



(12)发明专利

(10)授权公告号 CN 103577817 B

(45)授权公告日 2017.03.01

(21)申请号 201210258883.7

(22)申请日 2012.07.24

(65)同一申请的已公布的文献号

申请公布号 CN 103577817 A

(43)申请公布日 2014.02.12

(73)专利权人 阿里巴巴集团控股有限公司

地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 薛晖

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319

代理人 苏培华

(51)Int.Cl.

G06K 9/20(2006.01)

(56)对比文件

CN 101676930 A,2010.03.24,

US 5841905 A,1998.11.24,

CN 101908136 A,2010.12.08,

CN 101366020 A,2009.02.11,

Arturo Pizano等.A business form recognition system.《Proceedings of the Fifteenth Annual International Computer Software and Applications Conference》.1991,626-632.

Arturo Pizano等.A business form recognition system.《Proceedings of the Fifteenth Annual International Computer Software and Applications Conference》.1991,626-632.

审查员 邱川

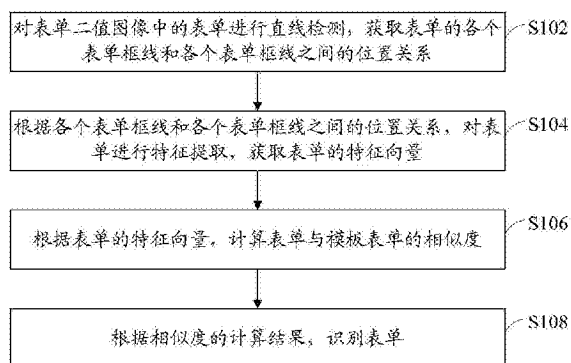
权利要求书4页 说明书12页 附图7页

(54)发明名称

表单识别方法与装置

(57)摘要

本申请提供了一种表单识别方法与装置,其中,表单识别方法包括:对表单二值图像中的表单进行直线检测,获取所述表单的各个表单框线和所述各个表单框线之间的位置关系;根据所述各个表单框线和所述各个表单框线之间的位置关系,对所述表单进行特征提取,获取所述表单的特征向量,其中,所述特征向量用于指示所述表单的结构特征;根据所述表单的特征向量,计算所述表单与模板表单的相似度;根据所述相似度的计算结果,识别所述表单。通过本申请,达到了能够准确识别图像文件中的表格或表格图像中的表格,以便将表格方便、快速地输入到计算机中的效果。



1. 一种表单识别方法,其特征在于,包括:

对表单二值图像中的表单进行直线检测,获取所述表单的各个表单框线和所述各个表单框线之间的位置关系;

根据所述各个表单框线和所述各个表单框线之间的位置关系,对所述表单进行特征提取,获取所述表单的特征向量,其中,所述特征向量用于指示所述表单的结构特征;

根据所述表单的特征向量,计算所述表单与模板表单的相似度;

根据所述相似度的计算结果,识别所述表单;

其中,所述对表单二值图像中的表单进行直线检测,获取所述表单的各个表单框线和所述各个表单框线之间的位置关系的步骤包括:

对于所述表单上的每一条直线,以本直线上的设定点为起点,对本直线上的相邻像素进行连通分量分析,获取本直线的连通分量;获取所述本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数;根据所述本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数,获取本直线的平均线宽;判断本直线的连通分量上的每一个像素所在处的线宽是否大于所述平均线宽,若是,则对大于所述平均线宽的像素所在处的线宽进行线宽去除处理;

根据进行了所述线宽去除处理的每一条直线,获取所述表单的各个表单框线和所述各个表单框线之间的位置关系。

2. 根据权利要求1所述的方法,其特征在于,所述对大于所述平均线宽的像素所在处的线宽进行线宽去除处理的步骤包括:

将所述大于平均线宽的像素从本直线的连通分量中去除;

判断去除的所述像素的位置是否在本直线的内部,若是,则对去除的所述像素两侧的线段进行线段合并。

3. 根据权利要求1所述的方法,其特征在于,在根据所述各个表单框线和所述各个表单框线之间的位置关系,对所述表单进行特征提取,获取所述表单的特征向量的步骤之前,还包括:

按照设定规则判断同一方向上相邻的两条表单框线是否需要线段合并,若是,则对所述相邻的两条表单框线进行线段合并。

4. 根据权利要求3所述的方法,其特征在于,所述按照设定规则判断同一方向上相邻的两条表单框线是否需要线段合并的步骤包括:

获取同一方向上相邻的两条表单框线的距离,根据所述距离与设定的第一阈值的比较结果判断是否需要线段合并;

和/或,

获取同一方向上相邻的两条表单框线的相似度,根据所述相似度判断是否需要线段合并。

5. 根据权利要求4所述的方法,其特征在于,所述获取同一方向上相邻的两条表单框线的相似度,根据所述相似度判断是否需要线段合并的步骤包括:

获取同一方向上相邻的两条表单框线之间的夹角,根据所述夹角与设定的第二阈值的比较结果判断是否需要线段合并;

和/或,

获取同一方向上相邻的两条表单框线之间的平行距离,根据所述平行距离与设定的第三阈值的比较结果判断是否需要线段合并。

6.根据权利要求1所述的方法,其特征在于,在根据所述各个表单框线和所述各个表单框线之间的位置关系,对所述表单进行特征提取,获取所述表单的特征向量的步骤之前,还包括:

对获取的所述表单的各个表单框线进行字符线段过滤。

7.根据权利要求6所述的方法,其特征在于,所述对获取的所述表单的各个表单框线进行字符线段过滤的步骤包括:

判断所述表单的各个表单框线的线段长度是否小于或等于设定的第四阈值,若是,则去除线段长度小于或等于所述第四阈值的表单框线。

8.根据权利要求7所述的方法,其特征在于,所述第四阈值通过以下方式设定:

获取所述表单的各个表单框线中的最大连通分量的长度,并将所述长度划分为多个长度区间;

将所述各个表单框线的每一个连通分量按照其长度划分到所述多个长度区间中的一个长度区间;

根据所述多个长度区间的长度值和每一个所述长度区间的连通分量的个数,生成线段长度分布直方图;

将所述直方图中纵坐标最高的点对应的长度值作为所述第四阈值。

9.根据权利要求1所述的方法,其特征在于,根据所述各个表单框线和所述各个表单框线之间的位置关系,对所述表单进行特征提取,获取所述表单的特征向量的步骤包括:

根据所述各个表单框线和所述各个表单框线之间的位置关系,提取所述表单框线的各个交叉点,获取所述各个交叉点的信息,以及,获取所述各个表单框线中水平方向表单框线的信息、和垂直方向表单框线的信息;

根据所述各个交叉点的信息、所述水平方向表单框线的信息、和所述垂直方向表单框线的信息,获取所述表单的特征向量。

10.根据权利要求9所述的方法,其特征在于,在根据所述各个表单框线和所述各个表单框线之间的位置关系,提取所述表单框线的各个交叉点的步骤之前,还包括:

确定所述表单的区域;

根据所述表单的区域,对所述表单做归一化处理。

11.根据权利要求10所述的方法,其特征在于,

在对所述表单做归一化处理的步骤之后,还包括:将归一化处理后的所述表单划分为多块;

所述根据所述各个表单框线和所述各个表单框线之间的位置关系,提取所述表单框线的各个交叉点,获取所述各个交叉点的信息,以及,获取所述各个表单框线中水平方向表单框线的信息、和垂直方向表单框线的信息的步骤包括:对所述多块表单中的每一块表单,分别根据本块表单内的各个表单框线和各个表单框线之间的位置关系,进行本块表单的交叉点提取,提取本块表单中的表单框线的各个交叉点,获取本块表单中的各个交叉点的信息,并且获取本块表单中的水平方向表单框线的信息、和垂直方向表单框线的信息。

12.根据权利要求1所述的方法,其特征在于,根据所述表单的特征向量,计算所述表单

与模板表单的相似度的步骤包括：

计算所述表单的特征向量与各个所述模板表单的特征向量之间的最相近距离和次相近距离的比值；

若所述比值大于或等于设定比值，则确定所述表单与所述最相近距离对应的模板表单的相似度最高；

若所述比值小于所述设定比值，则确定所述表单与所述各个模板表单均不相似。

13. 一种表单识别装置，其特征在於，包括：

直线检测模块，用于对于所述表单上的每一条直线，以本直线上的设定点为起点，对本直线上的相邻像素进行连通分量分析，获取本直线的连通分量；获取所述本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数；根据所述本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数，获取本直线的平均线宽；判断本直线的连通分量上的每一个像素所在处的线宽是否大于所述平均线宽，若是，则对大于所述平均线宽的像素所在处的线宽进行线宽去除处理；根据进行了所述线宽去除处理的每一条直线，获取所述表单的各个表单框线和所述各个表单框线之间的位置关系；

特征提取模块，用于根据所述各个表单框线和所述各个表单框线之间的位置关系，对所述表单进行特征提取，获取所述表单的特征向量，其中，所述特征向量用于指示所述表单的结构特征；

