WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

DOCUMENT MADE AVAILABLE UNDER THE PATENT COOPERATION TREATY (PCT)

International application number:		**PCT/US2016/024069**
International filing date:		**24 March 2016 (24.03.2016)**
Document type:		**Certified copy of priority document**
Document details:	Country/Office:	**CN**
	Number:	**201510133507.9**
	Filing date:	**25 March 2015 (25.03.2015)**
Date of receipt at the International Bureau:		**15 April 2016 (15.04.2016)**

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a),(b) or (b-*bis*)

34, chemin des Colombettes
1211 Geneva 20, Switzerland
www.wipo.int

US16/24069

⑦

中华人民共和国国家知识产权局
STATE INTELLECTUAL PROPERTY OFFICE
OF THE PEOPLE'S REPUBLIC OF CHINA

证　明

本证明之附件是向本局提交的下列专利申请文件副本。

申　　请　　号：201510133507.9

申 请 类 型：发明专利

发明 创造 名称：文本行分类器的生成方法及装置

申　　请　　日：2015年03月25日

申　　请　　人：阿里巴巴集团控股有限公司

发明人或设计人：金炫、王天舟、薛琴

局长
申长雨

2015 年 07 月 16 日

2015101335079

权 利 要 求 书

1、一种文本行分类器的生成方法，其特征在于，包括：

利用当前终端系统字库生成文本行样本；

对所述文本行样本和预存的标注样本进行特征提取；以及

根据提取到的特征进行模型训练，生成文本行分类器，以用于识别文本区域。

2、根据权利要求 1 所述的方法，其特征在于，还包括：

对待识别的图片进行检测，以获得检测结果；以及

使用所述文本行分类器针对所述检测结果输出分数，若所述分数大于预设阈值，则确认待识别的图片为文本区域，若所述分数小于等于预设阈值，则确认待识别的图片为非文本区域。

3、根据权利要求 1 或 2 所述的方法，其特征在于，所述利用当前终端系统字库生成文本行样本，包括：

利用当前终端系统字库生成文字样本，对所述文字样本进行处理，以生成不同类型的文本行样本，其中，所有文本行样本中包含的文字样本均满足以下条件：大小相同、旋转角度相同、字体相同、包含的常用字大于预设比例。

4、根据权利要求 1 或 2 所述的方法，其特征在于，所述对所述文本行样本和预存的标注样本进行特征提取，包括：

提取所述文本行样本对应图片的梯度方向直方图特征、梯度大小直方图特征、像素直方图特征和像素变化特征中的一种或几种；以及

获得所述文本行样本和所述标注样本的连通区域，并提取所述连通区域的特征。

5、根据权利要求 1 或 2 所述的方法，其特征在于，所述根据提取到的特征进行模型训练，生成文本行分类器，包括：

根据提取到的特征生成与文本行样本类型对应的模型，并利用所述标注样本设置所述模型的权重，以生成所述文本行分类器。

6、根据权利要求 4 所述的方法，其特征在于，所述获得所述标注样本的连通区域，包括：

使用第一预设算法获得所述标注样本的连通区域，所述第二预设算法包括最大稳定极值区域 MSER 算法和 MSER 算法的改进算法。

7、根据权利要求 4 所述的方法，其特征在于，所述获得所述连通区域的特征，包括：

使用第二预设算法获得所述连通区域中笔画的宽度特征，所述第二预设算法包括 SWT 算法和 SFT 算法。

8、一种文本行分类器的生成装置，其特征在于，包括：

生成模块，用于利用当前终端系统字库生成文本行样本；

提取模块，用于对所述生成模块生成的文本行样本和预存的标注样本进行特征提取；以及

训练模块，用于根据所述提取模块提取到的特征进行模型训练，生成文本行分类器，以用于识别文本区域。

9、根据权利要求 8 所述的装置，其特征在于，还包括：

检测模块，用于对待识别的图片进行检测，以获得检测结果；以及

识别模块，用于使用所述训练模块训练生成的所述文本行分类器针对所述检测结果输出分数，若所述分数大于预设阈值，则确认待识别的图片为文本区域，若所述分数小于等于预设阈值，则确认待识别的图片为非文本区域。

10、根据权利要求 8 或 9 所述的装置，其特征在于，所述生成模块，具体用于：

利用当前终端系统字库生成文字样本，对所述文字样本进行处理，以生成不同类型的文本行样本，其中，所有文本行样本中包含的文字样本均满足以下条件：大小相同、旋转角度相同、字体相同、包含的常用字大于预设比例。

11、根据权利要求 8 或 9 所述的装置，其特征在于，所述提取模块包括：

第一提取单元，用于提取所述文本行样本对应图片的梯度方向直方图特征、梯度大小直方图特征、像素直方图特征和像素变化特征中的一种或几种；以及

第二提取单元，用于获得所述文本行样本和所述标注样本的连通区域，并提取所述连通区域的特征。

12、根据权利要求 8 或 9 所述的装置，其特征在于，所述训练模块，具体用于：

根据提取到的特征生成与文本行样本类型对应的模型，并利用所述标注样本设置所述模型的权重，以生成所述文本行分类器。

13、根据权利要求 11 所述的装置，其特征在于，所述第二提取单元，具体用于：

使用第一预设算法获得所述标注样本的连通区域，所述第二预设算法包括最大稳定极值区域 MSER 算法和 MSER 算法的改进算法。

14、根据权利要求 11 所述的装置，其特征在于，所述第二提取单元，具体用于：

使用第二预设算法获得所述连通区域中笔画的宽度特征，所述第二预设算法包括 SWT 算法和 SFT 算法。

2015101335079

说　明　书

文本行分类器的生成方法及装置

技术领域

本申请涉及模式识别技术领域，尤其涉及一种文本行分类器的生成方法及装置。

背景技术

目前，很多图片例如淘宝网图片中含有大量违禁文字，为了识别这些违禁文字，可使用自然场景图片的光学字符识别（Optical Character Recognition，OCR）技术对文本检测、定位的结果进行筛选，滤除非文本的检测结果，筛选出候选文本送入识别装置，从而提高识别的准确度。

