(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112446351 B
(45)授权公告日 2022.08.09

(21)申请号 202011463078.9

(22)申请日 2020.12.09

(65)同一申请的已公布的文献号
申请公布号 CN 112446351 A

(43)申请公布日 2021.03.05

(73)专利权人 杭州米数科技有限公司
地址 310051 浙江省杭州市滨江区长河街道江陵路601号东方商务会馆225室
专利权人 杭州观图科技有限公司

(72)发明人 谭谞 高海东 沈旭辉 杨章岳

(74)专利代理机构 杭州融方专利代理事务所(普通合伙) 33266
专利代理师 沈相权

(51)Int.Cl.
G06V 30/413(2022.01)
G06V 30/146(2022.01)
G06V 30/18(2022.01)
G06V 10/82(2022.01)
G06F 16/245(2019.01)
G06N 3/04(2006.01)

(56)对比文件
CN 112016547 A,2020.12.01
CN 111738055 A,2020.10.02
CN 109886174 A,2019.06.14
CN 102332096 A,2012.01.25
CN 101770575 A,2010.07.07
US 2020226400 A1,2020.07.16

审查员 刘海艳

权利要求书2页 说明书7页 附图6页

(54)发明名称

医疗票据智能识别方法

(57)摘要

本发明公开了一种医疗票据智能识别方法。属于票据智能识别技术领域，提供一种易于对OCR结果进行高准确分行，可靠性高，方法的实现过程如下：N1、输入票据图像；N2、对票据图像进行方向检测和类型分类；N3、对图像进行翻正；N4、对票据文字区域进行检测；N5、对票据文字区域进行识别；N6、对票据版面进行分析；N7、对票据识别内容进行纠正；N8、结构化信息输出。

CN 112446351 B

1.医疗票据智能识别方法,其特征在于,方法的实现过程如下:

N1、输入票据图像;

N2、对票据图像进行方向检测和类型分类;

N3、对图像进行翻正;

N4、对票据文字区域进行检测;

N5、对票据文字区域进行识别;

N6、对票据版面进行分析;

N7、对票据识别内容进行纠正;

N8、结构化信息输出;

在对票据图像进行方向检测和类型分类时,采用深度学习分类网络对输入的票据图像进行方向识别和类型分类;票据图像的方向识别包括逆时针方向的0°、90°、180°、270°这4个方向识别;票据图像的类型分类包括门诊发票、住院清单和军医发票;

在对图像进行翻正时,根据票据图像的方向识别结果对票据图像转正方向,然后根据票据类型分类结果选择检测模型;

在对票据文字区域进行检测时,采用旋转候选矩形框的轮换区域生成网络RRPN方式来对倾斜的文字区域进行检测;RRPN通过在特征图上设置锚点,从而对输入图像进行密集采样,然后通过一个二分类任务来判断锚点是前景还是背景,并用一个回归模型来预测锚点的相对位置;RRPN对于锚点设置了不同的尺度、长宽比以及旋转角度;在对门诊发票进行文字区域检测时将发票上的固有文字与打印后的文字分别进行检测;在对住院清单进行文字区域检测时采用多个条目同时检测;

在对票据文字区域进行识别时,对文字区域检测得到的文字框,采取卷积循环神经网络CRNN对各个文字条目进行文字识别:先用CRNN提取图像卷积特征,然后使用双向长短期记忆人工神经网络LSTM进一步提取图像卷积特征中的序列特征,最后引入连接时序分类CTC损失解决字符对齐的问题;并利用医疗药品库信息,仿照医疗票据复印件图像的文字模糊实际情况,对文字条目进行数据增强,生成大量训练数据;

在对票据版面进行分析时,首先获取每个文字条目的行号:在得到各个文字条目之后,根据文字条目的纵坐标从小到大进行排序,先将当前行置为空,然后逐个取出文字条目,根据此文字条目的中心位置信息,判断其是否属于当前行;如果文字条目属于当前行,则加入当前行并且更新此行的相关信息;如果文字条目不属于当前行,则新起一行,作为当前行,行号加1;

然后获取每个文字条目的列号:对同属一行的文字条目,按照横坐标值,从小到大进行顺序,从而得到该文字条目所属的列号;

在对票据识别内容进行纠正时,在识别出的文字条目中,按照名称、金额的顺序,筛选出属于药品名称的药品文字条目;将得到的药品文字条目,与医药库中的标准药品名称信息进行比较,按照编辑距离、以及识别错误字典得到标准药品名称;

在结构化信息输出时,根据票据类型、以及识别出的文字信息,得到票据的票号、患者姓名信息,输出字典形式的结构化信息。

2.根据权利要求1所述的医疗票据智能识别方法,其特征在于,对票据版面进行分析的实现过程如下:

K1,对医疗票据的OCR结果框依次进行x、y坐标递增排序;

K2,按排序后的顺序依次遍历OCR框,其中第一个框必定是第一行,也是最新的一行,记作L1,读作第1行,以此类推,第N行记作L_n,读作第n行,最新的一行记作L_new;

K3,依次遍历接下来的OCR框,把遍历到的OCR框记作H框;尝试放到某一行中,尝试规则按K4执行;

K4,尝试把H框放到第L_(new-4)中,判定H框是否属于L(new-4),如果属于则放到该行,不属于则判定H框在L_(new-4)行的上一行还是在L_(new-4)行的下一行;如果是在L_(new-4)行的上一行则尝试放到L_(new-5)上,如果是在L_(new-4)行的下一行则尝试放到L_(new-3)上;