相似度计算模块，用于根据所述表单的特征向量，计算所述表单与模板表单的相似度；

表单识别模块，用于根据所述相似度的计算结果，识别所述表单。

14. 根据权利要求13所述的装置，其特征在於，所述直线检测模块在对大于所述平均线宽的像素所在处的线宽进行线宽去除处理时，将所述大于平均线宽的像素从本直线的连通分量中去除；判断去除的所述像素的位置是否在本直线的内部，若是，则对去除的所述像素两侧的线段进行线段合并。

15. 根据权利要求13所述的装置，其特征在於，所述特征提取模块，用于根据所述各个表单框线和所述各个表单框线之间的位置关系，提取所述表单框线的各个交叉点，获取所述各个交叉点的信息，以及，获取所述各个表单框线中水平方向表单框线的信息、和垂直方向表单框线的信息；根据所述各个交叉点的信息、所述水平方向表单框线的信息、和所述垂直方向表单框线的信息，获取所述表单的特征向量。

16. 根据权利要求15所述的装置，其特征在於，还包括：

表单处理模块，用于在所述特征提取模块根据所述各个表单框线和所述各个表单框线之间的位置关系，提取所述表单框线的各个交叉点之前，确定所述表单的区域；根据所述表单的区域，对所述表单做归一化处理。

17. 根据权利要求16所述的装置，其特征在於，

所述表单处理模块，还用于将归一化处理后的所述表单划分为多块；

所述特征提取模块，用于对所述多块表单中的每一块表单，分别根据本块表单内的各个表单框线和各个表单框线之间的位置关系，进行本块表单的交叉点提取，提取本块表单中的表单框线的各个交叉点，获取本块表单中的各个交叉点的信息，并且获取本块表单中的水平方向表单框线的信息、和垂直方向表单框线的信息；根据获取的所述每一块表单中的所述各个交叉点的信息、所述水平方向表单框线的信息、和所述垂直方向表单框线的信

息,获取所述表单的特征向量。

18.根据权利要求13所述的装置,其特征在于,所述相似度计算模块,用于计算所述表单的特征向量与各个所述模板表单的特征向量之间的最相近距离和次相近距离的比值;若所述比值大于或等于设定比值,则确定所述表单与所述最相近距离对应的模板表单的相似度最高;若所述比值小于所述设定比值,则确定所述表单与所述各个模板表单均不相似。

表单识别方法与装置

技术领域

[0001] 本申请涉及计算机技术领域,特别是涉及一种表单识别方法与装置。

背景技术

[0002] 模式识别是指对表征事物或现象的各种形式的(数值的、文字的和逻辑关系的)信息进行处理和分析,以对事物或现象进行描述、辨认、分类和解释的过程。随着计算机技术的发展,人们开始应用计算机进行模式识别,对一组事件或过程进行辨识和分类,所识别的事件或过程可以是文字、声音、图像等具体对象,也可以是状态、程度等抽象对象。

[0003] 例如,应用计算机进行文字识别。以OCR为例,OCR(Optical Character Recognition,光学字符识别)是指电子设备(例如扫描仪或数码相机)检查纸上打印的字符,通过检测暗、亮的模式确定其形状,然后用字符识别方法将形状翻译成计算机文字的过程。即,对文本资料进行扫描,生成图像文件,然后对图像文件进行分析处理,获取文字及版面信息的过程。通过OCR文字识别,能够识别大部分的文本文字。

[0004] 然而,目前应用计算机识别文字还存在着一些问题。例如,OCR不能够对文本资料扫描成的图像文件中的表格进行识别,当遇到表格时,OCR常会出现乱码,无法对表格进行正确识别。

[0005] 因此,如何正确对图像文件中的表格或表格图像进行识别,将表格方便、快速地输入到计算机中,已成为一个亟待解决的问题。

发明内容

[0006] 本申请提供了一种表单识别方法与装置,以解决现有技术无法对图像文件中的表格或表格图像进行正确识别的问题。

[0007] 为了解决上述问题,本申请公开了一种表单识别方法,包括:对表单二值图像中的表单进行直线检测,获取所述表单的各个表单框线和所述各个表单框线之间的位置关系;根据所述各个表单框线和所述各个表单框线之间的位置关系,对所述表单进行特征提取,获取所述表单的特征向量,其中,所述特征向量用于指示所述表单的结构特征;根据所述表单的特征向量,计算所述表单与模板表单的相似度;根据所述相似度的计算结果,识别所述表单。

[0008] 为了解决上述问题,本申请还公开了一种表单识别装置,包括:直线检测模块,用于对表单二值图像中的表单进行直线检测,获取所述表单的各个表单框线和所述各个表单框线之间的位置关系;特征提取模块,用于根据所述各个表单框线和所述各个表单框线之间的位置关系,对所述表单进行特征提取,获取所述表单的特征向量,其中,所述特征向量用于指示所述表单的结构特征;相似度计算模块,用于根据所述表单的特征向量,计算所述表单与模板表单的相似度;表单识别模块,用于根据所述相似度的计算结果,识别所述表单。

[0009] 与现有技术相比,本申请具有以下优点:

[0010] 本申请根据表单的结构特性,首先对表单二值图像中的表单进行直线检测,得到表单的各条线段和各条线段间的关系,也即,得到表单的各个表单框线和各个表单框线之间的位置关系;然后,根据这些信息对表单进行特征提取,获取表单的特征向量,以获取表单的结构特征;进而,根据表单的结构特征将该表单与模板表单进行匹配,以确定与该表单相似度最高的表单,也即结构最为相似的表单,根据匹配结果识别该表单的类别。可见,通过本申请,有效解决了现有技术无法对图像文件中的表格或表格图像进行正确识别的问题,达到了能够准确识别图像文件中的表格或表格图像中的表格,以便将表格方便、快速输入到计算机中的效果。

附图说明

- [0011] 图1是根据本申请实施例一的一种表单识别方法的步骤流程图;
- [0012] 图2是根据本申请实施例二的一种表单识别方法的步骤流程图;
- [0013] 图3是图2所示实施例中提取的一条水平直线的示意图;
- [0014] 图4是图2所示实施例中的一种表单特征点示意图;
- [0015] 图5是图2所示实施例中的一种通过距离相似度识别表单的示意图;
- [0016] 图6是图2所示实施例中的另一种通过距离相似度识别表单的示意图;
- [0017] 图7是根据本申请实施例三的一种表单直线检测的流程图;
- [0018] 图8是图7所示表单直线检测中的线段合并中的距离判断示意图;
- [0019] 图9是图7所示表单直线检测中的线段合并中的夹角判断示意图;
- [0020] 图10是图7所示表单直线检测中的线段合并中的平行距离判断示意图;
- [0021] 图11是图7所示表单直线检测中的字符线段过滤示意图;
- [0022] 图12是根据本申请实施例三的一种表单类型识别的流程图;
- [0023] 图13是图12所示表单类型识别中的一种多级分层示意图;
- [0024] 图14是图13所示表单进行直线检测后的输出结果图;
- [0025] 图15是图14所示表单输出结果进行特征点提取后的特征点示意图;
- [0026] 图16是根据本申请实施例三的一种表单模型建立的流程图;
- [0027] 图17是根据本申请实施例四的一种表单识别装置的结构框图。

具体实施方式

[0028] 为使本申请的上述目的、特征和优点能够更加明显易懂,下面结合附图和具体实施方式对本申请作进一步详细的说明。

[0029] 实施例一

[0030] 参照图1,示出了根据本申请实施例一的一种表单识别方法的步骤流程图。

[0031] 本实施例的表单识别方法包括以下步骤:

[0032] 步骤S102:对表单二值图像中的表单进行直线检测,获取表单的各个表单框线和各个表单框线之间的位置关系。

[0033] 对于表单来说,其线段多为规整的直线,对其进行直线检测可获得表单的各个表单框线和各个表单框线之间的位置关系。其中,表单框线表示表格中的每一条线段。

[0034] 步骤S104:根据各个表单框线和各个表单框线之间的位置关系,对表单进行特征

提取,获取表单的特征向量。

[0035] 其中,特征向量用于指示表单的结构特征。

[0036] 步骤S106:根据表单的特征向量,计算表单与模板表单的相似度。

[0037] 其中,模板表单是系统中存储的表单,可以通过对样本表单图像进行直线检测和特征提取后,建立相应的表单模型实现,也可以通过人工生成实现,当然,也可以由本领域技术人员通过其它适当方式实现。

[0038] 步骤S108:根据相似度的计算结果,识别表单。

[0039] 本实施例根据表单的结构特性,首先对表单二值图像中的表单进行直线检测,得到表单的各条线段和各条线段间的关系,也即,得到表单的各个表单框线和各个表单框线之间的位置关系;然后,根据这些信息对表单进行特征提取,获取表单的特征向量,以获取表单的结构特征;进而,根据表单的结构特征将该表单与模板表单进行匹配,以确定与该表单相似度最高的表单,也即结构最为相似的表单,根据匹配结果识别该表单的类别。可见,通过本实施例,有效解决了现有技术无法对图像文件中的表格或表格图像进行正确识别的问题,达到了能够准确识别图像文件中的表格或表格图像中的表格,以便将表格方便、快速地输入到计算机中的效果。为描述方便,下文中将图像文件中的表格或表格图像统称为表格图像。

