

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 4 月 30 日 (30.04.2020)



(10) 国际公布号
WO 2020/082562 A1

- (51) 国际专利分类号:
G06K 9/68 (2006.01)
- (21) 国际申请号: PCT/CN2018/122832
- (22) 国际申请日: 2018 年 12 月 21 日 (21.12.2018)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201811254944.6 2018 年 10 月 25 日 (25.10.2018) CN
- (71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。
- (72) 发明人: 周罡(ZHOU, Gang); 中国广东省深圳市
福田区福田街道福安社区益田路 5033 号平

安金融中心 23 楼, Guangdong 518000 (CN)。王彬(WANG, Bin); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市世纪恒程知识产权代理事务所(CENFO INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市南山区粤海街道高新技术产业园北区松坪山路 3 号奥特讯电力大厦 201, Guangdong 518057 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,

(54) Title: SYMBOL IDENTIFICATION METHOD, APPARATUS, DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 字符识别方法、装置、设备及存储介质

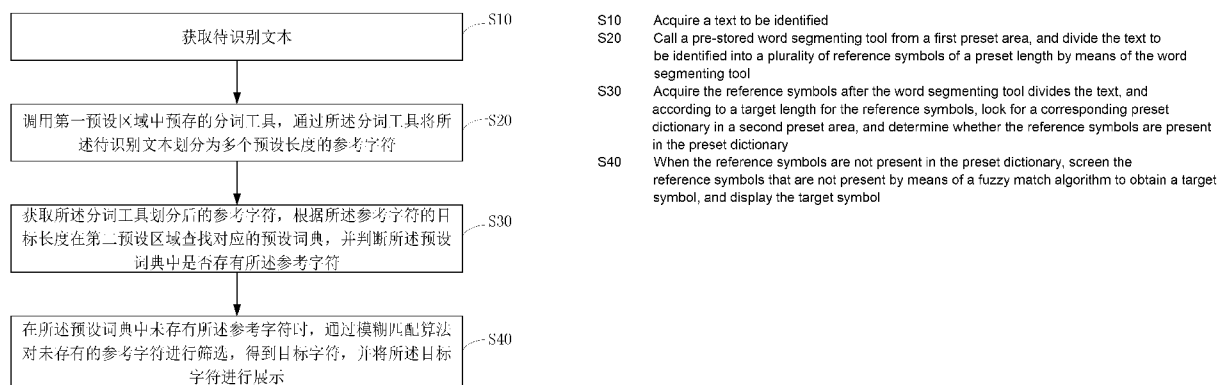


图 2

(57) Abstract: A big data processing-based symbol identification method, apparatus, device, and storage medium, the method comprising: acquiring a text to be identified (S10); calling a pre-stored word segmenting tool from a first preset area so that the word segmenting tool divides the text to be identified into a plurality of reference symbols of a preset length (S20); according to a target length for the reference symbols, looking for a corresponding preset dictionary in a second preset area, and determining whether the reference symbols are present in the preset dictionary (S30); and when the reference symbols are not present in the preset dictionary, screening a target symbol from the reference symbols that are not present by means of a fuzzy match algorithm (S40).

(57) 摘要: 一种基于大数据处理的字符识别方法、装置、设备及存储介质, 所述方法包括: 获取待识别文本 (S10); 调用第一预设区域中预存的分词工具, 以使所述分词工具将待识别文本划分为多个预设长度的参考字符 (S20); 根据所述参考字符的目标长度在第二预设区域查找对应的预设词典, 并判断所述预设词典中是否存有所述参考字符 (S30); 在所述预设词典中未存有所述参考字符时, 将未存有的参考字符通过模糊匹配算法筛选出目标字符 (S40)。

PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

- (84) 指定国 (除另有指明，要求每一种可提供的地区
保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

- 包括国际检索报告 (条约第21条(3))。

字符识别方法、装置、设备及存储介质

- [1] 本申请要求于2018年10月25日提交中国专利局、申请号为201811254944.6、发明名称为“字符识别方法、装置、设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在申请中。
- [2] 技术领域
- [3] 本申请涉及文本识别技术领域，尤其涉及一种字符识别方法、装置、设备及存储介质。
- [4] 背景技术
- [5] 目前，光学字符识别（Optical Character Recognition，OCR）主要是通过电子设备，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，在面对印刷体字符时，采用光学的方式将纸质文档中的文字转换成为黑白点阵的图像文件，并通过识别软件将图像中的文字转换成文本格式，供文字处理软件进一步编辑加工，但是，在字符识别过程中通常使用概率统计方法进行识别时，常常识别速度较低。
- [6] 发明内容
- [7] 本申请的主要目的在于提出一种字符识别方法、装置、设备及存储介质，旨在提高文本识别效率。
- [8] 为实现上述目的，本申请提供一种字符识别方法，所述字符识别方法包括以下步骤：
- [9] 获取待识别文本；
- [10] 调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符；
- [11] 获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符；
- [12] 在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考

字符进行筛选，得到目标字符，并将所述目标字符进行展示。

[13] 此外，为实现上述目的，本申请还提出一种字符识别装置，所述字符识别装置包括：

[14] 获取模块，用于获取待识别文本；

[15] 调用模块，用于调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符；

[16] 查找模块，用于获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符；

[17] 筛选模块，用于在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示。

[18] 此外，为实现上述目的，本申请还提出一种设备，所述设备包括：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可读指令，所述计算机可读指令配置为实现如上所述的字符识别方法的步骤。

[19] 此外，为实现上述目的，本申请还提出一种存储介质，所述存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如上文所述的字符识别方法的步骤。

