



(12) 发明专利

(10) 授权公告号 CN 109543525 B

(45) 授权公告日 2020.12.11

(21) 申请号 201811217691.5 CN 103258198 A, 2013.08.21

(22) 申请日 2018.10.18 CN 106156761 A, 2016.11.23

(65) 同一申请的已公布的文献号
申请公布号 CN 109543525 A CN 106897690 A, 2017.06.27

(43) 申请公布日 2019.03.29 CN 107958201 A, 2018.04.24

(73) 专利权人 成都中科信息技术有限公司
地址 610000 四川省成都市高新区天府大道北段1700号2栋1单元15楼1501号 CN 106446881 A, 2017.02.22

(72) 发明人 边赞 李天易 罗嘉礼 巫浩
李腾飞 倪浩原 刘艳顺等.《一种基于自适用结构元素的表格框线去除形态学算法》.《贵州大学学报(自然科学版)》.2008,第25卷(第4期),第350-353,361页.

(74) 专利代理机构 成都顶峰专利事务所(普通合伙) 51224 H. Conrad Cunningham等.《Designing a Flexible Framework for a Table Abstraction》.《Springer US 2009》.2009,第1-38页.

代理人 曾凯 刘昱.《印刷体表格识别的研究》.《中国优秀硕士学位论文全文数据库 信息科技辑》.2014,(第4期),第I138-979页.

(51) Int. Cl. 审查员 赵会玲
G06K 9/00 (2006.01)

(56) 对比文件
CN 108491788 A, 2018.09.04 权利要求书2页 说明书6页 附图6页

(54) 发明名称

一种通用表格图像的表格提取方法

(57) 摘要

本发明属于办公自动化领域,公开了一种通用表格图像的表格提取方法,包括如下步骤:S1、对通用表格图像中进行预处理;S2、使用形态学操作将预处理后图像中的文字进行过滤;S3、对过滤处理后图像进行重构操作;S4、进行表格重绘,实现表格提取;本发明解决了现有技术存在的表格提取的速度慢,精确度不高的问题。

1. 一种通用表格图像的表格提取方法,其特征在於:包括如下步骤:

S1:对通用表格图像中进行预处理;

S2:使用形态学操作将预处理后图像中的文字进行过滤;

S3:对过滤处理后图像进行重构操作,包括如下步骤:

S3-1:进行开运算图像重构,包括如下步骤:

A-1:将输入的过滤处理后图像即开运算图像初始化为原始标记图像;

A-2:创建第二结构元素;

A-3:根据原始标记图像和第二结构元素进行重构操作,其计算公式为:

$$h_{k+1} = (h_k \oplus B) \cap g$$

式中, h_{k+1} 为当前标记图像; h_k 为上一代标记图像; k 为迭代参数; B 为第二结构元素; \oplus 为膨胀操作函数; g 为掩模图像; \cap 为交集;

A-4:判断当前标记图像和上一代标记图像是否相同,若是则实现重构,输出得到的当前标记图像,否则更新迭代参数,并返回步骤A-3;

S3-2:提取获得的标记图像的线条特征坐标,包括如下步骤:

B-1:将二值化图像采用二维数据进行表示;

B-2:将二维数组的第一个元素作为投影原点,将其他元素进行水平和垂直方向的投影,并得到对应投影方向的累加值;

B-3:遍历累加值,筛选出所有符合条件的特征坐标,即特征坐标对应的累加值大于两投影方向坐标横线之间的最小距离;

B-4:对特征坐标进行进一步分析,当有一个以上符合条件的特征坐标时,提取其中间值;

B-5:遍历二维数组的其他元素,重复步骤B-2到B-4,得到特征坐标数组;

S3-3:根据特征坐标数组,对图像交点进行分类,并计算标记图像中表格的权重数组,包括如下步骤:

C-1:根据表格的规则线段与网格点相连的方式,划分成不同的类型;

C-2:根据得到的特征坐标数组即横纵线坐标数组,遍历其中每个交点;

所述交点为表格中的横线和纵线的交叉点;

C-3:以当前交点为起始点,依次提取四个方向的图像块,依次判断是否存在一条直线,若存在输出结果为1,否则输出结果为0;

四个方向为:横正向、横负向、纵正向以及纵负向;

C-4:根据输出结果将交点进行分类,并将输出结果保存在建立的对应三维数组中,并根据修改规则对所有三维数组进行修改和校验;

所述修改规则为:如果交点处存在横正向直线,则此交点右边交点必然存在横负向直线,其他方向同理;

C-5:根据三维数组中对应的每个方向的权重,得到每个交点的权重和,并保存在权重数组中,其计算公式为:

$$p(x, y) = q_1 * p(x, y, 0) + q_2 * p(x, y, 1) + q_3 * p(x, y, 2) + q_4 * p(x, y, 3)$$

式中, $p(x, y)$ 为权重数组; $q_1 * p(x, y, 0)$ 为横正向对应的权重和三维数组的乘积; $q_2 * p$

$(x, y, 1)$ 为横负向对应的权重和三维数组的乘积; $q_3 * p(x, y, 2)$ 为纵正向对应的权重和三维数组的乘积; $q_4 * p(x, y, 3)$ 为纵负向对应的权重和三维数组的乘积;

S4: 进行表格重绘, 实现表格提取;

根据交点的类型以及权重数组, 遍历每一行每一列, 查找对应的开始端点坐标和结束端点, 并将其连起来, 完成表格的重绘, 实现表格的提取。

2. 根据权利要求1所述的一种通用表格图像的表格提取方法, 其特征在于: 所述的步骤S1中, 预处理包括依次进行的缩小处理、灰度处理以及二值化处理。

3. 根据权利要求2所述的一种通用表格图像的表格提取方法, 其特征在于: 所述的步骤S1中, 预处理的计算公式为:

