(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 111353961 A
(43)申请公布日 2020.06.30

(21)申请号 202010172453.8

(22)申请日 2020.03.12

(71)申请人 上海合合信息科技发展有限公司
地址 200433 上海市杨浦区国定路335号8008－34室

(72)发明人 郭丰俊 李亚东 龙腾

(74)专利代理机构 上海恒锐佳知识产权代理事务所(普通合伙) 31286
代理人 殷晓雪

(51)Int.Cl.
G06T 5/00(2006.01)
G06T 7/13(2017.01)
G06K 9/32(2006.01)
G06K 9/38(2006.01)

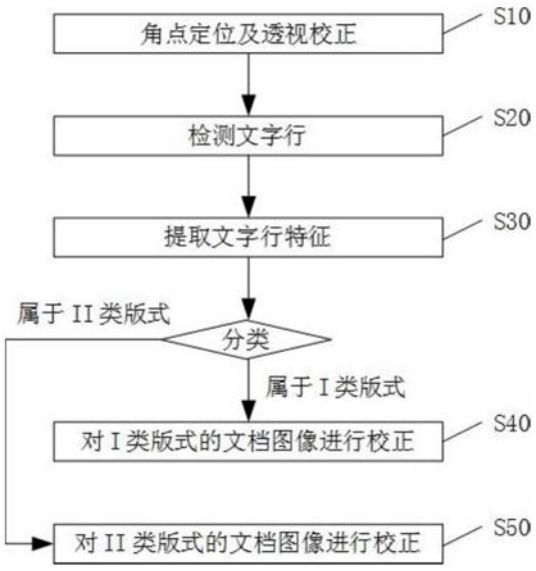
权利要求书3页 说明书8页 附图4页

(54)发明名称

一种文档曲面校正方法及装置

(57)摘要

本申请公开了一种文档曲面校正方法，包括如下步骤。步骤S10：对文档图像进行角点定位及透视校正。步骤S20：在文档图像中检测文字行。步骤S30：提取文字行特征，由分类器将文档图像分为I类版式和II类版式两类。步骤S40：对I类版式的文档图像采用I类校正方法。步骤S50：对II类版式的文档图像采用II类校正方法。本申请将文档图像根据版式分类，并自适应地采用不同的校正方法处理，这样可以提高文档曲面校正的鲁棒性以及最终校正质量。



CN 111353961 A

1.一种文档曲面校正方法,其特征是,包括如下步骤:

步骤S10:对文档图像进行角点定位及透视校正;

步骤S20:在文档图像中检测文字行;

步骤S30:提取文字行特征,由分类器将文档图像分为I类版式和II类版式两类;I类版式的文档图像进入步骤S40,II类版式的文档图像进入步骤S50;

步骤S40:对I类版式的文档图像采用I类校正方法;

步骤S50:对II类版式的文档图像采用II类校正方法。

2.根据权利要求1所述的文档曲面校正方法,其特征是,所述步骤S10中,文档角点定位方法包括基于直线检测、基于轮廓提取、基于跳变点检测的任一种。

3.根据权利要求2所述的文档曲面校正方法,其特征是,所述步骤S10中,采用基于直线检测的角点定位方法,具体包括如下步骤;

步骤S11:检测文档图像的边缘,得到文档图像的边缘图;

步骤S12:在边缘图上做直线检测;

步骤S13:组合四条直线形成四边形,对所有四条直线组合根据边缘响应强度、角度、边长中的一项或多项进行筛选,最终得到文档的角点。

4.根据权利要求1所述的文档曲面校正方法,其特征是,所述步骤S10中,对文档图像进行角点定位之后,根据文档角点坐标对文档图像做透视校正,并裁剪出文档区域。

5.根据权利要求1所述的文档曲面校正方法,其特征是,所述步骤S20中,采用基于轮廓提取的文字行检测方法,具体包括如下步骤;

步骤S21:对文档图像做二值化,并做反色操作,得到二值图;对二值图做横向膨胀,以将文字行连接在一起;然后做竖向腐蚀,去除图像中线段的干扰;

步骤S22:在步骤S21处理后的图片上找连通域,根据连通域的特征将连通域分类为文字区域和非文字区域两种;

步骤S23:将文字区域的连通域进行组合,拼接,得到最终文字行。

6.根据权利要求5所述的文档曲面校正方法,其特征是,所述步骤S22中,分类的规则包括如下一种或多种:连通域包围四边形宽度大于预设最小宽度,连通域包围四边形高度小于预设最大高度,连通域包围四边形宽高比大于预设最小宽高比。

7.根据权利要求5所述的文档曲面校正方法,其特征是,所述步骤S23中,组合、拼接具体包括如下步骤;

步骤S231:循环遍历所有连通域组合,判断是否将两个连通域组成连通域对;

步骤232:遍历所有连通域对,根据连通域对的信息,采用链表数据结构对连通域进行拼接组成连通域序列,重复步骤S231至步骤S232得到多个连通域序列;

步骤233:遍历连通域序列,若连通域序列的x轴方向长度大于预设最小长度,则对该连通域序列做离散采样,作为一个文字行。

8.根据权利要求7所述的文档曲面校正方法,其特征是,所述步骤S231中,判断规则包括如下一种或多种:两个连通域在x轴方向的重叠长度大于预设最小长度,两个连通域主方向之间的角度差小于预设最大角度,一个连通域上的中心点与过另一连通域中心点主方向的直线的距离小于预设最大距离。

9.根据权利要求7所述的文档曲面校正方法,其特征是,所述步骤S233中,文字行的处

理形式为一组散点序列;采样方式为按照预设间隔对连通域序列在x轴方向采样,该采样位置的y轴坐标设为连通域序列包围范围内y轴坐标的平均值,即若当前采样位置的x轴坐标为xi,则y轴坐标定义为公式一,其中,P_{xy}定义为公式二;

$$yi := \frac{\sum_y y P_{xiy}}{\sum_y P_{xiy}}$$ (公式一);

$$P_{xy} := \begin{cases} 1 \text{ 若点(x,y)在连通域序列包围范围内} \\ 0 \text{ 若点(x,y)不在连通域序列包围范围内} \end{cases}$$ (公式二)。