其中，自然场景的OCR技术一直是工业界和学术研究的热点之一，针对不同的语言，所使用的特征以及算法架构都会有所改变。目前国际上的OCR技术主要针对英文，相对于英文识别，由于中文汉字较为复杂且字符种类较多，汉字偏旁部首的存在也使得单个汉字并非连通区域，识别难度较大。

目前，对于自然场景中的中文OCR的文本区域识别方法分为三类：第一类，利用经验阈值进行分类；第二类，根据不同的应用场景标注大量样本，提取中文文本行经验特征，利用支持向量机（SVM）等分类器进行分类；第三类，利用更为大量的标注正负样本，并利用卷积神经网络（CNN）训练分类器进行分类。

在现有的中文OCR的文本区域识别中，使用经验阈值进行分类的方法最为简单，其进行判断的特征多来自于单字符验证提取的文字特征，但是该算法准确率较低且鲁棒性较差，容易出现过拟合现象；第二类方法是目前比较主流的方案，第三类方法的使用并不多见，主要原因在于CNN方法会消耗过多的计算资源，影响算法总体效率，但是，无论是第二类方法还是第三类方法，都需要标注大量样本，这必然会耗费大量的人力成本，且分类效果依赖于特征的提取以及样本的选取，因此对于不同的应用需求往往需要重新标注一批新的业务依赖数据，即新的样本，故现有的标注样本适用性差，不仅如此，中文文字的字体多样、样式复杂，还包括简体、繁体以及手写体等多个类型，使得文本行的多样性异常丰富，也无疑大大增加了中文文本区域的识别难度。

因此，迫切需要提供一种适用性强、简单、有效的中文OCR文本区域识别方法。

发明内容

本申请旨在至少在一定程度上解决相关技术中的技术问题之一。

为此，本申请的第一个目的在于提出一种文本行分类器的生成方法，该方法可以针对不同场景或不同需求进行文本区域识别，适用性强、应用范围广且实现简单。

本申请的第二个目的在于提出一种文本行分类器的生成装置。

为达上述目的，本申请第一方面实施例提出了一种文本行分类器的生成方法，该文本行分类器的生成方法包括：利用当前终端系统字库生成文本行样本；对文本行样本和预存的标注样本进行特征提取；以及根据提取到的特征进行模型训练，生成文本行分类器，以用于识别文本区域。

本申请实施例的文本行分类器的生成方法，基于系统字库生成文本行样本的方式，使得生成的文本行分类器可以针对不同场景或不同需求进行文本区域识别，适用性强、应用范围广且实现简单，同时结合标注样本进行文本行样本特征提取的方式使得生成的文本行分类器的准确率高。

为达上述目的，本申请第二方面实施例提出了一种文本行分类器的生成装置，该文本行分类器的生成装置包括：生成模块，用于利用当前终端系统字库生成文本行样本；提取模块，用于对生成模块生成的文本行样本和预存的标注样本进行特征提取；以及训练模块，用于根据提取模块提取到的特征进行模型训练，生成文本行分类器，以用于识别文本区域。

本申请实施例的文本行分类器的生成装置，通过生成模块生成文本行样本的方式，使得生成的文本行分类器可以针对不同场景或不同需求进行文本区域识别，适用性强、应用范围广且实现简单，通过提取模块结合标注样本进行文本行样本特征提取的方式使得生成的文本行分类器的准确率高。

附图说明

图 1a 是本申请一个实施例文本行分类器的生成方法的流程图。

图 1b 是本申请一个实施例的单字样本示意图。

图 1c 是本申请一个实施例的文本行样本示意图。

图 2a 是本申请一个实施例的生成文本行分类器的细化流程图。

图 2b 是本申请一个实施例的生成文本样本的流程图。

图 2c 是本申请一个实施例的利用文字样本生成文本行样本的流程图。

图 3 是本申请一个实施例利用 BP 神经网络进行特征训练的流程图。

图 4 是本申请另一个实施例文本区域的识别方法的流程图。

图 5 是本申请一个实施例文本行分类器的生成装置的结构示意图。

图 6 是本申请另一个实施例文本行分类器的生成装置的结构示意图。

具体实施方式

下面详细描述本申请的实施例，所述实施例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元件。下面通过参考附图描述的实施例是示例性的，旨在用于解释本申请，而不能理解为对本申请的限制。

下面参考附图描述本申请实施例的文本行分类器的生成方法及装置。

图 1a 是本申请一个实施例文本行分类器的生成方法的流程图。

如图 1a 所示，该文本行分类器的生成方法包括：

S101，利用当前终端系统字库生成文本行样本。

在该实施例中，利用当前终端系统字库生成文本行样本可以包括：利用当前终端系统字库生成文字样本，然后对文字样本进行处理，从而生成不同类型的文本行样本。

具体地，利用当前终端系统字库生成文字样本可以为：从字库提取不同字体的文字，加入间距、旋转、大小、噪声等扰动，从而生成文字样本；对文字样本进行处理，从而生成不同类型的文本行样本可以为：基于生成的文字样本，将同一字体的文字随机搭配，加入扰动后形成不同类型的文本行样本。

例如，从字库提取文字后，加入不同扰动生成的单字样本如图 1b 所示，需要说明的是，图 1b 仅为一个示例。又例如，基于文字样本可以生成的文本行样本如图 1c 所示。

S102，对文本行样本和预存的标注样本进行特征提取。

在该实施例中，在对文本行样本和预存的标注样本进行特征提取之前，还可以包括：保存标注样本。具体地，可以利用检测算法切出候选文本区域，人工对候选文本区域进行标注，即可以通过将候选文本区域标注为 1 或 0 来标识其是否为文本区域。

在生成文本行样本和保存标注样本之后，可以对这些样本进行特征提取，具体地，可以提取文本行样本对应图片的梯度方向直方图特征、梯度大小直方图特征、像素直方图特征和像素变化特征中的一种或几种；以及获得文本行样本和标注样本的连通区域，并提取上述连通区域的特征。由此可见，本发明实施例在提取特征时可以提取至少两个特征，即一组特征，从而有利于分类器的生成。