使用加权法进行灰度值处理, 计算公式为:

$$F(i, j) = 0.3R(i, j) + 0.59G(i, j) + 0.11B(i, j)$$

式中, $F(i, j)$ 为像素点的灰度值; $R(i, j)$ 、 $G(i, j)$ 、 $B(i, j)$ 对应为像素点的红绿蓝三种颜色的亮度值; (i, j) 为像素点坐标;

进行二值化处理的计算公式为:

$$g(x, y) = \begin{cases} 1 & f(x, y) \geq T \\ 0 & f(x, y) < T \end{cases}$$

式中, $g(x, y)$ 为二值化图像; $f(x, y)$ 为输入的灰度图像; T 为阈值。

4. 根据权利要求3所述的一种通用表格图像的表格提取方法, 其特征在于: 将类间方差最大的灰度值作为二值化处理的最佳阈值, 其计算公式为:

$$T^* = \arg \max [g(t)]$$

式中, T^* 为最佳阈值; $g(t)$ 为类间方差; t 为选取的灰度值变量, $t \in \{0, \dots, M-1\}$, 其中 M 为图像中灰度值数量; $\arg \max(\cdot)$ 为最大满足函数。

5. 根据权利要求4所述的一种通用表格图像的表格提取方法, 其特征在于: 所述的步骤S2中, 对预处理后图像即二值化图像进行的形态学操作为开运算操作, 即依次进行腐蚀操作和膨胀操作, 其计算公式为:

$$A \circ SE = (A \ominus SE) \oplus SE$$

式中, $A \circ SE$ 为开运算函数; SE 为第一结构元素; A 为输入的二值化图像; \oplus 为膨胀操作函数; \ominus 为腐蚀操作函数。

一种通用表格图像的表格提取方法

技术领域

[0001] 本发明属于办公自动化领域,具体涉及一种通用表格图像的表格提取方法。

背景技术

[0002] 表格,又称为表,即是一种可视化交流模式,又是一种组织整理数据的手段。人们在通讯交流、科学研究以及数据分析活动当中广泛采用着形形色色的表格。各种表格常常会出现在印刷介质、手写记录、计算机软件、建筑装饰、交通标志等许许多多地方。随着上下文的不同,用来确切描述表格的惯例和术语也会有所变化。此外,在种类、结构、灵活性、标注法、表达方法以及使用方面,不同的表格之间也炯然各异。在各种书籍和技术文章当中,表格通常放在带有编号和标题的浮动区域内,以此区别于文章的正文部分。

[0003] 表格作为一种高度精炼、集中的信息表达形式,在各个行业都得到广泛地应用。随着计算机的普及和企业信息化程度提高,利用计算机进行制作表格日渐兴起。在实际应用中,由于行业和应用领域的不同,表格的内容和格式差别很大,很难用几种特定的表格样式满足各种应用需求,并且人们已经越来越多的使用电子文档以取代纸质文档,例如,用户可以用智能手机拍摄纸质文档的图像,然后将图像发送给别人以完成信息的传递,但是,以拍摄或扫描得到的电子文档都是以图片格式进行存储的。

[0004] 综上所述,表格的多样性以及以图片格式进行存储的表格图像导致表格难以提取。

发明内容

[0005] 为了解决现有技术存在的上述问题,本发明目的在于提供一种快速、准确的通用表格图像的表格提取方法,从表格图像中进行表格的提取,提高了办公自动化中文档管理效率,解决了现有技术存在的表格提取的速度慢,精确度不高的问题。

[0006] 本发明所采用的技术方案为:

[0007] 一种通用表格图像的表格提取方法,包括如下步骤:

[0008] S1:对通用表格图像中进行预处理;

[0009] S2:使用形态学操作将预处理后图像中的文字进行过滤;

[0010] S3:对过滤处理后图像进行重构操作,包括如下步骤:

[0011] S3-1:进行开运算图像重构,包括如下步骤:

[0012] A-1:将输入的过滤处理后图像即开运算图像初始化为原始标记图像;

[0013] A-2:创建第二结构元素;

[0014] A-3:根据原始标记图像和第二结构元素进行重构操作,其计算公式为:

$$[0015] \quad h_{k+1} = (h_k \oplus B) \cap g$$

[0016] 式中, h_{k+1} 为当前标记图像; h_k 为上一代标记图像; k 为迭代参数; B 为第二结构元素; \oplus 为膨胀操作函数; g 为掩模图像; \cap 为交集;

[0017] A-4:判断当前标记图像和上一代标记图像是否相同,若是则实现重构,输出得到

的当前标记图像,否则更新迭代参数,并返回步骤A-3;

[0018] S3-2:提取获得的标记图像的线条特征坐标,包括如下步骤:

[0019] B-1:将二值化图像采用二维数据进行表示;

[0020] B-2:将二维数组的第一个元素作为投影原点,将其他元素进行水平和垂直方向的投影,并得到对应投影方向的累加值;

[0021] B-3:遍历累加值,筛选出所有符合条件的特征坐标,即特征坐标对应的累加值大于两投影方向坐标横线之间的最小距离;

[0022] B-4:对特征坐标进行进一步分析,当有一个以上符合条件的特征坐标时,提取其中间值;

[0023] B-5:遍历二维数组的其他元素,重复步骤B-2到B-4,得到特征坐标数组;