以此类推,如果判定到L_new的时候,H框仍 然处于L_new之下,则创建新行并把H框放到新创建的最新的行中,重复新的H框判定;判定一个H框是否属于某一行,按K5的规则判定执行;

K5,如果L_new的框个数不足两个或者该行的唯一一个框的宽高比不足5,则按K6处理,否则按K7步处理;

K6,取H框的往上和往下各15个OCR框的平均斜率作为对齐斜率,过H框的中点用此斜率虚拟出一条直线F,如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值,则H框属于该行,否则属于上一行或者下一行;

K7,取H框距离该行往上的一行最近的两个框的中点连接线斜率作为参考的直线F斜率,过H框的中点用此斜率虚拟出一条直线F,如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值,则H框属于该行,否则属于上一行或者下一行;

K8,经过上面K1-K7后,会得到第一版的分行结果,然后在分行结果中找出最优的一行,最优判定条件为,某行中两个邻近框的连接线斜率差最小,同时OCR框斜率差也为最小,则分行结果就为最优行;

K9,用最优行作为初始参考行,然后取最优行往上的所有OCR框和最优行往下的所有OCR框再次进行K1到K7后得到两份分行结果,然后再拼成一个完整分行结果,至此分行结束。

3.根据权利要求2所述的医疗票据智能识别方法,其特征在于,在K6或K7中,还包括,如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值时,并且在H框对应的OCR框的左下角顶点处画平行于直线F的直线J,如果直线J与H框所在行的上一行前一个OCR框相交或下一行后一个OCR框相交时,则判定该H框属于该行。

4.根据权利要求3所述的医疗票据智能识别方法,其特征在于,如果H框所在行的上一行前一个OCR框的左下角顶点落在直线J上,或者H框所在行的下一行后一个OCR框的左下角顶点落在直线J上时,则判定该H框属于该行,并且该H框处于最优行上。

医疗票据智能识别方法

技术领域

[0001] 本发明涉及票据智能识别技术领域，具体涉及一种医疗票据智能识别方法。

背景技术

[0002] 医疗票据图像识别在保险理赔鉴定等领域有着广泛的应用。医疗票据具有版式繁多(各地发票不同，各医院清单不同)、条目量大(一张清单有时达到几百条项目)、专有名词多、打印质量参差不齐、纸张放置不标准、用户揉捏导致的纸面不平整等情况。首先，采用人工校对核验的方式往往工作量巨大。例如，处理一张多达几百条目的住院费用清单，有经验的业务员一般也需要数分钟。

[0003] 其次，目前通用的光学字符识别OCR(OpticalCharacterRecognition)算法，对于医疗票据识别的准确率往往不尽如人意(单字准确率低于70%)，且未对识别的结果(例如项目、数量、金额等)进行结构化处理，因此无法在实际中得以应用。

[0004] 目前对于非常标准、用人眼看起来就是从左到右、从上到下的文本图像的光学字符识别，要对其OCR结果进行分行，则非常简单，一般按照正常分行方法，从左到右、从上到下按x坐标和y坐标进行排序遍历，新的框距离最新的一行的距离超过某个阈值时，则是新的一行，否则是当前的行，依次遍历所有OCR结果框，即可分行。

[0005] 但是对于实际场景中，图像往往大部分都存在扭曲、透视和旋转的问题，以上用于非常标准的图像的分行算法思路将不再适用。

发明内容

[0006] 本发明是为了解决现在票据智能识别未对识别的结果进行结构化处理无法在实际中得以应用的不足，提供一种易于对票据智能识别结果进行结构化处理，易于使用，可靠性好的医疗票据智能识别方法。

[0007] 为实现以上目的，本发明通过以下技术方案予以实现：

[0008] 医疗票据智能识别方法，方法的实现过程如下：

[0009] N1、输入票据图像；

[0010] N2、对票据图像进行方向检测和类型分类；

[0011] 采用深度学习分类网络对输入的票据图像进行方向识别和类型分类；票据图像的方向识别包括逆时针方向的0°、90°、180°、270°这4个方向识别；票据图像的类型分类包括门诊发票、住院清单和军医发票；

[0012] N3、对图像进行翻正；

[0013] 根据票据图像的方向识别结果对票据图像转正方向，然后根据票据类型分类结果选择检测模型；

[0014] N4、对票据文字区域进行检测；

[0015] 采用旋转候选矩形框的轮换区域生成网络RRPN方式来对倾斜的文字区域进行检测；RRPN通过在特征图上设置锚点，从而对输入图像进行密集采样，然后通过一个二分类任

务来判断锚点是前景还是背景，并用一个回归模型来预测锚点的相对位置；RRPN对于锚点设置了不同的尺度、长宽比以及旋转角度；在对门诊发票进行文字区域检测时将发票上的固有文字与打印后的文字分别进行检测；在对住院清单进行文字区域检测时采用多个条目同时检测；

[0016] N5、对票据文字区域进行识别；

[0017] 对文字区域检测得到的文字框，采取卷积循环神经网络CRNN对各个文字条目进行文字识别：先用CRNN提取图像卷积特征，然后使用双向长短期记忆人工神经网络LSTM进一步提取图像卷积特征中的序列特征，最后引入连接时序分类CTC损失解决字符对齐的问题；并利用医疗药品库信息，仿照医疗票据复印件图像的文字模糊实际情况，对文字条目进行数据增强，生成大量训练数据；

[0018] N6、对票据版面进行分析；

[0019] 首先获取每个文字条目的行号：在得到各个文字条目之后，根据文字条目的纵坐标从小到大进行排序，先将当前行置为空，然后逐个取出文字条目，根据此文字条目的中心位置信息，判断其是否属于当前行；如果文字条目属于当前行，则加入当前行并且更新此行的相关信息；如果文字条目不属于当前行，则新起一行，作为当前行，行号加1；

