(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109086714 A
(43)申请公布日 2018.12.25

(21)申请号 201810857609.9

(22)申请日 2018.07.31

(71)申请人 国科赛思(北京)科技有限公司
地址 100085 北京市海淀区安宁庄西路9号院29号楼5层507室

(72)发明人 李自豪

(74)专利代理机构 北京市商泰律师事务所 11255
代理人 黄晓军

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/34(2006.01)

权利要求书3页 说明书15页 附图8页

(54)发明名称

表格识别方法、识别系统及计算机装置

(57)摘要

本发明提供了一种表格识别方法和系统，属于表格识别技术领域，利用基于加权RC阈值迭代的非线性对比增强及高斯拉普拉斯LoG算子对符合格式的表格图像进行二值化处理，利用基于透视变化的倾斜校正算法进行倾斜校正；利用图像形态学处理方法提取表格框线，对单元格进行分割，获取最小单元格；建立最小单元格的字符数据库，进行神经网络训练，建立表格识别模型，对表格进行识别。本发明计算简、速度快，可精确识别对比度较弱、图像明暗分布不均且背景模糊的表格图像；倾斜校正速度快、效果好，建立专有的高频字符，训练专有的神经网络，进行模板匹配，提高了识别速度和精度，同时定制化神经网络结构简单，减少了训练和调优的时间和工作量。

CN 109086714 A

1.一种表格识别方法，首先对待识别表格图像的格式进行判别，将不符合格式的待识别表格图像转换为符合格式的表格图像，其特征在于，还包括如下步骤：

步骤S110：利用基于加权RC阈值迭代的非线性对比增强及高斯拉普拉斯LoG算子对所述符合格式的表格图像进行二值化处理，获取二值化表格图像；

步骤S120：利用基于透视变化的倾斜校正算法，对所述二值化表格图像进行倾斜校正；

步骤S130：利用图像形态学处理方法提取出校正后的二值化表格图像的表格框线，对单元格进行分割，提取单元格字符特征；

步骤S140：根据预先建立的字符数据库，针对所述单元格字符特征进行神经网络训练，识别表格。

2.根据权利要求2所述的表格识别方法，其特征在于，所述步骤S110具体包括：

步骤S111：通过非线性对比度增强分离所述符合格式的表格图像的前景字符和背景字符；具体的，定义拉伸率R，

$$R=\frac{y}{x-Min},x\in\left(\mathrm{Min},AVE\right],y\in\left(0,255\right] \qquad (2.1)$$

其中，x为原始像素灰度值，y为原始像素灰度值x经过映射拉伸后的灰度值，Min为原始像素最小灰度值，AVE为平均像素灰度值；

确定“S型”灰度值映射函数$y=\frac{255}{1+e^{-\eta\cdot x^{*}}}$，y∈[0,255]，η＞0，使得灰度值在(Min,T_{bMin}]内满足R＜1以突出前景像素，同时，灰度值在[T_{bmin},AVE]内满足R＞1以抑制背景象素，实现增强图像前景像素和背景像素对比度；其中，x^{*}为经过标准化处理后的原始像素灰度值、η为修正系数，T_{bMin}为背景像素的最小灰度值，T_{bMax}为背景像素的最大灰度值；

其中，利用均值-方差归一化方法对原始像素灰度值x进行标准化，

$$x^{*}=\frac{x-AVE}{\sigma_x}$$

$$AVE=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\sigma_x=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-AVE\right)^2$$

其中，x_i为第i个像素点的灰度值，n为像素点总数目，σ_x为像素灰度值方差。

步骤S112：利用LOG算子模板对所述前景字符及所述背景字符的边缘进行定位，确定前景像素和背景像素；

步骤S113：根据所述前景像素和所述背景像素，进行加权RC迭代阈值选取，获取所述二值化表格图像；具体的，对所述前景像素和所述背景像素的灰度值均值加权平均求阈值T_n，包括：

$$n=0$$

步骤3.1：令：$T_0=w_f\times g_{\min}+w_b\times g_{\max}$，

$$w_f+w_b=1$$

其中，g_{min}和g_{max}分别为所有象素点中的最小灰度值和最大灰度值，w_f和w_b分别为前景像素权值和背景像素权值；

步骤3.2：令：

$T_n = w_f m_f(T_n) + w_b m_b(T_n)$

$$m_f(T_n) = \frac{\sum_{g=0}^{T_n} g \cdot p(g)}{\sum_{g=0}^{T_n} p(g)}, m_b(T_n) = \frac{\sum_{g=T_n+1}^{G} g \cdot p(g)}{\sum_{g=T_n+1}^{G} p(g)},$$

其中，T_n为第n次迭代灰度阈值，g为像素点的灰度值(取值范围为从0到G)，p(g)为灰度值为g的像素点数目，$m_f(T_n)$，$m_b(T_n)$分别为图像第n次迭代后前景像素灰度值均值和背景像素的灰度值均值；

步骤3.3：重复步骤3.2，直到$|T_n - T_{n-1}| < \varepsilon$，算法结束。

3.根据权利要求2所述的表格识别方法，其特征在于，所述步骤S120具体包括：

步骤S121：对所述二值化表格图像进行平滑滤波和形态学处理，确定最小外接矩形，裁剪，获取最小外接矩形图像；

步骤S122：分别计算与所述最小外接矩形图像的四个角点距离最近的像素坐标，作为表格角点坐标；

步骤S123：检验所述表格角点坐标的对应直线的斜率之差是否满足要求，若满足要求，则利用透视变换法对二值化表格图像进行倾斜校正；若不满足要求，则遍历所述二值化表格图像，确定表格角点坐标；

步骤S124：检验所述步骤S123确定的所述表格角点坐标的对应直线的斜率之差是否满足要求，若满足要求，则利用透视变换法对二值化表格图像进行倾斜校正；若不满足要求，则调用表格角点人机交互模块，确定表格角点，再利用透视变换法对二值化表格图像进行倾斜校正。

4.根据权利要求3所述的表格识别方法，其特征在于，所述利用透视变化法对二值化表格图像进行倾斜校正包括：

根据所述表格角点确定两个灭点，对所述两个灭点依次进行透视变换，实现二值化表格图像的倾斜校正。

5.根据权利要求4所述的表格识别方法，其特征在于，所述步骤S130具体包括：

分别选择水平结构元素和竖直结构元素对所述校正后的二值化表格图像进行开运算，获取表格横线图像和表格竖线图像；

对所述表格横线图像和所述表格竖线图像进行与运算，获取表格框架图；

对所述表格框架图进行细化处理，提取表格框线骨架；具体的，由线条边缘开始一层一层向里腐蚀，直到线条剩下一个像素时为止，其中，细化运算由图像击中或击不中变换定义，集合A用结构元素B进行细化的表达式为$A \otimes B = A - (A \Theta B) = A \cap (A \Theta B)^c$；

对提取的表格框线骨架利用最小二乘法运算进行断裂合并，获取完整的表格框线；

根据完整的表格框线对所述校正后的二值化表格图像进行分割处理，得到最小单元格。

6.根据权利要求5所述的表格识别方法，其特征在于，所述步骤S140具体包括：

根据所需识别的表格确定其对应的专有领域，统计所述专有领域的相关高频字符，建立对应的字符数据库；依据所述字符数据库训练神经网络，利用训练好的神经网络进行模板匹配，识别表格字符。

7.一种表格识别系统，包括表格图像格式判别模块，用于将不符合格式的待识别表格图像转换为符合格式的表格图像，其特征在于，还包括：

表格图像二值化模块，用于利用基于加权RC阈值迭代的非线性对比增强及高斯拉普拉斯LoG算子对所述符合格式的表格图像进行二值化处理，获取二值化表格图像；

图像倾斜校正模块，用于利用基于透视变化的倾斜校正算法，对所述二值化表格图像进行倾斜校正；

表格框线提取模块，用于利用图像形态学处理方法提取出校正后的二值化表格图像的表格框线，对单元格进行分割，获取最小单元格；

表格识别模块，用于建立所述最小单元格的字符数据库，根据所述字符数据库进行神经网络训练，建立表格识别模型，对表格进行识别。

8.根据权利要求7所述的表格识别系统，其特征在于，所述表格图像二值化模块包括：

前景和背景分离单元，用于通过非线性对比度增强分离所述符合格式的表格图像的前景字符和背景字符；

像素确定单元，用于利用LOG算子模板对所述前景字符及所述背景字符的边缘进行定位，确定前景像素和背景像素；

加权RC迭代阈值选取单元，用于根据所述前景像素和所述背景像素，进行加权RC迭代阈值选取，获取所述二值化表格图像。

9.根据权利要求8所述的表格识别系统，其特征在于，所述图像倾斜校正模块包括：

表格角点确定单元，用于对所述二值化表格图像进行平滑滤波和形态学处理，确定最小外接矩形，裁剪，获取最小外接矩形图像，通过分别计算与所述最小外接矩形图像的四个角点距离最近的像素坐标，作为表格角点坐标，确定表格角点；或者，

遍历所述二值化表格图像，确定表格角点；

表格角点验证单元，用于检验所述表格角点坐标的对应直线的斜率之差是否满足要求，若满足要求，则判断表格角点为可用，若不满足要求，则判断表格角点为不可用；

倾斜校正单元，用于根据判断为可用的所述表格角点的坐标，利用透视变化法对所述二值化表格图像进行倾斜校正。

10.一种计算机装置，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时，实现如权利要求1至6任一项所述的表格识别方法的步骤。

表格识别方法、识别系统及计算机装置

技术领域

[0001] 本发明涉及表格图像识别处理技术领域，具体涉及一种计算简单，运算快，时间、空间成本低，对对比度较弱、图像明暗分布不均且背景模糊的表格识别精确的表格识别方法、识别系统及计算机装置。