[0024] S3-3:根据特征坐标数组,对图像交点进行分类,并计算标记图像中表格的权重数组,包括如下步骤:

[0025] C-1:根据表格的规则线段与网格点相连的方式,划分成不同的类型;

[0026] C-2:根据得到的特征坐标数组即横纵线坐标数组,遍历其中每个交点;

[0027] 交点为表格中的横线和纵线的交叉点;

[0028] C-3:以当前交点为起始点,依次提取四个方向的图像块,依次判断是否存在一条直线,若存在输出结果为1,否则输出结果为0;

[0029] 四个方向为:横正向、横负向、纵正向以及纵负向;

[0030] C-4:根据输出结果将交点进行分类,并将输出结果保存在建立的对应三维数组中,并根据修改规则对所有三维数组进行修改和校验;

[0031] 修改规则为:如果交点处存在横正向直线,则此交点右边交点必然存在横负向直线,其他方向同理;

[0032] C-5:根据三维数组中对应的每个方向的权重,得到每个交点的权重和,并保存在权重数组中,其计算公式为:

[0033]
$$p(x,y) = q_1 * p(x,y,0) + q_2 * p(x,y,1) + q_3 * p(x,y,2) + q_4 * p(x,y,3)$$

[0034] 式中, $p(x,y)$ 为权重数组; $q_1 * p(x,y,0)$ 为横正向对应的权重和三维数组的乘积; $q_2 * p(x,y,1)$ 为横负向对应的权重和三维数组的乘积; $q_3 * p(x,y,2)$ 为纵正向对应的权重和三维数组的乘积; $q_4 * p(x,y,3)$ 为纵负向对应的权重和三维数组的乘积;

[0035] S4:进行表格重绘,实现表格提取;

[0036] 根据交点的类型以及权重数组,遍历每一行每一列,查找对应的开始端点坐标和结束端点,并将其连起来,完成表格的重绘,实现表格的提取。

[0037] 进一步地,步骤S1中,预处理包括依次进行的缩小处理、灰度处理以及二值化处理。

[0038] 进一步地,步骤S1中,预处理的计算公式为:

[0039] 使用加权法进行灰度值处理,计算公式为:

[0040]
$$F(i,j) = 0.3R(i,j) + 0.59G(i,j) + 0.11B(i,j)$$

[0041] 式中, $F(i,j)$ 为像素点的灰度值; $R(i,j)$ 、 $G(i,j)$ 、 $B(i,j)$ 对应为像素点的红绿蓝三种颜色的亮度值; (i,j) 为像素点坐标;

[0042] 进行二值化处理的计算公式为:

$$[0043] \quad g(x,y) = \begin{cases} 1 & f(x,y) \geq T \\ 0 & f(x,y) < T \end{cases}$$

[0044] 式中, $g(x,y)$ 为二值化图像; $f(x,y)$ 为输入的灰度图像; T 为阈值。

[0045] 进一步地, 将类间方差最大的灰度值作为二值化处理的最佳阈值, 其计算公式为:

$$[0046] \quad T^* = \arg \max [g(t)]$$

[0047] 式中, T^* 为最佳阈值; $g(t)$ 为类间方差; t 为选取的灰度值变量, $t \in \{0, \dots, M-1\}$, 其中 M 为图像中灰度值数量; $\arg \max(\cdot)$ 为最大满足函数。

[0048] 进一步地, 步骤 $S2$ 中, 对预处理后图像即二值化图像进行的形态学操作为开运算操作, 即依次进行腐蚀操作和膨胀操作, 其计算公式为:

$$[0049] \quad A \circ SE = (A \ominus SE) \oplus SE$$

[0050] 式中, $A \circ SE$ 为开运算函数; SE 为第一结构元素; A 为输入的二值化图像; \oplus 为膨胀操作函数; \ominus 为腐蚀操作函数。

[0051] 本发明的有益效果为:

[0052] 1) 本发明基于数字图像处理技术, 对通用表格图像中的表格进行提取, 提高了提取方法的速率和实用性;

[0053] 2) 本发明使用形态学操作对通用表格图像中的文字进行过滤, 并使用开运算对图像中的文字进行过滤, 消除了细小的噪声, 达到对目标去噪的效果, 平滑了目标对象的轮廓, 消除了细小的突出物, 提高了表格提取的精确度;

[0054] 3) 本发明提高了办公自动化中文档管理效率, 解决了现有技术存在的表格提取的速度慢, 精确度不高的问题。

附图说明

[0055] 图1是通用表格图像的表格提取方法流程图;

[0056] 图2是重构操作的方法流程图;

[0057] 图3是进行开运算图像重构的方法流程图;

[0058] 图4是提取获得的标记图像的线条特征坐标的方法流程图;

[0059] 图5是对图像交点进行分类以及计算标记图像中表格的权重数组的方法流程图;

[0060] 图6是通用表格图像;

[0061] 图7为过滤处理后图像;

[0062] 图8为重绘表格图。

具体实施方式

[0063] 下面结合附图及具体实施例对本发明做进一步阐释。

[0064] 实施例1:

[0065] 如图1所示, 本实施例提供一种通用表格图像的表格提取方法, 包括如下步骤:

[0066] $S1$: 对如图6所示的通用表格图像中进行预处理, 预处理包括依次进行的缩小处理、灰度处理以及二值化处理;

[0067] 使用加权法进行灰度值处理, 计算公式为:

[0068] $F(i, j) = 0.3R(i, j) + 0.59G(i, j) + 0.11B(i, j)$

[0069] 式中, $F(i, j)$ 为像素点的灰度值; $R(i, j)$ 、 $G(i, j)$ 、 $B(i, j)$ 对应为像素点的红绿蓝三种颜色的亮度值; (i, j) 为像素点坐标;