[0020] 然后获取每个文字条目的列号：对同属一行的文字条目，按照横坐标值，从小到大进行顺序，从而得到该文字条目所属的列号；

[0021] N7、对票据识别内容进行纠正；

[0022] 在识别出的文字条目中，按照名称、金额的顺序，筛选出属于药品名称的药品文字条目；将得到的药品文字条目，与医药库中的标准药品名称信息进行比较，按照编辑距离、以及识别错误字典得到标准药品名称；

[0023] N8、结构化信息输出。

[0024] 根据票据类型、以及识别出的文字信息，得到票据的票号、患者姓名信息，输出字典形式的结构化信息。

[0025] 易于对票据智能识别结果进行结构化处理，易于使用，可靠性好，泛化能力强

[0026] 作为优选，对票据版面进行分析的实现过程如下：

[0027] K1，对OCR结果框依次进行x、y坐标递增排序；

[0028] K2，按排序后的顺序依次遍历OCR框，其中第一个框必定是第一行，也是最新的一行，记作L1，读作第1行，以此类推，第N行记作L_n，读作第n行，最新的一行记作L_new；

[0029] K3，依次遍历接下来的OCR框，把遍历到的OCR框记作H框；尝试放到某一行中，尝试规则按K4执行；

[0030] K4，尝试把H框放到第L_(new-4)中，判定H框是否属于L_(new-4)，如果属于则放到该行，不属于则判定H框在L_(new-4)行的上一行还是在L_(new-4)行的下一行；如果是在L_(new-4)行的上一行则尝试放到L_(new-5)上，如果是在L_(new-4)行的下一行则尝试放到L_(new-3)上；

[0031] 以此类推，如果判定到L_new的时候，H框仍然处于L_new之下，则创建新行并把H框放到新创建的最新的行中，重复新的H框判定；判定一个H框是否属于某一行，按K5的规则判定执行；

[0032] K5，如果L_new的框个数不足两个或者该行的唯一一个框的宽高比不足5，则按K6

处理，否则按K7步处理；

[0033] K6，取H框的往上和往下各15个OCR框的平均斜率作为对齐斜率，过H框的中点用此斜率虚拟出一条直线F，如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值，则H框属于该行，否则属于上一行或者下一行；

[0034] K7，取H框距离该行往上的一行最近的两个框的中点连接线斜率作为参考的直线F斜率，过H框的中点用此斜率虚拟出一条直线F，如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值，则H框属于该行，否则属于上一行或者下一行；

[0035] K8，经过上面K1-K7后，会得到第一版的分行结果，然后在分行结果中找出最优的一行，最优判定条件为，某行中两个邻近框的连接线斜率差最小，同时OCR框斜率差也为最小，则分行结果就为最优行；

[0036] K9，用最优行作为初始参考行，然后取最优行往上的所有OCR框和最优行往下的所有OCR框再次进行K1到K7后得到两份分行结果，然后再拼成一个完整分行结果，至此分行结束。

[0037] 本方案易于对OCR结果进行高准确分行，可靠性高，一是对扭曲图像仍然具备很高的分行准确率；二是对旋转不超过30度的图像具备很高的分行准确率；三是泛化能力强，只要本身图像中的文本按行排版，都可以用本算法进行分行。

[0038] 作为优选，在K6或K7中，还包括，如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值时，并且在H框对应的OCR框的左下角顶点处画平行于直线F的直线J，如果直线J与H框所在行的上一行前一个OCR框相交或下一行后一个OCR框相交时，则判定该H框属于该行。

[0039] 作为优选，如果H框所在行的上一行前一个OCR框的左下角顶点落在直线J上，或者H框所在行的下一行后一个OCR框的左下角顶点落在直线J上时，则判定该H框属于该行，并且该H框处于最优行上。

[0040] 本发明能够达到如下效果：

[0041] 本发明易于对票据智能识别结果进行结构化处理，易于使用，可靠性好，泛化能力强。

附图说明

[0042] 图1为本发明实施例1的一种流程示意图。

[0043] 图2为本发明实施例1票据方向、类型分类网络的一种流程示意图。

[0044] 图3为本发明实施例1行号的一种流程示意图。

[0045] 图4为本发明实施例2的一种流程示意图。

[0046] 图5为本发明实施例2的一种示意图。

[0047] 图6为本发明实施例2的一种流程示意图。

[0048] 图7为本发明实施例3的一种示意图。

具体实施方式

[0049] 下面结合附图与实施例对本发明作进一步的说明。

[0050] 实施例1，医疗票据智能识别方法，参见图1所示；方法的实现过程如下：

[0051] N1、输入票据图像；

[0052] N2、对票据图像进行方向检测和类型分类；参见图2所示；

[0053] 采用深度学习分类网络对输入的票据图像进行方向识别和类型分类；票据图像的方向识别包括逆时针方向的0°、90°、180°、270°这4个方向识别；票据图像的类型分类包括门诊发票、住院清单和军医发票；

[0054] 深度学习分类网络为卷积神经网络CNN(Convolutional Neural Network)，通过卷积神经网络对输入图像提取特征，然后分别针对图像方向和类型分类进行特征提取；