背景技术

[0002] 现有的利用OCR技术进行表格识别时，在对图像进行二值化处理时，主要采用的技术手段一般包括：全局阈值法、局部阈值法、区域增长的方法、水线算法、最小描述长度法、基于马尔科夫随机场的方法等。上述图像二值化处理方法存在种种缺陷。如，全局阈值法仅仅考虑了图像的灰度信息，而忽略了图像中的空间信息，对所有像素采用同一灰度阈值，只适合亮度处处均匀并且图像直方图具有较明显双峰的理想情况，当图像中不存在明显的灰度差异或各物体的灰度值范围有较大重叠时，通常难以获得令人满意的效果；局部阈值法虽能克服全局阈值法中存在的亮度分布不均的缺陷，但却在存在窗口大小设定的问题，即过小的窗口容易导致线条断裂，过大的窗口又容易使图像失去应有的局部细节。

[0003] 对于其它的图像二值化方法如最佳熵值分割二值化法，虽分割精度高，受目标大小影响小，但对噪声敏感。矩不变阈值分割二值化法运算速度较快，可以满足实时性的要求，但其受目标影响较大，目标大小的变化会影响分割的准确性。

[0004] 现有的图像倾斜校正技术手段一般包括基于投影图的方法、基于Hough变换的方法、最近邻簇方法以及矢量化方法等。上述方法存在一些不足，如，

[0005] 投影法需要计算每个倾斜角度的投影形状，要使倾斜估计精度较高的话，这种方法的计算量将非常大，且该方法一般适用于文字文档的倾斜校正，对于具有复杂结构的表格校正，该方法的效果较差；最近邻簇方法对于具有较多相互邻近的组成部分时，该方法十分费时，总体性能不理想；矢量化算法需要直接对光栅图像的各个像素进行处理，存储量大，而且其校正结果的好坏、算法的性能及图像处理的时间、空间成本均极大的依赖于矢量基元的选择；Hough变换计算量大，十分费时，且难以确定直线的起点和终点，只是对纯文本文档比较有效，而对带有图表的复杂结构的文档图像，由于图和表的干扰，无法得到满意的结果，因此在具体的工程实践中的应用却受到了限制。

[0006] 现有的表格识别技术方案在对表格中各个单元格分割并提取出单个字符后，一般是通过调用现有的字符识别工具或通过训练通用的神经网络分类器进行字符识别。该方法对于质量较差的图像或存在噪声的扫描文件其识别准确率一般较差，而且较为费时。此外，若通过训练神经网络来识别中文字符，由于中文字符数目较多且结构复杂，该方案将需要投入大量的人力、物力、财力和时间。

发明内容

[0007] 本发明的目的在于提供一种计算简单，运算快，时间、空间成本低，对对比度较弱、图像明暗分布不均且背景模糊的表格识别精确的表格识别方法、识别系统及计算机装置，

以解决上述背景技术中存在的技术问题。

[0008] 为了实现上述目的，本发明采取了如下技术方案：

[0009] 第一方面，本发明提供了一种表格识别方法，该方法包括如下步骤：

[0010] 首先对待识别表格图像的格式进行判别，将不符合格式的待识别表格图像转换为符合格式的表格图像，还包括如下步骤：

[0011] 步骤S110：利用基于加权RC阈值迭代的非线性对比增强及高斯拉普拉斯LoG算子对所述符合格式的表格图像进行二值化处理，获取二值化表格图像；

[0012] 步骤S120：利用基于透视变化的倾斜校正算法，对所述二值化表格图像进行倾斜校正；

[0013] 步骤S130：利用图像形态学处理方法提取出校正后的二值化表格图像的表格框线，对单元格进行分割，获取最小单元格；

[0014] 步骤S140：建立所述最小单元格的字符数据库，根据所述字符数据库进行神经网络训练，建立表格识别模型，对表格进行识别。

[0015] 进一步的，所述步骤S110具体包括：

[0016] 步骤S111：通过非线性对比度增强分离所述符合格式的表格图像的前景字符和背景字符；具体的，定义拉伸率R，

[0017] $$R=\frac{y}{x-Min},x\in\left(\mathrm{Min},AVE\right],y\in\left(0,255\right] \qquad (2.1)$$

[0018] 其中，x为原始像素灰度值，y为原始像素灰度值x经过映射拉伸后的灰度值，Min为原始像素最小灰度值，AVE为平均像素灰度值；

[0019] 确定“S型”灰度值映射函数$y=\frac{255}{1+e^{-\eta\cdot x^{*}}}$，y∈[0,255]，η＞0，使得灰度值在(Min，T_{bMin}]内满足R＜1以突出前景像素，同时，灰度值在[T_{bmin},AVE]内满足R＞1以抑制背景象素，实现增强图像前景像素和背景像素对比度；其中，x^{*}为经过标准化处理后的原始像素灰度值、η为修正系数，T_{bMin}为背景像素的最小灰度值，T_{bMax}为背景像素的最大灰度值；

[0020] 其中，利用均值-方差归一化方法对原始像素灰度值x进行标准化，

[0021] $$x^{*}=\frac{x-AVE}{\sigma_{x}}$$

[0022] $$AVE=\frac{1}{n}\sum_{i=1}^{n}x_{i}$$

[0023] $$\sigma_{x}=\frac{1}{n}\sum_{i=1}^{n}\left(x_{i}-AVE\right)^{2}$$

[0024] 其中，x_i为第i个像素点的灰度值，n为像素点总数目，σ_x为像素灰度值方差。

[0025] 步骤S112：利用LOG算子模板对所述前景字符及所述背景字符的边缘进行定位，确定前景像素和背景像素；

[0026] 步骤S113：根据所述前景像素和所述背景像素，进行加权RC迭代阈值选取，获取所述二值化表格图像；具体的，对所述前景像素和所述背景像素的灰度值均值加权平均求阈值T_n，包括：

[0027] 步骤3.1：令：

[0028] $n=0$

[0029] $T_0=w_f\times g_{min}+w_b\times g_{max}$，

[0030] $w_f+w_b=1$

[0031] 其中，g_{min}和g_{max}分别为所有象素点中的最小灰度值和最大灰度值，w_f和w_b分别为前景像素权值和背景像素权值；

[0032] 步骤3.2：令：

[0033] $T_n=w_f m_f(T_n)+w_b m_b(T_n)$

[0034] $$m_f(T_n)=\frac{\sum_{g=0}^{T_n} g\cdot p(g)}{\sum_{g=0}^{T_n} p(g)},m_b(T_n)=\frac{\sum_{g=T_n+1}^{G} g\cdot p(g)}{\sum_{g=T_n+1}^{G} p(g)},$$

[0035] 其中，T_n为第n次迭代灰度阈值，g为像素点的灰度值(取值范围为从0到G)，p(g)为灰度值为g的像素点数目，$m_f(T_n)$，$m_b(T_n)$分别为图像第n次迭代后前景像素灰度值均值和背景像素的灰度值均值；

[0036] 步骤3.3：重复步骤3.2，直到$|T_n-T_{n-1}|<\varepsilon$，算法结束。

[0037] 进一步的，所述步骤S120具体包括：

[0038] 步骤S121：对所述二值化表格图像进行平滑滤波和形态学处理，确定最小外接矩形，裁剪，获取最小外接矩形图像；

[0039] 步骤S122：分别计算与所述最小外接矩形图像的四个角点距离最近的像素坐标，作为表格角点坐标；

[0040] 步骤S123：检验所述表格角点坐标的对应直线的斜率之差是否满足要求，若满足要求，则利用透视变换法对二值化表格图像进行倾斜校正；若不满足要求，则遍历所述二值化表格图像，确定表格角点坐标；

[0041] 步骤S124：检验所述步骤S123确定的所述表格角点坐标的对应直线的斜率之差是否满足要求，若满足要求，则利用透视变换法对二值化表格图像进行倾斜校正；若不满足要求，则调用表格角点人机交互模块，确定表格角点，再利用透视变换法对二值化表格图像进行倾斜校正。

[0042] 进一步的，所述利用透视变化法对二值化表格图像进行倾斜校正包括：

[0043] 根据所述表格角点确定两个灭点，对所述两个灭点依次进行透视变换，实现二值化表格图像的倾斜校正。

[0044] 进一步的，所述步骤S130具体包括：

[0045] 分别选择水平结构元素和竖直结构元素对所述校正后的二值化表格图像进行开运算，获取表格横线图像和表格竖线图像；

[0046] 对所述表格横线图像和所述表格竖线图像进行与运算，获取表格框架图；

[0047] 对所述表格框架图进行细化处理，提取表格框线骨架；具体的，由线条边缘开始一层一层向里腐蚀，直到线条剩下一个像素时为止，其中，细化运算由图像击中或击不中变换定义，集合A用结构元素B进行细化的表达式为$A\otimes B=A-(A\Theta B)=A\cap(A\Theta B)^c$；

[0048] 对提取的表格框线骨架利用最小二乘法运算进行断裂合并，获取完整的表格框

线；

[0049] 根据完整的表格框线对对所述校正后的二值化表格图像进行分割处理，得到最小单元格。

[0050] 进一步的，所述步骤S140具体包括：

[0051] 根据所需识别的表格确定其对应的专有领域，统计所述专有领域的相关高频字符，建立对应的字符数据库；依据所述字符数据库训练神经网络，利用训练好的神经网络进行模板匹配，识别表格字符。

[0052] 第二方面，本发明还提供了一种表格识别系统，该系统包括

[0053] 表格图像格式判别模块，用于将不符合格式的待识别表格图像转换为符合格式的表格图像；

[0054] 表格图像二值化模块，用于利用基于加权RC阈值迭代的非线性对比增强及高斯拉普拉斯LoG算子对所述符合格式的表格图像进行二值化处理，获取二值化表格图像；