[0070] 进行二值化处理的计算公式为:

$$[0071] \quad g(x, y) = \begin{cases} 1 & f(x, y) \geq T \\ 0 & f(x, y) < T \end{cases}$$

[0072] 式中, $g(x, y)$ 为二值化图像; $f(x, y)$ 为输入的灰度图像; T 为阈值;

[0073] 将类间方差最大的灰度值作为二值化处理的最佳阈值, 其计算公式为:

$$[0074] \quad T^* = \arg \max [g(t)]$$

[0075] 式中, T^* 为最佳阈值; $g(t)$ 为类间方差; t 为选取的灰度值变量, $t \in \{0, \dots, M-1\}$, 其中, M 为图像中灰度值数量, $\arg \max(\cdot)$ 为最大满足函数;

[0076] S2: 使用形态学操作将预处理后图像中的文字进行过滤;

[0077] 对预处理后图像即二值化图像进行的形态学操作为开运算操作, 即依次进行腐蚀操作和膨胀操作;

[0078] 膨胀操作就是图像中的高亮部分进行膨胀的“领域扩张”, 效果图拥有比原图更大的高亮区域, 在数学上, 膨胀定义为集合运算, 则SE对A的膨胀定义为:

$$[0079] \quad A \oplus SE = \{z | ((SE)_z \cap A \neq \emptyset)\}$$

[0080] 该式表明SE对A膨胀的结果是使 $(SE)_z$ 与A至少有一个像素重叠的点z的集合, 膨胀操作用来将与目标区域的背景点合并到该目标物中, 可以填补图像上物体中的细小空间;

[0081] 腐蚀就是原图中的高亮部分被腐蚀的“领域被蚕食”, 效果图拥有比原图更小的高亮区域, 按数学方面来说, 膨胀或者腐蚀操作就是将图像(或图像的一部分区域)与核进行卷积, 则SE对A的腐蚀定义为:

$$[0082] \quad A \ominus SE = \{z | (SE)_z \subseteq A\}$$

[0083] 该式表明SE对A膨胀的结果是使 $(SE)_z$ 包含A中的点z的集合, 腐蚀操作会去除图像中的小于结构元的毛刺凸起;

[0084] 在实际操作中, 对二值化图像进行处理时, 会把膨胀操作跟腐蚀操作组合起来一起使用, 开运算的操作就是由SE先对A进行腐蚀操作, 然后在进行膨胀操作。则SE对A进行开运算的计算公式为:

$$[0085] \quad A \circ SE = (A \ominus SE) \oplus SE$$

[0086] 式中, $A \circ SE$ 为开运算函数; SE为第一结构元素; A为输入的二值化图像; \oplus 为膨胀操作函数; \ominus 为腐蚀操作函数。

[0087] 开运算可以消除细小的噪声, 达到对目标去噪的效果, 平滑了目标对象的轮廓, 消除了细小的突出物;

[0088] S3: 对如图7所示的过滤处理后图像进行重构操作;

[0089] 重构操作, 如图2所示, 包括如下步骤:

[0090] S3-1: 进行开运算图像重构, 重构是一种涉及到两幅图像和一个结构元素的形态学变换, 一幅图像是标记图像, 即变换的开始点, 另一幅图像是掩模图像, 用来约束变换过

程,重构操作的标记图像获得方法,如图3所示,包括如下步骤:

[0091] A-1:将输入的过滤处理后图像即开运算图像初始化为原始标记图像;

[0092] A-2:创建第二结构元素;

[0093] A-3:根据原始标记图像和第二结构元素进行重构操作,其计算公式为:

$$[0094] \quad h_{k+1} = (h_k \oplus B) \cap g$$

[0095] 式中, h_{k+1} 为当前标记图像; h_k 为上一代标记图像; k 为迭代参数; B 为第二结构元素; \oplus 为膨胀操作函数; g 为掩模图像; \cap 为交集;

[0096] A-4:判断当前标记图像和上一代标记图像是否相同,若是则实现重构,输出得到的当前标记图像,否则更新迭代参数,并返回步骤A-3;

[0097] S3-2:提取获得的标记图像的线条特征坐标,如图4所示,一幅二值化图像可以用一个二维数组来表示,包括如下步骤:

[0098] B-1:将二值化图像采用二维数据进行表示;

[0099] B-2:将二维数组的第一个元素作为投影原点,将其他元素进行水平和垂直方向的投影,并得到对应投影方向的累加值;

[0100] B-3:遍历累加值,筛选出所有符合条件的特征坐标,即特征坐标对应的累加值大于两投影方向坐标横线之间的最小距离;

[0101] B-4:对特征坐标进行进一步分析,当有一个以上符合条件的特征坐标时,提取其中间值;

[0102] B-5:遍历二维数组的其他元素,重复步骤B-2到B-4,得到特征坐标数组;

[0103] S3-3:根据特征坐标数组,对图像交点进行分类,并计算标记图像中表格的权重数组,如图5所示,包括如下步骤:

[0104] C-1:根据表格的规则线段与网格点相连的方式,划分成不同的类型;

[0105] C-2:根据得到的特征坐标数组即横纵线坐标数组,遍历其中每个交点;

[0106] 交点为表格中的横线和纵线的交叉点;

[0107] C-3:以当前交点为起始点,依次提取四个方向的图像块,依次判断是否存在一条直线,若存在输出结果为1,否则输出结果为0;

