获取的图片携带了颜色特征，所以首先需要对图像进行二值化操作。除此之外，为了取得更好的字

符定位效果和字符识别精度，还需要对图像进行倾斜校正等操作。

1.2.1 图像二值化

由于文档图片的颜色特征比较单一，因此本文采用参考文献[2]提出的基于简单阈值的 Ridler 和 Calvard 的聚类算法，取得了较好的二值化效果。

1.2.2 图像的倾斜校正

答题卷在被照相机拍照的过程中，或多或少都会存在一定的倾斜。倾斜会给字符的分割和识别带来很大的影响，所以处理表格图像很关键的一步是图像的倾斜校正。

由参考文献[3]可知，传统的校正方法可通过 Hough 变化求出水平线的角度后，采用仿射变换调节图像的角度。该方法因为采用了 Hough 变化，需要耗费大量的时间。参考文献[4]提出了一种基于纵坐标投影峰值的表格图像倾斜调整算法，通过计算峰值之间的距离来度量表格的倾斜度。本文采用这种方法来进行倾斜矫正，该方法适用于带有表格信息的二值图像，其算法计算量较小，在实际中有一定的使用价值。

1.3 表格数据定位与拆分

往往一张答题卷分为多个部分，而客观题部分只是其中一个模块，所以表格数据定位拆分模块主要负责表格边框的定位，并且有效地对表格中的数据进行拆分。

1.4 表格数据识别

答题卷的客观题部分一般主要由答题序号（也就是印刷体数字）和答案（手写体 A、B、C、D）构成。在进行数字和字符识别之前，通过参考文献[4]提供的方法可以使每个字符边缘没有空白区域，如图 3 所示。

图 3 字符预处理后

1.4.1 手写字符特征提取算法

字符特征提取的好坏直接影响着最终的识别效果。参考文献[5]提供了一些字符的特征提取应遵循的规则。参考文献[6]提出了一种针对手写数字适应性较强的模板法——13 点特征提取法，该方法具有实时、快速和准确等特点，但是该方法遇到大小不一的图片时就会难以识别。在此基础上，本文采用此方法归一化的结果来提取特征以适用于手写英文字符的识别。

1.4.2 手写体字符的识别

本文采用模板匹配进行字符识别，将标准的书写方式作为模板录入模板库中，利用图像之间的最短距离作为判别函数。对于一个待测试的样本 $X=(a_1,a_2,\cdots,a_n)^T$，计算 X 和训练集中的某个样本 $X_j(0<j<N,\ N$ 为训练集中样本的个数）之间的距离循环求出待测样本和。训练集中各个已知样本之间的距离为 d_j，比较所有的 d_j 值，找出最小的 d 作为 X 所属的训练集类别。

$$d_j=[\sum_{i=1}^{n}(a_i-a_{ji})^2]^{1/2} \tag{1}$$

2 表格数据定位算法优化

目前，表格框线检测算法有很多种，如 Hough 变换算法、投影法和交叉点特征法等。这些算法在传统的表格识别中都能运用于单一表格的拆分和框线的识别，但是对于表格外噪声数据的抗干扰能力较差。本文结合了 Hough 变换算法与交叉点特征法，实现了一种基于区域生长的表格定位算法。

2.1 算法假设

(1)图像中存在的表格是封闭的矩阵单元。

(2)图像中的待识别区域只存在单一表格，不存在多个表格混合识别的情况。

(3)图像中识别区域以外的数据应该与表格有一定的距离。

(4)图像中表格数据的直线检测允许出现断线，但不允许出现整个表格完全断裂。

2.2 算法设计

要对表格边框进行定位，首先需要对表格线进行检测。在答题卷中，经过倾斜校正后，常见的表格线分为水平横线和垂直竖线两种。

目前，Hough 变换是在图像中寻找直线的最佳方法。OpenCV 提供了基本 Hough 变换和累计 Hough 变换[7] 两种 Hough 变换算法。设定一定阈值 T，采用比较直线起点与终点的横坐标与纵坐标来对线条进行分类。设一条直线 $l(x_1,x_2)$ 起点为 $x_1(i_1,j_1)$，终点为 $x_2(i_2,j_2)$，则分类规则为：

$$\begin{cases} l(x_1,x_2)\in \text{横线},\ (|j_1-j_2|\leqslant T) \\ l(x_1,x_2)\in \text{竖线},\ (|i_1-i_2|\leqslant T) \end{cases} \tag{2}$$

提取表格边线后，可以得到大量的横线与竖线的线条集合，如图 4 所示。可以看出这些集合不仅包含图像中待识别的表格部分的线条，同时也包含表格外的线条。

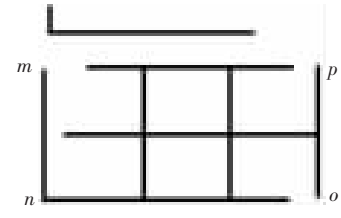


图 4 通过 Hough 算法得到的直线集

2.3 算法描述

本文针对横线集和竖线集中包含非目标区域的表格线的问题以及横线集与竖线集断线的问题，改进了表格框线检测提取算法中的 Hough 变换算法和交叉点特征法[8]，提出了一种基于区域生长的表格定位算法。算法具体流程如下。