[0055] 图像倾斜校正模块，用于利用基于透视变化的倾斜校正算法，对所述二值化表格图像进行倾斜校正；

[0056] 表格框线提取模块，用于利用图像形态学处理方法提取出校正后的二值化表格图像的表格框线，对单元格进行分割，获取最小单元格；

[0057] 表格识别模块，用于建立所述最小单元格的字符数据库，根据所述字符数据库进行神经网络训练，建立表格识别模型，对表格进行识别。

[0058] 进一步的，所述表格图像二值化模块包括：

[0059] 前景和背景分离单元，用于通过非线性对比度增强分离所述符合格式的表格图像的前景字符和背景字符；

[0060] 像素确定单元，用于利用LOG算子模板对所述前景字符及所述背景字符的边缘进行定位，确定前景像素和背景像素；

[0061] 加权RC迭代阈值选取单元，用于根据所述前景像素和所述背景像素，进行加权RC迭代阈值选取，获取所述二值化表格图像。

[0062] 进一步的，所述图像倾斜校正模块包括：

[0063] 表格角点确定单元，用于对所述二值化表格图像进行平滑滤波和形态学处理，确定最小包围矩形，裁剪，获取最小包围距图像，通过分别计算与所述最小包围距图像的四个角点距离最近的像素坐标，作为表格角点坐标，确定表格角点；或者，

[0064] 遍历所述二值化表格图像，确定表格角点；

[0065] 表格角点验证单元，用于检验所述表格角点坐标的对应直线的斜率之差是否满足要求，若满足要求，则判断表格角点为可用，若不满足要求，则判断表格角点为不可用；

[0066] 倾斜校正单元，用于根据判断为可用的所述表格角点的坐标，利用透视变化法对所述二值化表格图像进行倾斜校正。

[0067] 第三方面，本发明提供一种计算机装置，该装置包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时，实现如第一方面所述的表格识别方法的步骤。

[0068] 本发明有益效果：计算简单，运算速度较快，尤其在对比度较弱、图像明暗分布不均且背景模糊的情况下较传统方法相比取得了令人满意的效果；可精确选择表格角点，确

定灭点，对灭点依次透视变换，对表格文档进行倾斜校正，校正速度快、效果好，且十分适用于类似表格结构的文档图片的倾斜校正；提出针对特定领域的表格字符识别，建立专有的数据库存储该领域特有的高频字符，并根据字符数据库训练专有的神经网络，在表格字符识别时进行模板匹配，提高了识别速度和精度，同时定制化神经网络与通用字符识别网络相比结构简单，大大减少了训练和调优的时间和工作量。

[0069] 本发明附加的方面和优点将在下面的描述中部分给出，这些将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0070] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0071] 图1为本发明实施例所述的表格识别方法流程图。

[0072] 图2为本发明实施例所述的表格识别方法中图像格式判别转换流程图。

[0073] 图3为本发明实施例所述的表格识别方法中基于加权RC阈值迭代非线性对比增强及LoG算子二值化图像方法流程图。

[0074] 图4为本发明实施例所述的表格识别方法中修正系数取值与灰度映射函数曲线关系示意图。

[0075] 图5为本发明实施例所述的表格识别方法中表格倾斜角度较小时表格角点定位示意图。

[0076] 图6为本发明实施例所述的表格识别方法中表格倾斜角度较大时表格角点定位示意图。

[0077] 图7为本发明实施例所述的表格识别方法中表格角点确定流程示意图。

[0078] 图8为本发明实施例所述的表格识别方法中两个灭点透视示意图。

[0079] 图9为本发明实施例所述的表格识别方法中单灭点透视校正示意图。

[0080] 图10为本发明实施例所述的表格识别方法中表格框线提取流程示意图。

[0081] 图11为本发明实施例所述的神经网络训练识别表格的方法流程图。

具体实施方式

[0082] 下面详细描述本发明的实施方式，所述实施方式的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的模块。下面通过参考附图描述的实施方式是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。

[0083] 为便于对本发明实施例的理解，下面将结合附图以具体实施例为例做进一步的解释说明，且实施例并不构成对本发明实施例的限定。

[0084] 本领域普通技术人员应当理解的是，附图只是一个实施例的示意图，附图中的部件或装置并不一定是实施本发明所必须的。

[0085] 实施例一

[0086] 如图1所示，本发明实施例一提供的一种表格识别方法，首先对输入的文档格式进

行判别，若为PDF文件则通过格式转换模块将PDF文件转换为JPG格式的图片，并保存。然后利用基于加权RC阈值迭代的非线性对比增强及LoG算子二值化方法，将RGB图片转化为二值图并保存。再利用基于透视变化的倾斜校正算法，根据选取的四个透视角点，对图像进行倾斜校正。同时，利用图像形态学处理的方法提取出表格的框线，对个单元格进行分割。最后结合该表格应用领域的特点，建立专有的字符数据库，并训练定制化的神经网络对字符进行识别。

[0087] 在实际应用中，由于其应用场景或需求等方面的区别，待识别的文档部分为扫描的PDF格式文件，而部分为JPG等格式的照片。而现有的表格识别技术方案中并未对输入文档的格式进行区分，一般只是针对图片进行处理。

[0088] 如图2所示，在本发明实施例中，利用PDF转PNG格式模块，将输入文件中的扫描文件格式进行批量处理，统一为照片格式，便于后续的处理和识别，格式转换。

[0089] 该模块能批量将PDF文档格式转换为PNG照片格式。在实际操作中我们只需要输入待识别文件的保存文件夹地址，该模块将自主判别文件夹中PDF格式的文档，并将多页PDF文档拆分为单页进行格式转换。将转换完成后的图片按顺序保存，以便后续的处理。

[0090] 如图3所示，本发明实施例一提供一种二值化处理方式，通过非线性对比度拉伸改善图像前景、背景象素在直方图中的分布，尽量使直方图出现双峰或近似双峰的特性，从而突出前景字符；然后用高斯型拉普拉斯算子模板对图像中的字符边缘进行定位，并利用LoG，找出字符的内部象素。最后，我们利用加权RC迭代阈值选取对图像进行二值化处理。

[0091] 图像二值化是图像处理的关键步骤，二值化质量的好坏将直接影响图像处理和后续步骤所取得的效果。与线性对比度增强的方法相比，非线性对比度增强克服线性灰度拉伸不充分的缺点，进一步突出前景象素并抑制背景象素。

[0092] 在本发明一个具体实施例一中，为衡量图像灰度拉伸的程度，这里定义拉伸率R，如式(2.1)所示。

[0093] $$R=\frac{y}{x-Min},x\in\left(\mathrm{Min},AVE\right],y\in\left(0,255\right] \qquad (2.1)$$

[0094] 上式中，x为原始像素灰度值；

[0095] y为原始像素灰度值x经过映射拉伸后的灰度值；

[0096] Min为原始像素最小灰度值；

[0097] AVE为平均像素灰度值。

[0098] 显然，有当R＝1即灰度值映射函数为y＝x时，映射后的图像等于原图像，我们称这样的映射为恒等映射；当R＞1时，灰度值映射的权重偏向较高(较亮)的灰度值；反之，当R＜1时，灰度值映射的权重偏向较低(较暗)的灰度值。这里记T_{bMin}，T_{bMax}分别为背景像素的最小灰度值和最大灰度值。

[0099] 在本发明具体实施例中，通过灰度拉伸，应尽可能的减少像素灰度值落在[T_{bmin}，AVE]区间内，即通过某一灰度映射函数使得灰度值在(Min，T_{bMin}]内，其拉伸率满足R＜1以突出前景象素，同时灰度值在[T_{bmin}，AVE]内拉伸率又满足R＞1以抑制背景象素，从而真正达到增强图像前景和背景象素对比度的目的。

[0100] 由此，在本发明具体实施例中，设计“S型”映射函数。如式(2.2)所示。

[0101] $$y=\frac{255}{1+e^{-\eta\cdot x^{*}}},\ y\in[0,255],\ \eta>0 \qquad (2.2)$$

[0102] 上式中，x^{*}为经过标准化处理后的原始像素灰度值，η为修正系数，其取值与原始图像的像素分布及平均像素灰度值有关。

[0103] 我们利用均值-方差归一化方法对原始像素灰度值x进行标准化，如下式所示。

[0104] $$x^{*}=\frac{x-AVE}{\sigma_x}$$

[0105] $$AVE=\frac{1}{n}\sum_{i=1}^{n}x_i$$

[0106] $$\sigma_x=\frac{1}{n}\sum_{i=1}^{n}(x_i-AVE)^2 \qquad (2.3)$$

[0107] 上式中，x_i为第i个像素点的灰度值，n为像素点总数目，σ_x为像素灰度值方差。

[0108] 当修正系数η取不同值时，其非线性灰度映射函数曲线示意图如图4所示。

[0109] 如图4所示，图中序号①-⑭所示的曲线，分别为当η分别取值为0.01、0.1、0.5、0.2、0.25、0.3、0.5、0.7、0.9、3、5、7、9、100时的非线性映射曲线，序号⑮所示的曲线为线性变换恒等映射曲线。由图4可知，当$x\in(Min,T_{bMin}]$时，非线性映射曲线位于线性变换恒等映射曲线的下方，此时显然有$R<1$；当$x\in[T_{bmin},AVE]$时，非线性映射曲线位于恒等映射直线的上方，此时拉伸率$R>1$。我们根据文档图片的实际灰度均值和分布情况调整修正系数η的取值。由图4可知，出当η取值较小时其非线性灰度映射函数退化为线性映射函数，且灰度值的分布被压缩。当η取值在0.15附近时，其非线性映射函数曲线与恒等映射曲线大致重合，此时能达到恒等变换的效果。当η取值较大时，非线性映射函数灰度分布区域两极化，此时对比度区将十分明显。

[0110] 通过选择合理的η值能使非线性灰度映射函数曲线呈“S型”，与线性对比度增强相比，非线性对比度增强使文档图像的前景象素灰度更暗，同时使背景象素灰度更亮，从而更能有效增大图像对比度。

[0111] 在本发明的具体实施例中，将拉普拉斯算子引入至图像二值化过程中。具体的，一个

[0112] 二元函数f(x,y)的拉普拉斯变换定义为：

[0113] $$\nabla^2 f=\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2} \qquad (2.4)$$

[0114] 在数字图像的滤波处理中，我们将其离散化。分别在x方向和y方向定义其二阶偏微分，如下所示：

[0115] $$\frac{\partial^2 f}{\partial x^2}=f(x+1,y)+f(x-1,y)-2f(x,y) \qquad (2.5)$$