10.根据权利要求1所述的文档曲面校正方法,其特征是,所述步骤S30中,提取的文字行特征包括以下一项或多项:所有文字行的平均长度、所有文字行的长度中位值、所有文字行左边界x轴坐标的平均值、所有文字行左边界x轴坐标的中位值、所有文字行右边界x轴坐标的平均值、所有文字行右边界x轴坐标的中位值,长度大于图像宽度的一定比例的文字行中y轴坐标的最小值与最大值。

11.根据权利要求1所述的文档曲面校正方法,其特征是,所述步骤S30中,分类器是事先通过训练建立的;这包括收集文档图像样本集,对样本集内所有文档图像分别用I类校正方法和II校正方法进行曲面校正;对校正后图片进行人工分类,如果I类校正方法的校正质量优于II类校正方法的校正质量,则将该文档图像分类为I类版式;否则,将该文档图像分类为II类版式;这被称为建立训练数据集,人工分类的结果作为数据标注。

12.根据权利要求11所述的文档曲面校正方法,其特征是,所述步骤S30中,由分类器根据训练数据集的数据标注、以及待校正的文档图像所提取的特征,训练分类器将待校正的文档图像分为I类版式和II类版式两种。

13.根据权利要求1所述的文档曲面校正方法,其特征是,所述步骤S40中,I类校正方法具体包括如下步骤;

步骤S41:提取上文字行和下文字行;

步骤S42:对上文字行、下文字行进行多项式曲线拟合;

步骤S43:横向遍历列像素,逐列进行校正。

14.根据权利要求13所述的文档曲面校正方法,其特征是,所述步骤S41中,将文字行按y轴坐标升序排序;记排序后的y轴坐标落在前50%的文字行中长度最大值为l_{max},对这些文字行进行遍历,找出满足长度大于0.8倍的l_{max}且y轴坐标值最小的文字行作为上文字行;

记排序后的y轴坐标落在后50%的文字行中长度最大值为l_{max2},对这些文字行进行遍历,找出满足长度大于0.8倍的l_{max2}且y轴坐标值最大的文字行作为下文字行。

15.根据权利要求13所述的文档曲面校正方法,其特征是,所述步骤S43中,令当前列像素穿越的所有上文字行的y轴坐标变为所有上文字行的y轴坐标平均值,令当前列像素穿越的所有下文字行的y轴坐标变为所有下文字行的y轴坐标平均值,得到当前列的线性变换关系,以将曲面文字校正为水平。

16.根据权利要求1所述的文档曲面校正方法,其特征是,所述步骤S50中,II类校正方法具体包括如下步骤;

步骤S51:记第i个文字行的第j个采样点为P_{ij};

步骤S52:计算与P_{ij}对应的校正后点坐标P_{ij}';

步骤S53:优化曲面参数和投影参数;

步骤S54:通过图像重映射的方法,根据曲面参数和投影参数,得到校正后的图像。

17.根据权利要求16所述的文档曲面校正方法,其特征是,所述步骤S51中,记第i个文字行的第j个采样点为P_{ij};所述步骤S52中,将采样点P_{ij}的y轴坐标改为该采样点所在的文字行的所有采样点的y轴坐标平均值my,即得到与采样点P_{ij}对应的校正后点坐标P_{ij}'。

18.根据权利要求17所述的文档曲面校正方法,其特征是,所述步骤S53中,假设文档表面z轴方向为二次样条函数构成的曲面,则优化过程如下;首先将P_{ij}'转换为齐次坐标表示H_{ij}';假设H_{ij}'的z轴坐标是随x轴坐标变化的二次样条函数;其次通过投影变换,将H_{ij}'投影到二维平面上,得到Q_{ij};最后将所有Q_{ij}和P_{ij}的欧氏距离之和,也就是投影误差,作为目标函数;优化投影变换参数以及二次样条函数中的参数,以最小化目标函数,这样就得到了曲面参数和投影参数。

19.根据权利要求18所述的文档曲面校正方法,其特征是,所述步骤S54中,重映射的方法是遍历目标图上的像素坐标,用映射关系计算得到对应于原图中的像素坐标,通过差值方法得到像素值;若记目标图上像素坐标为(x,y),二次样条函数为f(x),将目标图坐标记为齐次坐标形式dst=$(x,y,f(x))^T$;投影变换参数为H=$(h1^T,h2^T,h3^T)$,其中$h1^T$,$h2^T$,$h3^T$为矩阵H的行元素;对应原图中坐标为(x',y'),如公式三、公式四所示;

$x' = (h1^T * dst) / (h3^T * dst)$ (公式三);

$y' = (h2^T * dst) / (h3^T * dst)$ (公式四)。

20.一种文档曲面校正装置,其特征是,包括初步处理单元、检测单元、分类单元、I类校正单元和II类校正单元;

所述初步处理单元用来对文档图像进行角点定位和透视校正;

所述检测单元用来在文档图像中检测文字行;

所述分类单元用来提取文字行特征,还用来将文档图像分为I类版式和II类版式两类;

所述I类校正单元用来对I类版式的文档图像采用I类校正方法进行校正;

所述II类校正单元用来对II类版式的文档图像采用II类校正方法进行校正。

一种文档曲面校正方法及装置

技术领域

[0001] 本申请涉及一种数字图像处理方法，特别是涉及一种文档图像的校正方法。

背景技术

[0002] 随着高质量摄像头在手机等移动设备上的普及，利用移动设备对文档进行数字化采集已经非常普遍。通过图像校正技术，移动设备采集的文档图像质量甚至可以与专用的文档扫描仪相当。然而，一些文档(例如书页)中存在的形变无法通过简单的透视变换进行校正。

[0003] 为提升存在曲面形变文档的校正质量，现在普遍采用的方法可以大致分为两类。

[0004] 第一类是利用多目相机、结构光或者激光雷达等专用设备对文档进行扫描，获得文档表面的3D结构信息，进而对文档校正展平。授权公告号为CN102592124B、授权公告日为2013年11月27日的中国发明专利《文本图像的几何校正方法、装置和双目立体视觉系统》公开了一种利用双目立体视觉系统对文本图像进行校正的方法。授权公告号为CN102801894B、授权公告日为2014年10月1日的中国发明专利《一种变形书页展平方法》公开了一种利用左右两台相机对变形书页进行展平校正的方法。这类方法一般可以得到比较好的校正效果，但依赖专用设备的特点限制了其使用场景。