S103，根据提取到的特征进行模型训练，生成文本行分类器，以用于识别文本区域。

在该实施例中，可以利用反向传播（BP，Back Propagation）神经网络对提取的特征进行单模型训练，由于每种类型的文本行样本可以训练出一个模型，故多种类型的文本行样本可以训练出多个模型，每个模型可以作为一棵决策树，初始设置每棵决策树的权重，然后利用一部分标注样本对决策树进行权重训练，使得每棵决策树可以获得合适的权重以保证分类的准确性，通过上述过程，可以生成文本行分类器。利用上述文本行分类器可以识

别文本区域，进而可以识别出含有违禁文字的图片。

假设一个文本行样本提取的特征可以用一个向量表示，即一个文本行样本对应的图片可以生成一维向量 X，则所有文本行样本可以生成向量集合{X（i）}，针对所有文本行样本进行模型训练的过程可以为：向量输入 BP 神经网络进行训练，得到训练模型 Model（X），即文本行分类器。

使用上述文本行分类器进行文本区域识别的过程可以为：针对当前待识别别图片提取特征向量 Y，然后将 Y 输入训练模型（即文本行分类器），上述文本行分类器输出分数 score=Model（Y），若 score 大于预设阈值，则判定为文本行，否则，判定为背景图片。

由于该实施例是利用系统字库生成文本行样本，而非通过人工标注大量文本行样本，因此，使得训练样本覆盖更全面，可以适用于不同的场景或应用需求，适用性强；而结合标注样本提取特征以生成文本行分类器，是为了提高分类的有效性和准确性。

上述文本行分类器的生成方法，利用当前终端系统字库生成文本行样本，并对文本行样本和预存的标注样本进行特征提取，然后根据提取到的特征进行模型训练，以生成用于识别文本区域的文本行分类器；上述基于系统字库生成文本行样本的方式，使得生成的文本行分类器可以针对不同场景或不同需求进行文本区域识别，适用性强、应用范围广且实现简单，同时结合标注样本进行文本行样本特征提取的方式使得生成的文本行分类器的准确率高。

图 2a 是本申请一个实施例的生成文本行分类器的细化流程图。

如图 2a 所示，生成文本行分类器包括：

S201，利用当前终端系统字库生成文字样本。

其中，文字样本可以为汉字样本。

具体地，如图 2b 所示，S201 可以包括以下几个步骤：S2011，从 windows 自带字库、方正字库、手写字库等字库提取单个汉字；S2012，每个汉字取用{5,5.5,6,6.5,7,7.5,8,9,10,10.5,11,12,14,16,18,20,22,24,26,28,36,48,72}（磅）21 个大小；S2013，每个汉字生成斜体的形变和非形变；S2014，每个汉字生成-10°至 10°的旋转扰动。通过上述操作，可以生成汉字样本。

S202，利用文字样本生成文本行样本。

如图 2c 所示，利用 S201 生成的文字样本生成文本行样本的过程可以包括以下几个步骤：S2021、每个文本行留出 5 到 10 个像素的边缘间隔；S2022、每个文本行中的每个汉字间隔取用汉字宽度的{1,2，..,10}/10；S2023、每个文本行取用汉字个数为 2 至 10；S2024、每个文本行分为水平文本行、竖直文本行、45°斜文本行和-45°斜文本行；S2025、每个文本

行加入-10°至10°的旋转扰动；S2026、每个文本行采用多组高斯低通滤波器进行模糊处理。经过上述步骤可以生成文本行样本，但是文本行样本需要满足以下几个限制：每个文本行样本中汉字的大小必须相同；每个文本行样本中汉字的旋转角度必须相同；每个文本行样本中的汉字必须同时满足斜体或非斜体；每个文本行样本中汉字的字体必须相同；每个文本行样本中包含的常用汉字必须占预设比例例如80%以上。

S203，结合预存的标注样本进行文本行特征提取。

具体地，S203可以包括：

（1）提取文本行样本对应图片的梯度方向直方图特征、梯度大小直方图特征、像素直方图特征和像素变化特征中的一种或几种，下面详细介绍如何提取这些特征：

A）梯度方向直方图特征

该特征计算方式描述如下：

$$\theta(x,y) = \arctan\left(\frac{d_y(x,y)}{d_x(x,y)}\right)$$

其中，$d_x(x,y) = p(x+1,y) - p(x,y)$，$d_y(x,y) = p(x,y+1) - p(x,y)$，$p(x,y)$为对应像素位置的像素值，$d_x$为对应像素位置的$x$方向梯度，$d_y$为对应像素位置的$y$方向梯度。

B）梯度大小直方图特征

$$Gv(x,y) = \sqrt{d_x(x,y)^2 + d_y(x,y)^2}$$

d_x和d_y与之前描述相同。

C）像素直方图

对文本行样本对应区域内的像素进行统计，将像素按照像素值大小等分为8个区间段，每段大小为32像素，以区间段的像素个数大小作为特征，共输出8维特征。

D）像素变化特征

在该实施例中，像素变化特征可以通过以下两种方式提取：

a.利用投影法计算文本的主轴方向，根据主轴方向将文本等分为“田”字区域，计算每个区域的连通区域的像素个数以及方差；

b.以水平线为基线，计算基线经过像素的差，即d_x，统计$d_x > 0$的个数。

（2）获得文本行样本和标注样本的连通区域，并提取连通区域的特征。

具体地，对于文本行样本，可以直接利用阈值二值化，得到连通区域；对于标注样本，可以通过最大稳定极值区域（MSER）算法或MSER改进算法提取得到其连通区域。然后

利用笔画宽度变换（Stroke Width Transform，SWT）算法或笔画特征变换（Stroke FeatureTransform，SFT）算法计算出连通区域的笔画宽度，并可取用笔画宽度的平均值、最小值、最大值和方差等维度。另外，也可以计算每个连通区域的转折点个数以及连通区域内的孔洞个数。其中，SFT 算法为 SWT 算法的改进算法，相对于 SWT 算法而言，其引入了颜色通道，加强了边界的约束机制，对于背景的干扰鲁棒性更强。由于本申请实施例中的文本行样本是基于字库生成，而非通过人工进行标注的样本，且提取的特征与现有技术不同，具体地，本发明实施例在提取特征时可以提取至少两个特征，即一组特征，从而有利于分类器的生成。