[0116] $$\frac{\partial^2 f}{\partial y^2}=f(x,y+1)+f(x,y-1)-2f(x,y) \qquad (2.6)$$

[0117] 故：

[0118] $$\nabla^2 f = f(x+1,y)+f(x-1,y)+f(x,y+1)+f(x,y-1)-4f(x,y) \quad (2.7)$$

[0119] 根据(2.7)式对图像进行平滑处理,我们可以通过滤波模板$\begin{bmatrix}0 & 1 & 0\\ 1 & -4 & 1\\ 0 & 1 & 0\end{bmatrix}$来实现。

[0120] 此外,也可将式(2.5)或(2.6)离散拉普拉斯变换引入至滤波模板中,即两个对角线方向各增加一个新添项,其滤波模板为$\begin{bmatrix}1 & 1 & 1\\ 1 & -8 & 1\\ 1 & 1 & 1\end{bmatrix}$。

[0121] 在本发明的一个具体实施例中,在应用拉普拉斯算子之前先用高斯函数对图像进行模糊处理以减少噪声的影响,该二维高斯函数为:

[0122] $$h(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2.8)$$

[0123] 利用式(2.8)与图像做卷积运算以模糊图像,图像模糊的程度由σ值决定。

[0124] 由式(2.7)知二阶导数为线性运算,故我们可以交换二阶偏导与卷积运算的次序,即先根据高斯型滑函数模糊图像然后应用拉普拉斯算子,与先应用拉普拉斯算子再利用高斯函数模糊图像所取得的效果相同。故高斯函数h的二阶导数如式(2.9)。

[0125] $$\nabla^2 h(x,y) = -\left[\frac{x^2+y^2-\sigma^2}{\sigma^4}\right] e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2.9)$$

[0126] (2.9)式即为高斯型拉普拉斯算子(Laplacian of a Gaussian,LoG),其图像的形状形如一个墨西哥草帽。将式(2.9)离散处理,所得滤波模板为$\begin{bmatrix}0 & 0 & -1 & 0 & 0\\ 0 & -1 & -2 & -1 & 0\\ -1 & -2 & 16 & -2 & -1\\ 0 & -1 & -2 & -1 & 0\\ 0 & 0 & -1 & 0 & 0\end{bmatrix}$。

[0127] 在实际应用中,该滤波模板并不唯一,与$\nabla^2 h$图像形状相似的滤波模板,即满足滤波模板矩阵中心元素为正值,其周围相邻元素均为负值,外围元素取值为0,且模板的系数总和为零,以满足在灰度级不变的区域中模板响应为零的模板矩阵均可认为是LoG滤波模板或其变式。

[0128] LoG算子在拉普拉斯算子的基础上引入高斯滤波,该方法降低了且抵消了由拉普拉斯算子中的二阶导数引起的噪声的影响,在应用中为了能使处理所得结果达到预期要求,即背景灰度值较亮而前景灰度值较暗,我们一般将LoG算子的运算分开进行:即先进行高斯滤波,然后用复合拉普拉斯算子模板$\begin{bmatrix}-1 & -1 & -1\\ -1 & 9 & -1\\ -1 & -1 & -1\end{bmatrix}$进行滤波处理。

[0129] 复合拉普拉斯算子模板实际上是模板$\begin{bmatrix}1 & 1 & 1\\ 1 & -8 & 1\\ 1 & 1 & 1\end{bmatrix}$取反后与原图相加而来，即$g(x,y)=f(x,y)+\nabla^2 f(x,y)$，这相当于模板中心系数比原来增加了单位1。实际上该过程也是图像增强的过程。通过上述处理使图像整体显的更加锐利，背景与前景象素的对比度也进一步拉大。

[0130] 在本发明的一个具体实施例中，为增加图像的对比度，新阈值T_n通过前景和背景两类象素各自的灰度值均值加权平均来求出。算法步骤如下所示：

[0131] Step1.初始化，令n＝0，$T_0=w_f g_{min}+w_b g_{max}$，($w_f+w_b=1$)。其中，$g_{min}$和$g_{max}$分别为图像所有象素点中的最小和最大灰度值，$w_f$和$w_b$分别为图像前景和后景像素权值。

[0132] Step2.令：

[0133] $T_n=w_f m_f(T_n)+w_b m_b(T_n)$

[0134] $$m_f(T_n)=\frac{\sum_{g=0}^{T_n} g\cdot p(g)}{\sum_{g=0}^{T_n} p(g)},m_b(T_n)=\frac{\sum_{g=T_n+1}^{G} g\cdot p(g)}{\sum_{g=T_n+1}^{G} p(g)} \qquad (2.10)$$

[0135] 上式中，T_n为第n次迭代灰度阈值；

[0136] g为像素点的灰度值(取值范围为从0到G)；

[0137] p(g)为灰度值为g的像素点数目；

[0138] $m_f(T_n)$，$m_b(T_n)$分别为图像第n次迭代后前景和背景象素的灰度值均值。

[0139] Step3.重复Step2，直到$|T_n-T_{n-1}|<\varepsilon$，ε取值一般较小。

[0140] 由于(2.10)式结果可能为非整数，每次迭代后我们把右边的结果四舍五入为整数处理(事实上灰度图像的阈值只能为整数)，当迭代结果满足$|T_n-T_{n-1}|<\varepsilon$时，我们认为算法达到收敛。

[0141] 经过基于加权RC阈值迭代非线性对比增强及LoG算子二值化处理方法，我们即可获得较为满意的结果。

[0142] 在本发明的一个具体实施例一中，利用透视变化对表格进行倾斜校正。该方法的实际校正结果的质量主要取决于灭点即表格四个角点选取的好坏，对此本专利通过距离计算与边缘扫描相结合的方法，精确确定表格角点坐标。

[0143] 方法一：距离计算方法。

[0144] 如图5所示，当表格倾斜角度较小时其表格角点A'，B'，C'，D'与图像角点A，B，C，D的距离较近，此时我们可以通过距离计算寻找出表格的角点。具体方法如下：

[0145] Step1.对图像进行平滑滤波、形态学处理，寻找包围二值图像的最小矩形框。根据矩形框对图像进行剪裁。

[0146] Step2.获取图片的尺寸大小H×W，以图片的左上角点为原点建立笛卡尔直角坐标系，则四个角点A，B，C，D的坐标分别为(0，0)，(W，0)，(W，H)，(0，H)；

[0147] Step3.分别寻找与图像角点A，B，C，D距离最近的四个角点，并将其作为表格角点

A',B',C',D',其计算公式如下：

[0148] $$\min\ d_i=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2} \quad (2.11)$$

[0149] 上式中，x_i,y_i分别为图像角点A,B,C,D的横、纵坐标，x_j,y_j分别为其它像素点的横、纵坐标；

[0150] 通过计算图像中其它像素点与图像角点A,B,C,D的距离d_i，并将距离取得最小值时的像素点作为其对应的表格角点。

[0151] 方法二：边缘扫描。

[0152] 如图6所示，当表格倾斜程度较大时，利用上述方法求取的表格角点将存在较大误差，此时本专利采用如下方式确定表格角点A',B',C',D'的坐标。

[0153] 具体的，对二值图进行遍历，分别寻找出x,y坐标取最大、最小值时的像素点，将其作为表格的角点A',B',C',D'。但是在实际的应用过程中，该方法受噪声等影响较大，因此我们首先需要对图像进行平滑滤波、形态学等处理，以消除噪声，同时寻找包围二值图像的最小矩形框，对图像进行剪裁。

[0154] 在表格角点的确定过程，最初我们一般并不知道表格的具体倾斜角度，而且表格倾斜程度的大小具有一定的模糊和不确定性，难以通过数学模型将其量化。对此，我们首先尝试使用方法一，即通过距离计算寻找表格角点。在得到表格角点的坐标后对其进行检验。方法如下：

[0155] $$\left|k_1-k_2\right|=\begin{cases}\left|\dfrac{y_A-y_B}{x_A-x_B}-\dfrac{y_D-y_C}{x_D-x_C}\right|<\varsigma, & \text{角点坐标选取合理}\\ \left|\dfrac{y_A-y_B}{x_A-x_B}-\dfrac{y_D-y_C}{x_D-x_C}\right|\geq\varsigma, & \text{角点坐标选取不合理}\end{cases} \quad (2.12)$$

[0156] 从式(2.12)中可以看出，当上、下角点所确定的直线斜率差别较小时，由方法一所得的角点为表格角点的可信度较高，此时我们可以将其作为表格角点A',B',C',D'。否则根据方法二，即遍历二值图，寻找x,y坐标的最小值确定表格角点。对方法二确定的表格角点坐标进行同样的检验，若通过检验则将其作为表格角点坐标A',B',C',D'，否则调用人机交互角点确定模块，手工选取表格角点坐标。其中ς值的大小与图片的具体尺寸有关，当图像的横纵比W/H较大且大于1时，ς值可适当的减小，当图像的横纵比W/H较小较且小于1时，ς值可适当的增加。

[0157] 综合上述两种方法，表格角点确定过程的流程图如图7所示。

[0158] 图形的两点透视图如图8所示，其中m1,m2为两灭点，A,B,C,D为图像的四个控制点A',B',C',D'。本专利首先根据多种方法确定表格角点A',B',C',D'，并根据所得角点确定灭点，对两个灭点依次进行透视变换，通过两次单灭点透视完成对图像的校正过程。

[0159] 如图9所示，将图像旋转至ab边与x轴平行(即以ab边为x轴方向建立直角坐标系)，同时确定单灭点e(mx,my)坐标，对图像acdb进行x方向的校正。校正后的图形为a'c'd'b'。此时ac边和bd边经投影后被校正为平行于y轴的竖直线。然后，我们再对y方向进行单点透视校正，即可获得二次校正的正方形。