[0055] N3、对图像进行翻正；

[0056] 根据票据图像的方向识别结果对票据图像转正方向，然后根据票据类型分类结果选择检测模型；

[0057] N4、对票据文字区域进行检测；

[0058] 采用旋转候选矩形框的轮换区域生成网络RRPN(Rotation Region Proposal Networks)方式来对倾斜的文字区域进行检测；RRPN通过在特征图上设置锚点，从而对输入图像进行密集采样，然后通过一个二分类任务来判断锚点是前景还是背景，并用一个回归模型来预测锚点的相对位置；RRPN对于锚点设置了不同的尺度、长宽比以及旋转角度；在对门诊发票进行文字区域检测时将发票上的固有文字与打印后的文字分别进行检测；在对住院清单进行文字区域检测时采用多个条目同时检测；

[0059] N5、对票据文字区域进行识别；

[0060] 对文字区域检测得到的文字框，采取卷积循环神经网络CRNN(Convolutional Recurrent Neural Network)对各个文字条目进行文字识别：先用CRNN提取图像卷积特征，然后使用双向长短期记忆人工神经网络LSTM进一步提取图像卷积特征中的序列特征，最后引入连接时序分类CTC(Connectionist Temporal Classification)损失解决字符对齐的问题；并利用医疗药品库信息，仿照医疗票据复印件图像的文字模糊实际情况，对文字条目进行数据增强，生成大量训练数据；

[0061] 通过数据增强方式让票据文字识别模型对于实际情况有了更多的兼容性；

[0062] N6、对票据版面进行分析；参见图2所示；

[0063] 首先获取每个文字条目的行号：在得到各个文字条目之后，根据文字条目的纵坐标从小到大进行排序，先将当前行置为空，然后逐个取出文字条目，根据此文字条目的中心位置信息，判断其是否属于当前行；如果文字条目属于当前行，则加入当前行并且更新此行的相关信息；如果文字条目不属于当前行，则新起一行，作为当前行，行号加1；

[0064] 然后获取每个文字条目的列号：对同属一行的文字条目，按照横坐标值，从小到大进行顺序，从而得到该文字条目所属的列号；

[0065] N7、对票据识别内容进行纠正；

[0066] 在识别出的文字条目中，按照名称、金额的顺序，筛选出属于药品名称的药品文字条目；将得到的药品文字条目，与医药库中的标准药品名称信息进行比较，按照编辑距离、以及识别错误字典得到标准药品名称；

[0067] 比如，识别成“氧化钠注射液”的条目，通过编辑距离可在标准名称库里发现“氯化钠注射液”此条目。再加上“氯”识别成”氧”字的情况亦存在于识别错误字典中，因此可将“氧化钠注射液”修正为“氯化钠注射液”。

[0068] N8、结构化信息输出。

[0069] 根据票据类型、以及识别出的文字信息(如姓名、社保号码等固有文字),得到票据的票号、患者姓名信息,输出字典形式的结构化信息。

[0070] 实施例2,实施例2与实施例1不同在于,参见图4所示;

[0071] 对票据版面进行分析的实现过程如下:

[0072] K1,对医疗票据的OCR结果框依次进行x、y坐标递增排序;

[0073] K2,按排序后的顺序依次遍历OCR框,其中第一个框必定是第一行,也是最新的一行,记作L1,读作第1行,以此类推,第N行记作L_n,读作第n行,最新的一行记作L_new;

[0074] K3,依次遍历接下来的OCR框,把遍历到的OCR框记作H框;尝试放到某一行中,尝试规则按K4执行;

[0075] K4,尝试把H框放到第L_(new-4)中,判定H框是否属于L_(new-4),如果属于则放到该行,不属于则判定H框在L_(new-4)行的上一行还是在L_(new-4)行的下一行;如果是在L_(new-4)行的上一行则尝试放到L_(new-5)上,如果是在L_(new-4)行的下一行则尝试放到L_(new-3)上;

[0076] 以此类推,如果判定到L_new的时候,H框仍然处于L_new之下,则创建新行并把H框放到新创建的最新的行中,重复新的H框判定;判定一个H框是否属于某一行,按K5的规则判定执行;

[0077] K5,如果L_new的框个数不足两个或者该行的唯一一个框的宽高比不足5,则按K6处理,否则按K7步处理;

[0078] K6,取H框的往上和往下各15个OCR框的平均斜率作为对齐斜率,过H框的中点用此斜率虚拟出一条直线F,如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值,则H框属于该行,否则属于上一行或者下一行;

[0079] K7,取H框距离该行往上的一行最近的两个框的中点连接线斜率作为参考的直线F斜率,过H框的中点用此斜率虚拟出一条直线F,如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值,则H框属于该行,否则属于上一行或者下一行;

[0080] K8,经过上面K1-K7后,会得到第一版的分行结果,然后在分行结果中找出最优的一行,最优判定条件为,某行中两个邻近框的连接线斜率差最小,同时OCR框斜率差也为最小,则分行结果就为最优行;

[0081] K9,用最优行作为初始参考行,然后取最优行往上的所有OCR框和最优行往下的所有OCR框再次进行K1到K7后得到两份分行结果,然后再拼成一个完整分行结果,至此分行结束。

[0082] 参见图5所示,开始,进行第一遍分行,取第一次分行的最优行,然后基于找到的最优行进行往上和往下再次分行,然后结合往上和往下分行的结果进行合并,最后得到分好行后的结果。

[0083] 在判定某个框是处于某行之上还是某行之下,如图4所示,假设判断H13是否属于L_4,则虚拟出一条虚线,该虚线斜率是通过本专利使用的一种最优参考邻近算法得到的H8和H10的中心连接线斜率,过H13中点,如果此虚线穿过L_4的距离H13最邻近点H11,则H13属于L_4,否则属于其他行,在其他行进行同样的判断,直到找到H13所处的行。