需要说明的是，除了可以采用以上方式提取特征，还可以采用其他方式例如卷积神经网络（CNN）的特征学习过程进行提取。

S204，利用提取的特征进行模型训练。

可以采用 BP 神经网络根据提取到的特征生成与文本行样本类型对应的模型，也即生成的模型的个数与文本行样本类型的个数相同。

S205，结合预存的标注样本设置模型的权重，以生成文本行分类器。

在该实施例中，首先为生成的多个模型分配一个权重，然后利用标注样本修改模型的权重，以生成文本行分类器。

具体地，上述 S204 和 S205 的实现过程可参见图 3：

1）利用 BP 神经网络对提取的特征进行训练，设置输入层特征维度 64，隐藏层 128，每一种类型的文本行训练出一个模型；

2）假设训练完成后共有 N 个模型，每个模型单独作为一个决策树分类器输出，设置每个模型的初始权重为 1/N；

3）利用标注样本进行权重训练：针对每一次分类，若分类错误，对应的决策树分类器权重减去分类器输出的分值；若分类正确，对应决策树分类器权重加上分类器输出的分值。

4）对分类器权重归一化，使分类器权重和为 1。

通过图 3 的过程可以为生成的多个模型设置合适的权重，故可以提高分类器的分类准确率。

由此可见，通过图 2 所示的生成过程可以简单、有效地生成适用性强且准确率高的文本行分类器。

图 4 是本申请另一个实施例文本区域的识别方法的流程图。

如图 4 所示，该文本区域的识别方法包括：

S401，对待识别的图片进行检测，以获得检测结果。

具体地，可以对待识别的图片进行 OCR 检测，获得检测结果。

S402，使用文本行分类器针对上述检测结果输出分数，若分数大于预设阈值，则确认待识别的图片为文本区域，若分数小于等于预设阈值，则确认待识别的图片为非文本区域。

具体地，可将检测结果输入采用本申请图 2 所示实施例生成的文本行分类器中，由文本行分类器输出分数，分数越高代表其为文本的可能性越大，也可以根据经验设置一个阈值，若分数大于预设阈值，则确认待识别的图片为文本区域，否则，确认待识别的图片为非文本区域。

在确认待识别的图片为文本区域后，可以进一步识别图片中的文字，以识别图片中是否存在违禁文字。

由于通过图 2 生成的文本行分类器可以针对不同场景或不同需求进行文本区域识别，适用性强、应用范围广且实现简单，同时结合标注样本进行文本行样本特征提取的方式使得生成的文本行分类器的准确率高，因此，使用图 2 所示方法生成的文本行分类器可以简单、有效、准确地识别出不同场景下的图片是否为文本区域，进而可以识别出对应图片中是否包含违禁文字，为如何处理对应图片提供依据。

图 5 是本申请一个实施例文本行分类器的生成装置的结构示意图。

如图 5 所示，该文本行分类器的生成装置包括生成模块 51、提取模块 52 和训练模块 53，其中：

生成模块 51 用于利用当前终端系统字库生成文本行样本；提取模块 52 用于对生成模块 51 生成的文本行样本和预存的标注样本进行特征提取；训练模块 53 用于根据提取模块 52 提取到的特征进行模型训练，生成文本行分类器，以用于识别文本区域。

具体地，生成模块 51 可以用于：利用当前终端系统字库生成文字样本，对上述文字样本进行处理，以生成不同类型的文本行样本。其中，利用当前终端系统字库生成文字样本可以为：从字库提取不同字体的文字，加入间距、旋转、大小、噪声等扰动，从而生成文字样本；对文字样本进行处理，从而生成不同类型的文本行样本可以为：基于生成的文字样本，将同一字体的文字随机搭配，加入扰动后形成不同类型的文本行样本。

更具体地，生成文字样本可以包括以下几个步骤：1）利用 windows 自带字库、方正字库、手写字库等，提取单个汉字；2）每个汉字取用{5,5.5,6,6.5,7,7.5,8,9,10,10.5,11,12,14,16,18,20,22,24,26,28,36,48,72}（磅）21 个大小；3）每个汉字生成斜体的形变和非形变；4）每个汉字生成-10°至 10°的旋转扰动。利用上述文字样本生成文本行样本的过程可以包括以下几个步骤：1）每个文本行留出边缘间隔 5 到 10 个像素；2）每个文本行中的每个汉字间隔取用汉字宽度的{1,2，..,10}/10；3）每个文本行

取用汉字个数为 2 至 10；4）每个文本行分为水平文本行、竖直文本行、45°斜文本行和-45°斜文本行；5）每个文本行加入-10°至 10°的旋转扰动；6）每个文本行采用多组高斯低通滤波器进行模糊处理。经过上述步骤可以生成文本行样本，但是文本行样本需要满足以下几个限制：每个文本行样本中汉字的大小必须相同；每个文本行样本中汉字的旋转角度必须相同；每个文本行样本中的汉字必须同时满足斜体或非斜体；每个文本行样本中汉字的字体必须相同；每个文本行样本中包含的常用汉字必须占预设比例例如 80%以上。

在该实施例中，在提取模块 52 对文本行样本和预存的标注样本进行特征提取之前，还需要保存标注样本。具体地，人工标注样本的过程为：可以利用检测算法切出候选文本区域，人工对候选文本区域进行标注，即可以通过将候选文本区域标注为 1 或 0 来标识其是否为文本区域。

为了提取样本特征，提取模块 52 可以包括第一提取单元 521 和第二提取单元 522，其中：第一提取单元 511 用于提取文本行样本对应图片的梯度方向直方图特征、梯度大小直方图特征、像素直方图特征和像素变化特征中的一种或几种；第二提取单元 522 用于获得上述文本行样本和上述标注样本的连通区域，并提取连通区域的特征。