[0160] 根据图像中a,b,c,d四点坐标求得单灭点e坐标(mx,my)，然后根据透视缩小效应，

x方向对其进行反运算，实现x方向的校正。在x方向校正的具体操作中，在根据原始图像尺寸，在图像高度范围内，任选一水平直线(该直线高度需大于被校正图像的最大高度，即该水平直线不能与被校正图形相交)作为标准直线，这里我们选择纵坐标为H的直线作为标准线。该直线ea，eb分别交于点q_1，q_2，此时我们将ac边和bd边在x方向分别投影至a'q_1和b'q_2上。这里我们以ac边上任一点p为例说明透视变换过程。

[0161] 由图9可知，Δpqp'～Δpep"，因此有

[0162] $\frac{\Delta x}{pp"}=\frac{q_1p'}{ep"}$，$\Delta x=\frac{q_1p'.pp"}{ep"}=\frac{(H-i)\bullet(mx-j)}{my-i}$，因此校正后的点p'的x坐标为$x_{p'}=x_p+\Delta x=j+\frac{(H-i)\bullet(mx-j)}{my-i}$，y坐标保持不变，因此经过单灭点透视变换后p"的坐标为$\left(j+\frac{(H-i)\bullet(mx-j)}{my-i},\ i\right)$，其x方向的校正公式如下所示。

[0163] $$\begin{cases}x=j+\dfrac{(H-i)\bullet(mx-j)}{my-i}\\ y=i\end{cases}\qquad(2.14)$$

[0164] 根据该方法依次对线段ac和bd上的每一点进行透视变换，最终得到校正后的线段为a'c'，b'd'。由透视原理知，图像在y方向的缩放比例和x放向的缩放比例相同，即$\frac{u}{X}=\frac{v}{Y}$，因此我们可以根据x方向变换时得到的比例关系对y方向进行同样的比例校正，校正公式如下所示。

[0165] $$\begin{cases}x=j_0\\ y=\dfrac{i}{\dfrac{mx}{mx-\dfrac{(H-1)\bullet mx}{my-i}}}\end{cases}\qquad(2.15)$$

[0166] 通过两次透视校正，即可将倾斜图片变换为正视图即矩形acbd。

[0167] 如上所述，利用透视法对图像进行倾斜矫正，需要依赖四个控制点a，b，c，d，该控制点即为我们在表格矫正中的四个角点A'，B'，C'，D'。通过四个角点确定双灭点，本专利将双灭点透视变换转换为两次单灭点的透视变换，完成对表格图片的倾斜校正。

[0168] 如图10所示，本发明一个具体实施例一中，采用数学形态学处理图像关键在于结构元素SE的选择，所以如果要提取水平线则选择的结构元素SE为水平结构，如果要提取竖直线则选择的结构元素SE为垂直结构元素。

[0169] 数学形态学检测直线和特征提取的算法如下：

[0170] Step1.求出水平直线。对图像作开运算，得到图像F_1。该图像保留了水平表格线上的几乎全部象素，而垂直表格线和文字图像以及大部分噪声点都被去除。

[0171] Step2.求出垂直直线。对图像作开运算，得到图像F_2。

[0172] Step3.对求得的表格横线图像F_1和表格竖线图像F_2作与运算，求出水平直线与垂直直线的所有交点。

[0173] 运用数学形态学进行表格直线提取的关键在于结构元素SE的选取。结构元素SE的选择，一般应大于文字直线的长度，小于表格的行高。这样动态选取表格线，一方面可避免由于结构元素过小，将文字的横竖线当作表格线提取，另一方面可防止由于结构元素过大，漏掉部分表格线。同时，在用数学形态学对横线和竖线进行提取时，适当选取结构元素SE，能滤去由文字或噪声干扰等产生的伪直线，因此，不必在表格识别之前作去噪与去除文字的预处理。

[0174] 对图像进行细化处理，提取源图像的骨架，即将原图像中线条宽度大于1个像素的线条细化成只有一个像素宽，形成“骨架”。细化的过程即从线条边缘开始一层一层向里腐蚀，直到线条剩下一个像素时为止。细化运算可由图像击中或击不中变换定义，集合A用结构元素B进行细化的表达式如下所示：

[0175] $$A \otimes B = A - (A \Theta B) = A \cap (A \Theta B)^c \quad (2.16)$$

[0176] 经细化处理后得到表格框线的骨架，此时利用最小二乘法合并断裂线段。记线段的各个像素点的坐标为(x_i,y_i)，拟合直线表达式为$y=ax+b$，则由最小二乘法有：

[0177] $$\frac{\partial F}{\partial a} = -2\sum_{i=1}^{n} x_i (y_i - a x_i - b) = 0$$

[0178] $$\frac{\partial F}{\partial b} = -2\sum_{i=1}^{n} (y_i - a x_i - b) = 0 \quad (2.17)$$

[0179] 联立方程解得：

[0180] $$\begin{cases} a = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ b = \left(\dfrac{1}{n}\right)\sum_{i=1}^{n} y_i - \left(\dfrac{a}{n}\right)\sum_{i=1}^{n} x_i \end{cases} \quad (2.18)$$

[0181] 经过上述过程我们即可提取出表格框线。

[0182] 实施例二

[0183] 如图11所示，本发明实施例二提供的采用训练专有神经网络对字符进行识别的方法。具体来说，

[0184] 首先，统计专有领域中待识别表格所包含的高频字符及字符串，并收集字符图片，将其作为神经网络的样本集。然后，对图片进行二值化处理，同时分割出每个字符，并对其进行标准化处理，统一图片格式及大小。接下来，对预处理后的图片进行特征提取，提取出字符结构点特征、字符投影特征等。最后，根据十折交叉训练，训练网络，并利用调优网络识别字符，根据识别结果计算其与字符串数据库中各个字符串的编辑距离，同时比较最小编辑距离与可信度阈值间的大小关系，若编辑距离小于阈值则将最小编辑距离字符串作为识别结果输出，否则直接将识别结果输出。

[0185] (1)字符串分割

[0186] 字符串的分割方法为:对字符串二值图从上至下,从左至右进行扫描,若满足式$Y=\sum_{j=0}^{n}f(j,m_1)>2$,则$m_1$为字符的左边界。其中j为像素点的纵坐标,m为像素点的横坐标,f(j,m_1)为像素点的取值(0或1),n为图片高度。同理,当$Y=\sum_{j=0}^{n}f(j,m_2)\le 2 \&Y=\sum_{j=0}^{n}f(j,m_2+1)=0$,时$m_2$为字符的右边界。利用此方法分割出每个字符。

[0187] (2)图片标准化处理

[0188] 对分割出的单个字符图片进行标注化处理,将其大小统一为32×64点阵。

[0189] (3)特征提取

[0190] 为提高字符的识别精度,本专利根据组合特征描述子提取字符特征进行字符识别,分别为:投影特征及网格特征。

[0191] 字符投影特征提取。字符各个方向的投影可以反映出字符的特点。而且不同方向的投影反映出的特点不同,如横向(纵向)投影突出反映了字符中横笔竖笔的特征。据统计由于横、竖笔为字符的主要构成结构,故本专利只对字符横纵向进行投影,不考虑135°和45°方向,最终将投影结果保存为特征向量。

[0192] 字符网络特征提取。字符的网格特征是指将字符点阵平均分成m×m份,求出每份网格中黑点数所占整个文字黑点数的百分比,这样组成的m×m维矩阵即为该字符的网格特征,本专利将其转换为一位向量表示。网格特征体现了字符整体形状的分布。具体过程如下:

[0193] Step1.将字符点阵分成8×8份;

[0194] Step2.求出每份中的黑点数,用$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{18} \\ p_{21} & p_{22} & \cdots & p_{28} \\ \vdots & \vdots & \vdots & \vdots \\ p_{81} & p_{82} & \cdots & p_{88} \end{bmatrix}$表示;

[0195] Step3.求出文字的总黑点数,p=$p_{11}+p_{12}+\ldots+p_{18}+p_{21}+\ldots+p_{88}$;

[0196] Step4.求出每份中黑点数所占整个文字黑点数的百分比$p_{ij}=p_{ij}*100/p$,则特征向量p=(p_{11},p_{12},..,p_{18},p_{21},...,p_{88})即为字符的网格特征。

[0197] (4)网络结构设计

[0198] 本发明实施例二设计BP神经网络对字符进行识别。由于提取的是提一种混合字符特征,它结合了字符网格特征和投影特征。对于标准化为32×64点阵大小的字符,我们用8×8的网格对它进行划分,得到的网格特征向量为64维,此外投影特征向量为32+64=96维,故一共为160维混合特征向量,因此需要160个输入层神经元。此外,由于本实施例二为特定领域的字符识别,其统计的高频字符数目与通用字符相比数目较少,一般在300~500个间,故网络的输出神经元数目也在300~500间。网络的激活函数选择Logistic函数($f(x)=\frac{1}{1+e^{-x}}$)。一般在没有特定的理由要求使用多个隐藏层时,仅仅使用一个隐藏层进行处理是最简单的。利用Nelson和Illingwnrth的结论:$n_1=\sqrt{n+m}+a$,其中,n_1为隐含

层节点数，m为输出节点数目，n为输入节点数目，a为1～10之间的常数，根据输入、输出神经元数目估计隐层神经元数目为22～36个。

[0199] (5)网络参数设计

[0200] 网络连接权值初始值设置。考虑到发明实施例二采用的是Logistic函数作为节点的激活函数，其输出范围在0和1之间，对于一个有d个输入神经元的网络，输入层到隐含层的权值的初始值应该在$\left(0,1/\sqrt{d}\right)$的范围内。对于从隐含层到输出层的权值，若隐藏层神经元数目为n_h，则隐藏层到输出层的权值初始值应该在$\left(0,1/\sqrt{n_h}\right)$范围内。在初始化输入层到隐含层的权值时采用的是在$\left(0,1/\sqrt{160}\right)$的范围内随机选取数据的方法，而在初始化隐藏层到输出层的权值时这个范围被设定为$\left(0,1/\sqrt{36}\right)$。

[0201] 网络学习参数设置。针对不同的应用，我们不可能事先知道最优学习率η_{out}，也就不能初始化最优学习率参数η_{out}。因此，本发明实施例二设置α＝0.9和η的初值0.1，然后利用公式η(n)＝η(n-1)(1-c/p)(其中，c为一常数2，p为学习周期，即样本总数)的方法对学习参数进行设置。