[0108] 四个方向为:横正向、横负向、纵正向以及纵负向;

[0109] C-4:根据输出结果将交点进行分类,并将输出结果保存在建立的对应三维数组中,并根据修改规则对所有三维数组进行修改和校验;

[0110] 修改规则为:如果交点处存在横正向直线,则此交点右边交点必然存在横负向直线,其他方向同理;

[0111] C-5:根据三维数组中对应的每个方向的权重,得到每个交点的权重和,并保存在权重数组中,其计算公式为:

$$[0112] \quad p(x, y) = q_1 * p(x, y, 0) + q_2 * p(x, y, 1) + q_3 * p(x, y, 2) + q_4 * p(x, y, 3)$$

[0113] 式中, $p(x, y)$ 为权重数组; $q_1 * p(x, y, 0)$ 为横正向对应的权重和三维数组的乘积; $q_2 * p(x, y, 1)$ 为横负向对应的权重和三维数组的乘积; $q_3 * p(x, y, 2)$ 为纵正向对应的权重和三维数组的乘积; $q_4 * p(x, y, 3)$ 为纵负向对应的权重和三维数组的乘积;

[0114] S4:进行表格重绘,实现表格提取,如图8所示,即根据交点的类型以及权重数组,遍历每一行每一列,查找对应的开始端点坐标和结束端点,并将其连起来,完成表格的重

绘,实现表格的提取。

[0115] 本发明目的在于提供一种快速、准确的通用表格图像的表格提取方法,从表格图像中进行表格的提取,提高了办公自动化中文档管理效率,解决了现有技术存在的表格提取的速度慢,精确度不高的问题。

[0116] 本发明不局限于上述可选的实施方式,任何人在本发明的启示下都可得出其他各种形式的产品。上述具体实施方式不应理解成对本发明的保护范围的限制,本发明的保护范围应当以权利要求书中界定的为准,并且说明书可以用于解释权利要求书。