[20] 本申请提出的字符识别方法，通过获取待识别文本，调用分词工具，以使所述分词工具将所述待识别文本划分为多个预设长度的字符，根据所述预设长度的字符查找对应的预设词典，判断所述预设词典中是否存有所述字符，在所述预设词典中未存有所述字符时，说明所述字符存在识别异常的情况，在这种情况下，将未存有的字符通过模糊匹配算法筛选出目标字符，从而通过模糊匹配算法实现文字识别，提高文字识别效率。

[21] 附图说明

[22] 图1是本申请实施例方案涉及的硬件运行环境的设备结构示意图；

[23] 图2为本申请字符识别方法第一实施例的流程示意图；

[24] 图3为本申请字符识别方法第二实施例的流程示意图；

[25] 图4为本申请字符识别方法第三实施例的流程示意图；

[26] 图5为本申请字符识别装置第一实施例的功能模块示意图。

[27] 本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

[28] 具体实施方式

[29] 应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[30] 参照图1，图1为本申请实施例方案涉及的硬件运行环境的设备结构示意图。

[31] 如图1所示，该设备可以包括：处理器1001，例如CPU，通信总线1002、用户接口1003，网络接口1004，存储器1005。其中，通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏（Display）、输入单元比如按键，可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口（如WI-FI接口）。存储器1005可以是高速RAM存储器，也可以是稳定的存储器（non-volatile memory），例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。

[32] 本领域技术人员可以理解，图1中示出的设备结构并不构成对设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

[33] 如图1所示，作为一种存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及计算机可读指令。

[34] 在图1所示的设备中，网络接口1004主要用于连接外网，与其他网络设备进行数据通信；用户接口1003主要用于连接用户设备，与设备进行数据通信；本申请设备通过处理器1001调用存储器1005中存储的计算机可读指令，并执行本申请实施例提供的字符识别的实施方案。

[35] 基于上述硬件结构，提出本申请字符识别方法实施例。

[36] 参照图2，图2为本申请字符识别方法第一实施例的流程示意图。

[37] 在第一实施例中，所述字符识别方法包括以下步骤：

[38] 步骤S10，获取待识别文本。

[39] 需要说明的是，在本实施例中，首先通过OCR获取历史识别文本，将所述历史识别文本作为所述待识别文本，在具体实现中，主要通过输入设备将识别文档输入到计算机中，所述输入设备可为扫描仪，还可为其他可实现相同功能的设

备，通过扫描一幅简单的印刷文档的图像，对测量文档放置的倾斜角，对文档进行版面分析，对选出的文字域进行排版确认，对横、竖排版的文字行进行切分，实现每一行的文字图像的分离，标点符号的判别等，从而进行对图像的预处理，将处理后的每一个文字图像分检出来交给识别模块识别，其中，版面分析是对文本图像的总体分析，是将文档中的所有文字块分检出来，区分出文本段落及排版顺序，以及图像、表格的区域。将各文字块的域界，包括域在图像中的始点、终点坐标，还包括域内的属性，即横、竖排版方式以及各文字块的连接关系作为一种数据结构，提供给识别模块自动识别，对于文本区域直接进行识别处理，对于表格区域进行专用的表格分析及识别处理，对于图像区域进行压缩或简单存储。行字切分是将大幅的图像先切割为行，再从图像行中分离出单个字符的过程。

- [40] 需要说明的是，在对文本进行识别时，从扫描文本中分检出的文字图像，由计算机将其图形、图像转变成文字的标准代码，根据文字的笔画、特征点、投影信息、点的区域分布等进行分析，从而提供文本中识别的每个字符的top10的结果，并从结果中选取top1作为基础文本，例如，对于中文文本中对“我是中国人”通过OCR进行识别之后将top1中的识别结果“我是中园人”对于基础文本，将所述基本文本作为所述待识别文本，从而实现对识别文档的初始识别。
- [41] 步骤S20，调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符。
- [42] 在本实施中，设有分词工具，通过所述分析工具将所述待识别文本进行分析，其中，所述分词工具可为如jieba、SnowNLP、THULAC、NLPIR，还可为其他分词工具，本实施例对此不作限制，通过分词工具将所述待识别文本分为预设词长度的词组，例如通过分词工具将“我是中国人”分为“我”、“是”以及“中园人”，或者“我是”、“中园”以及“人”等。对于中文字符来说，所述预设长度可为字的个数，例如“我是”为长度为2的字符，“人”为长度为1的字符，从而实现不同规则的分词，提高分词的精度。
- [43] 需要说明的是，为了提高识别的效率，在本实施例中，将所述预设长度大于2的词组进行列出，即“我是”、“中国”，从而实现对词组进行分析，还可列出符合

其他规则的词组，本实施例对此不作限制，在本实施例中，以将待识别文本分成长度为2的词组，从而提高文本识别的效率。

[44] 步骤S30，获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符。

[45] 需要说明的是，所述参考字符为通过分词工具进行分词后的若干词组，例如将“我是中国人”分为长度为2的若干个词组，例如“我是”、“中园”以及“人”，其中，所述第一预设区域和第二预设区域用于区分所述分词工具的存储地址与所述预设字典的存储地址。

[46] 在本实施例中，所述预设词典为根据预设字段进行分类后的词典，例如对于词长度为2的词典，词长度为3的词典等，对于词长度为2的词典中包含例如“中国”，对于词长度为3的词典中包含例如“中国人”等，从而根据词的长度将常用的词组进行分类，从而实现对常用词组的管理。

[47] 在具体实现中，通过所述预设词典可检查分词后的目标词组是否为常见的词组，例如本实施例中分词后长度为2的词组包括“我是”以及“中园”，将“我是”以及“中园”在长度为2的词典中查找是否存在，在不存在时，表明识别有异常，例如未查到“中园”这个词组，可查找到“我是”这个词组，表明“我是”识别正常，“中园”识别异常。