[0202] (6)网络的训练

[0203] 训练样本、测试样本的选择。本发明实施例二采用十折交叉训练确定训练样本和测试样本。

[0204] 网络损失函数。网络损失函数选择欧几里德范数：$Loss=\frac{1}{2N}\sum_{j=1}^{N}\left\|d_j-F\left(x_j\right)\right\|^2$，以使经验风险最小化。其中，||•||是所含向量的欧几里德范数，N为输入网络训练的样本数目，d_j为实际值，$F(x_j)$为识别结果。

[0205] 训练停止条件。当误差小于我们设定的可接受的值时，或者网络已经达到最大迭代次数时，系统就结束网络的训练。当系统是因为第一个条件，即误差足够小时结束的网络训练，那么它将认为网络已经达到收敛，可以使用；而当系统是因为网络学习的迭代次数达到上限而结束的网络学习，它将会提示“网络无法收敛，请修改网络参数并重新训练”。

[0206] 网络的训练过程。本专利中BP网络的训练过程设计如下：

[0207] Step1，设置变量和参数，其中包括训练样本，权值矩阵和学习参数。

[0208] Step2，初始化，给各个权值矩阵一个较小的随机非零向量。

[0209] Step3，输入训练样本。

[0210] Step4，对输入样本，前向计算BP网络每层神经元的输入信号和输出信号。

[0211] Step5，由实际输出和期望输出求得误差。判断是否训练完所有样本，如果是转到第6步，否则转到第3步。

[0212] Step6，计算全局误差并判断是否满足要求，如果是转到第9步，否则转到第7步。

[0213] Step7，判断是否已经到了最大迭代次数，若到，转到第9步，否则反向计算每层神经元的局部梯度。

[0214] Step8，根据局部梯度修正各个矩阵的权值，更新学习参数，并转到第3步。

[0215] Step9，判断是否达到最大迭代次数，是则提示网络无法收敛，否则提示网络已经

收敛，训练结束。

[0216] (7)字符串的识别

[0217] 对需要识别的字符串，首先利用图像处理功能，分割字符串。根据已调优的网络对字符进行识别，得到字符串的识别结果R，根据字符串的识别结果计算其与数据库中各个字符串Ω的编辑距离d，判断最小编辑距离d_{min}是否小于设定的阈值θ，即判断字符串匹配结果是否合理。若字符串最小编辑距离小于阈值，则匹配成功，将最终匹配结果M输出，否则直接将识别结果R输出。

[0218] $$\begin{cases} d_{\min} < \theta, Output = R \\ Others, Output = M \end{cases}$$

[0219] 通过定制化神经网络的设计，使得字符识别精度和速度得到提高。

[0220] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求的保护范围为准。

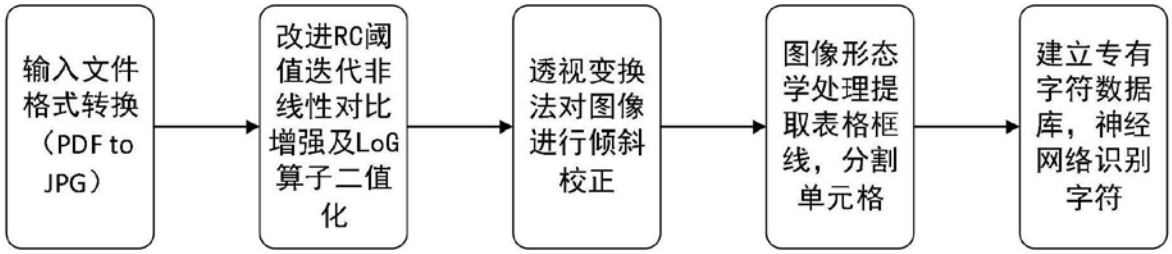
输入文件格式转换（PDF to JPG）
改进RC阈值迭代非线性对比增强及LoG算子二值化
透视变换法对图像进行倾斜校正
图像形态学处理提取表格框线，分割单元格
建立专有字符数据库，神经网络识别字符

图1

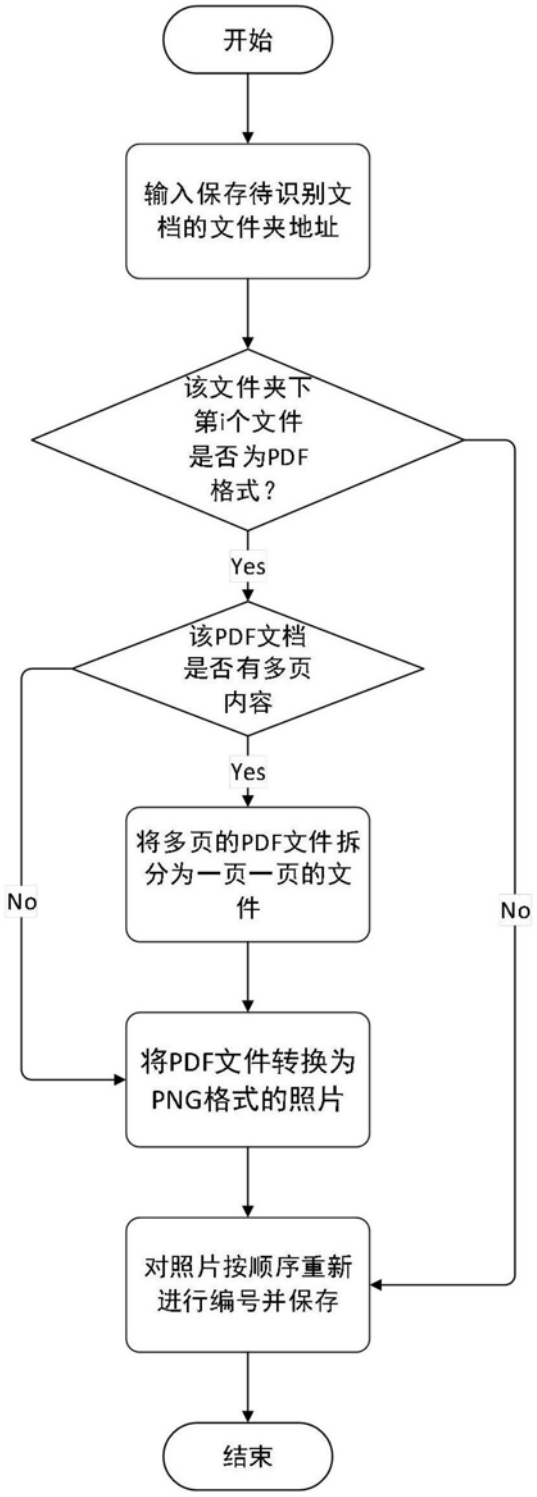
开始
输入保存待识别文档的文件夹地址
该文件夹下第i个文件是否为PDF格式？
Yes
No
该PDF文档是否有多页内容
Yes
No
将多页的PDF文件拆分为一页一页的文件
将PDF文件转换为PNG格式的照片
对照片按顺序重新进行编号并保存
结束

图2

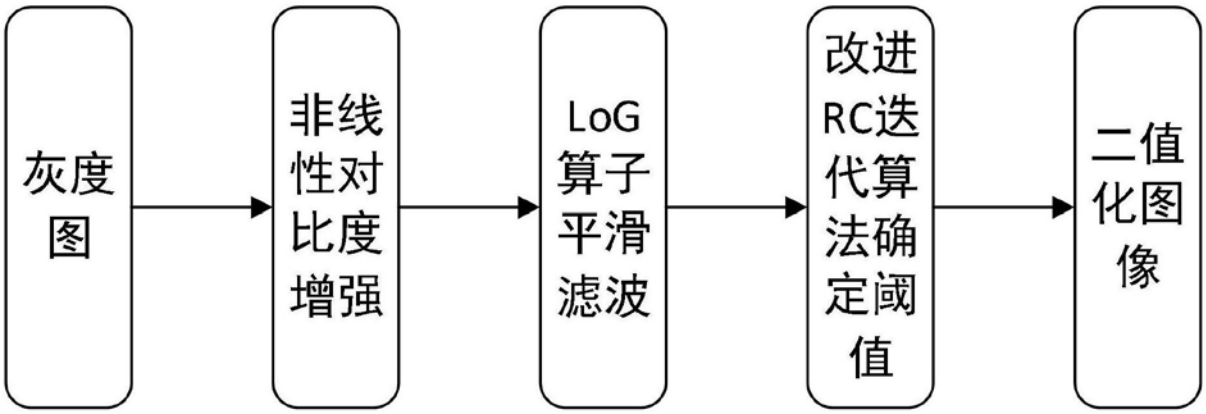
灰度图
非线性对比度增强
LoG算子平滑滤波
改进RC迭代算法确定阈值
二值化图像