[0005] 第二类是完全依靠图像信息以及文档形变的先验知识对图像进行校正。授权公告号为CN102208025B、授权公告日为2013年2月27日的中国发明专利《一种文本图像几何畸变的矫正方法》公开了一种利用拟合的文本行曲线对文本图像进行几何畸变的校正方法。申请公布号为CN102254171A、申请公布日为2011年11月23日的中国发明专利申请《一种基于文本边界的中文文档图像畸变校正方法》公开了一种利用上下文本边界线进行文档图像畸变校正的方法。这类方法一般需要进行文字行或者表格线的检测，并假设曲面符合特定的几何约束，如曲面是柱面。这类方法可以在普通的移动设备上实现，但是其校正效果受文字行检测准确度的限制，对文档版式比较敏感，无法处理存在大量图表的文档，且误检的文字行有可能会对校正造成严重干扰。

发明内容

[0006] 本申请所要解决的技术问题是提出一种基于机器学习的自适应文档曲面校正方法，属于前述的第二类文档曲面校正方法。本申请只依赖采集的图像信息，利用机器学习的方法自适应地对不同版式文档采取不同的校正策略，提高了对复杂版式文档(例如多栏排版文档以及存在大量图标的文档等)的曲面校正质量与鲁棒性(robustness)。

[0007] 为解决上述技术问题，本申请提供了一种文档曲面校正方法，包括如下步骤。步骤S10：对文档图像进行角点定位及透视校正。步骤S20：在文档图像中检测文字行。步骤S30：提取文字行特征，由分类器将文档图像分为I类版式和II类版式两类。I类版式的文档图像进入步骤S40，II类版式的文档图像进入步骤S50。步骤S40：对I类版式的文档图像采用I类校正方法。步骤S50：对II类版式的文档图像采用II类校正方法。上述方法将文档图像根据

版式分类，并自适应地采用不同的校正方法处理，这样可以提高文档曲面校正的鲁棒性以及最终校正质量。

[0008] 进一步地，所述步骤S10中，文档角点定位方法包括基于直线检测、基于轮廓提取、基于跳变点检测的任一种。这些角点定位方法均可用于本申请。

[0009] 优选地，所述步骤S10中，采用基于直线检测的角点定位方法，具体包括如下步骤。步骤S11：检测文档图像的边缘，得到文档图像的边缘图。步骤S12：在边缘图上做直线检测。步骤S13：组合四条直线形成四边形，对所有四条直线组合根据边缘响应强度、角度、边长中的一项或多项进行筛选，最终得到文档的角点。这是步骤S10中的角点定位的一种优选实现方式。

[0010] 进一步地，所述步骤S10中，对文档图像进行角点定位之后，根据文档角点坐标对文档图像做透视校正，并裁剪出文档区域。这是对步骤S10中的透视校正的详细描述。

[0011] 优选地，所述步骤S20中，采用基于轮廓提取的文字行检测方法，具体包括如下步骤。步骤S21：对文档图像做二值化，并做反色操作，得到二值图；对二值图做横向膨胀，以将文字行连接在一起；然后做竖向腐蚀，去除图像中线段的干扰。步骤S22：在步骤S21处理后的图片上找连通域，根据连通域的特征将连通域分类为文字区域和非文字区域两种。步骤S23：将文字区域的连通域进行组合，拼接，得到最终文字行。这是步骤S20的一种优选实现方式。

[0012] 优选地，所述步骤S22中，分类的规则包括如下一种或多种：连通域包围四边形宽度大于预设最小宽度，连通域包围四边形高度小于预设最大高度，连通域包围四边形宽高比大于预设最小宽高比。这些规则均可用于本申请。

[0013] 优选地，所述步骤S23中，组合、拼接具体包括如下步骤。步骤S231：循环遍历所有连通域组合，判断是否将两个连通域组成连通域对。步骤232：遍历所有连通域对，根据连通域对的信息，采用链表数据结构对连通域进行拼接组成连通域序列，重复步骤S231至步骤S232得到多个连通域序列。步骤233：遍历连通域序列，若连通域序列的x轴方向长度大于预设最小长度，则对该连通域序列做离散采样，作为一个文字行。这是步骤S23的一种优选实现方式。

[0014] 优选地，所述步骤S231中，判断规则包括如下一种或多种：两个连通域在x轴方向的重叠长度大于预设最小长度，两个连通域主方向之间的角度差小于预设最大角度，一个连通域上的中心点与过另一连通域中心点主方向的直线的距离小于预设最大距离。这些规则均可用于本申请。

[0015] 优选地，所述步骤S233中，文字行的处理形式为一组散点序列；采样方式为按照预设间隔对连通域序列在x轴方向采样，该采样位置的y轴坐标设为连通域序列包围范围内y轴坐标的平均值，即若当前采样位置的x轴坐标为xi，则y轴坐标定义为公式一，其中，P_{xy}定义为公式二。

$$yi := \frac{\sum_y y P_{xiy}}{\sum_y P_{xiy}} \text{（公式一）。} P_{xy} := \begin{cases} 1 \text{ 若点(x,y)在连通域序列包围范围内} \\ 0 \text{ 若点(x,y)不在连通域序列包围范围内} \end{cases} \text{（公式二）。}$$

这是步骤S233的一种优选实现方式。

[0016] 进一步地，所述步骤S30中，提取的文字行特征包括以下一项或多项：所有文字行的平均长度、所有文字行的长度中位值、所有文字行左边界x轴坐标的平均值、所有文字行

左边界x轴坐标的中位值、所有文字行右边界x轴坐标的平均值、所有文字行右边界x轴坐标的中位值，长度大于图像宽度的一定比例的文字行中y轴坐标的最小值与最大值。这是文字行特征的一些可能的形式，均可用于本申请。

[0017] 进一步地，所述步骤S30中，分类器是事先通过训练建立的；这包括收集文档图像样本集，对样本集内所有文档图像分别用I类校正方法和II校正方法进行曲面校正；对校正后图片进行人工分类，如果I类校正方法的校正质量优于II类校正方法的校正质量，则将该文档图像分类为I类版式；否则，将该文档图像分类为II类版式；这被称为建立训练数据集，人工分类的结果作为数据标注。这是对分类器的详细描述。

[0018] 优选地，所述步骤S30中，由分类器根据训练数据集的数据标注、以及待校正的文档图像所提取的特征，训练分类器将待校正的文档图像分为I类版式和II类版式两种。这是步骤S30的一种优选实现方式。