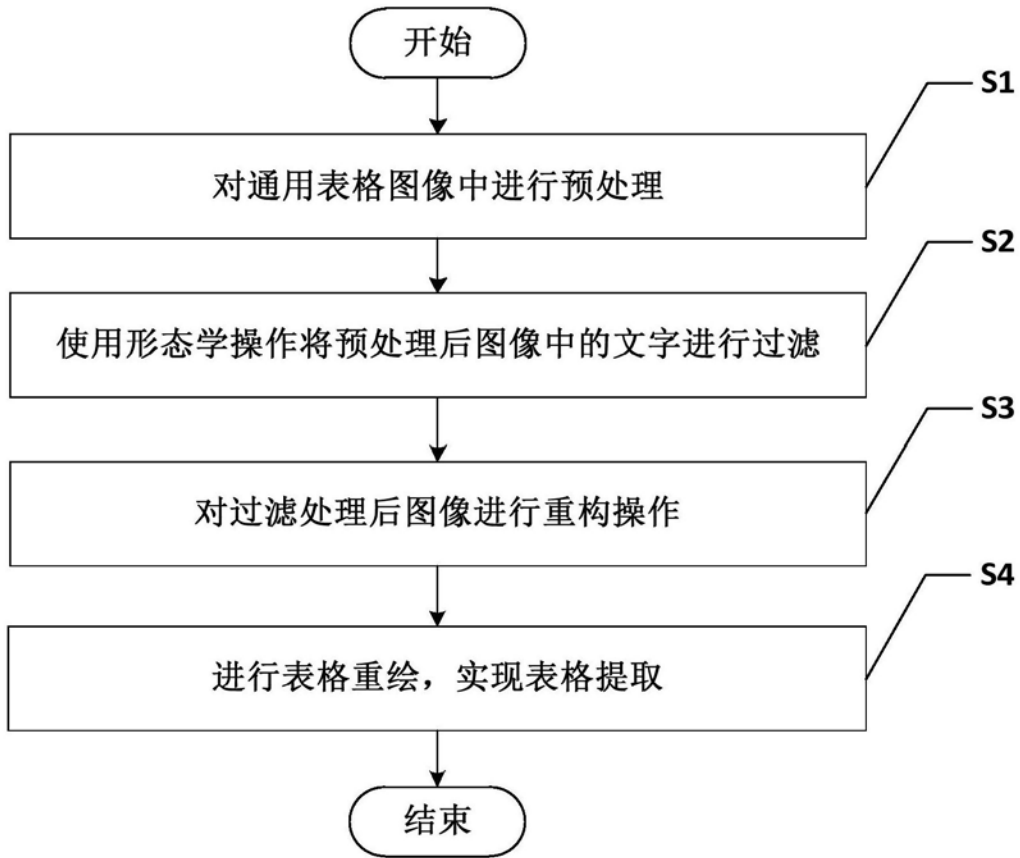


图1

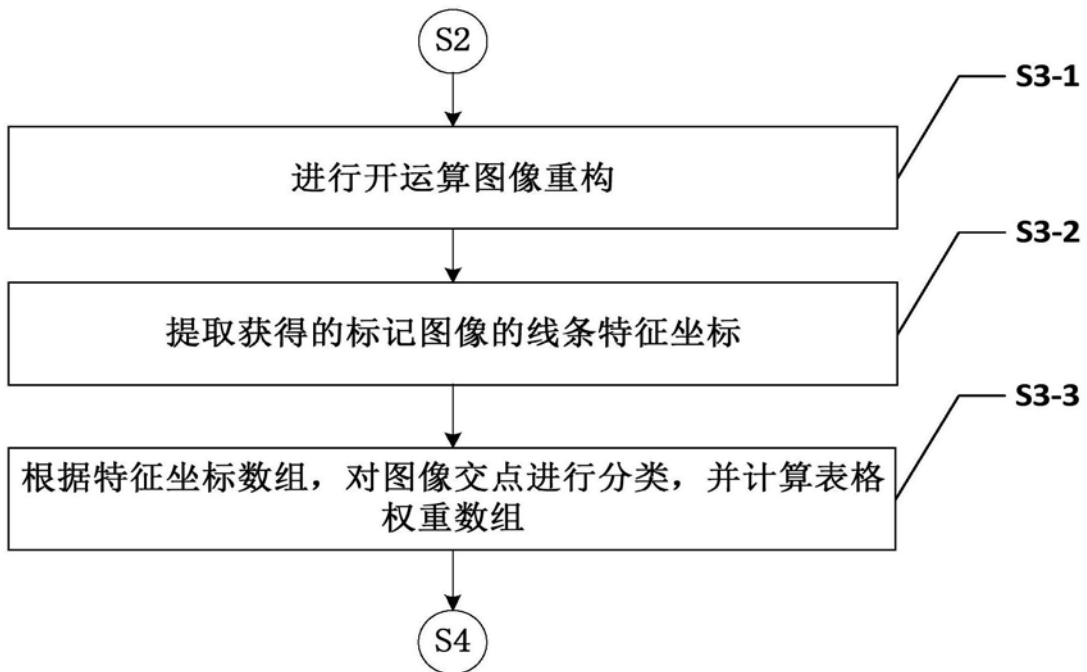


图2

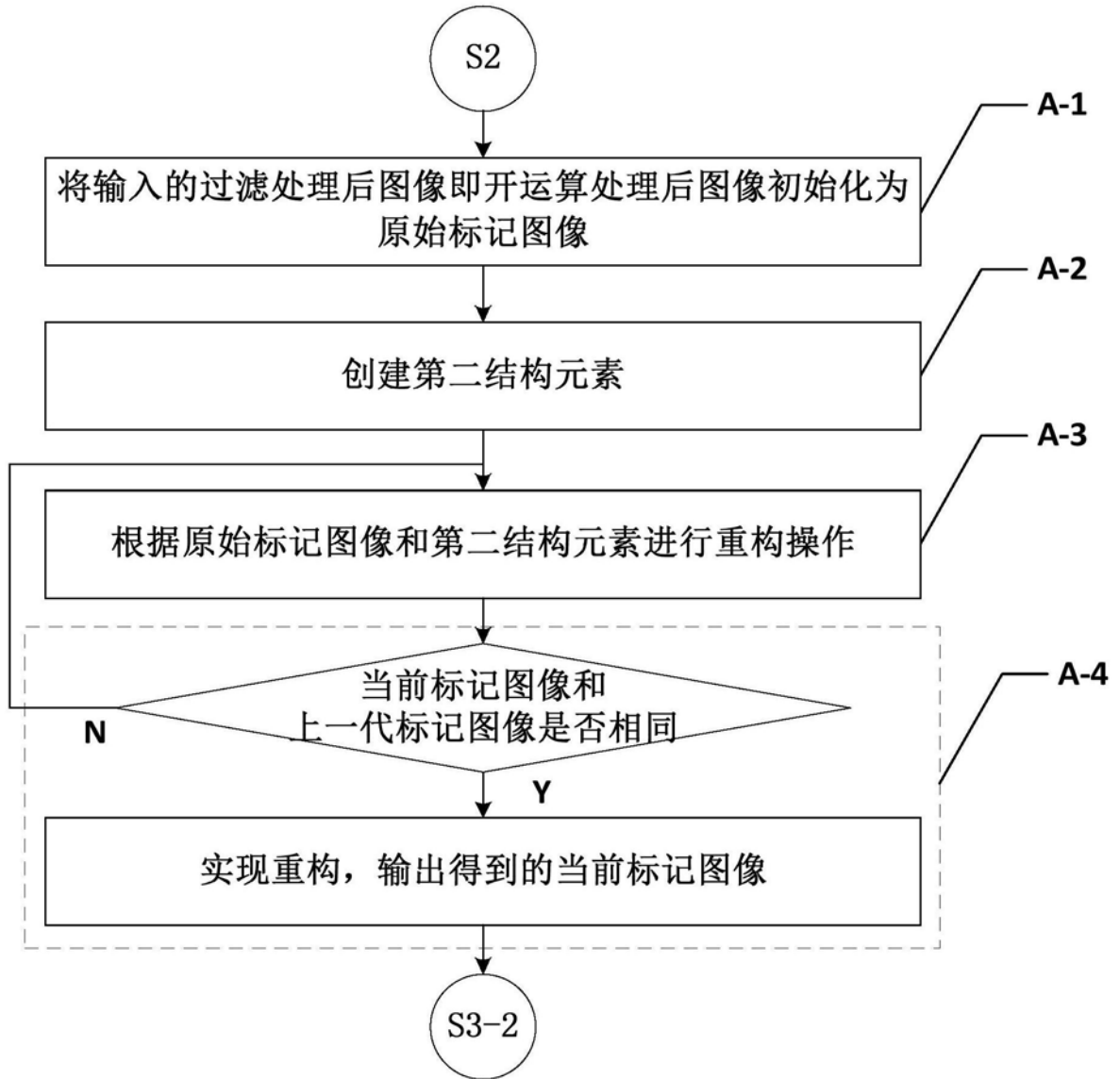


图3

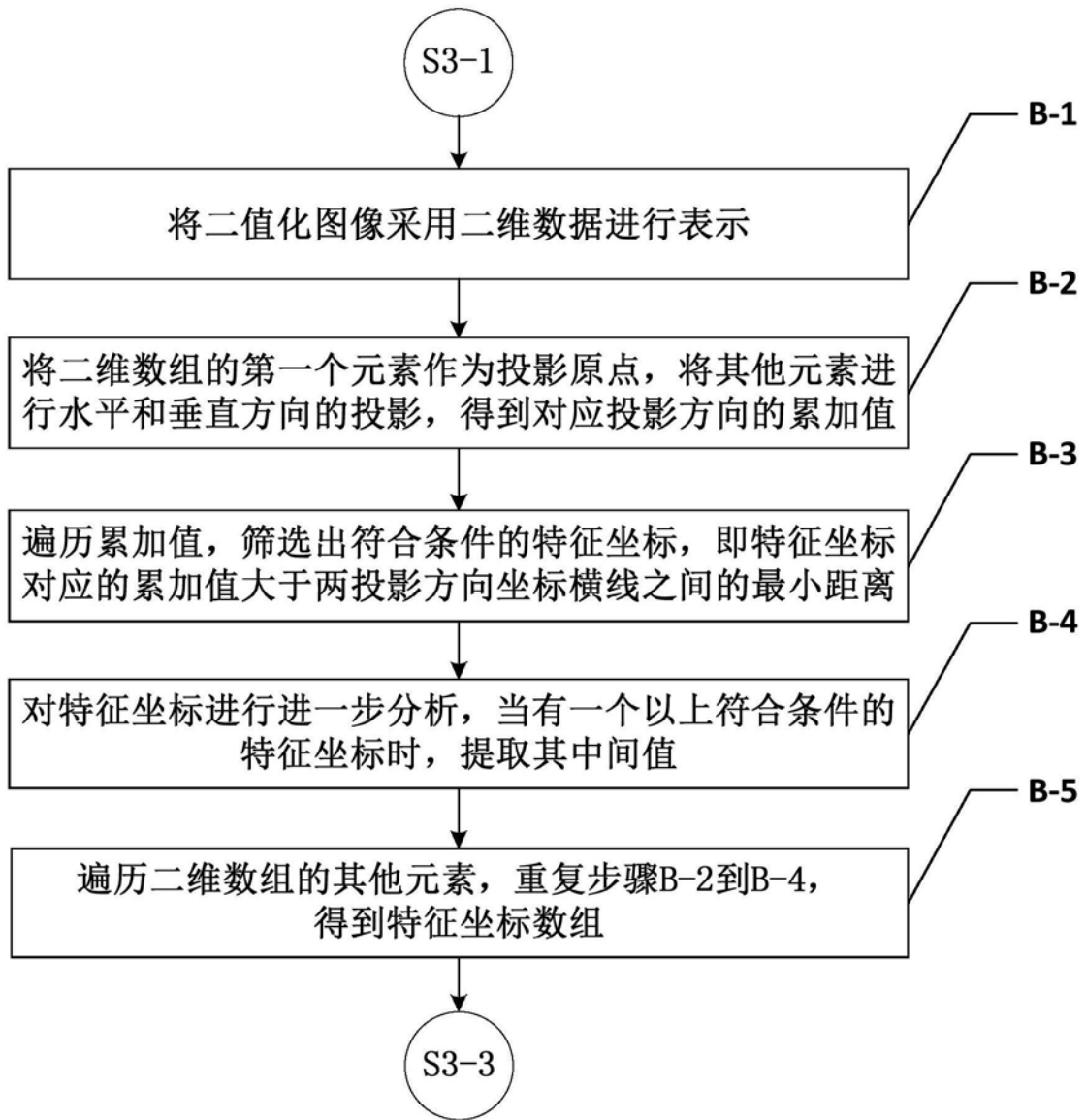


图4