[48] 步骤S40，在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示。

[49] 在本实施例中，将未存有的字符通过模糊匹配算法进行筛选，所述模糊匹配算法为BK-tree (Burkhard-Keller-tree) 算法，由Burkhard和Keller提出的，通过模糊匹配算法是指两个字串之间，由一个转成另一个所需的最少编辑操作次数，将所述操作次数作为所述编辑距离，在编辑距离越小的两个字符串越相似，当编辑距离为0时，两字符串相等，从而实现对字符的识别。

[50] 本实施例通过上述方案，通过获取待识别文本，调用分词工具，以使所述分词工具将所述待识别文本划分为多个预设长度的字符，根据所述预设长度的字符查找对应的预设词典，判断所述预设词典中是否存有所述字符，在所述预设词

典中未存有所述字符时，说明所述字符存在识别异常的情况，在这种情况下，将未存有的字符通过模糊匹配算法筛选出目标字符，从而通过模糊匹配算法实现文字识别，提高文字识别效率。

[51] 进一步地，如图3所示，基于第一实施例提出本申请字符识别方法第二实施例，在本实施例中，所述步骤S20之前，所述方法还包括：

[52] 步骤S201，接收工具写入指令，提取所述工具写入指令中的分词工具和分词写入地址信息，根据所述分词写入地址信息将所述分词工具写入所述第一预设区域并进行保存。

[53] 可以理解的是，为了实现对待识别文本的比对分析，首先写入分词工具在预设区域中，在获取到待识别文本后通过调用预设区域中的分词工具对所述待识别文本进行更细化的分析，其中，所述分词工具可为一段小程序，还可为其他形式的分词工具，本实施例对此不作限制。

[54] 需要说明的是，所述工具写入指令可为通过写入平台界面进行写入操作，还可可通过数据串口进行写入，本实施例对此不作限制。

[55] 进一步地，所述步骤S20，包括：

[56] 步骤S202，调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本与各个预设长度的关键词进行比较，根据比较结果提取所述待识别文本中的各个预设长度的目标关键词，将所述目标关键词作为所述预设长度的参考字符。

[57] 在具体实现中，所述分词工具可设有各个关键词，通过将待识别文本与各个关键词进行比较，从而实现对所述待识别文本中的各个关键词的识别，例如将待识别文本“武汉风景好”通过所述分词工具进行分词，可将“武汉风景好”与各个关键词进行比较，从而得到“武汉”、“风景”以及“好”这几个关键词，从而实现对待识别文本的处理。

[58] 本实施例提供的方案，通过接收写入指令，根据所述写入指令预先写入所述分词工具，通过所述分词工具对所述待识别文本进行分词处理，从而实现更细化的文本识别。

[59] 进一步地，如图4所示，基于第一实施例或第二实施例提出本申请字符识别方

法第三实施例，在本实施例中，基于第一实施例进行说明，所述步骤S30之前，所述方法还包括：

- [60] 步骤S301，接收字典写入指令，提取所述字典写入指令中的预设字典和字典写入地址信息，根据所述字典写入地址信息将所述预设字典写入所述第二预设区域。
- [61] 需要说明的是，为了提高识别的准确性，首先需要写入所述预设词典，具体为接收写入指令，提取写入指令中的预设词典，将所述预设词典保存在预设区域，由于之前保存有分词工具，可将分词工具的存储地址与预设词典的分词地址保存在不同区域，并标上不同的识别标签，即通过第一预设区域和第二预设区域进行区分，从而实现对数据的有效管理。
- [62] 进一步地，所述步骤S30，包括：
- [63] 步骤S302，获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在预设地址关系映射表中查找对应的存储地址。
- [64] 需要说明的是，所述存储地址为预设字典的存储地址，在数据库中存在多个词典，例如长度为2的词典以及长度为3的词典，还存有其他形式的词典，为了实现对词典的管理，可将词典使用不同的存储地址进行储存，并将储存地址与词典的长度的对应关系建立所述预设地址关系映射表，通过获取字符的长度在所述预设地址关系映射表即可查找到对应的字典的地址，例如在所述参考字符长度为2时，根据字符长度2在所述预设地址关系映射表中查找长度2的字典存储的地址信息，从而实现对地址的有效管理。
- [65] 步骤S303，根据所述存储地址在预设区域查找对应的预设词典，并提取所述参考字符的特征信息，将所述特征信息与查找到的词典中的字符的特征信息进行比较，根据比较结果判断所述词典中是否存有所述参考字符。
- [66] 为了判断识别的参考字符的准确性，通过将参考字符与词典中的字符进行比较，判断所述词典中是否存有所述参考字符，在所述词典中未存有所述参考字符时，说明当前参考字符有异常，在所述词典中存有所述参考字符时，说明当前参考字符识别正确，例如判断所述词典是否存有“我是”、“中园”以及“人”，可知在所述词典中存有“我是”，但是并未存有“中园”，从而可判断出“中园”存在异常

- 。
- [67] 在具体实现中，通过提取参考字符的特征信息，所述特征信息可为所述参考字符的点的区域分布，各个点的几何分布状态，还可为其他形式的特征信息，本实施例对此不作限制。
- [68] 进一步地，所述步骤S40，包括：
- [69] 步骤S401，在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符，将所述目标字符进行展示。
- [70] 在具体实现中，通过BK-tree算法查找编辑距离不大于该词长度的词，例如不存在“中园”，则从过BK-tree查找编辑距离不大于该词长度的词可为“中国”，其中所述编辑距离为字符串A到B的编辑距离，通过只用插入、删除和替换三种操作，最少需要多少步可以把A变成B。例如，从FAME到GATE需要两步，两次替换，从GAME到ACM则需要三步，包括删除G和E再添加C，将筛选出的“中国”作为所述目标字符进行展示，从而通过模糊匹配算法实现文本的识别，提高文本识别的准确性。
- [71] 进一步地，所述步骤S10之后，所述方法还包括：将待识别文本中的各个初始识别字符建立初始识别列表，所述步骤S401，包括：
- [72] 步骤S402，在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符。
- [73] 需要说明的是，通过OCR在对文本进行识别时，从扫描文本中分检出的文字图像，由计算机将其图形、图像转变成文字的标准代码，根据文字的笔画、特征点、投影信息、点的区域分布等进行分析，从而提供文本中识别的每个字符的top10的结果，将所述每个字符的top10的结果建立对应每个字符的初始识别列表。
- [74] 步骤S403，判断所述目标字符的数量，在所述数量为多个时，判断所述目标字符是否存在所述初始识别列表中，将存在所述初始识别列表中的字符对应的目标字符进行展示。
- [75] 需要说明的是，在通过BK-tree进行筛选时，可能出现多个词的情况，例如上述