具体地，第一提取单元 511 和第二提取单元 522 提取特征的详细过程可以参见图 2 所示实施例中步骤 S203 中的相应部分，此处不赘述。

其中，训练模块 53 用于：根据提取到的特征生成与文本行样本类型对应的模型，并利用上述标注样本设置上述模型的权重，以生成上述文本行分类器。具体地，可以利用反向传播（BP，Back Propagation）神经网络对提取的特征进行单模型训练，由于每种类型的文本行样本可以训练出一个模型，故多种类型的文本行样本可以训练出多个模型，每个模型可以作为一棵决策树，初始设置每棵决策树的权重，然后利用一部分标注样本对决策树进行权重训练，使得每棵决策树可以获得合适的权重以保证分类的准确性，通过上述过程，可以生成文本行分类器，具体的实现过程可参见图 3。

由于该实施例是利用系统字库生成文本行样本，而非通过人工标注大量文本行样本，因此，使得训练样本覆盖更全面，可以适用于不同的场景或应用需求，适用性强；而结合标注样本提取特征以生成文本行分类器，是为了提高分类的有效性和准确性。

另外，该装置还可以包括检测模块 54 和识别模块 55，如图 6 所示，检测模块 54 用于对待识别的图片进行检测，以获得检测结果；识别模块 55 用于使用上述训练模块 53 训练生成的上述文本行分类器针对检测模块 54 输出的检测结果输出分数，若上述分数大于预设阈值，则确认待识别的图片为文本区域，若上述分数小于等于预设阈值，则确认待识别的图片为非文本区域。在确认待识别的图片为文本区域后，可以进一步识别图片中的文字，以识别图片中是否存在违禁文字。

由于经过生成模块51、提取模块52和训练模块53生成的文本行分类器可以针对不同场景或不同需求进行文本区域识别，适用性强、应用范围广且实现简单，同时结合标注样本进行文本行样本特征提取的方式使得生成的文本行分类器的准确率高，因此，使用本申请实施例生成的文本行分类器可以简单、有效、准确地识别出不同场景下的图片是否为文本区域，进而可以识别出对应图片中是否包含违禁文字，为如何处理对应图片提供依据。

上述文本行分类器的生成装置，通过生成模块生成文本行样本的方式，使得生成的文本行分类器可以针对不同场景或不同需求进行文本区域识别，适用性强、应用范围广且实现简单，通过提取模块结合标注样本进行文本行样本特征提取的方式使得生成的文本行分类器的准确率高。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本申请的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本申请的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本申请的实施例所属技术领域的技术人员所理解。

在流程图中表示或在此以其他方式描述的逻辑和/或步骤，例如，可以被认为是用于实现逻辑功能的可执行指令的定序列表，可以具体实现在任何计算机可读介质中，以供指令执行系统、装置或设备（如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、装置或设备取指令并执行指令的系统）使用，或结合这些指令执行系统、装置或设备而使用。就本说明书而言，"计算机可读介质"可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用

的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为可以例如通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

应当理解，本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本申请各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

上述提到的存储介质可以是只读存储器，磁盘或光盘等。尽管上面已经示出和描述了本申请的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本申请的限制，本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