[0040] 实施例二

[0041] 参照图2,示出了根据本申请实施例二的一种表单识别方法的步骤流程图。

[0042] 本实施例对实施例一的表单识别方法进行了进一步地优化,首先,通过连通分量分析实现表单的直线检测,并对获取的表单框线进行了进一步地优化处理;其次,在进行表单的特征提取前,对表单框线进行字符线段过滤;再者,通过特征交叉点提取实现了表单的特征提取;此外,通过表单与模板表单的最相近距离和次相近距离的比值判定,实现了表单的精确识别。

[0043] 本实施例中,以简单的“田”字形表单为例,对本申请的表单识别方法进行说明。

[0044] 本实施例的表单识别方法包括以下步骤:

[0045] 步骤S202:通过对表单二值图像中的表单的每条线段进行连通分量分析实现直线检测,获取表单的各个表单框线,以及获取各个表单框线之间的位置关系。

[0046] 具体地,在对每条线段,也即表单的每一条直线,进行连通分量分析时,可以以本直线上的设定点为起点(如本直线的起始点或结束点或任意其它设定点),对本直线上的相邻像素进行连通分量分析,获取本直线的连通分量;然后,获取本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数;根据本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数,获取本直线的平均线宽;判断本直线的连通分量上的每一个像素所在处的线宽是否大于平均线宽,若是,则对大于平均线宽的像素所在处的线宽进行线宽去除处理。在进行了线宽去除处理后,系统可以根据进行了线宽去除处理的每一条直线,获取表单的各个表单框线。同时,还可以依据连通分量间的位置关系获取各个表单框线之间的位置关系,如表单框线A在线段起始端与表单框线B垂直等。

[0047] 优选地,在对大于平均线宽的像素所在处的线宽进行线宽去除处理时,可以将大于平均线宽的像素从本直线的连通分量中去除;进而判断去除的像素的位置是否在本直线的内部,若是,则对去除的像素两侧的线段进行线段合并。

[0048] 通过上述方式进行直线检测,能够有效处理质量较差的表单图像,尤其是存在模糊、毛刺的情况。

[0049] 以下,以“田”字形表单中第二条水平直线为例,结合图3,对通过连通分量分析实现表单的直线检测进行说明。

[0050] 以第二条水平直线的起始点(如图3中的像素1)为起点,进行该直线的连通分量分析。例如,设定当前像素为第二条水平直线的起始点像素(即像素1),判断在水平方向上(左侧或右侧,本实施例中为右侧)是否存在下一个像素与当前像素邻接,如果存在,则将其添加至连通域,将该下一个像素作为当前像素,继续判断是否存在与其相邻接的像素,直至没有像素与当前像素邻接,获得该直线的连通分量(如图3中的像素1-2-3-4-5-6-7-8)。需要说明的是,在表单简单且清楚的情况下,如本实施例中,第二条水平直线的连通分量可能覆盖整条直线,但在某些复杂表单的情况下,对一条直线进行连通分量分析,可能获得多个连通分量。

[0051] 对于连通分量上的每一个像素,将与本直线垂直方向上连通的像素统一作为直线上的一个点来处理(如图3中像素1垂直方向上的像素均采用像素1的标号),以兼容不同线宽。通过该方式,可以获取本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数。需要说明的是,本步骤也可以在对直线进行连通分量分析之前进行,这时,可以获取本直线上的每一个像素在与本直线垂直方向上的连通分量的像素个数。第二条水平直线的连通分量在获取了垂直方向上的连通分量后,如图3所示。

[0052] 根据与本直线垂直方向上的连通分量的像素个数,计算本直线的平均线宽(本实施例中为水平直线在纵向的平均像素个数),假设为 N (像素)。针对连通分量上的每一个像素点,如果其所在处即当前位置的线宽(纵向像素个数)大于 N ,或者,与 N 的比例超过一定的阈值,则对该像素点所在处的线宽进行线宽去除处理。如,将造成多余线宽的像素点去除,或者,将该像素点从连通分量中去除;如果这个像素点的位置在直线的内部,则该直线将被打断为两段,再对该两段直线进行线段合并。以图3为例, N 为水平直线的平均线宽,在本实施例中为 $(2+3+3+4+2+3+3+1)/8=21/8=2.625$;假设阈值取为1.2,则 $1.2 \times 2.625=3.15$,即垂直方向上的像素数目如果超过3.15,则将其剔除(比如图3中的像素4);如果剔除点出现在直线的中间(比如图3中的像素4),则将直线打断为两条直线(1-2-3)和(5-6-7-8)。通过本步骤,可以消除噪音、文字以及垂直线的影响。

[0053] 对垂直直线的直线检测可以仿照上述水平直线的直线检测进行,在此不再赘述。

[0054] 步骤S204:对获取的表单的各个表单框线进行字符线段过滤。

[0055] 包括:判断表单的各个表单框线的线段长度是否小于或等于设定的第四阈值,若是,则去除线段长度小于或等于第四阈值的表单框线。

[0056] 其中,第四阈值可以通过多种方式设定,如根据试验结果,或者根据经验值,或者其它适当方式设定。优选地,本实施例的第四阈值通过以下方式设定:获取表单的各个表单框线中的最大连通分量的长度,并将长度划分为多个长度区间;将各个表单框线的每一个连通分量按照其长度划分到多个长度区间中的一个长度区间;根据多个长度区间的长度值和每一个长度区间的连通分量的个数,生成线段长度分布直方图;将直方图中纵坐标最高的点对应的长度值作为第四阈值。

[0057] 获取的表单的表单框线中可能会存在许多字符线段,一般说来,这类线段的长度

(对应字符的笔画)往往比较小,可以通过设定阈值的方式加以过滤。通过字符线段过滤,能够有效提高表单识别精度,减少字符对表单识别的影响。

[0058] 需要说明的是,本步骤为可选步骤,对于某些纯表格图像中的表格(仅有表格,表格中无文字或字符),该步骤可省略。

[0059] 步骤S206:根据进行了字符线段过滤后的各个表单框线和各个表单框线之间的位置关系,对表单进行特征提取,获取表单的特征向量。

[0060] 优选地,可以根据各个表单框线和各个表单框线之间的位置关系,提取表单框线的各个交叉点,获取各个交叉点的信息,以及,获取各个表单框线中水平方向表单框线的信息、和垂直方向表单框线的信息;根据各个交叉点的信息、水平方向表单框线的信息、和垂直方向表单框线的信息,获取表单的特征向量。

[0061] 仍以简单的“田”字形表单为例,如图4所示,常见的表单总共有九类交叉点也即特征点可供提取,根据表单的特性,可以提取的九类特征点包括:西北部特征点(图4中的NW类特征点)、北部特征点(图4中的N类特征点)、东北部特征点(图4中的NE类特征点)、东部特征点(图4中的E类特征点)、东南部特征点(图4中的SE类特征点)、南部特征点(图4中的S类特征点)、西南部特征点(图4中的SW类特征点)、西部特征点(图4中的W类特征点)、中部特征点(图4中的M类特征点)。

[0062] 除了上述提取的九类特征点,本实施例还提取水平方向和垂直方向直线(也即表单框线)的信息如数量信息,共提取11类特征,将这些特征作为描述该表单的结构特征的特征向量输出。

[0063] 特征点不是坐标点,而是各类框线特征的统计值。经过直线检测之后的表单图像,所有的元素必定是以下11种情况之一:西北部元素、北部元素、东北部元素、东部元素、东南部元素、南部元素、西南部元素、西部元素、中部元素以及水平、垂直直线。不必关心具体的坐标值,只需统计11类情况的出现频次,即可获取表单图像中表格的结构特征。

[0064] 步骤S208:计算表单的特征向量与各个模板表单的特征向量之间的最相近距离和次相近距离的比值,根据计算结果识别表单。

[0065] 其中,最相近距离是指表单的特征向量与各个模板表单的特征向量的所有距离数值中,数值最小的那个距离,该距离对应的模板表单在所有模板表单中与待识别的表单最为相似;而次相近距离是指所有距离数值中,数值第二小的那个距离,只有最相近距离小于次相近距离。

[0066] 若最相近距离和次相近距离的比值大于或等于设定比值,则确定表单与最相近距离对应的模板表单的相似度最高,可以认为该表单属于模板表单类别;若比值小于设定比值,则确定表单与各个模板表单均不相似。

[0067] 使用最相近距离和次相近距离的比值识别表单是因为:假设已经提取了“最相近距离”,如何衡量输入表单确实是属于该类别的?因为即使输入表单不属于样本,也会计算出一个最相近的值。一种方法是,设定一个阈值 s ,如果最相近距离小于 s ,说明确实是属于该类别表单的;反之则说明不属于表单库中任何一个类目。该方法的缺点是,较难给出明确的 s 值,需要通过大量人工测试才能确定。另一种方法是,设定一个阈值 s ,同时求出次相近,并计算最相近点和次相近点比例;如果该比例大于 s ,则认为属于“最相近”类别的表单;反之,反之则说明不属于表单库中任何一个类目,如图5和图6所示。如果最相近和次相近差别