中可筛选出除了“中国”，还包括“中文”以及“家园”等，针对这种情况，可从筛选词中找出改变的字为之前top10中出现的字对应的词作为所述目标字符进行展示，从而提高文本识别的准确性。

[76] 本实施例提供的方案，通过模糊匹配算法加入文本识别中进行识别，根据编辑距离查找出相似的字符，将筛选出的字符作为所述目标字符，从而提高文本识别的准确性。

[77] 本申请进一步提供一种字符识别装置。

[78] 参照图5，图5为本申请字符识别装置第一实施例的功能模块示意图。

[79] 本申请字符识别装置第一实施例中，该字符识别装置包括：

[80] 获取模块10，获取待识别文本。

[81] 需要说明的是，在本实施例中，首先通过OCR获取历史识别文本，将所述历史识别文本作为所述待识别文本，在具体实现中，主要通过输入设备将识别文档输入到计算机中，所述输入设备可为扫描仪，还可为其他可实现相同功能的设备，通过扫描一幅简单的印刷文档的图像，对测量文档放置的倾斜角，对文档进行版面分析，对选出的文字域进行排版确认，对横、竖排版的文字行进行切分，实现每一行的文字图像的分离，标点符号的判别等，从而进行对图像的预处理，将处理后的每一个文字图像分检出来交给识别模块识别，其中，版面分析是对文本图像的总体分析，是将文档中的所有文字块分检出来，区分出文本段落及排版顺序，以及图像、表格的区域。将各文字块的域界，包括域在图像中的始点、终点坐标，还包括域内的属性，即横、竖排版方式以及各文字块的连接关系作为一种数据结构，提供给识别模块自动识别，对于文本区域直接进行识别处理，对于表格区域进行专用的表格分析及识别处理，对于图像区域进行压缩或简单存储。行字切分是将大幅的图像先切割为行，再从图像行中分离出单个字符的过程。

[82] 需要说明的是，在对文本进行识别时，从扫描文本中分检出的文字图像，由计算机将其图形、图像转变成文字的标准代码，根据文字的笔画、特征点、投影信息、点的区域分布等进行分析，从而提供文本中识别的每个字符的top10的结果，并从结果中选取top1作为基础文本，例如，对于中文文本中对“我是中国人”

通过OCR进行识别之后将top1中的识别结果“我是中国人”对于基础文本，将所述基本文本作为所述待识别文本，从而实现对识别文档的初始识别。

[83] 调用模块20，用于调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符。

[84] 在本实施中，设有分词工具，通过所述分析工具将所述待识别文本进行分析，其中，所述分词工具可为如jieba、SnowNLP、THULAC、NLPIR，还可为其他分词工具，本实施例对此不作限制，通过分词工具将所述待识别文本分为预设词长度的词组，例如通过分词工具将“我是中国人”分为“我”、“是”以及“中国人”，或者“我是”、“中国”以及“人”等。对于中文字符来说，所述预设长度可为字的个数，例如“我是”为长度为2的字符，“人”为长度为1的字符，从而实现不同规则的分词，提高分词的精度。

[85] 需要说明的是，为了提高识别的效率，在本实施例中，将所述预设长度大于2的词组进行列出，即“我是”、“中国”，从而实现对词组进行分析，还可列出符合其他规则的词组，本实施例对此不作限制，在本实施例中，以将待识别文本分成长度为2的词组，从而提高文本识别的效率。

[86] 查找模块30，用于获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符。

[87] 需要说明的是，所述参考字符为通过分词工具进行分词后的若干词组，例如将“我是中国人”分为长度为2的若干个词组，例如“我是”、“中国”以及“人”。

[88] 在本实施例中，所述预设词典为根据预设字段进行分类后的词典，例如对于词长度为2的词典，词长度为3的词典等，对于词长度为2的词典中包含例如“中国”，对于词长度为3的词典中包含例如“中国人”等，从而根据词的长度将常用的词组进行分类，从而实现对常用词组的管理。

[89] 在具体实现中，通过所述预设词典可检查分词后的目标词组是否为常见的词组，例如本实施例中分词后长度为2的词组包括“我是”以及“中国”，将“我是”以及“中国”在长度为2的词典中查找是否存在，在不存在时，表明识别有异常，例如未查到“中国”这个词组，可查找到“我是”这个词组，表明“我是”识别正常，“中国”