[0019] 进一步地，所述步骤S40中，I类校正方法具体包括如下步骤。步骤S41：提取上文字行和下文字行。步骤S42：对上文字行、下文字行进行多项式曲线拟合。步骤S43：横向遍历列像素，逐列进行校正。这是步骤S40的一种优选实现方式。

[0020] 优选地，所述步骤S41中，将文字行按y轴坐标升序排序；记排序后的y轴坐标落在前50％的文字行中长度最大值为l_{max}，对这些文字行进行遍历，找出满足长度大于0.8倍的l_{max}且y轴坐标值最小的文字行作为上文字行。记排序后的y轴坐标落在后50％的文字行中长度最大值为l_{max2}，对这些文字行进行遍历，找出满足长度大于0.8倍的l_{max2}且y轴坐标值最大的文字行作为下文字行。这是步骤S41的一种优选实现方式。

[0021] 优选地，所述步骤S43中，令当前列像素穿越的所有上文字行的y轴坐标变为所有上文字行的y轴坐标平均值，令当前列像素穿越的所有下文字行的y轴坐标变为所有下文字行的y轴坐标平均值，得到当前列的线性变换关系，以将曲面文字校正为水平。这是步骤S43的一种优选实现方式。

[0022] 进一步地，所述步骤S50中，II类校正方法具体包括如下步骤。步骤S51：记第i个文字行的第j个采样点为P_{ij}。步骤S52：计算与P_{ij}对应的校正后点坐标P_{ij}’。步骤S53：优化曲面参数和投影参数。步骤S54：通过图像重映射的方法，根据曲面参数和投影参数，得到校正后的图像。这是步骤S50的一种优选实现方式。

[0023] 优选地，所述步骤S51中，记第i个文字行的第j个采样点为P_{ij}；所述步骤S52中，将采样点P_{ij}的y轴坐标改为该采样点所在的文字行的所有采样点的y轴坐标平均值my，即得到与采样点P_{ij}对应的校正后点坐标P_{ij}’。这是步骤S51和步骤S52的一种优选实现方式。

[0024] 优选地，所述步骤S53中，假设文档表面z轴方向为二次样条函数构成的曲面，则优化过程如下；首先将P_{ij}’转换为齐次坐标表示H_{ij}’；假设H_{ij}’的z轴坐标是随x轴坐标变化的二次样条函数；其次通过投影变换，将H_{ij}’投影到二维平面上，得到Q_{ij}；最后将所有Q_{ij}和P_{ij}的欧氏距离之和，也就是投影误差，作为目标函数；优化投影变换参数以及二次样条函数中的参数，以最小化目标函数，这样就得到了曲面参数和投影参数。这是步骤S53的一种优选实现方式。

[0025] 优选地，所述步骤S54中，重映射的方法是遍历目标图上的像素坐标，用映射关系计算得到对应于原图中的像素坐标，通过差值方法得到像素值；若记目标图上像素坐标为(x,y)，二次样条函数为f(x)，将目标图坐标记为齐次坐标形式$dst=(x,y,f(x))^T$；投影变

换参数为H＝(h1^T,h2^T,h3^T)，其中h1^T,h2^T,h3^{T}为矩阵H的行元素；对应原图中坐标为(x’,y’)，如公式三、公式四所示。x′＝(h1^T*dst)/(h3^T*dst)(公式三)。y′＝(h2^T*dst)/(h3^T*dst)(公式四)。这是步骤S54的一种优选实现方式。

[0026] 本申请还提供了一种文档曲面校正装置，包括初步处理单元、检测单元、分类单元、I类校正单元和II类校正单元。所述初步处理单元用来对文档图像进行角点定位和透视校正。所述检测单元用来在文档图像中检测文字行。所述分类单元用来提取文字行特征，还用来将文档图像分为I类版式和II类版式两类。所述I类校正单元用来对I类版式的文档图像采用I类校正方法进行校正。所述II类校正单元用来对II类版式的文档图像采用II类校正方法进行校正。上述装置将文档图像根据版式分类，并自适应地采用不同的校正方法处理，这样可以提高文档曲面校正的鲁棒性以及最终校正质量。

[0027] 本申请取得的技术效果包括如下几个方面。第一，不需要使用多目相机等专用设备获取文档表面深度信息，只需要由一台摄影设备从一个角度拍摄的文档图像即可进行曲面校正。第二，适用于复杂版式文档校正，可以处理多栏、图文混排等复杂版式文档。第三，处理快速，可以在移动设备中实现几乎实时的曲面文档校正。

附图说明

[0028] 图1是本申请提供的文档曲面校正方法的流程图。

[0029] 图2是步骤S10中基于直线检测的角点定位方法的流程图。

[0030] 图3是步骤S20中基于轮廓提取的文字行检测方法的流程图。

[0031] 图4是步骤S40中I类校正方法的流程图。

[0032] 图5是步骤S50中II类校正方法的流程图。

[0033] 图6是I类版式的文档图像校正前后的对比示意图。

[0034] 图7是II类版式的文档图像校正前后的对比示意图。

[0035] 图8是本申请提供的文档曲面校正装置的结构示意图。

[0036] 图中附图标记说明：10为初步处理单元；20为检测单元；30为分类单元；40为I类校正单元；50为II类校正单元。

具体实施方式

[0037] 请参阅图1，本申请提供的文档曲面校正方法包括如下步骤。

[0038] 步骤S10：对待校正的文档图像进行角点定位(corner detection)及透视校正(perspective correction)。

[0039] 步骤S20：在待校正的文档图像中检测文字行。

[0040] 步骤S30：提取文字行特征，由分类器将待校正的文档图像分为I类版式和II类版式两类。I类版式的文档图像进入步骤S40，II类版式的文档图像进入步骤S50。

[0041] 步骤S40：对属于I类版式的待校正的文档图像进行校正，称为I类校正方法。I类校正方法对文字行的y轴方向跨度(即高度)较大、且x轴方向跨度(即长度)占文档图像宽度较大的文档图像有较好的曲面校正效果，但不适用于版式复杂的文档图像。

[0042] 步骤S50：对属于II类版式的待校正的文档图像进行校正，称为II类校正方法。II类校正方法适用于复杂版式的文档图像，对文字行误检有一定鲁棒性，但不适用于有较大