说 明 书 附 图

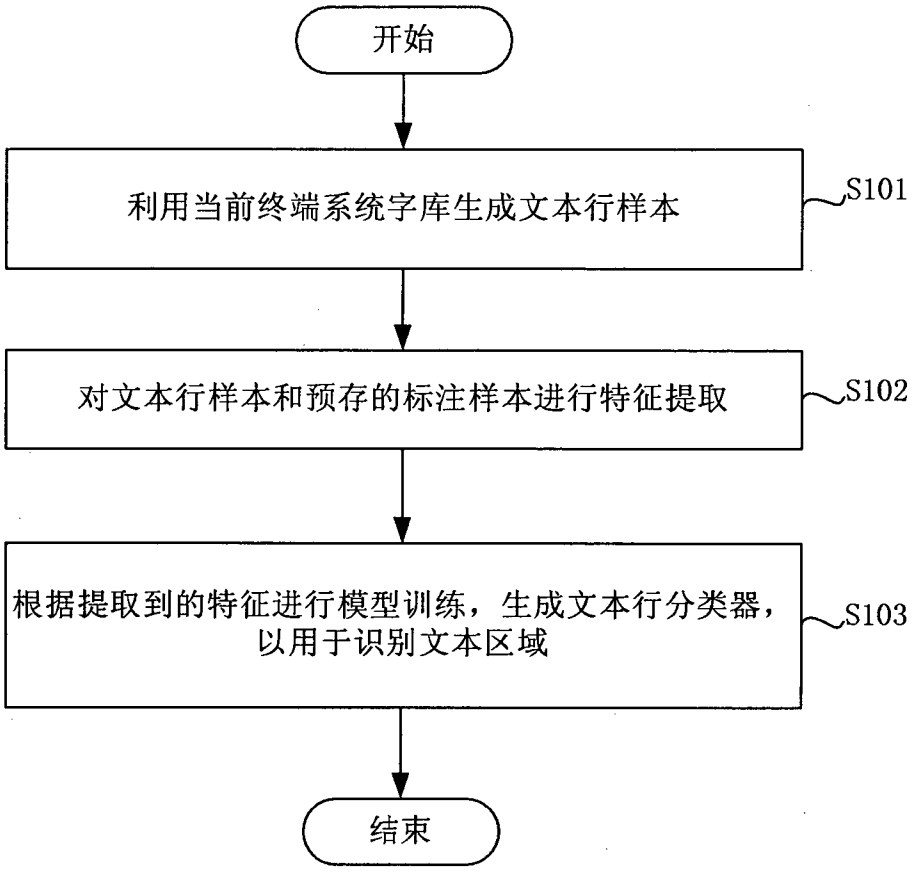

图 1a

图 1b

图盂诵悛军汛雷

图 1c

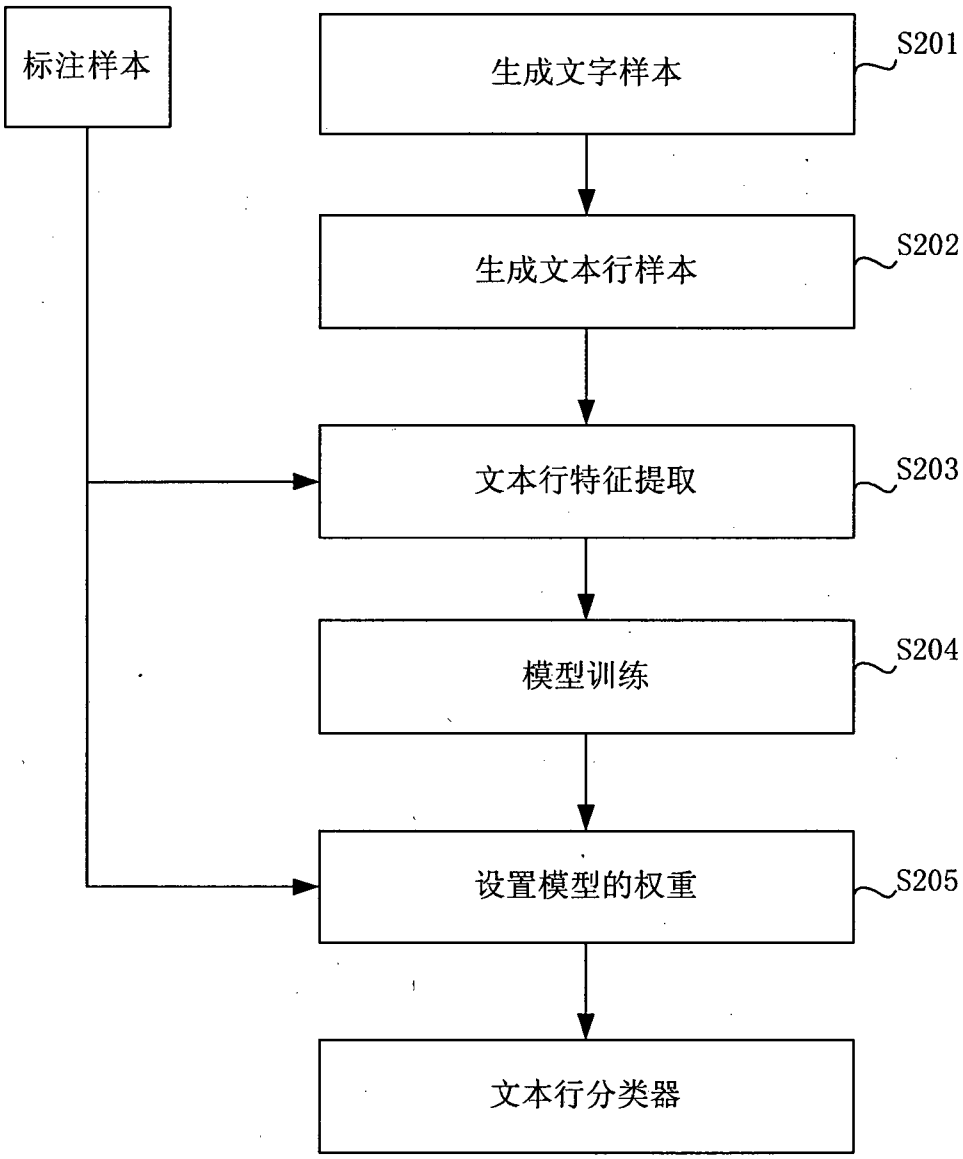

图 2a

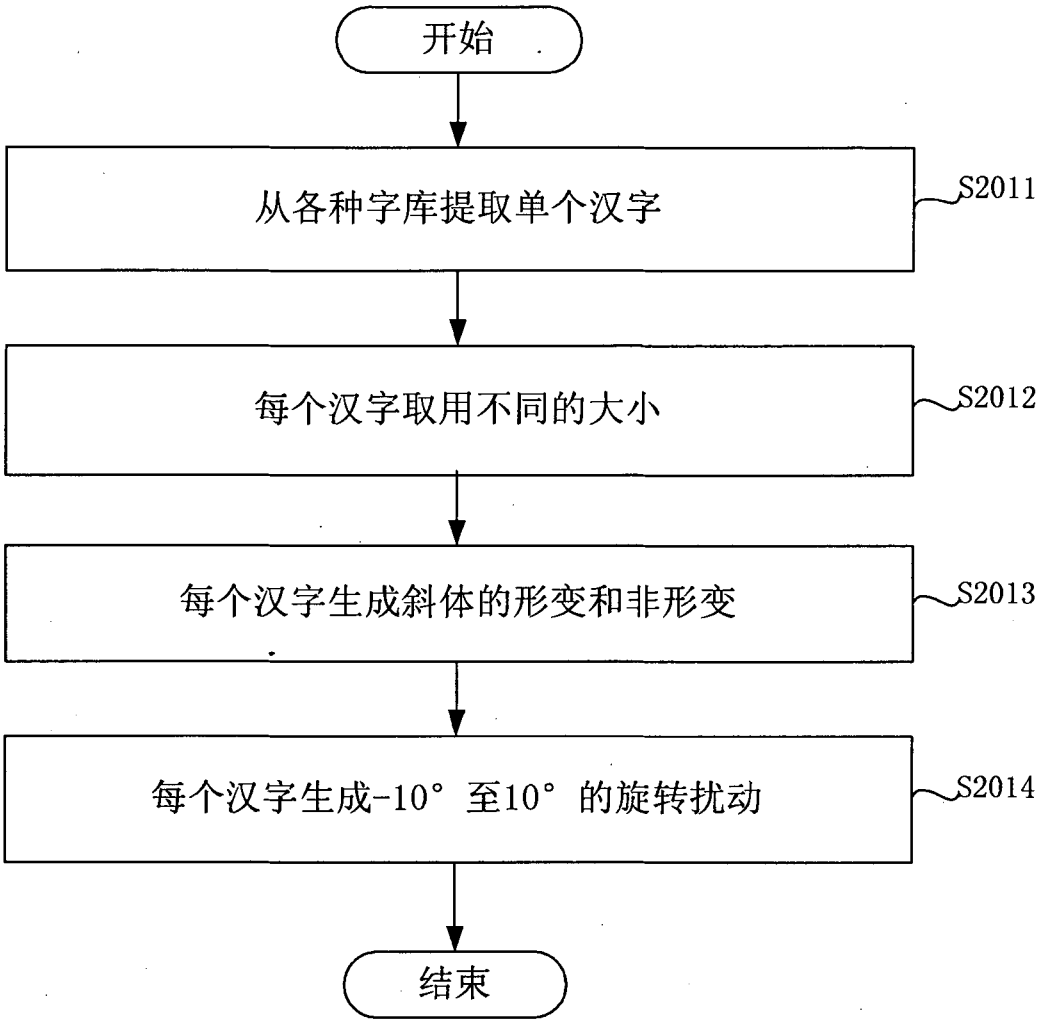
开始
从各种字库提取单个汉字
S2011
每个汉字取用不同的大小
S2012
每个汉字生成斜体的形变和非形变
S2013
每个汉字生成-10°至10°的旋转扰动
S2014
结束