识别异常。

- [90] 筛选模块40，用于在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示。
- [91] 在本实施例中，将未存有的字符通过模糊匹配算法进行筛选，所述模糊匹配算法为BK-tree（Burkhard-Keller-tree）算法，由Burkhard和Keller提出的，通过模糊匹配算法是指两个字串之间，由一个转成另一个所需的最少编辑操作次数，将所述操作次数作为所述编辑距离，在编辑距离越小的两个字符串越相似，当编辑距离为0时，两字符串相等，从而实现对字符的识别。
- [92] 本实施例通过上述方案，通过获取待识别文本，调用分词工具，以使所述分词工具将所述待识别文本划分为多个预设长度的字符，根据所述预设长度的字符查找对应的预设词典，判断所述预设词典中是否存有所述字符，在所述预设词典中未存有所述字符时，说明所述字符存在识别异常的情况，在这种情况下，将未存有的字符通过模糊匹配算法筛选出目标字符，从而通过模糊匹配算法实现文字识别，提高文字识别效率。
- [93] 此外，为实现上述目的，本申请还提出一种设备，所述设备包括：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可读指令，所述计算机可读指令配置为实现如上文所述的字符识别方法的步骤。
- [94] 此外，本申请实施例还提出一种存储介质，所述计算机可读存储介质可以为非易失性可读存储介质。
- [95] 本申请存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行如上文所述的字符识别方法的步骤。
- [96] 其中，该计算机可读指令被执行时所实现的方法可参照本申请发票开具方法的各个实施例，此处不再赘述。
- [97] 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包

括一个.....”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[98] 上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

[99] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个计算机可读存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台智能终端设备(可以是手机，计算机，终端设备，空调器，或者网络终端设备等)执行本申请各个实施例所述的方法。

[100] 以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权利要求书

- [权利要求 1] 一种字符识别方法，其特征在于，所述字符识别方法包括：
获取待识别文本；
调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符；
获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符；
在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示。
- [权利要求 2] 如权利要求1所述的字符识别方法，其特征在于，所述调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符之前，所述方法包括：
接收工具写入指令，提取所述工具写入指令中的分词工具和分词写入地址信息，根据所述分词写入地址信息将所述分词工具写入所述第一预设区域并进行保存。
- [权利要求 3] 如权利要求1所述的字符识别方法，其特征在于，所述调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符，包括：
调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本与各个预设长度的关键词进行比较，根据比较结果提取所述待识别文本中的各个预设长度的目标关键词，将所述目标关键词作为所述预设长度的参考字符。
- [权利要求 4] 如权利要求1所述的字符识别方法，其特征在于，所述获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符之前，所述方法还包括：

接收字典写入指令，提取所述字典写入指令中的预设字典和字典写入地址信息，根据所述字典写入地址信息将所述预设字典写入所述第二预设区域。

[权利要求 5] 如权利要求1所述的字符识别方法，其特征在于，所述获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符，包括：

获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在预设地址关系映射表中查找对应的存储地址；

根据所述存储地址在预设区域查找对应的预设词典，并提取所述参考字符的特征信息，将所述特征信息与查找到的词典中的字符的特征信息进行比较，根据比较结果判断所述词典中是否存有所述参考字符。

[权利要求 6] 如权利要求1所述的字符识别方法，其特征在于，所述在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示，包括：

在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符，将所述目标字符进行展示。

[权利要求 7] 如权利要求6所述的字符识别方法，其特征在于，所述获取待识别文本之后，所述方法还包括：

将待识别文本中的各个初始识别字符建立初始识别列表；

所述在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符，将所述目标字符进行展示，包括：

在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符；

判断所述目标字符的数量，在所述数量为多个时，判断所述目标字符

是否存在所述初始识别列表中，将存在所述初始识别列表中的字符对应的目标字符进行展示。

[权利要求 8]

一种字符识别装置，其特征在于，所述字符识别装置包括：

获取模块，用于获取待识别文本；

调用模块，用于调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符；

查找模块，用于获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符；

筛选模块，用于在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示。

[权利要求 9]

一种设备，其特征在于，所述设备包括：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可读指令，所述计算机可读指令配置为实现以下步骤：

获取待识别文本；

调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符；

获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符；

在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示。

[权利要求 10]

如权利要求9所述的设备，其特征在于，所述计算机可读指令还配置为实现以下步骤：

接收工具写入指令，提取所述工具写入指令中的分词工具和分词写入地址信息，根据所述分词写入地址信息将所述分词工具写入所述第一

预设区域并进行保存。

[权利要求 11] 如权利要求9所述的设备，其特征在于，所述计算机可读指令还配置为实现以下步骤：

调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本与各个预设长度的关键词进行比较，根据比较结果提取所述待识别文本中的各个预设长度的目标关键词，将所述目标关键词作为所述预设长度的参考字符。

[权利要求 12] 如权利要求9所述的设备，其特征在于，所述计算机可读指令还配置为实现以下步骤：

接收字典写入指令，提取所述字典写入指令中的预设字典和字典写入地址信息，根据所述字典写入地址信息将所述预设字典写入所述第二预设区域。

[权利要求 13] 如权利要求9所述的设备，其特征在于，所述计算机可读指令还配置为实现以下步骤：

获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在预设地址关系映射表中查找对应的存储地址；

根据所述存储地址在预设区域查找对应的预设词典，并提取所述参考字符的特征信息，将所述特征信息与查找到的词典中的字符的特征信息进行比较，根据比较结果判断所述词典中是否存有所述参考字符。

[权利要求 14] 如权利要求9所述的设备，其特征在于，所述计算机可读指令还配置为实现以下步骤：

在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符，将所述目标字符进行展示。

[权利要求 15] 如权利要求14所述的设备，其特征在于，所述计算机可读指令还配置为实现以下步骤：

将待识别文本中的各个初始识别字符建立初始识别列表；

在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在

所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符；

判断所述目标字符的数量，在所述数量为多个时，判断所述目标字符是否存在所述初始识别列表中，将存在所述初始识别列表中的字符对应的目标字符进行展示。

[权利要求 16] 一种存储介质，其特征在于，所述存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时实现以下步骤：

获取待识别文本；

调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本划分为多个预设长度的参考字符；

获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在第二预设区域查找对应的预设词典，并判断所述预设词典中是否存有所述参考字符；