很大,则输入表单很有可能属于S0类别,如图5所示,T0和T1差别很大,输入表单属于S0类别;如果最相近和次相近差别很小,则输入表单很有可能不属于样本集中的任何一个类别,如图6所示,T0和T1差别不大,输入表单很有可能不属于样本集中的任何一个类别。通过这种方式,能够简单有效地识别出表单所属类别。

[0068] 通过本实施例,对表单图像线段进行连通分量分析获取表单框线,对表单框线进行线宽去除处理以及字符线段过滤,以使表单框线更为精准有效;通过提取表单特征点获取表单有效的结构特征,进而通过相似度分析确定与表单最为相似的模板表单,实现表单识别。可见,通过本实施例,有效解决了现有技术无法对图像文件中的表格或表格图像进行正确识别的问题,达到了能够准确识别图像文件中的表格或表格图像中的表格,以便将表格方便、快速地输入到计算机中的效果。

[0069] 实施例三

[0070] 本实施例以实施例二为基础,从表单直线检测和表单类型识别方面对本申请的表单识别方法进行说明。其中,在表单直线检测过程中,进一步进行了表单框线合并的优化处理;在表单类型识别过程中,进一步进行了表单的多分辨率处理的优化。同时,对表单模型的建立也进行了说明。

[0071] 以下,从上述方面对本实施例的表单识别方法进行说明。

[0072] 首先,本实施例的表单直线检测流程如图7所示,包括:输入表单图像、表单图像预处理、表单图像连通分量提取、表单图像线段合并、表单图像字符线段过滤和表单图像直线检测结果输出几部分。

[0073] 下面分别对表单直线检测流程的各部分进行详细描述:

[0074] 步骤S302:输入表单图像。

[0075] 本申请支持输入的图像来源包括屏幕截图、照片上传、扫描仪上传或者其它方式,输入表单的类型可以是电子表单、纸质票据或者工业图纸。

[0076] 步骤S304:表单图像预处理。

[0077] 主要是对输入的表单图像进行灰度化和二值化,经过预处理的输出图像仅包含黑、白两种颜色。

[0078] 步骤S306:表单图像连通分量提取。

[0079] 采用连通分量分析的方式提取直线可以有效地解决倾斜直线的问题。由于表格框线一般呈近似水平或者垂直的结构,因此本实施例采用的连通分量提取方案如下:

[0080] 以检测水平直线为例,

[0081] (a1)垂直方向上连通的像素统一作为直线上的一个点来处理(兼容不同线宽);

[0082] 考虑到处理的表单图像可能质量较差,存在模糊、毛刺等情况,所以在检测水平直线的时候,将每一个像素连同垂直方向上的像素都作为直线的元素。

[0083] (b1)如果在水平方向上(左侧或右侧)存在一个像素与当前像素邻接,则将其添加至连通域。

[0084] (c1)针对上述提取出来的连通分量,计算平均线宽(即水平直线在纵向的平均像素个数),假设为N。针对区域的每一个横坐标,如果当前位置的线宽(纵向像素个数)与N的比例超过一定的阈值,则从连通分量中去除;如果这个位置在直线的内部,则将直线打断为两段。该步骤主要是用来消除噪音、文字以及垂直线的影响。

[0085] 垂直直线检测的方法与此相似,可仿照该水平直线检测方法进行。

[0086] 步骤S308:表单图像线段合并。

[0087] 本步骤主要是针对原本是同一条,但经各种原因断裂的直线进行合并。在合并时,需要按照设定规则判断同一方向上相邻的两条表单框线是否需要线段合并,若是,则对所述相邻的两条表单框线进行线段合并。优选地,可以获取同一方向上相邻的两条表单框线的距离,根据所述距离与设定的第一阈值的比较结果判断是否需要线段合并;和/或,获取同一方向上相邻的两条表单框线的相似度,根据所述相似度判断是否需要线段合并。优选地,在获取同一方向上相邻的两条表单框线的相似度,根据所述相似度判断是否需要线段合并时,获取同一方向上相邻的两条表单框线之间的夹角,根据所述夹角与设定的第二阈值的比较结果判断是否需要线段合并;和/或,获取同一方向上相邻的两条表单框线之间的平行距离,根据所述平行距离与设定的第三阈值的比较结果判断是否需要线段合并。

[0088] 也即,两段直线合并要求满足以下条件至少之一:

[0089] 条件A:两段线段间的最近距离(即左边直线的右端和右边直线的左端)小于一定的阈值。优选地,该阈值设置为12个像素,低于这个阈值则认为是正常的直线断裂,两条线段参与合并,否则认为是原本分离的两条直线,不参与合并。如图8所示,当T1小于一定的阈值时,两条线段(左侧连通分量1-2-3-4-5-6-7-8和右侧连通分量1-2-3-4-5-6-7-8)参与合并,否则不参与合并。

[0090] 条件B:计算两条线段间的相似度,即它们为同一直线的可能性。

[0091] 具体方法为,计算各直线的中心像素位置,然后通过直线拟合的方法(例如最小二乘法)获得各直线的逼近表达式。本实施例通过两个指标来衡量两条线段为同一直线的可能性。第一个指标是两者之间的夹角(可通过余弦定理获得),如图9所示,针对提取到的任意两个连通分量,计算其纵向的中心点,根据这些中心点可以拟合出最接近的直线,然后通过几何知识计算出两者之间的夹角,实现连通分量间夹角的计算,夹角越小表示两条直线属于同一直线的可能性越大;第二个指标是两者之间的平行距离,可以通过计算第二条线段各像素中心相对于第一条线段的点线距的均方和来表示,如图10所示,针对提取到的任意两个连通分量,针对左边区域的中心拟合出直线,并计算其右边连通分量各中心点相对于延长线距离的平方和,从而实现连通分量间平行距离的计算,该值越小说明两条直线为同一直线的可能性越大。优选地,第一个指标的阈值为5度,第二个指标的阈值为第二连通分量平均线宽平方值的两倍,即只有同时满足这两个条件的两条线段,才被合并为同一直线。

[0092] 线段合并主要是针对一些复杂的、图像质量较差的表单图片出现“原本是同一条表单直线由于各种原因断裂”的现象,而出现的一种补救措施。对于简单、并且较清晰的表单图像,该步骤为可选步骤。

[0093] 步骤S310:表单图像字符线段过滤。

[0094] 经过上述的线段合并后,结果中可能仍然会存在许多字符线段,一般说来,这类线段的长度(对应字符的笔画)往往比较小,可以通过设定阈值的方式加以过滤。

[0095] 本实施例采用了一种自动化提取阈值的方式,针对前述步骤的输出结果,统计其直线长度的直方图,具体如下:

[0096] (a2)针对前面的输出结果,首先获得最长连通域的长度;

[0097] (b2)将其按照相等的间隔分成M个区间;

[0098] (c2)针对每一条线段,根据其长度将其划分到某一个区间中去,并将该区间的计数器累加一;

[0099] (d2)对各线段进行如上操作即可形成线段长度分布的直方图,取最高点对应的长度作为字符线段的长度。在水平和垂直两个方向可以分别得到字符宽度和字符高度,作为筛选线段的阈值;

[0100] (e2)只保留高于阈值的线段(也即,被划分到阈值右侧的区间的线段)。

[0101] 以下,结合图11,对上述字符线段过滤进行说明。

[0102] 本实施例中所有长度均以像素来表示,假设最长连通分量(代表检测出的最长的直线)的长度为500,取M=100,即将整个空间划分为0、5、10、...、495、500总共100个区间。将整个表单图像中检测到的直线,按照像素长度统计成直方图,如图11所示,其中,横坐标为像素区间,纵坐标为出现的次数。需要说明的是,M=100仅为示例性说明,在实际应用中,M的取值可以按照实际情况来定,如果图片大,像素自然就多,M可以相应调大,反之则可以调小。

[0103] 在前述的直线检测中,将会输出许多直线段(包含真实的表格框线以及字符笔划线段)。绝大多数情况下,表格的框线长度较长并且数量较少,而字符笔划长度较短并且数量较多(特别是包含密集文字的情况)。因此通过本步骤检测出峰值情况下的阈值(如图11中椭圆线框部分对应的字符线段的长度),并且只保留高于阈值的线段(真正的表格框线)。

[0104] 字符线段过滤主要是针对以下的情况:即训练过程或者待识别过程中的输入表单图片包含了大量文字性输入内容。除非输入表单包含极少的输入,否则那些输入的文字也会产生细小的直线,对后续步骤造成影响。通过字符线段过滤可以有效提高精度,减小文字性输入内容对表单识别的影响。