图 2b

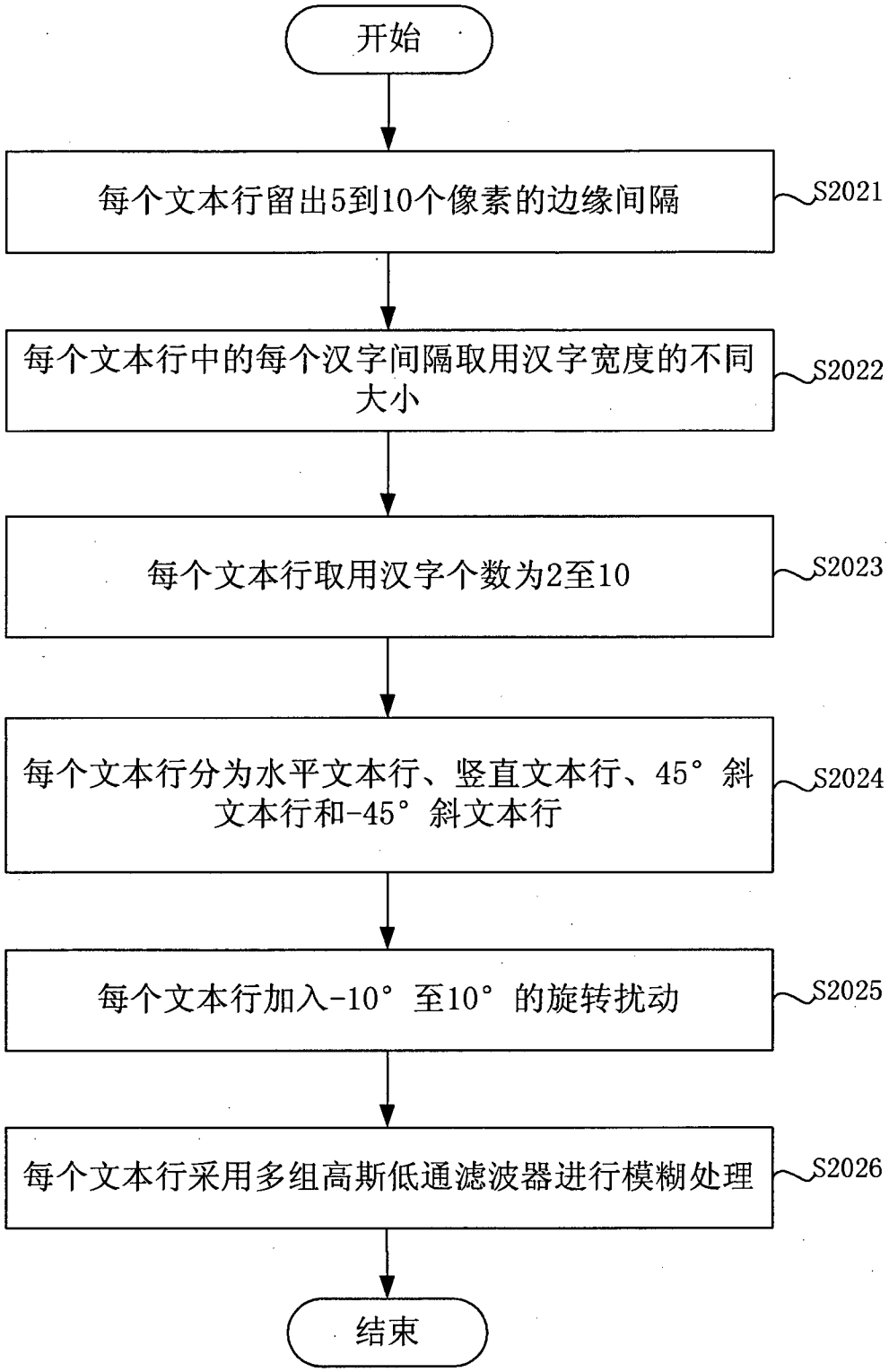
开始
每个文本行留出5到10个像素的边缘间隔
S2021
每个文本行中的每个汉字间隔取用汉字宽度的不同大小
S2022
每个文本行取用汉字个数为2至10
S2023
每个文本行分为水平文本行、竖直文本行、45°斜文本行和-45°斜文本行
S2024
每个文本行加入-10°至10°的旋转扰动
S2025
每个文本行采用多组高斯低通滤波器进行模糊处理
S2026
结束