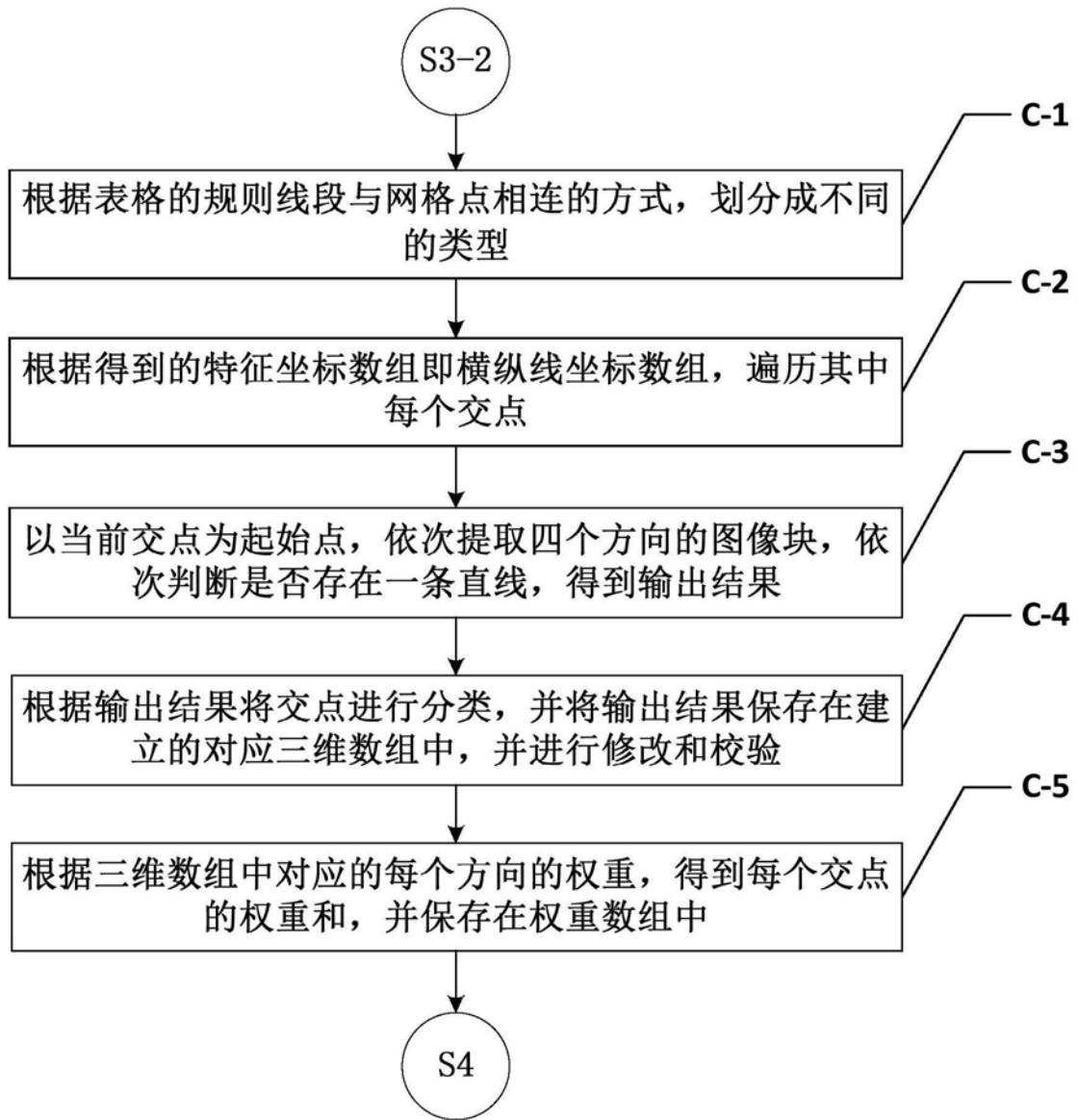


图5

社区第九届社区居委会选举选票												
	主任			副主任			委员					
候选人												

说明：
①应选主任1人；副主任1人；委员5人。
②同意的在候选人姓名下方空格内划○，不同意的不做任何符号，另选他人空格栏内填写其他选民姓名，并在其姓名下方空格内划○。