图3

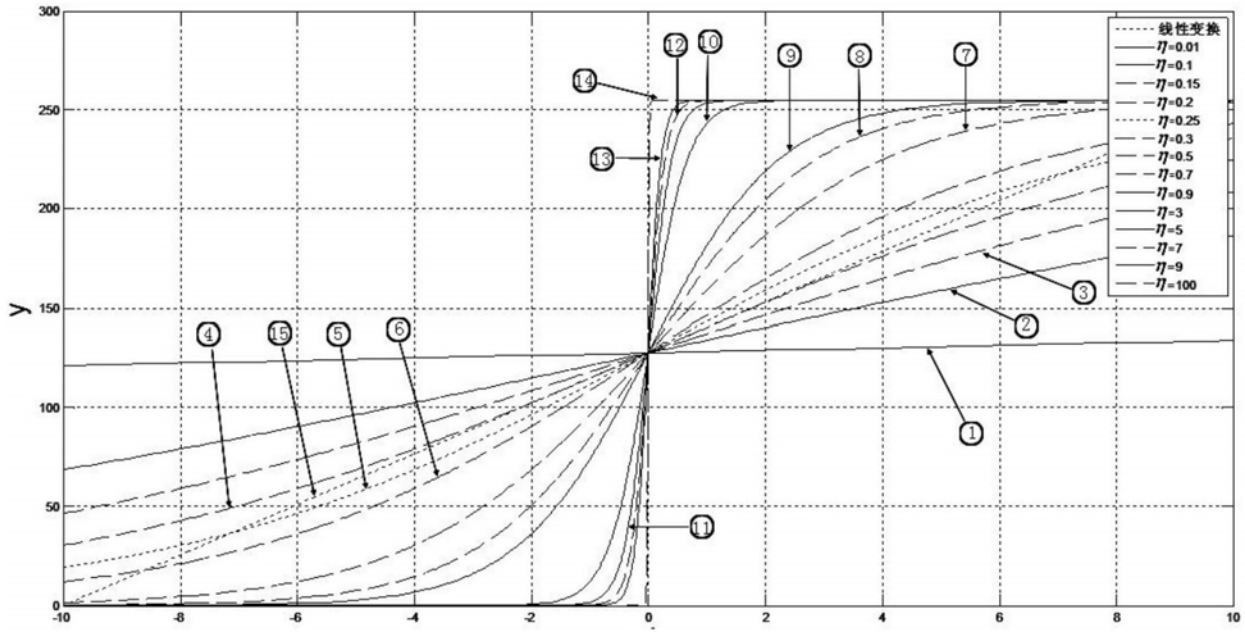
线性变换
η=0.01
η=0.1
η=0.15
η=0.2
η=0.25
η=0.3
η=0.5
η=0.7
η=0.9
η=3
η=5
η=7
η=9
η=100
y

图4

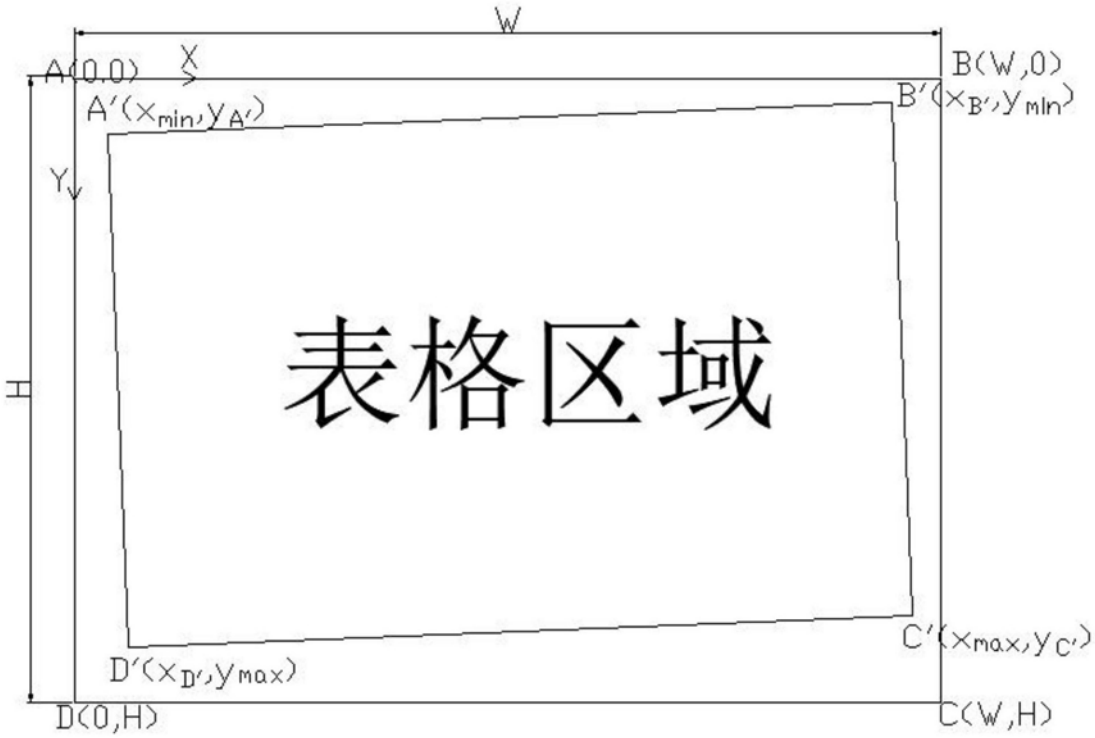
W
X
A(0,0)
B(W,0)
A′(xmin,yA′)
B′(xB′,ymin)
Y
表格区域
H
C′(xmax,yC′)
D′(xD′,ymax)
D(0,H)
C(W,H)

图5

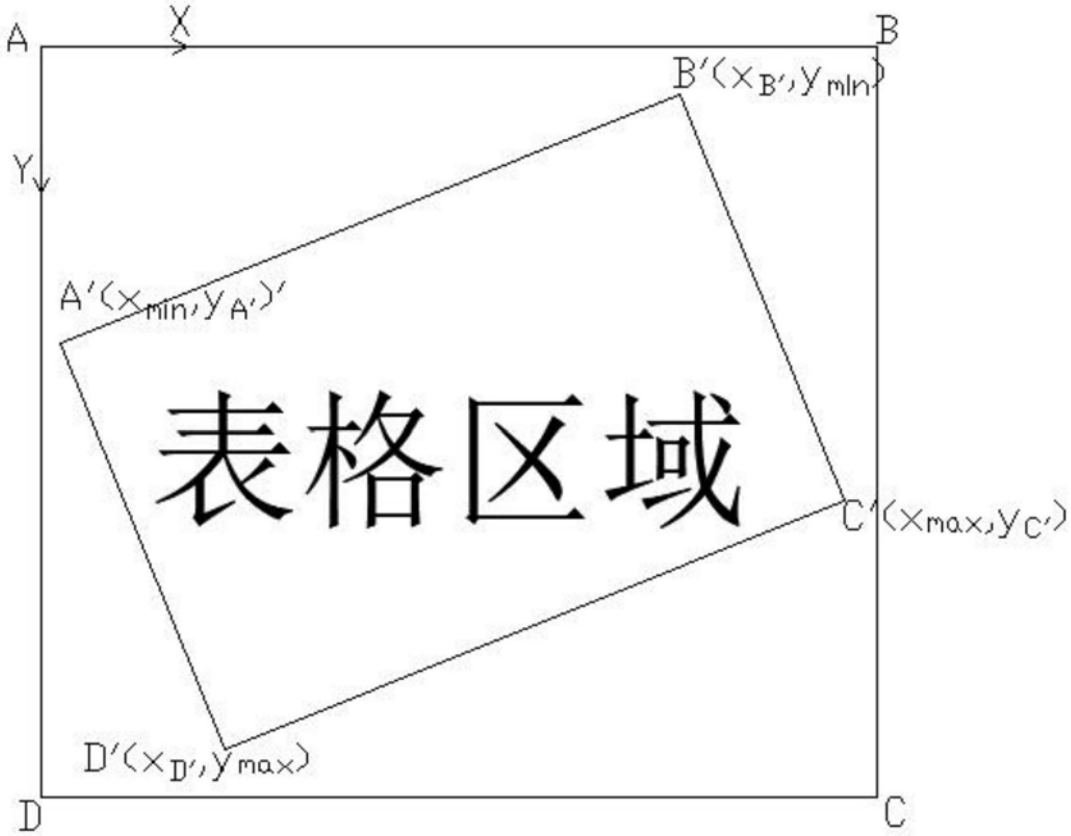
A
X
B
B′(xB′,ymin)
Y
A′(xmin,yA′)
表格区域
C′(xmax,yC′)
D′(xD′,ymax)
D
C

图6

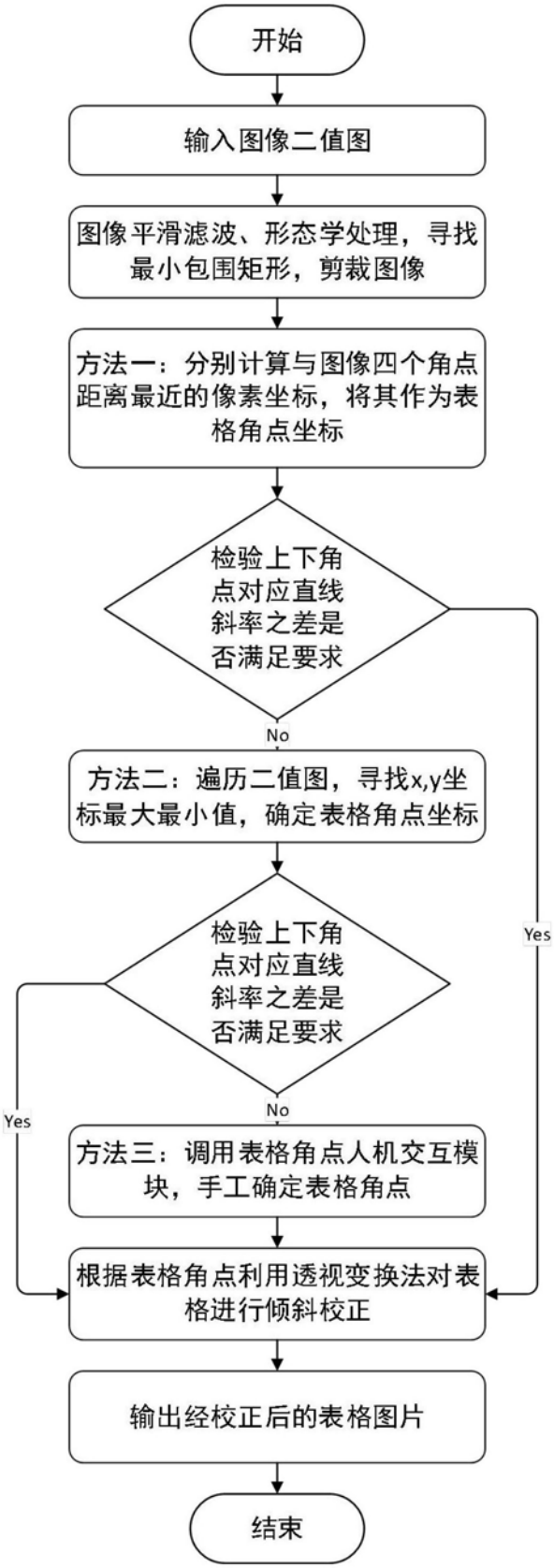
开始
输入图像二值图
图像平滑滤波、形态学处理，寻找最小包围矩形，剪裁图像
方法一：分别计算与图像四个角点距离最近的像素坐标，将其作为表格角点坐标
检验上下角点对应直线斜率之差是否满足要求
No
Yes
方法二：遍历二值图，寻找x,y坐标最大最小值，确定表格角点坐标
检验上下角点对应直线斜率之差是否满足要求
Yes
No
方法三：调用表格角点人机交互模块，手工确定表格角点
根据表格角点利用透视变换法对表格进行倾斜校正
输出经校正后的表格图片
结束