图 2c

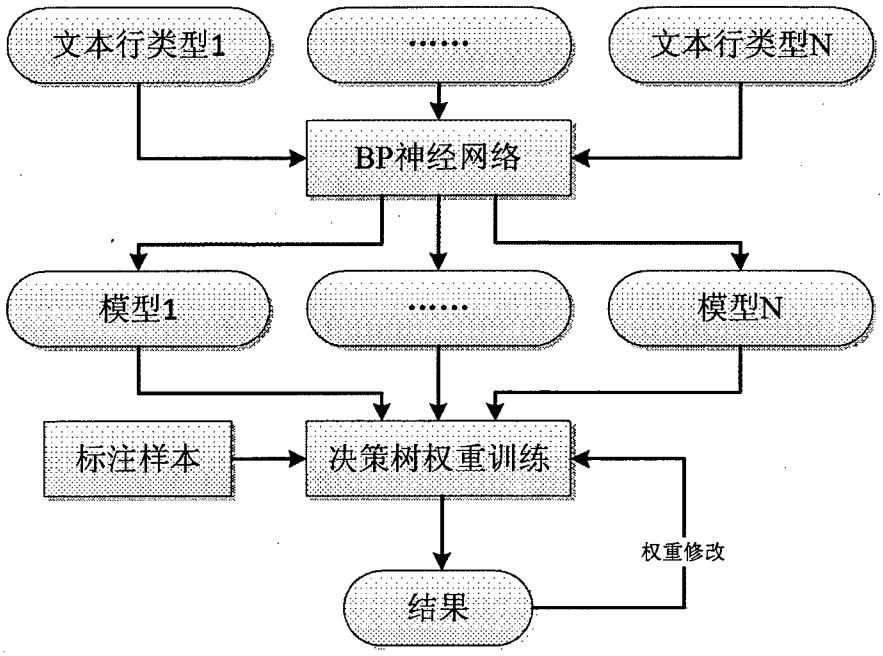
文本行类型1
……
文本行类型N
BP神经网络
模型1
……
模型N
标注样本
决策树权重训练
权重修改
结果

图 3

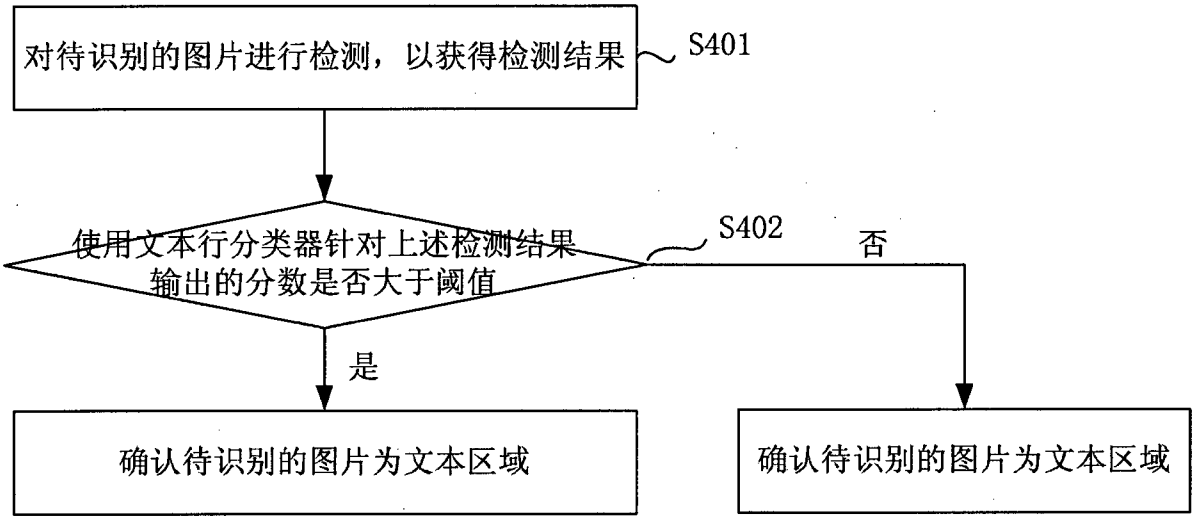
对待识别的图片进行检测，以获得检测结果
S401
使用文本行分类器针对上述检测结果输出的分数是否大于阈值
S402
否
是
确认待识别的图片为文本区域
确认待识别的图片为文本区域

图 4

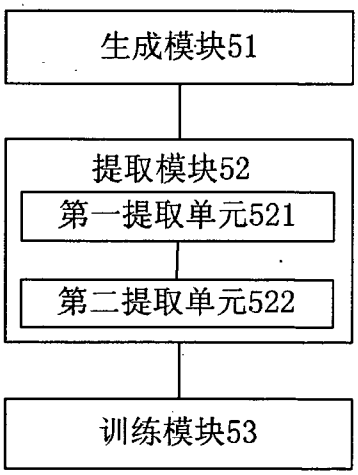
生成模块51
提取模块52
第一提取单元521
第二提取单元522
训练模块53

图 5

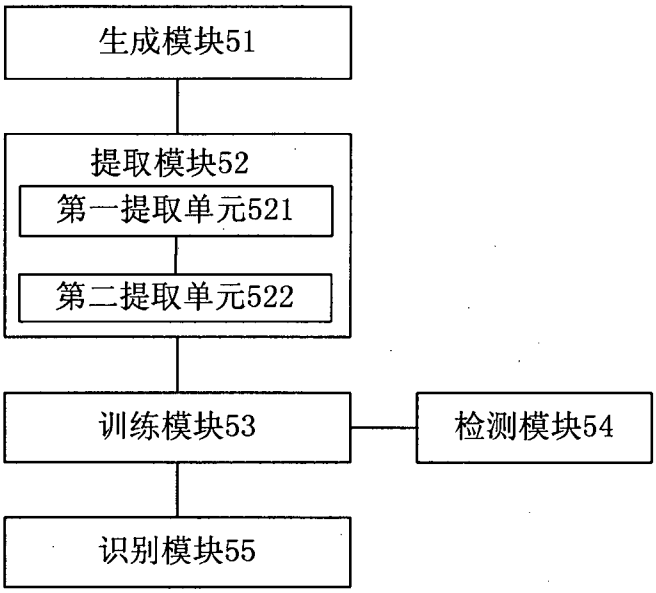
生成模块51
提取模块52
第一提取单元521
第二提取单元522
训练模块53
检测模块54
识别模块55

图 6