[0105] 步骤S312:表单图像直线检测结果输出。

[0106] 此外,针对每条线段,可以保留以下信息:Pstart:起点坐标;Pend:终点坐标;L:长度;W:平均线宽。

[0107] 在对表单进行直线检测后,进行表单类型识别。本实施例的表单类型识别流程如图12所示,包括:表单特征提取、计算与各模板表单的相似度、判断表单所属的表单类别几部分。

[0108] 下面分别对表单类型识别流程的各部分进行详细描述:

[0109] 步骤S314:表单特征提取。

[0110] 不同种类的表单具有不同的结构,本实施例通过一种通用的方式来描述表单,获得表单的结构特征,并以此作为判断表单类型的依据。

[0111] 具体方法如下所述:

[0112] (a3)交叉点(特征点)提取:常见的表单总共有九类交叉点可供提取,之前提取的每条线段都包含有起点坐标和终点坐标,可以用于进行特征点的提取。再次参考图4,对特征点的提取可以根据表单的特性,提取九类特征点:西北部特征点、北部特征点、东北部特征点、东部特征点、东南部特征点、南部特征点、西南部特征点、西部特征点、中部特征点。

[0113] 以图13的表单为例,进行了直线检测后输出的表单如图14所示,对图14所示的表单进行交叉点提取,如图15所示,其中,NW表示西北部特征点,N表示北部特征点,NE表示东

北部特征点,E表示东部特征点,SE表示东南部特征点,S表示南部特征点,SW表示西南部特征点,W表示西部特征点,M表示中部特征点。另外,水平直线条数 $H=8$,垂直直线 $V=12$,统计各维度(N、NE、E、SE、S、SW、W、NW、M、H、V)的出现次数,得出该表单在此维度下的11维特征(10,1,6,1,10,1,6,1,13,8,12)。

[0114] 为了求得更高的精度,对该表单进行下述多分辨率处理。在多分辨率处理下,该表单的特征向量就是在不同子窗口下上述各维度特征的统计次数。

[0115] (b3)表单外框区域确定:根据上面提取到的四个角上的特征点,可以确定出表单的区域。

[0116] (c3)多分辨率处理:得到表单所在区域之后,首先对表单的大小作归一化的处理,比如统一缩放到 640×480 。对归一化之后的图像进行多级分层的 $N \times N$ 划块。如图13所示,比如3层的图像,每层进行 2×2 划块,总共可以划分为 $1+4+16=21$ 个子区域(窗口)。

[0117] (d3)特征提取:本实施例总共提取11类特征,包括(a3)中提取的九个特征的数量,以及水平和垂直直线的数量。也即,对多块表单中的每一块表单,分别根据本块表单内的各个表单框线和各个表单框线之间的位置关系,进行本块表单的交叉点提取,提取本块表单中的表单框线的各个交叉点,获取本块表单中的各个交叉点的信息,并且获取本块表单中的水平方向表单框线的信息、和垂直方向表单框线的信息。

[0118] (e3)特征直方图统计:针对图13中的多级分层,提取各图像子窗口的特征数量,然后合并为直方图。对于图13中的三层直方图,总共可以提取 $(1+4+16) \times 11=232$ 维特征。

[0119] (f3)特征输出:将上述232维特征作为描述该类表单结构的特征向量输出。

[0120] 由上述可见,如果用11维向量用于表单识别容易产生较大的误差(维度太少,信息不够丰富),而通过多分辨率处理可以提高识别的精度,例如,采用文中的 $1 \times 2 \times 2$ 、 4×4 三种分辨率,最后的结果就成为了 $11 \times 21=231$ 维,识别的精度更高了。

[0121] 步骤S316:计算与各模型(模板表单)的相似度及判断所属的表单类别。

[0122] 例如,计算表单的特征向量与各模板表单的特征向量之间的欧氏距离。如果已知输入表单一定属于表单模板库,可以直接返回欧氏距离最短的表单类型;否则,可以采用以下的方法来判断:

[0123] (a4)计算各模板表单与输入表单图片中的表单的特征向量欧氏距离次相近以及最相近的比值;

[0124] (b4)如果该值大于或等于T,则返回与输入表单图片距离最相近的模板表单;

[0125] (c4)如果该值小于T,则认为输入表单图片不符合任一模板表单。

[0126] 优选地,T取3。

[0127] 此外,本实施例中表单模型(也即模板表单)可以通过图16的流程建立,包括:样本图片(表单模板)输入、样本图片直线检测和提取、样本图片特征提取和表单模型建立。其中,样本图片(表单模板)输入和样本图片直线检测和提取可以参照本实施例中表单直线检测流程的相应部分进行,而样本图片特征提取则可以参照本实施例中表单类型识别流程中的特征提取部分进行,在对样本图片中的表单进行了特征提取,获取了特征向量后,将该特征向量作为描述该类表单的模型输出,从而生成模板表单。例如,准备不同的表单模板(比如发票A、表格B、报销单C),采用本实施例中的方式进行直线检测和特征提取,最后训练输出的结果就是每个表单模型对应一个高维向量(如232维),如下所示:A:{232维向量};B:

{232维向量};C:{232维向量}。

[0128] 通过本实施例,提供了一种表单图片的结构自动识别方案,其对表单图像的直线检测对光照亮度、倾斜、断裂、噪音、不同线段宽度等都具有较强的抵抗性,输入表单图的获取方式可以是屏幕截屏、照片拍摄、或者扫描仪上传,来源可能是电子表格或是手绘图表,并且对图片质量没有很严格的要求;当样本库中含有多个模板表单时,提供了一种表单特征的描述方法,可以通过实现训练的方式,自动识别输入表单各自归属类别;并且,除去训练过程,整个识别过程都可以自动化,不需要任何人工的干预。

[0129] 实施例四

[0130] 参照图17,示出了根据本申请实施例四的一种表单识别装置的结构框图。

[0131] 本实施例的表单识别装置包括:直线检测模块402,用于对表单二值图像中的表单进行直线检测,获取表单的各个表单框线和各个表单框线之间的位置关系;特征提取模块404,用于根据各个表单框线和各个表单框线之间的位置关系,对表单进行特征提取,获取表单的特征向量,其中,特征向量用于指示表单的结构特征;相似度计算模块406,用于根据表单的特征向量,计算所述表单与模板表单的相似度;表单识别模块408,用于根据相似度的计算结果,识别所述表单。

[0132] 优选地,直线检测模块402,用于对于所述表单上的每一条直线,以本直线上的设定点为起点,对本直线上的相邻像素进行连通分量分析,获取本直线的连通分量;获取本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数;根据本直线的连通分量上的每一个像素在与本直线垂直方向上的连通分量的像素个数,获取本直线的平均线宽;判断本直线的连通分量上的每一个像素所在处的线宽是否大于平均线宽,若是,则对大于平均线宽的像素所在处的线宽进行线宽去除处理;根据进行了线宽去除处理的每一条直线,获取表单的各个表单框线和各个表单框线之间的位置关系。

[0133] 优选地,直线检测模块402在对大于平均线宽的像素所在处的线宽进行线宽去除处理时,将大于平均线宽的像素从本直线的连通分量中去除;判断去除的像素的位置是否在本直线的内部,若是,则对去除的像素两侧的线段进行线段合并。

[0134] 优选地,特征提取模块404,用于根据各个表单框线和各个表单框线之间的位置关系,提取表单框线的各个交叉点,获取各个交叉点的信息,以及,获取各个表单框线中水平方向表单框线的信息、和垂直方向表单框线的信息;根据各个交叉点的信息、水平方向表单框线的信息、和垂直方向表单框线的信息,获取表单的特征向量。

[0135] 优选地,本实施例的表单识别装置还包括:表单处理模块410,用于在特征提取模块404根据各个表单框线和各个表单框线之间的位置关系,提取表单框线的各个交叉点之前,确定表单的区域;根据表单的区域,对表单做归一化处理。

[0136] 优选地,表单处理模块410,还用于将归一化处理后的表单划分为多块;特征提取模块404,用于对多块表单中的每一块表单,分别根据本块表单内的各个表单框线和各个表单框线之间的位置关系,进行本块表单的交叉点提取,提取本块表单中的表单框线的各个交叉点,获取本块表单中的各个交叉点的信息,并且获取本块表单中的水平方向表单框线的信息、和垂直方向表单框线的信息;根据获取的每一块表单中的各个交叉点的信息、水平方向表单框线的信息、和垂直方向表单框线的信息,获取表单的特征向量。

[0137] 优选地,相似度计算模块406,用于计算表单的特征向量与各个模板表单的特征向

量之间的最相近距离和次相近距离的比值;若比值大于或等于设定比值,则确定表单与最相近距离对应的模板表单的相似度最高;若比值小于设定比值,则确定表单与各个模板表单均不相似。

[0138] 优选地,本实施例的表单识别装置还包括:线段合并模块412,用于在特征提取模块404根据各个表单框线和各个表单框线之间的位置关系,对表单进行特征提取,获取所述表单的特征向量之前,按照设定规则判断同一方向上相邻的两条表单框线是否需要线段合并,若是,则对所述相邻的两条表单框线进行线段合并。