图7

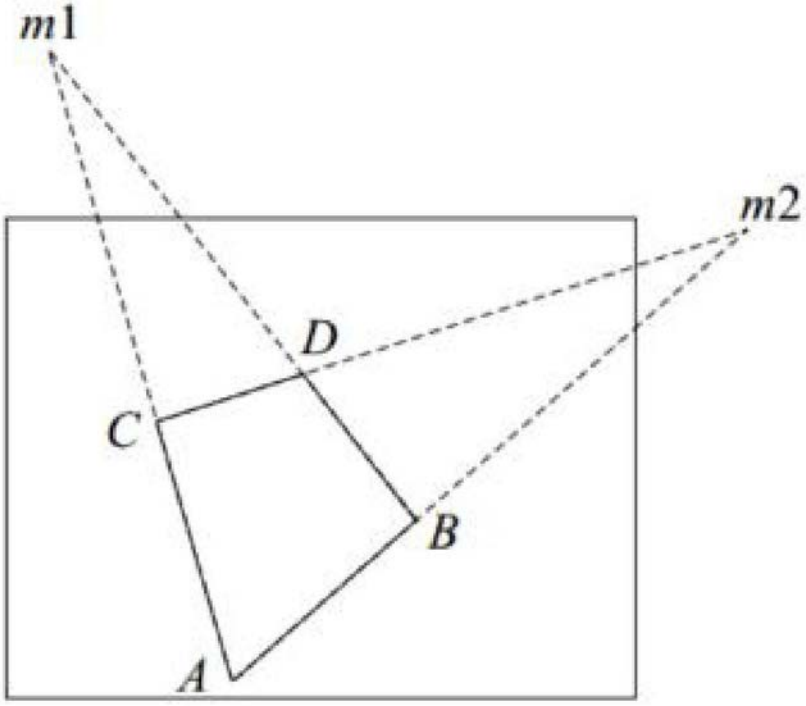

图8

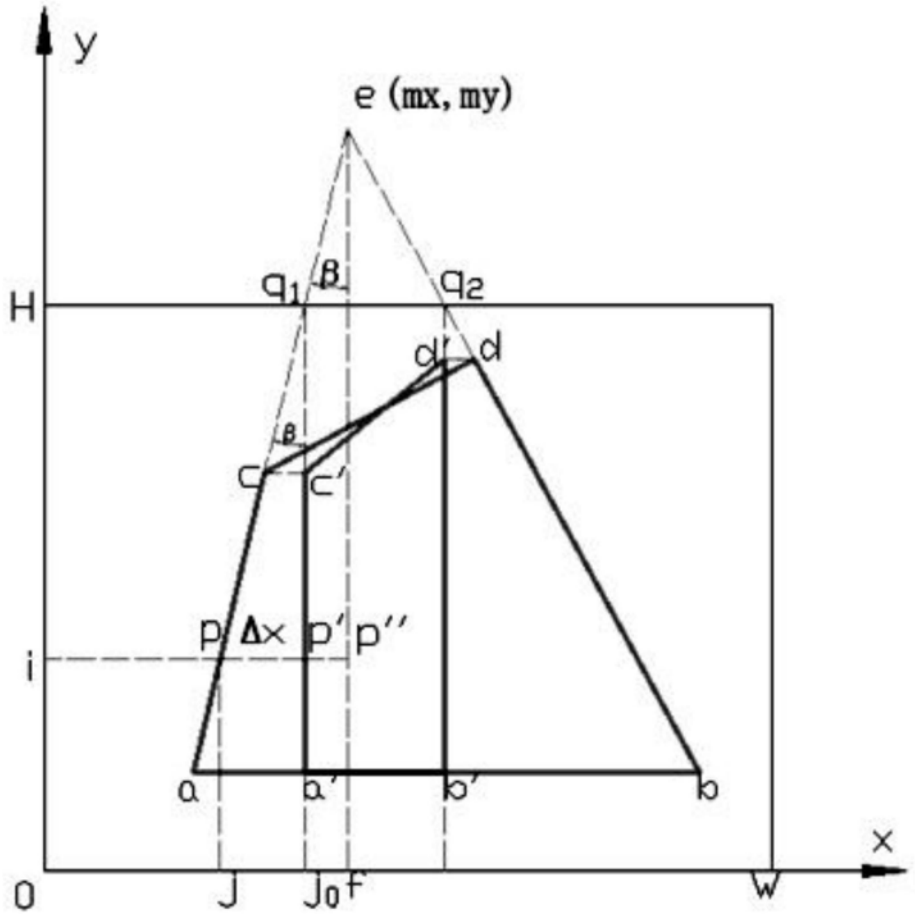
y
e (mx, my)
H
q1
β
q2
d'
d
β
c
c'
p
Δx
p'
p''
i
a
a'
b'
b
0
J
J0
f'
w
x

图9

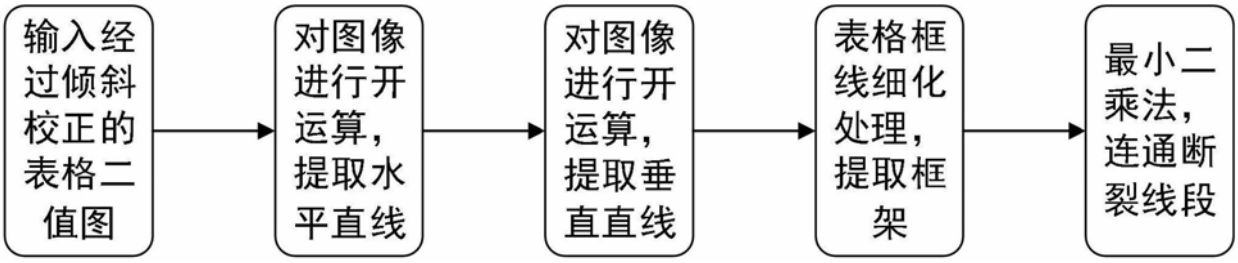
输入经过倾斜校正的表格二值图
对图像进行开运算，提取水平直线
对图像进行开运算，提取垂直直线
表格框线细化处理，提取框架
最小二乘法，连通断裂线段

图10

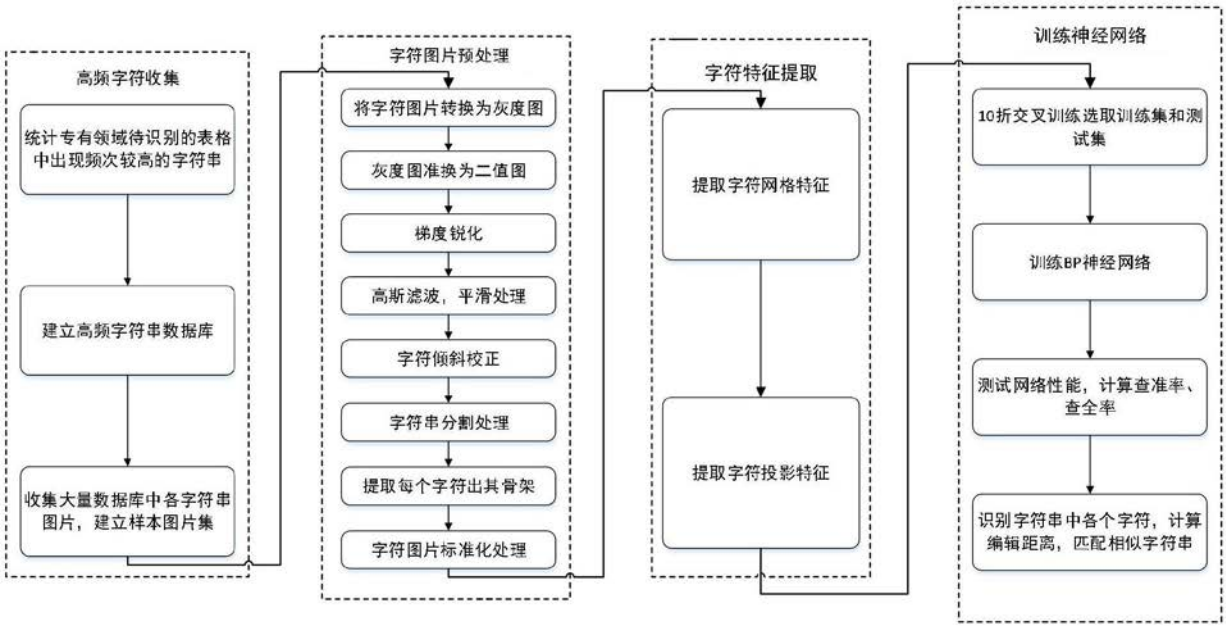
高频字符收集
统计专有领域待识别的表格中出现频次较高的字符串
建立高频字符串数据库
收集大量数据库中各字符串图片，建立样本图片集
字符图片预处理
将字符图片转换为灰度图
灰度图准换为二值图
梯度锐化
高斯滤波，平滑处理
字符倾斜校正
字符串分割处理
提取每个字符出其骨架
字符图片标准化处理
字符特征提取
提取字符网格特征
提取字符投影特征
训练神经网络
10折交叉训练选取训练集和测试集
训练BP神经网络
测试网络性能，计算查准率、查全率
识别字符串中各个字符，计算编辑距离，匹配相似字符串

图11