[0084] 实施例1的核心思想就是在判定一个新的OCR框是否属于某一行时,会取某行的上

几行和该OCR框的x坐标最近的线段的斜率作为该OCR框的对齐斜率，同时结合当前框和上几行的距离进行综合判定，提升分行准确率。

[0085] 实施例1一是对扭曲图像仍然具备很高的分行准确率；二是对旋转不超过30度的图像具备很高的分行准确率；三是泛化能力强，只要本身图像中的文本按行排版，都可以用本算法进行分行。本实施例能高精度地对OCR结果进行准确分行，并且存在很高的泛化能力。

[0086] 开始分行处理，对OCR进行y坐标从上到下排序，遍历每个OCR框；参见图6所示。

[0087] 然后判断当前框是否是第一个框；

[0088] 如果当前框是第一个框，则取当前框前后15个框的平均斜率作为斜率，然后用斜率画过参考框中心点的直线；

[0089] 如果当前框不是第一个框，则判断上行是否有两个以上框或上行这个框的长度够长；

[0090] 如果上行有两个以上框或上行一个框的长度够长，则取上一行x坐标最邻近的两个框中点连接线斜率，然后用斜率画过参考框中心点的直线；

[0091] 如果上行只有一个框或这个框的长度不够长，则判断当前行是否有两个以上的框，如果当前行没有两个以上的框，则取当前框前后15个框的平均斜率作为斜率；如果当前行有两个以上的框，则取当前行x坐标最邻近的两个框中点连接线斜率，然后用斜率画过参考框中心点的直线；

[0092] 当用斜率画过参考框中心点的直线后，再判断直线是否经过当前框，如果直线经过当前框，则属于当前行，如果直线没有经过当前框，则判断当前框在直线之下还是在直线之上，如果是在直线之下，则属于下一行，如果是在直线之上则属于上一行；然后对结果进行整合后结束。

[0093] 例如，对于一张医疗票据的实现过程如下：

[0094] S1、在OCR识别返回的结果中，会给出每个文本框的信息，文本框的信息包括位置、大小、角度和具体识别的文本。

[0095] S2、对OCR结果进行排版，也就是分行；排版实现过程如下：

[0096] S2.1、对整体OCR结果进行y坐标从上到下排序。

[0097] S2.2、取OCR结果框相对于该框前后15个框的斜率(每个框旋转角度算出的斜率)平均值作为平均斜率，消除起始框斜率异常。

[0098] S2.3、以平均斜率过第一个框的中点画出一条直线，然后往下一一遍历OCR框，过这条直线的属于同一行，在这条直线之上的属于上一行，在这条直线之下的属于下一行。

[0099] S2.4、在遍历过程中，每次遍历过一个新的OCR框的时候，斜率都需要更新，更新按如下的规则遍历：

[0100] S2.4.1、如果该框是第一个OCR框，或者上一行只有一个框，同时框的宽度过小时，则使用用平均斜率。

[0101] S2.4.2、上一行的框宽度够长或者上一行的框个数大于等于两个，则取当前框相对于上一行最近的两个框的中点连接线的斜率作为该框的搜索斜率。

[0102] S2.4.3、如果上一行没有有效的参考行，但是该行已经存在了两个以上的框，则使用该行的两个框的中点连接线的延长线的斜率作为斜率。

[0103] S5、按照不同斜率不断更新，按照第“S2.3”步中的规则不断遍历搜索，即可得到按行分好的OCR结果。

[0104] S6、在OCR结果预处理完成后，将得到一份按行区分好的OCR结果。

[0105] S7、提取字段关键字列表，提取的内容分为两个类型，规则也对应分为两个类型的规则。

[0106] S7.1、内容类型提取规则，包括三个关键字“姓名”、“性别”“医院”，如果OCR结果中有这样的一行内容“姓名:张三，性别:男，医院:市中医院”，那么使用规则库定义，算法将会使用规则关键字“姓名”“性别”“医院”对OCR行进行搜索和分割成为“姓名:张三”、“性别:男”和“医院:市中医院”，那么再在每个分割后的单元中去掉关键字，将得到“张三”、“男”和“市中医院”这三个内容，把所需要的数据从每行中提取出来。

[0107] S7.2、表格类型成行列分布的数据，这种在表格中分好行以后，建立的规则只包括表头标记和内容行结尾关键词，表头标记包括项目名称、金额、单价和数量这些关键字，在算法从上往下搜索，找到表头标记的内容开始，然后继续往下搜索所有行，直到找到结尾关键字，这些关键字都在数据库中定义好，然后指定表头行开始往下、结尾行标记往上都是有效内容行，然后对这些有效内容行进行分列处理，然后对分好列的内容进行表头对齐和正则判断，即可格式化输出整张表的有效内容。

[0108] 实施例3，实施例3与实施例2不同在于，参见图7所示，在K6或K7中，还包括，如果这条直线F过待判定的行距离H框最近的一个框的距离低于某个阈值时，并且在H框对应的OCR框的左下角顶点处画平行于直线F的直线J，如果直线J与H框所在行的上一行前一个OCR框相交或下一行后一个OCR框相交时，则判定该H框属于该行。

[0109] 如果H框所在行的上一行前一个OCR框的左下角顶点落在直线J上，或者H框所在行的下一行后一个OCR框的左下角顶点落在直线J上时，则判定该H框属于该行，并且该H框处于最优行上。