在所述预设词典中未存有所述参考字符时，通过模糊匹配算法对未存有的参考字符进行筛选，得到目标字符，并将所述目标字符进行展示。

[权利要求 17] 如权利要求16所述的存储介质，其特征在于，所述计算机可读指令还配置为实现以下步骤：

调用第一预设区域中预存的分词工具，通过所述分词工具将所述待识别文本与各个预设长度的关键词进行比较，根据比较结果提取所述待识别文本中的各个预设长度的目标关键词，将所述目标关键词作为所述预设长度的参考字符。

[权利要求 18] 如权利要求16所述的存储介质，其特征在于，所述计算机可读指令还配置为实现以下步骤：

获取所述分词工具划分后的参考字符，根据所述参考字符的目标长度在预设地址关系映射表中查找对应的存储地址；

根据所述存储地址在预设区域查找对应的预设词典，并提取所述参考字符的特征信息，将所述特征信息与查找到的词典中的字符的特征信

息进行比较，根据比较结果判断所述词典中是否存有所述参考字符。

[权利要求 19] 如权利要求16所述的存储介质，其特征在于，所述计算机可读指令还配置为实现以下步骤：

在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符，将所述目标字符进行展示。

[权利要求 20] 如权利要求19所述的存储介质，其特征在于，所述计算机可读指令还配置为实现以下步骤：

将待识别文本中的各个初始识别字符建立初始识别列表；

在所述预设词典中未存有所述参考字符时，通过所述模糊匹配算法在所述预设词典中查找出编辑距离小于所述参数字符对应的目标长度的目标字符；

判断所述目标字符的数量，在所述数量为多个时，判断所述目标字符是否存在所述初始识别列表中，将存在所述初始识别列表中的字符对应的目标字符进行展示。

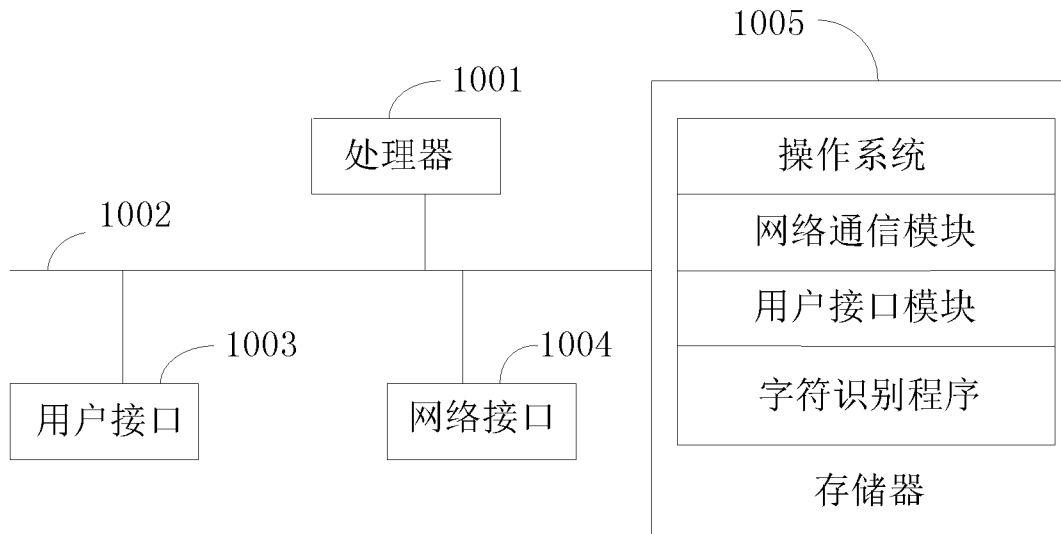


图 1

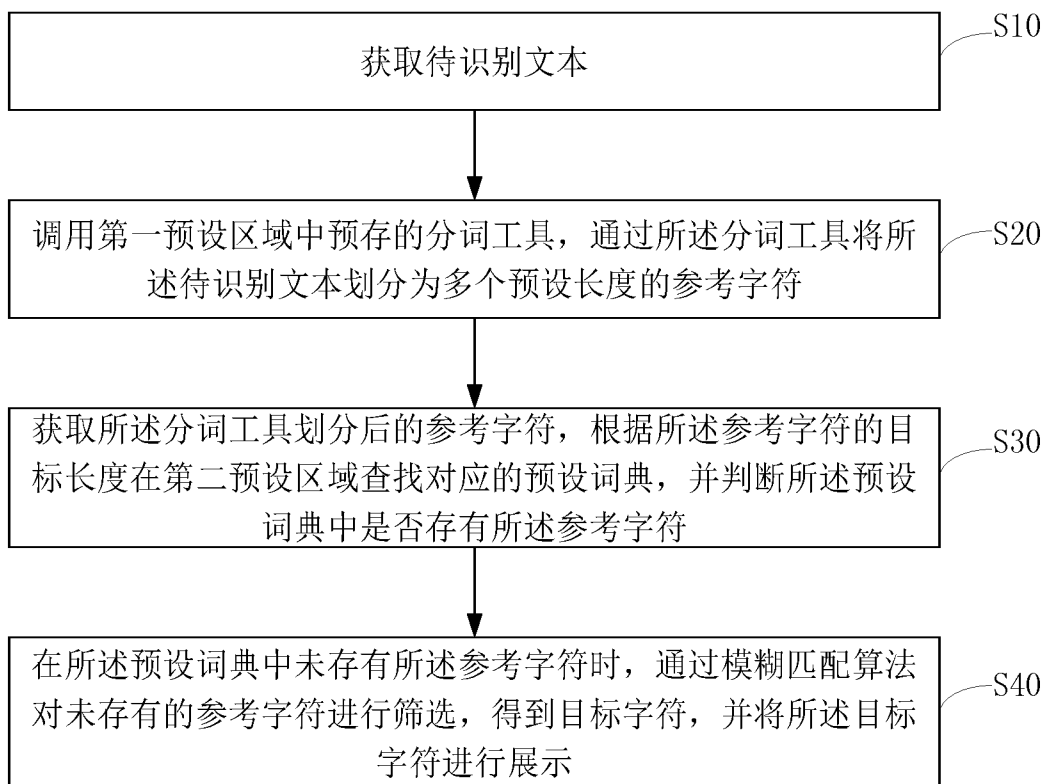


图 2

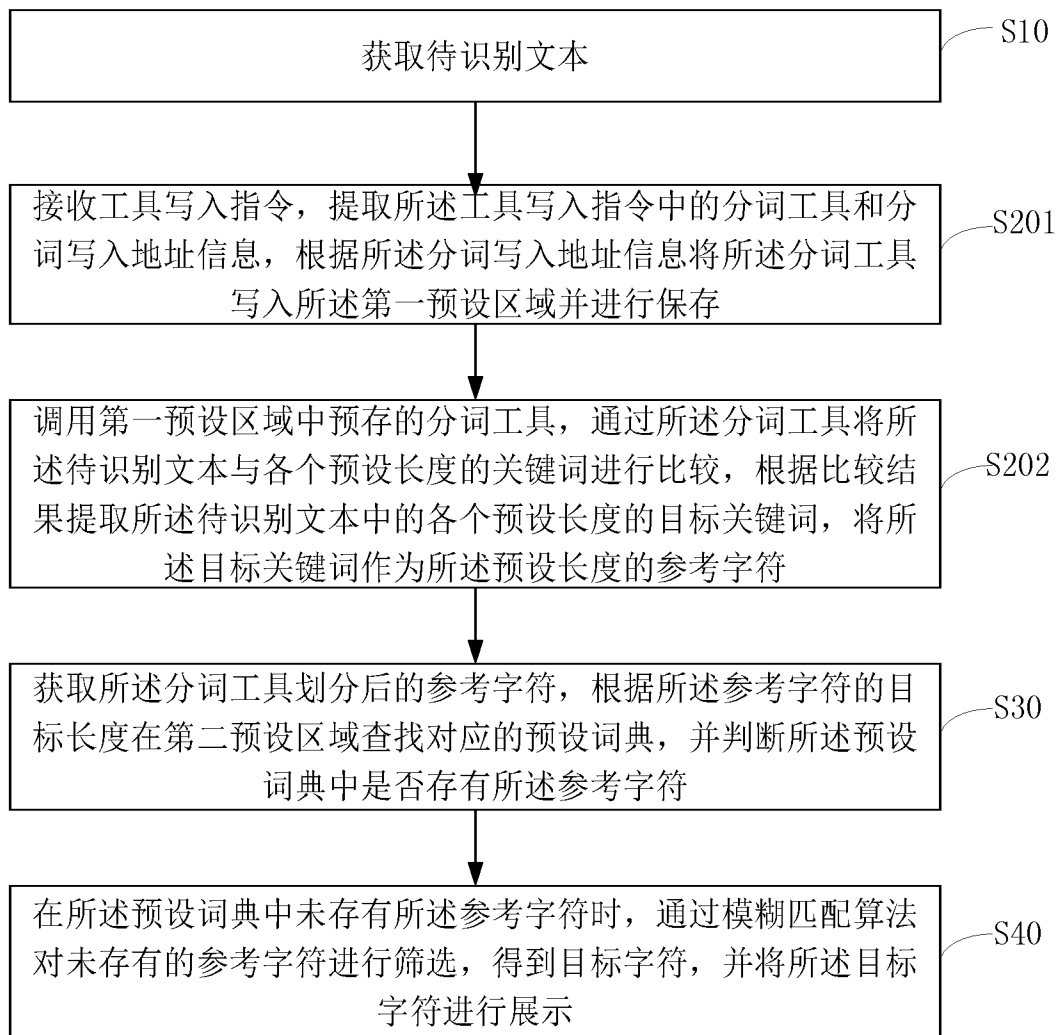


图 3

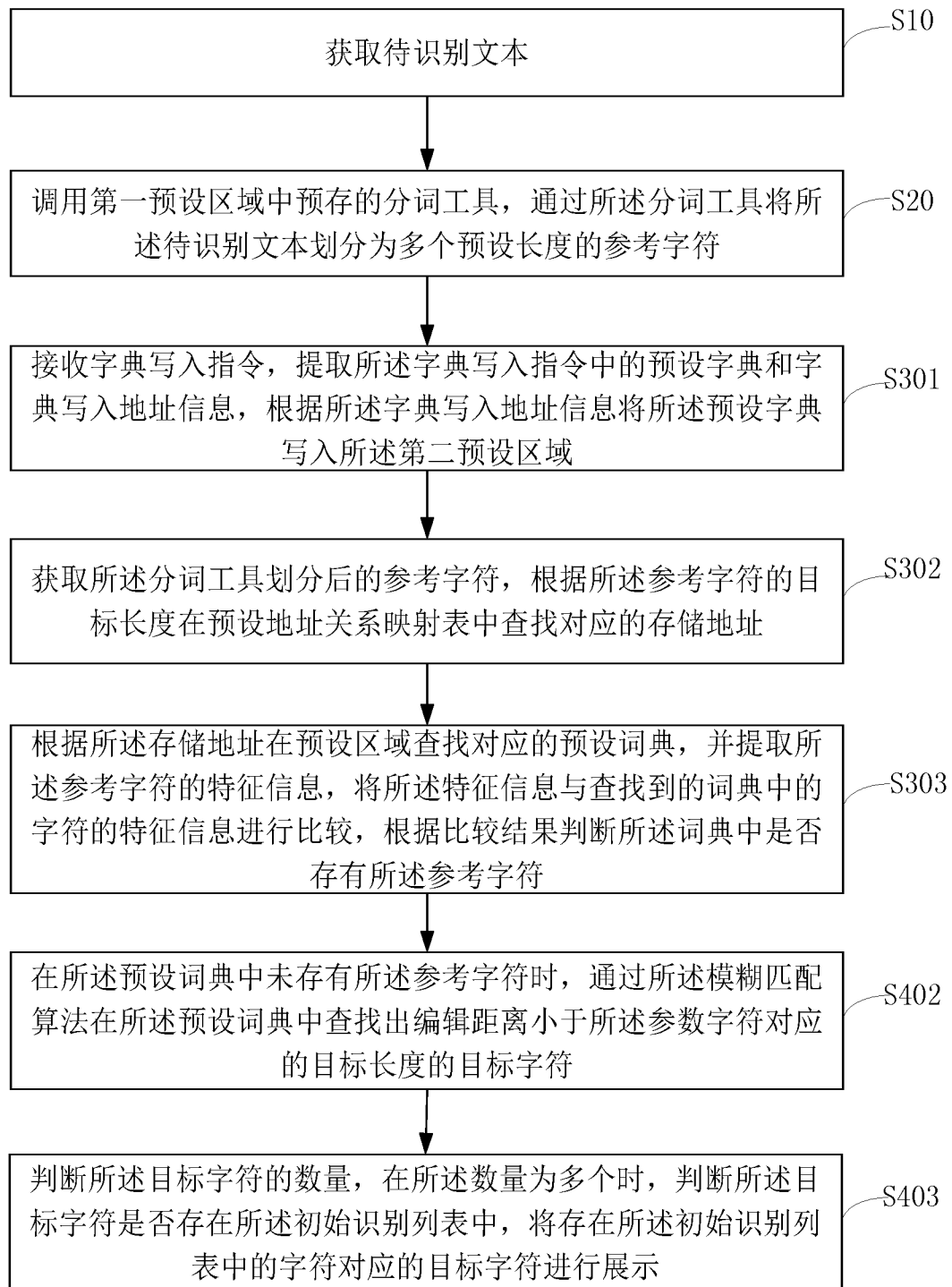


图 4

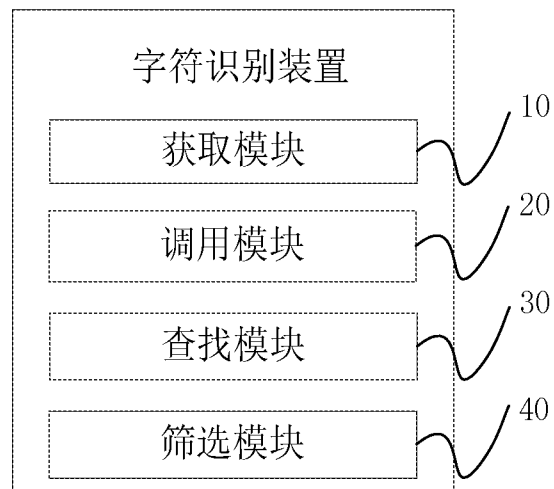


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/122832

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/68(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, IEEE: 字符, 识别, 文本, 分词, 字典, 词典, 模糊, 匹配, character, recognition, text, word, segmentation, dividing, dictionary, fuzzy, matching