[0139] 优选地,线段合并模块412,用于获取同一方向上相邻的两条表单框线的距离,根据所述距离与设定的第一阈值的比较结果判断是否需要线段合并,和/或,获取同一方向上相邻的两条表单框线的相似度,根据所述相似度判断是否需要线段合并;若是,则对所述相邻的两条表单框线进行线段合并。

[0140] 优选地,线段合并模块412在获取同一方向上相邻的两条表单框线的相似度,根据所述相似度判断是否需要线段合并时,获取同一方向上相邻的两条表单框线之间的夹角,根据所述夹角与设定的第二阈值的比较结果判断是否需要线段合并;和/或,获取同一方向上相邻的两条表单框线之间的平行距离,根据所述平行距离与设定的第三阈值的比较结果判断是否需要线段合并。

[0141] 优选地,本实施例的表单识别装置还包括:字符线段过滤模块414,用于在特征提取模块404根据各个表单框线和各个表单框线之间的位置关系,对表单进行特征提取,获取所述表单的特征向量之前,对获取的表单的各个表单框线进行字符线段过滤。

[0142] 优选地,字符线段过滤模块414,用于判断表单的各个表单框线的线段长度是否小于或等于设定的第四阈值,若是,则去除线段长度小于或等于第四阈值的表单框线。

[0143] 优选地,第四阈值通过以下方式设定:获取表单的各个表单框线中的最大连通分量的长度,并将所述长度划分为多个长度区间;将各个表单框线的每一个连通分量按照其长度划分到多个长度区间中的一个长度区间;根据多个长度区间的长度值和每一个长度区间的连通分量的个数,生成线段长度分布直方图;将直方图中纵坐标最高的点对应的长度值作为第四阈值。

[0144] 本实施例的表单识别装置用于实现前述多个方法实施例中相应的表单识别方法,并具有相应的方法实施例的有益效果,在此不再赘述。

[0145] 本申请提供了一种表单图像的自动识别方案,其主要包括直线检测和输入表单图片中的表单类型判断两部分。在直线检测中,提供了一种通过连通分量分析提取直线的优选方案,通过该方案,能够对表单框线进行准确、清晰地提取;在表单类型判断中,提供了一种对表单进行多分辨率特征提取的优选方案,通过该方案,能够提高表单识别的精度。

[0146] 本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。对于装置实施例而言,由于其与方法实施例基本相似,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0147] 以上对本申请所提供的一种表单识别方法和装置进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体