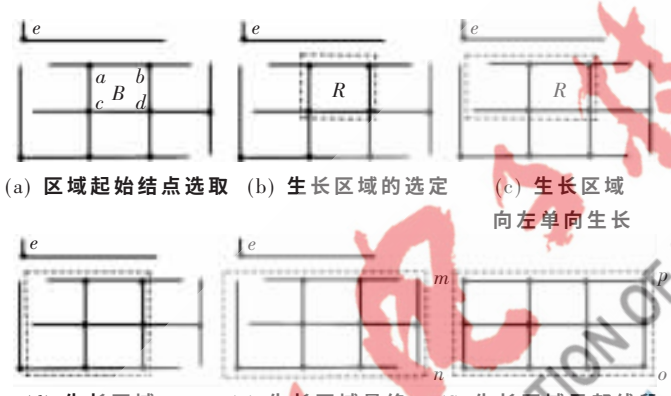
(1)对倾斜校正过后的图像表格线进行检测，将检测到的线条进行归类，得到横线集与竖线集。设横线集为 $U1$，竖线集为 $U2$，通过与操作可以粗略得到交点集 C，而最终需要得到的表格线外区域由 4 点集 B 构成。同时假设 C 中包含孤立交点为 e。

(2)从 C 中选取 4 个相邻交点作为 R 的起始节点或叫种子节点,如图 5(a)所示,区域应尽量小,且不包含其他交点,并且 a 与 d 和 b 与 c 形成横线关系,a 与 b 和 c 与 d 形成竖线关系。

(3)生长区域选定,扩充区域 R 进行生长,这样可以填充掉断线区域。选定阈值 T,将 R 区域向 4 个方向扩充 T 个像素,如图 5(b)所示(T 的选择应该适中,选择过大,会导致区域错误生长;选择过小,区域不能正常生长)。

(4)迭代横线集 $U1$ 与竖线集 $U2$ 中的线段,若线段 L 与生长区域相交,则 R 区域向线段 L 扩充直至 L 末端为止,形成新的 R 区域,返回步骤(3)进行新生长区域的选定;若不存在 L 与生长区域相交,则跳转至步骤(5)。图 5(c)、图 5(d)、图 5(e)所示为表格生长区域的生长过程。

(5)对最终得到的 R 区域运用局部 Hough 算法,得到线段集合,线段集合将 R 区域分割开来形成最终的识别区域。将新生成的线段集合中的线段延长可以得到所有的交点,如图 5(f)所示。



(a) 区域起始结点选取 (b) 生长区域的选定 (c) 生长区域向左单向生长

(d) 生长区域向下单向生长 (e) 生长区域最终生长结果 (f) 生长区域局部线段延长的结果

图 5 表格识别过程举例

由区域算法的特性可知,如果选择适当的阈值 T,可以在生长过程中找到区域的边界 $\{m,n,o,p\}$,同时不会包含表格外的直线和交点 e。

2.4 算法原理

(1) 算法中,生长区域 R 由 4 个满足一定条件的相邻交叉点包围的区域 B 向外扩充阈值 T 所构成。由于 B 为 4 个相邻交叉点构成的区域,则 $R\cap(U1\cup U2)!=\Phi$,则必定可以选择一条相交直线进行延伸。

(2) 假如在延伸过程中,一条直线 l 与表格的其他直线之间存在断线,但是断裂距离小于 T,则在生长过程中,必定有 $l\cap R_n!=\Phi$(R_n 为 R 的第 N 次生长结果),生长区域 R 则可以沿该直线扩充。

(3) 由于算法假设中整个表格不可能出现完全断裂(这里完全断裂是指小模块与表格其他线段的最小距离大于 T),因此通过线段连通性可以证明区域多次扩展后,最后的生长区域 R 在包含所有的表格区域后停止生长。

(4) 由于假设中表格数据与表格外的数据之间存在一定的距离(假设距离大于 T),则最终的生长区域不包含表格外数据。

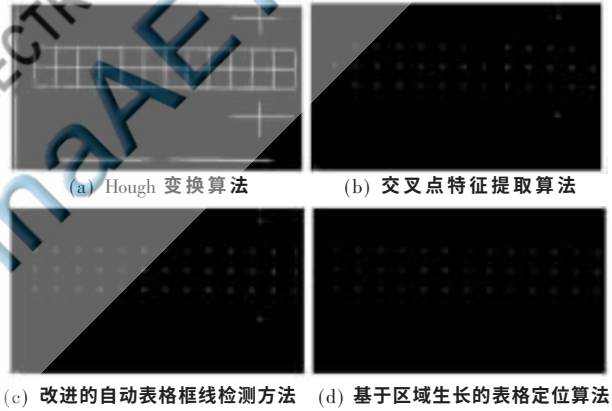
3 测试结果及分析

本文采用三星 i9300 作为测试平台,Android 4.0.4 的系统版本。本文采用 OpenCV2(版本号为 2.4.5)作为图像处理的库。

3.1 几种常见定位算法的对比

对带有表格外噪声数据的考试试卷(如图 6 所示)运用不同表格框线检测算法进行表格数据定位。分别通过 Hough 变换表格框线检测算法[8]、交点特征提取算法[9]、由郑秀清等人提出的一种改进的自动表格框线检测算法[10]和本文提出的基于区域生长的表格定位算法得到的结果如图 7 所示。