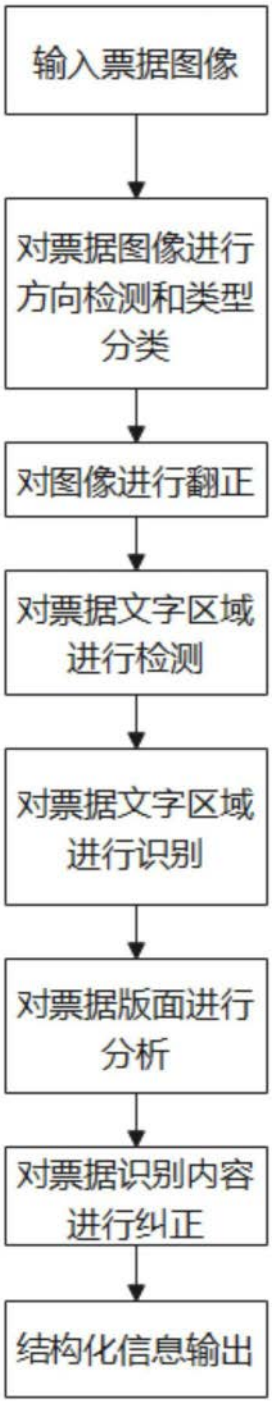
输入票据图像
对票据图像进行方向检测和类型分类
对图像进行翻正
对票据文字区域进行检测
对票据文字区域进行识别
对票据版面进行分析
对票据识别内容进行纠正
结构化信息输出

图1

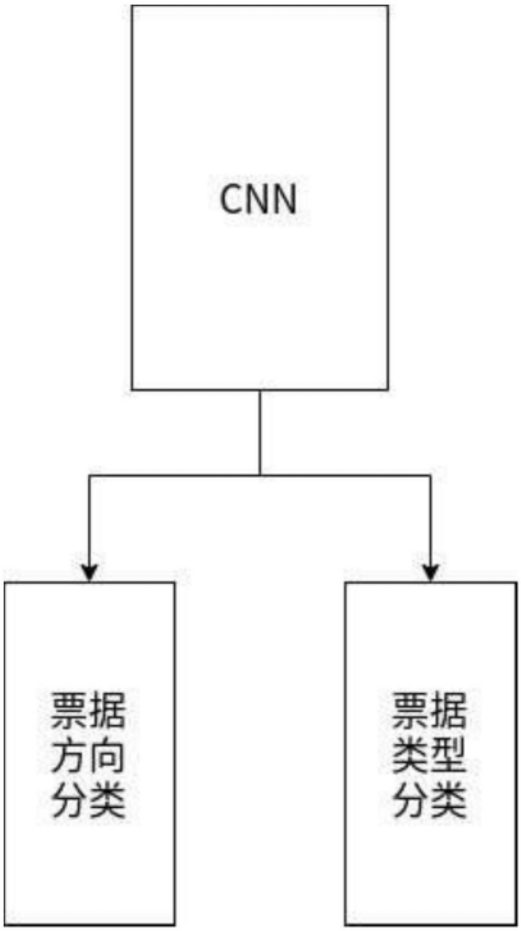

图2

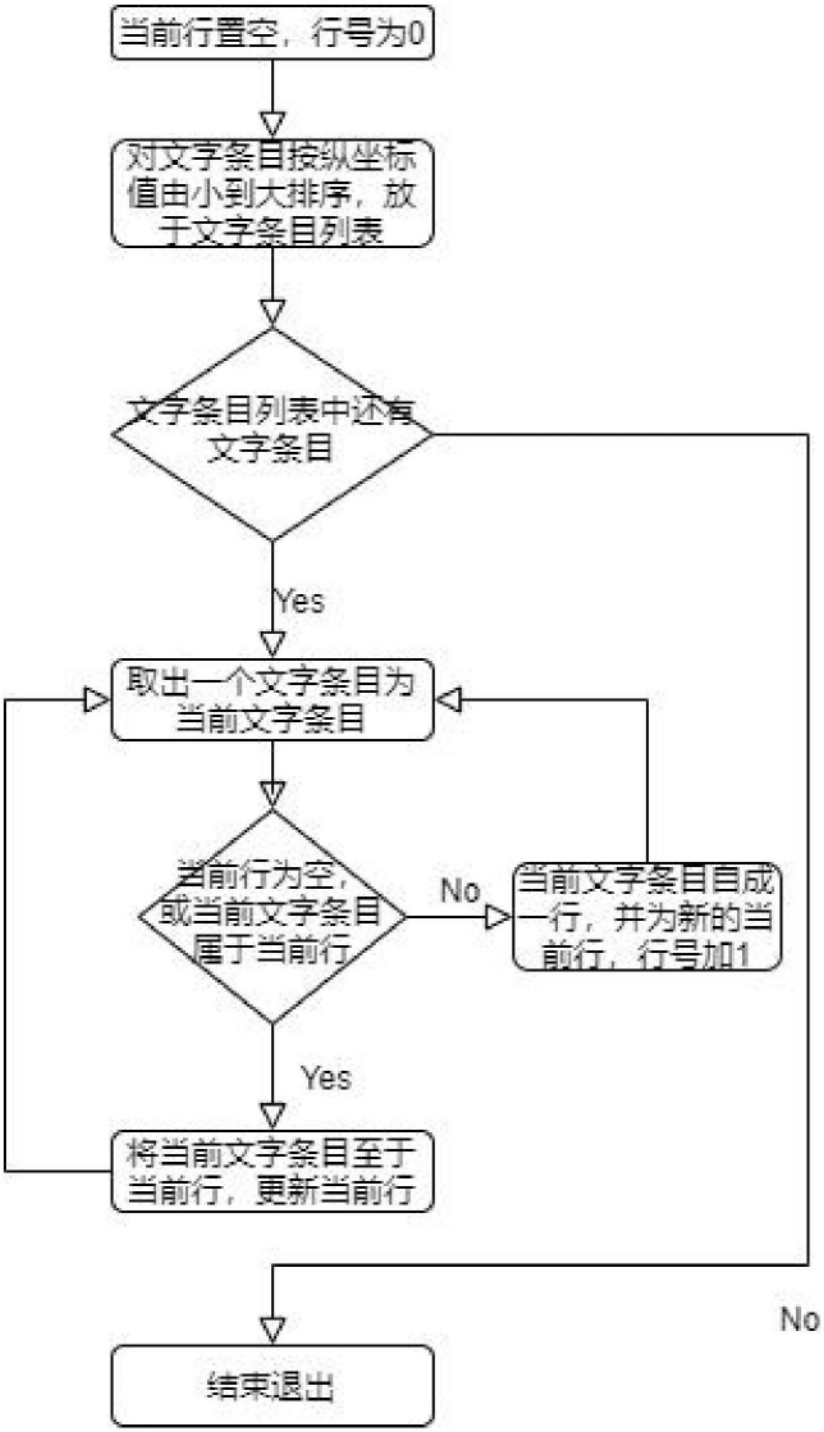
当前行置空，行号为0
对文字条目按纵坐标值由小到大排序，放于文字条目列表
文字条目列表中还有文字条目
Yes
取出一个文字条目为当前文字条目
当前行为空，或当前文字条目属于当前行
No
当前文字条目自成一行，并为新的当前行，行号加1
Yes
将当前文字条目至于当前行，更新当前行
No
结束退出