实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

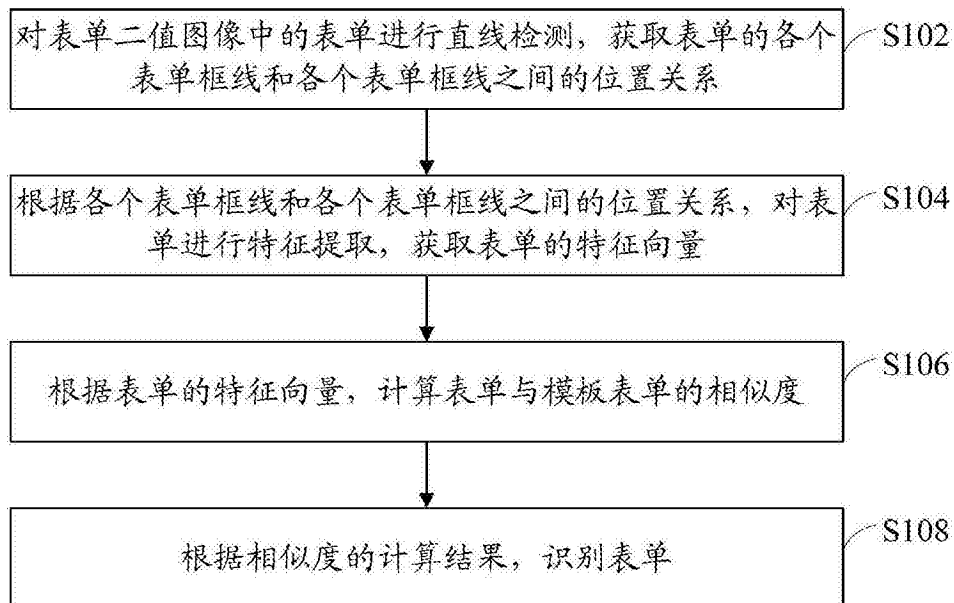


图1

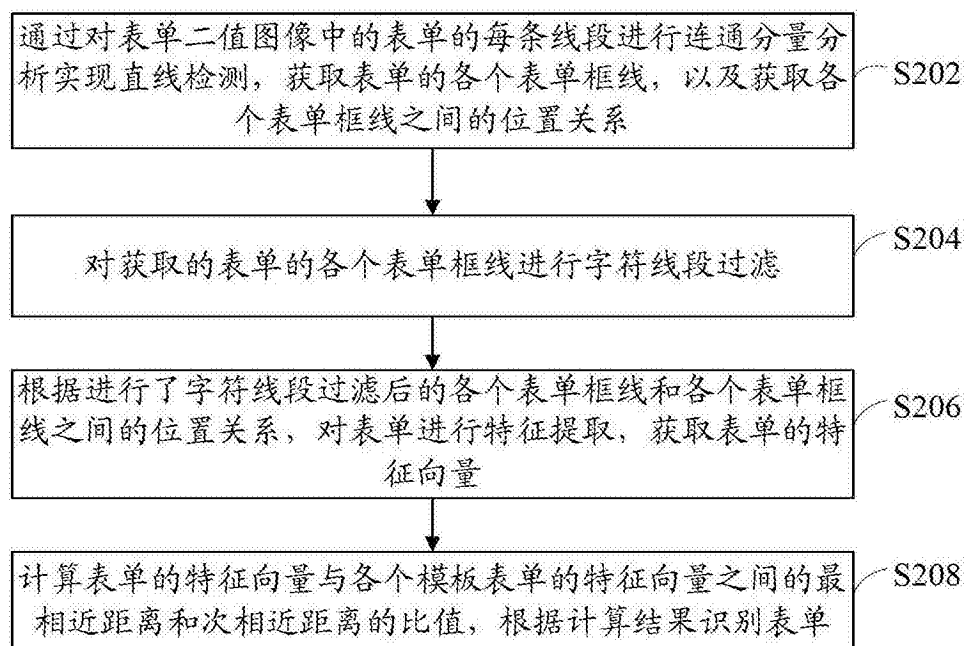


图2

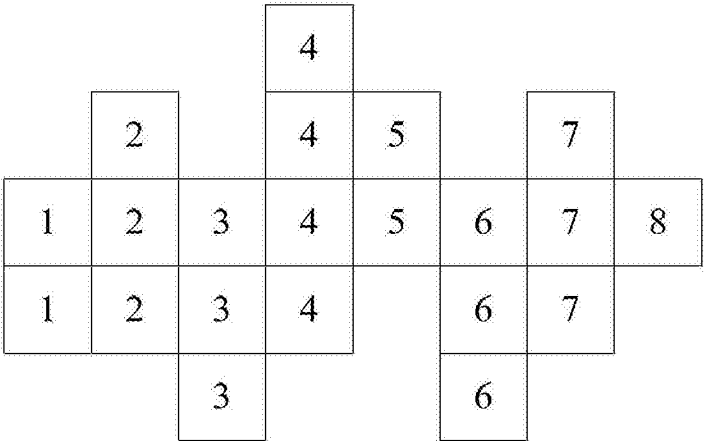


图3

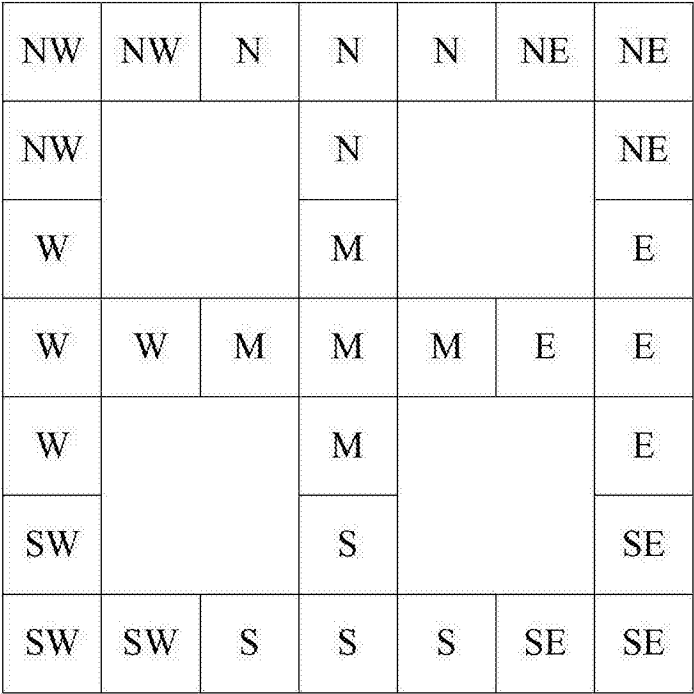


图4

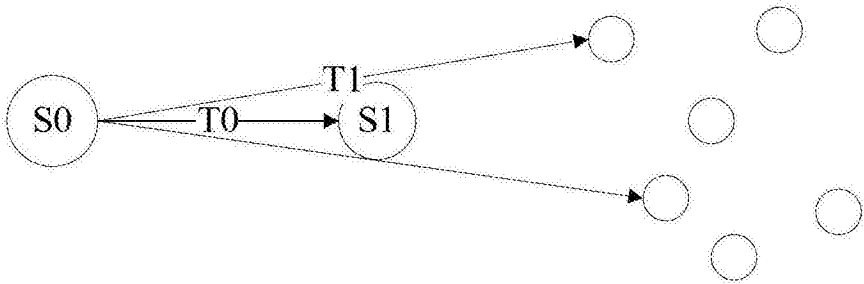


图5

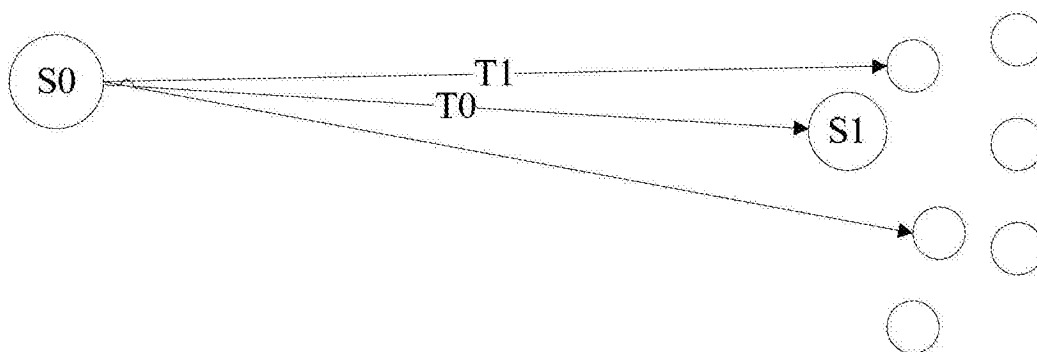


图6

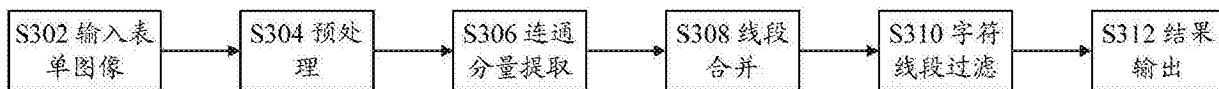


图7

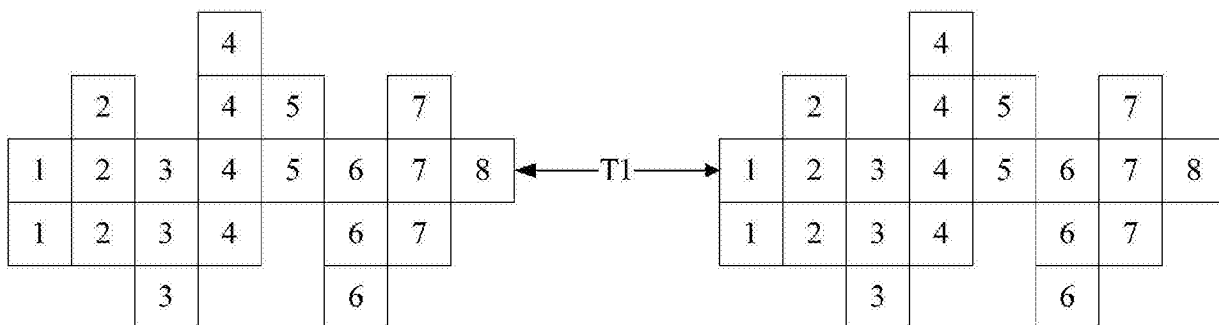


图8