图 6 带有表格外噪声数据的考试试卷



(a) Hough 变换算法 (b) 交叉点特征提取算法

(c) 改进的自动表格框线检测方法 (d) 基于区域生长的表格定位算法

图 7 各种不同的表格框线检测算法的测试结果

Hough 变换检查直线算法输入原始图片后,可以检测到图像中的所有直线。将这些直线中的横线与竖线选择出来,可以构成检测结果,如图 7(a)所示。由结果可以看出,Hough 算法得到的线段除表格框线外,还存在表格外的直线,利用算法得到的直线集很难区分出哪些是表格线,哪些是非表格线,因此很难对表格区域进行有效的定位。

交点特征提取算法将横线与竖线的交点作为特征,通过交点可以大体描绘出表格区域。然而由于图像采集中光线不均匀与图像预处理的误差,容易出现漏检短框线的问题,从而会直接影响到交点的检测,如图 7(b)中出现部分交点漏检的情况。

郑秀清等人提出的改进的自动表格框线检测算法主要应用数学形态学来对表格框线进行处理,对表格线

进行了断点补偿，很好地去除噪声带来的困扰，但是同样不能过滤掉表格外的孤立交点，如图 7(c)所示。

通过区域定位算法则可以很好地进行断线补偿并且过滤掉表格区域以外的孤立交点，如图 7(d)所示，采用该方法可以得到表格数据的交点集，对表格数据进行有效的定位。

随机选取 20 种常见的不同类型的表格型试卷，通过图像采集得到 80 张图片作为图像库，分别用不同方法进行答题区域的分割，结果如表 1 所示。对比结果表明，基于区域生长的表格定位算法在图片较清晰、满足算法假设条件的情况下有较高的定位准确率。

表 1 几种常见的表格定位算法的对比

	样本数	正确定位样本数	正确定位率/%
交叉点特征提取算法	80	23	29.8
改进的自动表格框线检测算法	80	61	76.2
基于区域生长的表格定位算法	80	73	90.1

3.2 常见答题卷的定位结果

由于考试科目的不同，答题卷的类型有所不同，客观题的数量也有所不同。本文对不同格式的几种试卷进行了测试。用 Android 手机采集其中两种类型的试卷照片，经过图像的二值化、图像去噪、倾斜校正、Hough 直线检测和区域定位等算法后，可以得到有效的识别区域，如图 8 所示。

(a) 第一类答题卷定位结果 (b) 第二类答题卷定位结果

图 8 学生答题卷区域定位结果

传统的考试答题卷客观题部分主要存在于表格中，针对这种答题卷，本文针对 Android 移动终端设计了一种高效易用的自动阅卷方案，包括图像的采集、预处理、表格图像的识别和字符识别等。针对表格图像定位与手写字符识别算法进行了优化，实验结果表明，该自动阅卷方案具有高效、简单易用、便于推广的优点。

参考文献

[1] 张站,刘政怡.基于图像识别的阅卷系统的设计与实现[J].微型机与应用，2011,30(4):44-47.

[2] RIDLER T W, CALVARD S. Picture thresholding using an iterative threshold selection method[J]. IEEE Transactions on Systems, Man and Cybernetics, 1978,8(8):630-632.

[3] 谢亮. 表格识别预处理技术与表格字符提取算法的研究[D].广州：中山大学，2005.

[4] 巨志勇，郑应平. 二值表格图像倾斜校正算法[C]. 第一届中国高校通信类院系学术研讨会论文集，2007.

[5] 许雁飞，陈春玲，陈夏梅. 基于 OpenCV 的脱机手写字符识别技术[J]. 信息与电脑(理论版)，2011(8):39.

[6] 钟乐海，胡伟. 手写体数字识别系统中一种新的特征提取方法[J]. 四川大学学报（自然科学版），2007(5):15.

[7] KIRYATI N, ELDAR Y, BRUCKSTEIN A M. A probabilistic Hough transform[J]. Pattern Recognition,1991, 24(4):303-316.

[8] ILLINGWORTH J, KITTLER J. A survey of the Hough transform[J].Computer Vision, Graphics,and Image Processing, 1988,44(1):87-116.

[9] WENYIN L, DORI D. From raster to vectors: extracting visual information from line drawings[J]. Pattern Analysis & Applications, 1999,2(1):10-21.

[10] 郑秀清,付茂名.一种改进的自动表格框线检测方法[J].中国民航飞行学院学报，2004,15(4):30-32.

（收稿日期：2013-11-25）

作者简介：

夏禾，男，1989 年生，硕士研究生，主要研究方向：数字图像处理，手机移动开发。

冯军焕，男，1961 年生，博士，教授，主要研究方向：移动通信，数字图像处理等。