曲面形变的文档图像。

[0043] 本申请将文档图像根据版式分类，并自适应地采用不同的校正方法处理，这样可以提高文档曲面校正的鲁棒性以及最终校正质量。

[0044] 所述步骤S10中，不依赖具体的文档角点定位方法。目前广泛采用的文档角点定位方法包括基于直线检测、基于轮廓提取、基于跳变点(changing point)检测等，本申请均可采用。作为一个示例，所述步骤S10采用基于直线检测的角点定位方法，如图2所示，具体包括如下步骤。

[0045] 步骤S11：检测文档图像的边缘，得到文档图像的边缘图(edge map)。例如采用图像处理中常用的Canny边缘检测方法。

[0046] 步骤S12：在边缘图上做直线检测。例如采用霍夫变换(Hough transform)直线检测方法。

[0047] 步骤S13：组合四条直线形成四边形，对所有四条直线组合根据边缘响应强度、角度、边长等信息进行筛选，最终得到文档的角点(corner point)。

[0048] 在步骤S11至步骤S13进行角点定位之后，根据文档角点坐标对文档图像做透视校正，并裁剪出文档区域。所述步骤S10中，对文档图像做透视校正将简化后续步骤中的曲面校正复杂度，提高后续步骤的处理效率和缩短处理时间。

[0049] 所述步骤S20中，不依赖具体的文字行检测方法。作为一个示例，考虑到手机等移动设备的计算能力限制，所述步骤S20采用基于轮廓提取的文字行检测方法，如图3所示，具体包括如下步骤。

[0050] 步骤S21：对文档图像做二值化(binarization)，并做反色操作，得到二值图。对二值图做横向膨胀(dilate)，以将文字行连接在一起；然后做竖向腐蚀(erode)，去除图像中线段的干扰。膨胀、腐蚀等图像形态学处理可以把断开的线连起来，把孤立的噪声去除。

[0051] 步骤S22：在图片上找连通域(Connected Component)，根据连通域的特征将连通域分类为文字区域和非文字区域两种。本步骤中不依赖特定规则，作为示例，采用的规则例如包括如下一种或多种：连通域包围四边形宽度大于预设最小宽度，连通域包围四边形高度小于预设最大高度，连通域包围四边形宽高比大于预设最小宽高比。在不同的应用场景下，各项规则可以得到文字区域、非文字区域两种判定结果。

[0052] 步骤S23：将文字区域的连通域根据规则进行组合，拼接，得到最终文字行。作为示例，所述组合、拼接具体包括如下步骤：

[0053] 步骤S231：循环遍历所有连通域组合，根据一定规则判断是否将两个连通域组成连通域对。所述规则例如包括如下一种或多种：两个连通域在x轴方向(即横方向)的重叠长度大于预设最小长度，两个连通域主方向(可以通过图像矩(Image Moment)计算得到)之间的角度差小于预设最大角度，一个连通域上的中心点(可以通过图像矩计算得到)与过另一连通域中心点主方向的直线的距离小于预设最大距离。

[0054] 步骤232：对连通域进行拼接。遍历所有连通域对，根据连通域对的信息，采用链表(Linked List)数据结构对连通域进行拼接。当一个连通域与多个其他连通域组成连通域对时，将该连通域与所在连通域对中的长度最长的其他连通域采用链表数据结构拼接组成连通域序列；重复步骤S231至步骤S232得到多个连通域序列。每个连通域序列是由两个或多个连通域组成。

[0055] 步骤233:遍历连通域序列,若连通域序列的x轴方向长度大于预设最小长度,则对该连通域序列做离散采样,作为一个文字行。本申请中,文字行的处理形式为一组散点序列。采样方式为按照预设间隔对连通域序列在x轴方向采样,该采样位置的y轴坐标设为连通域序列包围范围内y轴坐标的平均值,即若当前采样位置的x轴坐标为xi,则y轴坐标定义为公式一,其中,P_{xy}定义为公式二。

[0056] $$yi := \frac{\sum_y y P_{xiy}}{\sum_y P_{xiy}}$$ (公式一)。

[0057] $$P_{xy} := \begin{cases} 1 \text{ 若点(x,y)在连通域序列包围范围内} \\ 0 \text{ 若点(x,y)不在连通域序列包围范围内} \end{cases}$$ (公式二)。

[0058] 文字行的y轴坐标定义为该文字行散点序列的y轴坐标的平均值。

[0059] 所述步骤S30中,由于曲面校正需要利用文字行几何信息作为线索,不同校正方法的校正质量主要依赖于文字行分布。本申请提出通过提取文字行特征并用机器学习方法对图像进行分类,自动化选择最优的校正方法。本申请不依赖于具体特征选择,作为示例,提取的文字行特征例如包括以下一项或多项:所有文字行的平均长度、所有文字行的长度中位值(median)、所有文字行左边界x轴坐标的平均值、所有文字行左边界x轴坐标的中位值、所有文字行右边界x轴坐标的平均值、所有文字行右边界x轴坐标的中位值,长度大于整个文档图像宽度30%的文字行中y轴坐标的最小值与最大值,类似的,长度大于整个文档图像宽度40%、50%、60%的文字行中y轴坐标的最小值与最大值。优选地,所有涉及长度、坐标的特征都通过文档图像的宽高值进行归一化处理。

[0060] 所述步骤S30中,分类器是事先通过训练建立的。这包括收集文档图像样本集,对样本集内所有文档图像分别用I类校正方法和II校正方法进行曲面校正。对校正后图片进行人工分类,如果I类校正方法的校正质量优于II类校正方法的校正质量,则将该文档图像分类为I类版式;否则,将该文档图像分类为II类版式。这被称为建立训练数据集,人工分类的结果作为数据标注。

[0061] 所述步骤S30中,由分类器根据训练数据集的数据标注、以及待校正的文档图像所提取的特征,训练分类器将待校正的文档图像分为I类版式和II类版式两种。优选地,所述分类器为决策树(decision tree)模型。

[0062] 所述步骤S40中,如图4所示,I类校正方法具体包括如下步骤。

[0063] 步骤S41:提取上文字行和下文字行。基于I类版式的文档图像的特点,筛选出上文字行和下文字行。以上文字行为例:将文字行按y轴坐标升序排序;记排序后的y轴坐标落在前50%的文字行中长度最大值为l_{max},对这些文字行进行遍历,找出满足长度大于0.8倍的l_{max}且y轴坐标值最小的文字行作为上文字行。上文字行指长度满足条件(大于0.8倍的l_{max})的最靠上的文字行。下文字行提取规则与之类似:将文字行按y轴坐标升序排序;记排序后的y轴坐标落在后50%的文字行中长度最大值为l_{max2},对这些文字行进行遍历,找出满足长度大于0.8倍的l_{max2}且y轴坐标值最大的文字行作为下文字行。下文字行指长度满足条件(大于0.8倍的l_{max2})的最靠下的文字行。