③等于或少于应选名额的选票有效，多于应选名额的选票无效，字迹模糊无法辨认的选票为废票。
④三种职位，哪一种职位划错，该种职位票作废，其他划对的职位视为有效。

图6

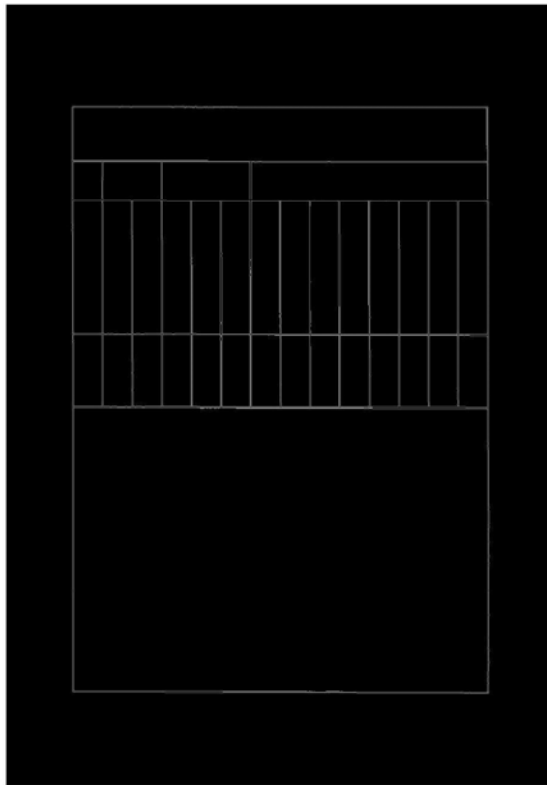


图7

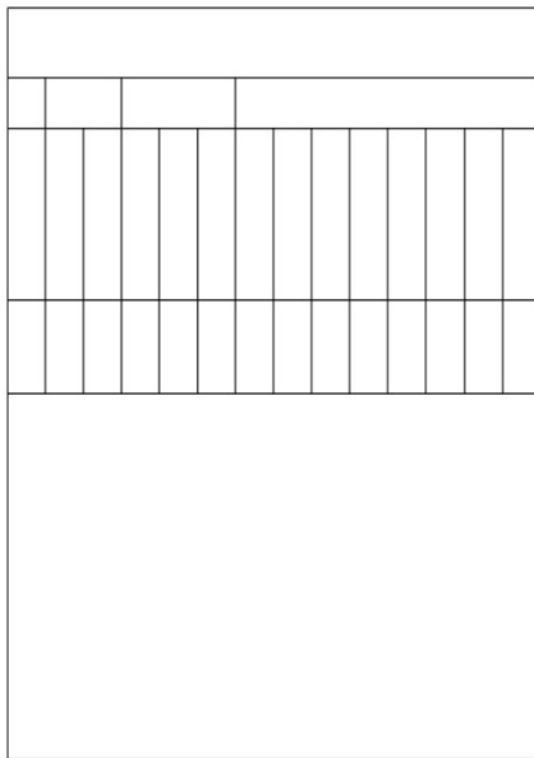


图8