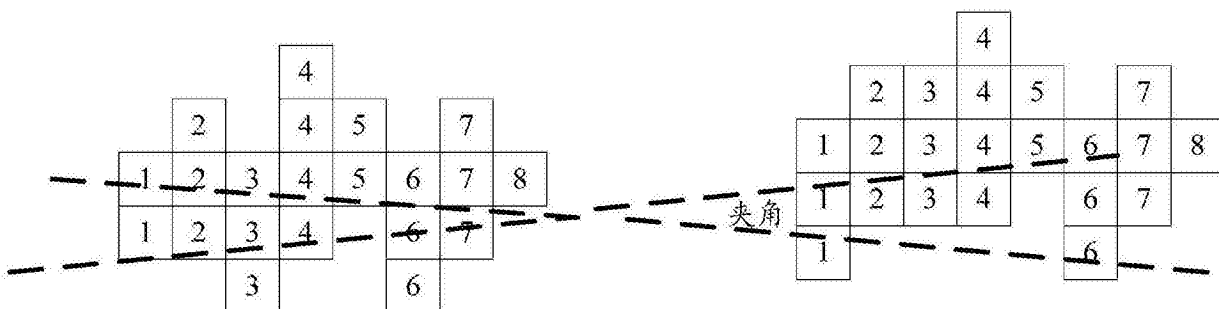


图9

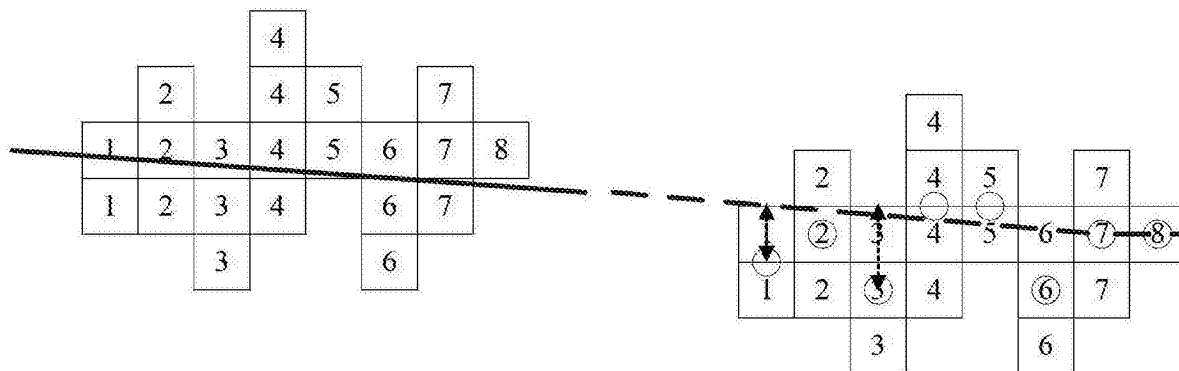


图10

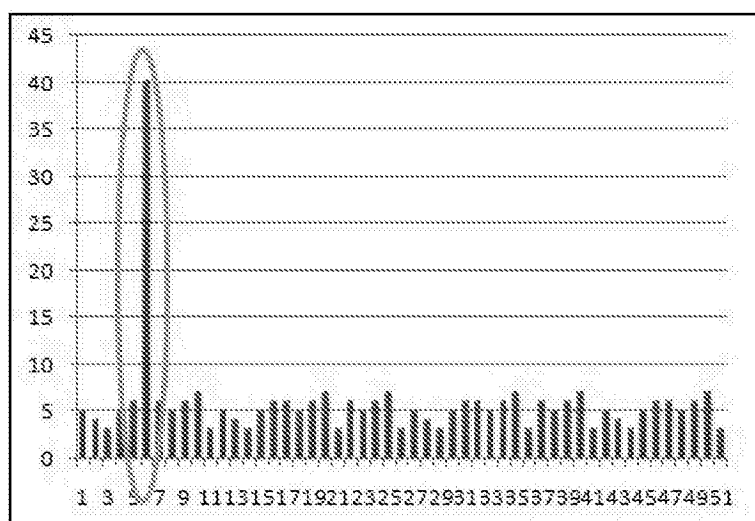


图11

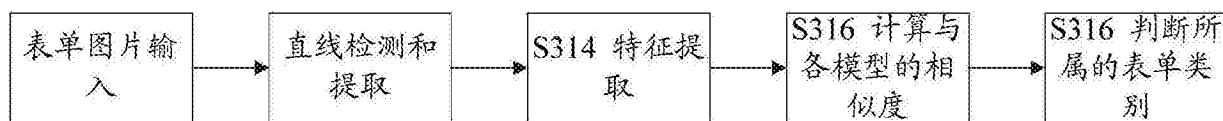


图12

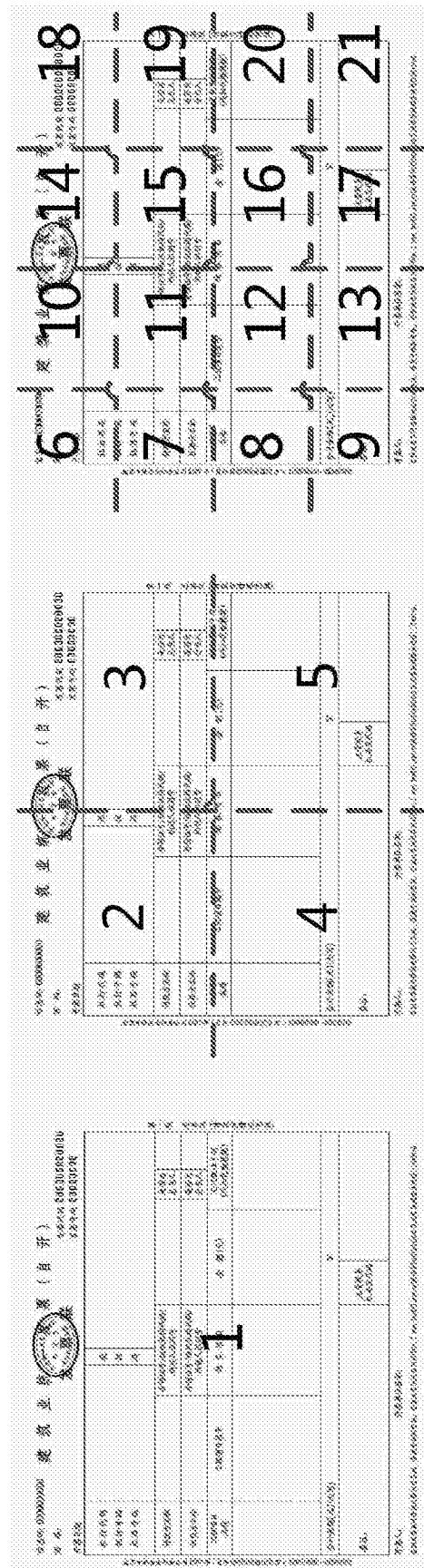


图13

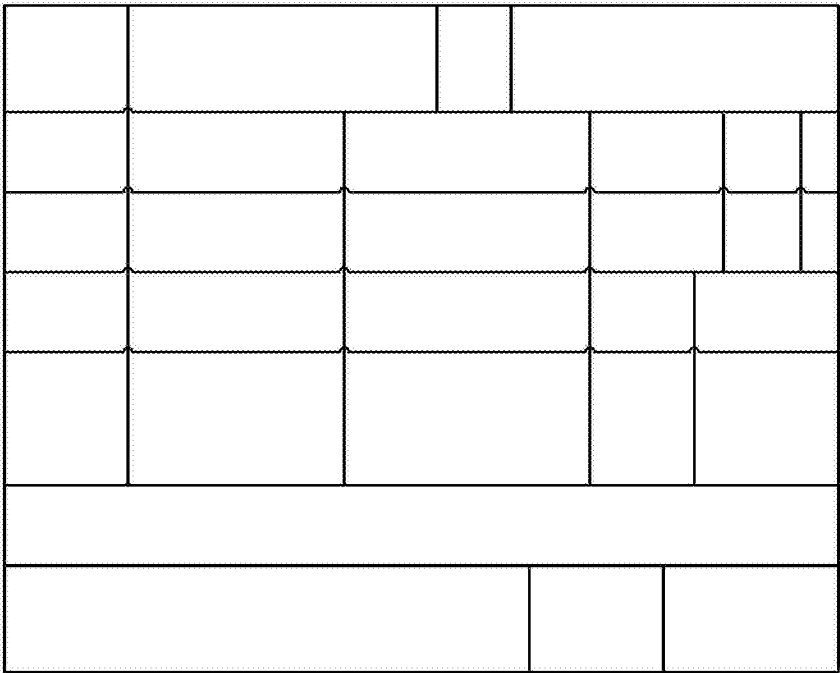


图14

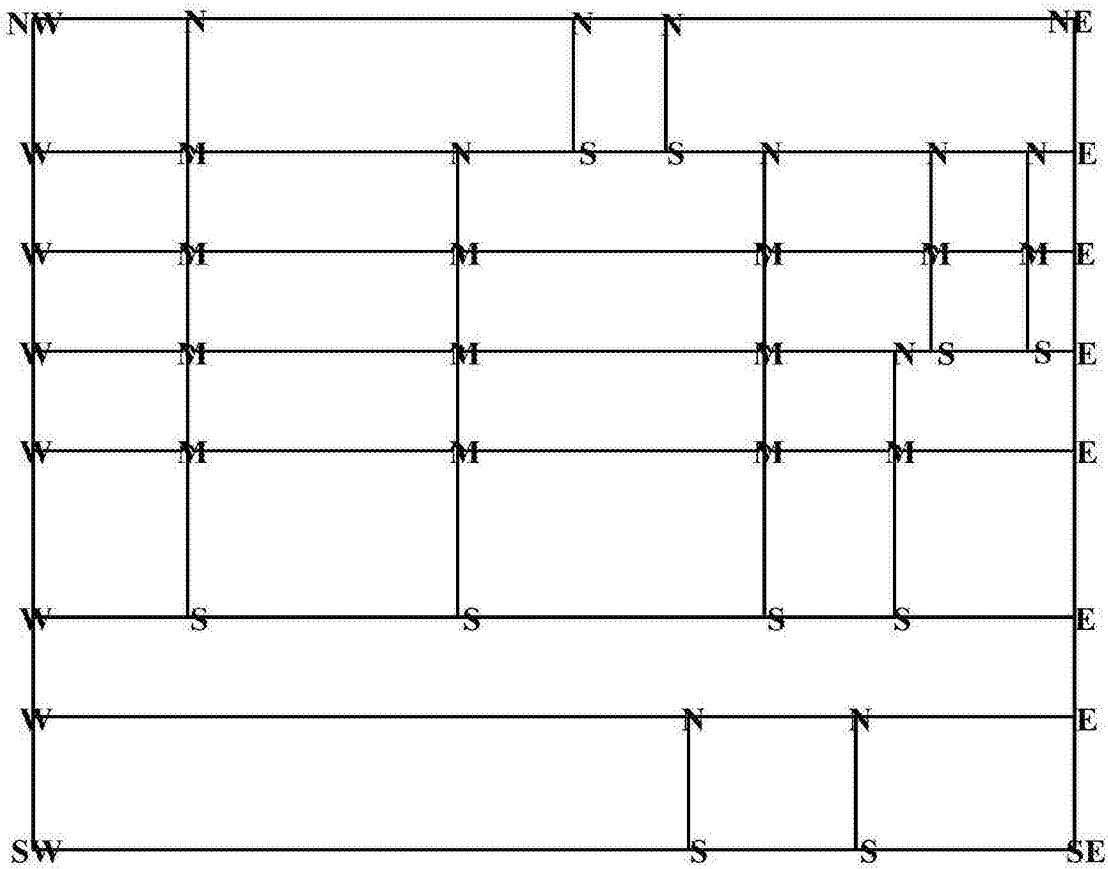


图15

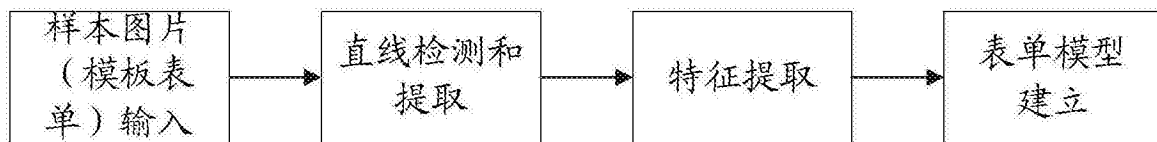


图16

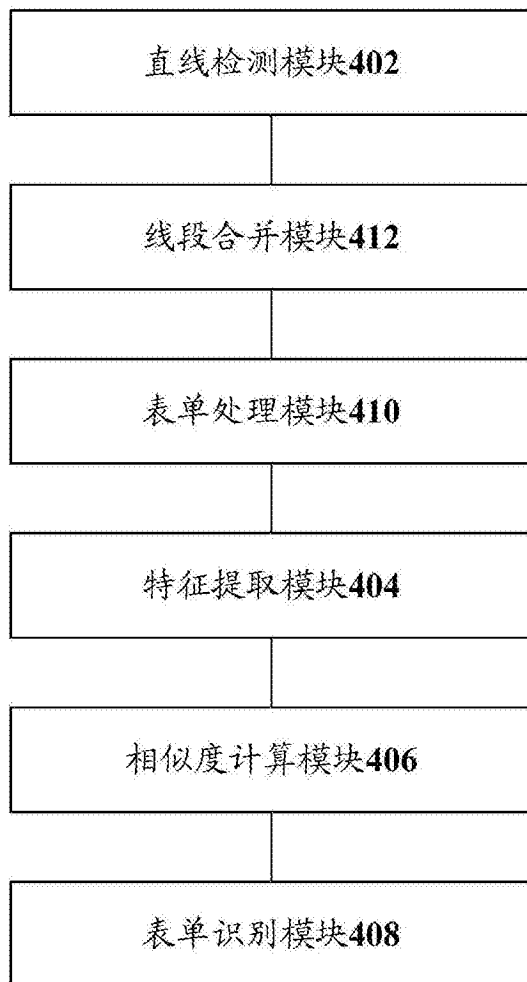


图17