[0064] 步骤S42:对上文字行、下文字行进行多项式曲线拟合(Polynomial Curve Fitting)。例如采用四次多项式拟合。步骤S41获取的多个上文字行的中心连线可看作为一条曲线,步骤S41获取的多个下文字行的中心连线也可看作为一条曲线。这一步是认为所有

的上文字行的y轴坐标本应相同，只是由于文档图像发生了曲面变形而使得上文字行的y轴坐标存在多个，因此后续会将所有上文字行的y轴坐标统一。同样地，这一步是认为所有的下文字行的y轴坐标本应相同，只是由于文档图像发生了曲面变形而使得下文字行的y轴坐标存在多个，因此后续会将所有下文字行的y轴坐标统一。

[0065] 步骤S43：横向遍历列像素，逐列进行校正。记上文字行的所有y轴坐标平均值为my_{top}，记下文字行的所有y轴坐标平均值为my_{bottom}。记当前列像素与上文字行交叉点y轴坐标为y_{top}。记当前列像素与下文字行交叉点y轴坐标为y_{bottom}。计算对y坐标值的线性变换，令当前列像素与上文字行交叉点的y轴坐标变为所有上文字行的y轴坐标平均值，令当前列像素与下文字行交叉点的y轴坐标变为所有下文字行的y轴坐标平均值，也就是令$y_{top}=my_{top}$、$y_{bottom}=my_{bottom}$。通过my_{top}与y_{top}、my_{bottom}与y_{bottom}之间的对应关系可以得到当前列的线性变换关系，所述线性变换关系是一个两变量的f(x)＝ax+b形式，从而可以将曲面文字校正为水平。由于在步骤S10中已经对文档图像做过透视校正，所以在这一步逐列变换过程中，不需要将y轴坐标转换为齐次坐标(homogeneous coordinates)，可以在原坐标上做线性变换，降低了计算复杂度，提高了处理效率。逐列进行校正的原理在于：将展平后文档的文字行方向认为是x轴坐标方向，假设文档的起伏只依赖于x轴坐标，即文档的弯曲方式是柱面弯曲，那么每一列像素都可以独立的计算投影关系进而进行校正。

[0066] 所述步骤S50中，如图5所示，II类校正方法具体包括如下步骤。

[0067] 步骤S51：记第i个文字行的第j个采样点为P_{ij}。在所述步骤S233中，对文字行在x轴方向的采样间隔例如设为20像素。

[0068] 步骤S52：计算与P_{ij}对应的校正后点坐标P_{ij}'。具体计算方法为：计算第i个文字行所有采样点P_i的y轴坐标平均值，记为my。将采样点P_{ij}的y轴坐标改为该采样点所在的文字行的所有采样点的y轴坐标平均值my，即得到P_{ij}'。

[0069] 步骤S53：优化曲面参数和投影参数。例如假设文档表面z轴方向为二次样条(quadric spline)函数构成的曲面，则优化过程如下。首先将P_{ij}'转换为齐次坐标表示H_{ij}'。假设H_{ij}'的z轴坐标是随x轴坐标变化的二次样条函数。其次通过投影变换，将H_{ij}'投影到二维平面上，得到Q_{ij}。最后将所有Q_{ij}和P_{ij}的欧氏距离(Euclidean distance)之和，也就是投影误差，作为目标函数。优化投影变换参数以及二次样条函数中的参数，以最小化目标函数，这样就得到了曲面参数(即二次样条函数中的参数)和投影参数(即投影变换参数)。例如采用了拟牛顿法(Quasi-Newton Methods)进行优化。

[0070] 步骤S54：根据步骤S53中的优化参数，可以得到从校正后图像到曲面图像的映射关系。通过图像重映射(remap)的方法，根据曲面参数和投影参数，得到校正后的图像。重映射是图像处理中常用的图像变换手段，方法是遍历目标图上的像素坐标，用映射关系计算得到对应于原图中的像素坐标，通过差值方法得到像素值。具体在本申请中，若记目标图(即校正后的文档图像)上像素坐标为(x,y)，二次样条函数为f(x)，将目标图坐标记为齐次坐标形式$dst=(x,y,f(x))^T$。投影变换参数为$H=(h1^T,h2^T,h3^T)$，其中$h1^T,h2^T,h3^T$为矩阵H的行元素。对应原图中坐标为(x',y')，如公式三、公式四所示。

[0071] $x'=(h1^T*dst)/(h3^T*dst)$ (公式三)。

[0072] $y'=(h2^T*dst)/(h3^T*dst)$ (公式四)。

[0073] 典型的I类版式文档图像如图6所示。I类版式的文档图像的特点是：文字行的长度

占整个文档图像宽度的比例较大，大量文字行的长度都是水平贯穿了整个文档图像。

[0074] 典型的II类版式文档图像如图7所示。II类版式的文档图像的特点是：文字行的长度占整个文档图像宽度的比例较小，大量文字行的长度都只有整个文档图像的一半宽度；文字和图片之间呈现出分栏、混杂的复杂版式。

[0075] 从图6、图7可以发现，本申请进行文档曲面校正的目的是改善透视校正无法处理的文档存在弯曲的情况。图6、图7中左边原始文档图像的文字行均呈现出一定的弯曲，右边校正后文档图像的文字行呈水平，能够说明曲面校正效果理想。

[0076] 请参阅图8，本申请提供的文档曲面校正装置包括初步处理单元10、检测单元20、分类单元30、I类校正单元40和II类校正单元50；与图1所示的文档曲面校正方法相对应。