图3

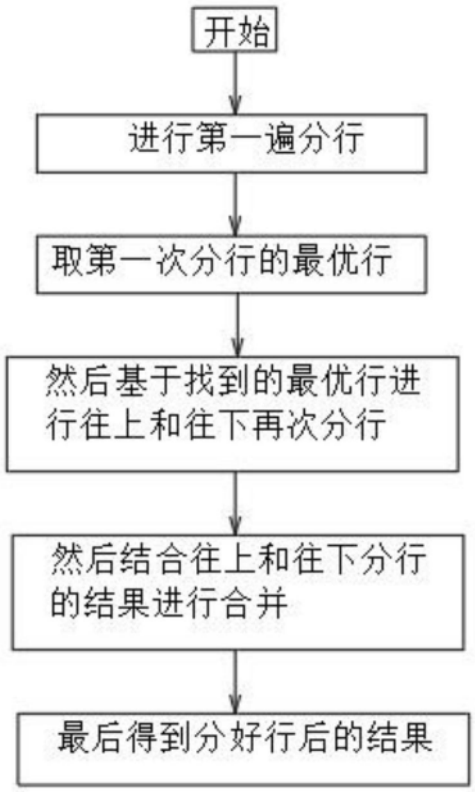
开始
进行第一遍分行
取第一次分行的最优行
然后基于找到的最优行进行往上和往下再次分行
然后结合往上和往下分行的结果进行合并
最后得到分好行后的结果

图4

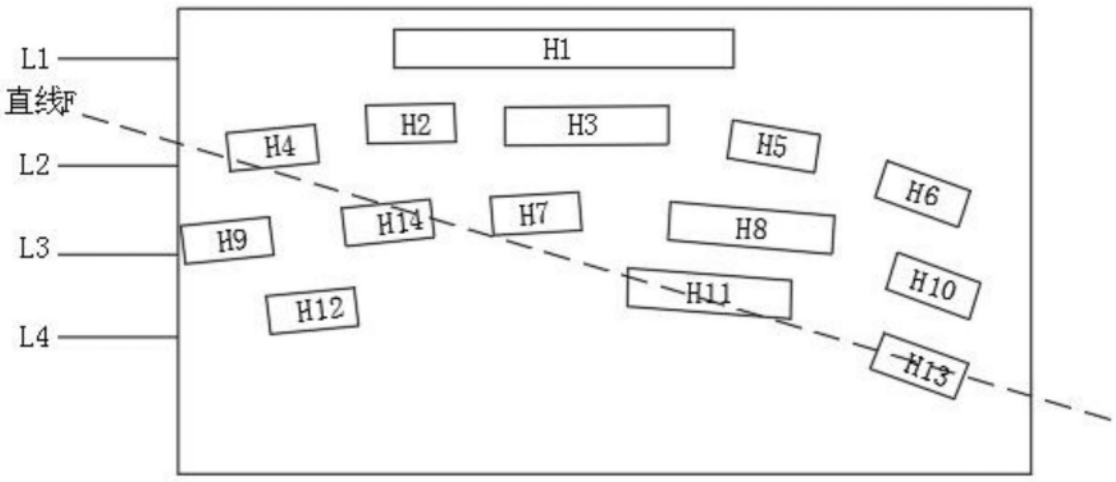
L1
直线F
L2
L3
L4
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12
H13
H14