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104991889 A (JIANGSU UNIVERSITY OF SCIENCE AND TECHNOLOGY) 21 October 2015 (2015-10-21) description, paragraphs [0043]-[0081], and claims 1-8	1-20
A	CN 107622044 A (ALIBABA GROUP HOLDING LIMITED) 23 January 2018 (2018-01-23) entire document	1-20
A	CN 108304484 A (BEIJING URBAN NETWORK NEIGHBOR INFORMATION TECHNOLOGY CO., LTD.) 20 July 2018 (2018-07-20) entire document	1-20
A	CN 105068994 A (EBAO INTERCONNECTED MEDICAL INFORMATION TECHNOLOGY BEIJING CO., LTD.) 18 November 2015 (2015-11-18) entire document	1-20

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

04 July 2019

Date of mailing of the international search report

24 July 2019

Name and mailing address of the ISA/CN

**China National Intellectual Property Administration (ISA/
CN)**
**No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
 100088**
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/122832

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	104991889	A	21 October 2015	None			
CN	107622044	A	23 January 2018	TW	201804341	A	01 February 2018
				WO	2018010579	A1	18 January 2018
CN	108304484	A	20 July 2018	None			
CN	105068994	A	18 November 2015	None			

国际检索报告

国际申请号

PCT/CN2018/122832

A. 主题的分类 G06K 9/68(2006.01) i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类		
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G06K 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) WPI, EPDOC, CNPAT, CNKI, IEEE: 字符, 识别, 文本, 分词, 字典, 词典, 模糊, 匹配, character, recognition, text, word, segmentation, dividing, dictionary, fuzzy, matching		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 104991889 A (江苏科技大学) 2015年 10月 21日 (2015 - 10 - 21) 说明书第[0043]-[0081]段, 权利要求1-8	1-20
A	CN 107622044 A (阿里巴巴集团控股有限公司) 2018年 1月 23日 (2018 - 01 - 23) 全文	1-20
A	CN 108304484 A (北京城市网邻信息技术有限公司) 2018年 7月 20日 (2018 - 07 - 20) 全文	1-20
A	CN 105068994 A (易保互联医疗信息科技北京有限公司) 2015年 11月 18日 (2015 - 11 - 18) 全文	1-20
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期		国际检索报告邮寄日期
2019年 7月 4日		2019年 7月 24日
ISA/CN的名称和邮寄地址		受权官员
中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		王晓敏 电话号码 86-(10)-53961359

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/122832

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	104991889	A	2015年 10月 21日	无			
CN	107622044	A	2018年 1月 23日	TW	201804341	A	2018年 2月 1日
				WO	2018010579	A1	2018年 1月 18日
CN	108304484	A	2018年 7月 20日	无			
CN	105068994	A	2015年 11月 18日	无			