[0077] 所述初步处理单元10用来对待校正的文档图像进行角点定位和透视校正。

[0078] 所述检测单元20用来在待校正的文档图像中检测文字行。

[0079] 所述分类单元30用来提取文字行特征，还用来将待校正的文档图像分为I类版式和II类版式两类。

[0080] 所述I类校正单元40用来对I类版式的待校正的文档图像进行校正。

[0081] 所述II类校正单元50用来对II类版式的待校正的文档图像进行校正。

[0082] 综上所述，本申请提供了一种文档曲面校正方法及装置。依次进行文档图像的角点定位及透视校正；文字行检测；根据文字行检测结果提取特征，构建分类器对文档进行分类，分为I类版式文档和II类版式文档两类。对于I类版式文档，采用I类校正方法，具体为提取上、下文字行，根据上、下文字行的对应关系，用一维线性变换逐列像素对文档图像进行校正。对II类版式文档，采用II类校正方法，具体为构建弯曲文字行与平直文字行之间的对应关系，通过优化曲面参数和投影参数使平直文字行映射后与弯曲文字行重合，最后通过图像重映射将文档图像进行校正。

[0083] 以上仅为本申请的优选实施例，并不用于限定本申请。对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

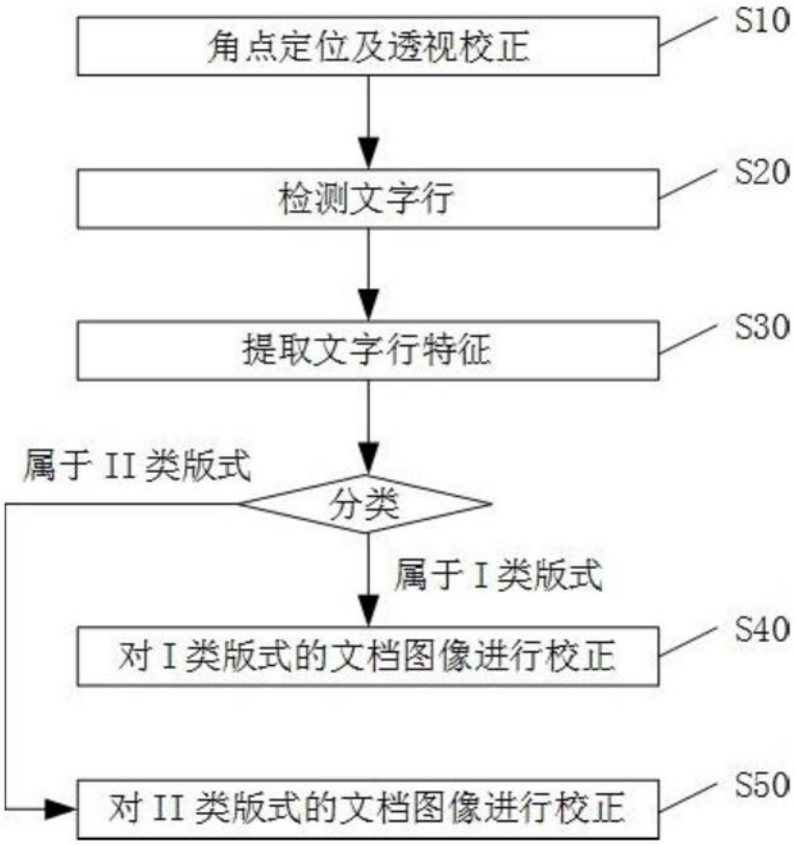

角点定位及透视校正
S10
检测文字行
S20
提取文字行特征
S30
属于 II 类版式
分类
属于 I 类版式
对 I 类版式的文档图像进行校正
S40
对 II 类版式的文档图像进行校正
S50