图5

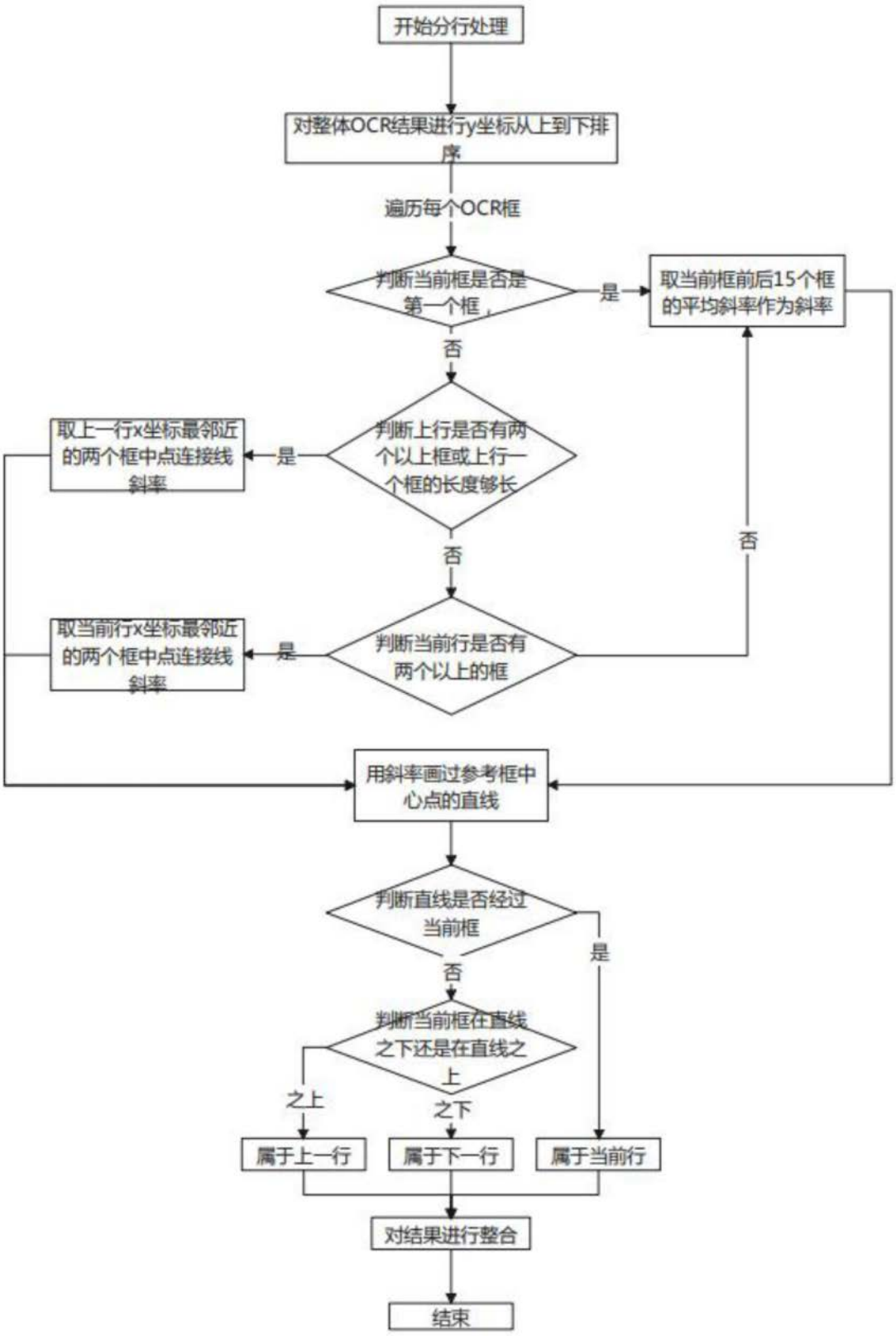
开始分行处理
对整体OCR结果进行y坐标从上到下排序
遍历每个OCR框
判断当前框是否是第一个框，
是
取当前框前后15个框的平均斜率作为斜率
否
判断上行是否有两个以上框或上行一个框的长度够长
是
取上一行x坐标最邻近的两个框中点连接线斜率
否
判断当前行是否有两个以上的框
是
取当前行x坐标最邻近的两个框中点连接线斜率
否
用斜率画过参考框中心点的直线
判断直线是否经过当前框
是
否
判断当前框在直线之下还是在直线之上
之上
之下
属于上一行
属于下一行
属于当前行
对结果进行整合
结束

图6

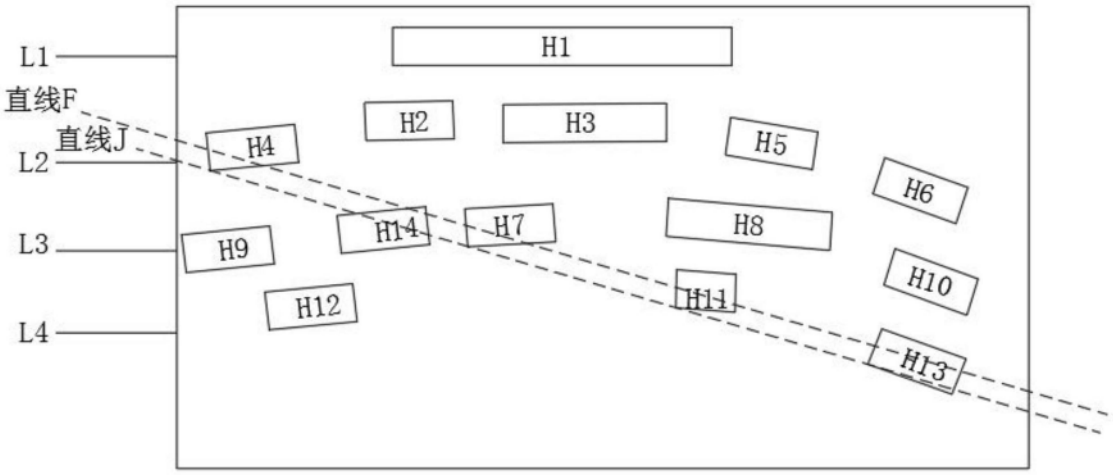

图7