图1

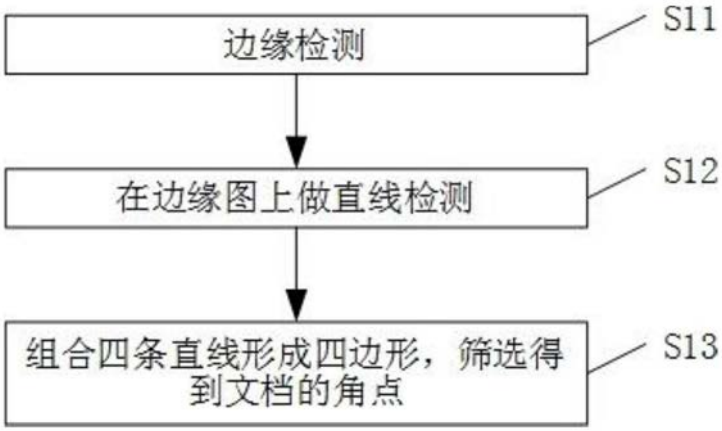

边缘检测
S11
在边缘图上做直线检测
S12
组合四条直线形成四边形，筛选得到文档的角点
S13

图2

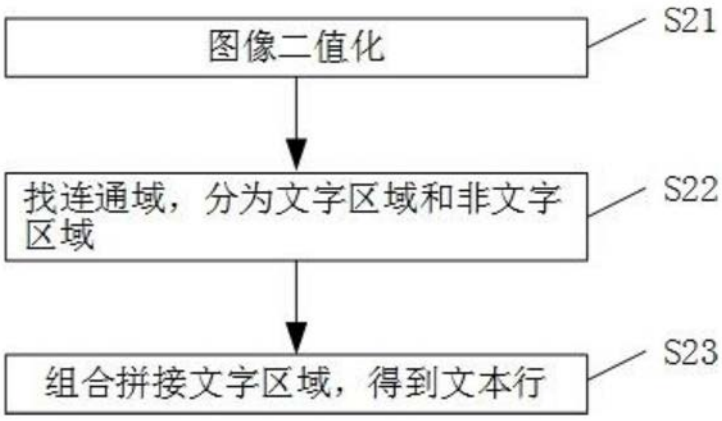

图像二值化
S21
找连通域，分为文字区域和非文字区域
S22
组合拼接文字区域，得到文本行
S23

图3

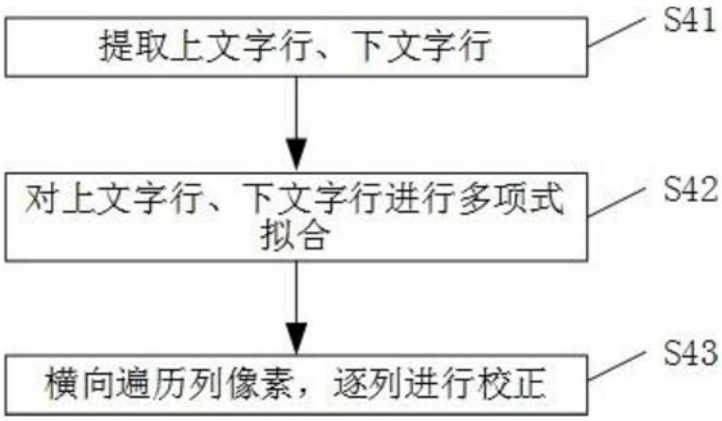

提取上文字行、下文字行
S41
对上文字行、下文字行进行多项式拟合
S42
横向遍历列像素，逐列进行校正
S43

图4

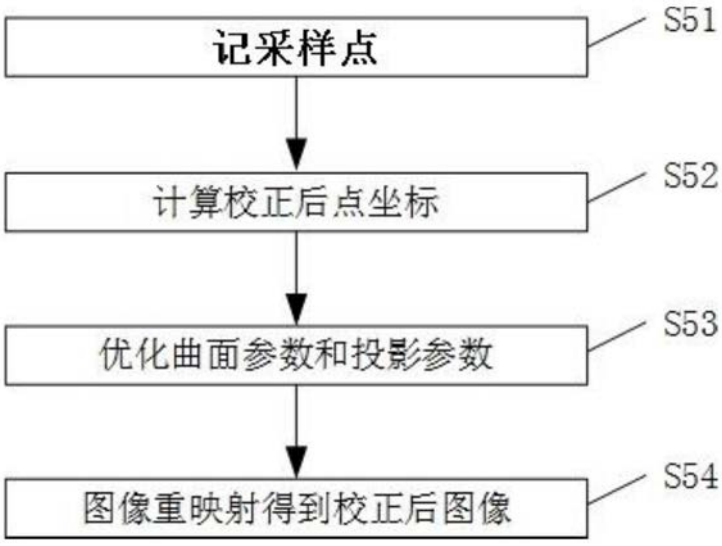
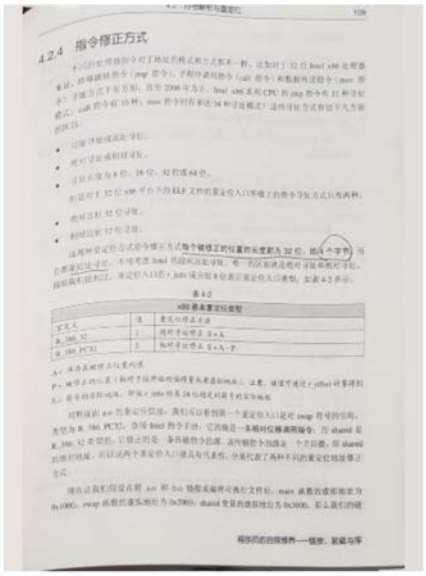
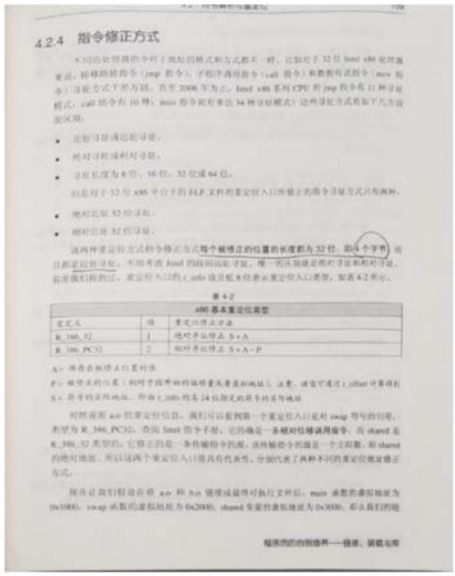

记采样点
S51
计算校正后点坐标
S52
优化曲面参数和投影参数
S53
图像重映射得到校正后图像
S54

图5



校正前



校正后

图6

校正前

校正后

图7

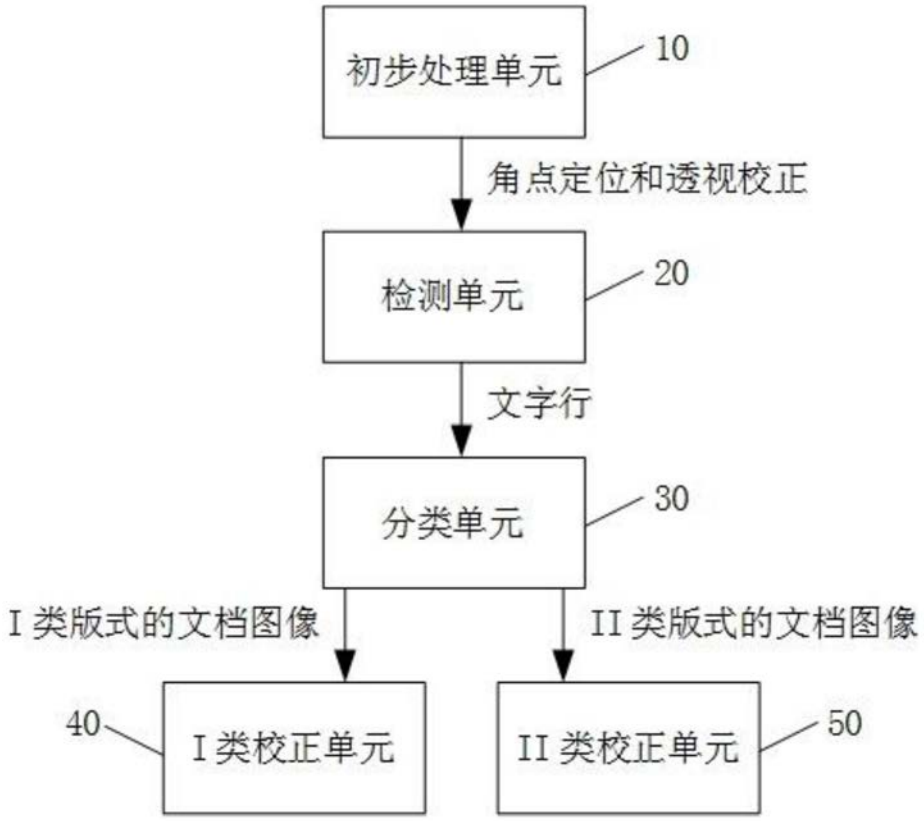

初步处理单元
10
角点定位和透视校正
检测单元
20
文字行
分类单元
30
I 类版式的文档图像
II 类版式的文档图像
40
I 类校正单元
II 类校正单元
50

图8