(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
2018 年 2 月 8 日 (08.02.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/024243 A1

(51) 国际专利分类号:
G06K 9/20 (2006.01)

(21) 国际申请号: PCT/CN2017/095992

(22) 国际申请日: 2017 年 8 月 4 日 (04.08.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610641373.6 2016年8月5日 (05.08.2016) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(72) 发明人: 韩盛(HAN, Sheng); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 王红法 (WANG, Hongfa); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 周龙沙 (ZHOU, Longsha); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 宋辉 (SONG, Hui); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人: 北京德琦知识产权代理有限公司 (DEQI INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区知春路1号学院国际大厦7层, Beijing 100083 (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,

(54) **Title:** METHOD AND DEVICE FOR VERIFYING RECOGNITION RESULT IN CHARACTER RECOGNITION

(54) **发明名称**: 字符识别中识别结果的校验方法和装置

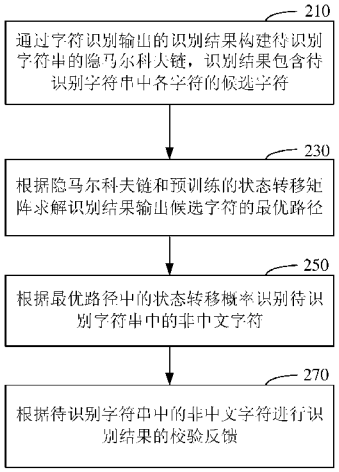

图 2

210 Construct, via a recognition result output by a character recognition process, a hidden Markov chain of a character string to be recognized, wherein the recognition result comprises candidate characters of all characters to be recognized in the character string to be recognized
230 Solve, according to the hidden Markov chain and a pre-trained state transition matrix, the recognition result, so as to output an optimal path of the candidate characters
250 Recognize, according to a state transition probability in the optimal path, a non-Chinese character in the character string to be recognized
270 Verify, according to the non-Chinese character in the character string to be recognized, the recognition result, so as to feed back a verification result

(57) Abstract: A method and device for verifying a recognition result in character recognition. The method comprises the following steps: constructing, via a recognition result output by a character recognition process, a hidden Markov chain of a character string to be recognized, wherein the recognition result comprises candidate characters of all characters to be recognized in the character string to be recognized, and each of the characters to be recognized corresponds to at least one candidate character (210); solving, according to the hidden Markov chain and a pre-trained state transition matrix, the recognition result, so as to output an optimal path of the candidate characters, wherein a candidate character string comprises at least one candidate character corresponding to each character to be recognized (230); recognizing, according to a state transition probability in the optimal path, a non-Chinese character in the character string to be recognized (250); and verifying, according to the non-Chinese character in the character string to be recognized, the recognition result, so as to feed back a verification result (270).

(57) 摘要: 一种字符识别中识别结果的校验方法和装置。所述方法包括：通过字符识别过程输出的识别结果构建待识别字符串的隐马尔科夫链，所述识别结果包含所述待识别字符串中各待识别字符的候选字符（210）；每个待识别字符对应至少一个候选字符；根据隐马尔科夫链和预训练的状态转移矩阵求解识别结果输出候选字符的最优路径（230）；所述候选字符串包含分别对应各待识别字符的至少一个所述候选字符；根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符（250）；根据所述待识别字符串中的非中文字符进行所述识别结果的校验反馈（270）。

WO 2018/024243 A1

[见续页]

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, 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))。

字符识别中识别结果的校验方法和装置

本申请要求于 2016 年 8 月 5 日提交中国专利局、申请号为 201610641373.6、发明名称为“字符识别中识别结果的校验方法和装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及字符识别技术领域，特别涉及一种字符识别中识别结果的校验方法和装置。

背景

随着计算机技术的发展，各种票据、报刊、书籍、文稿以及其它印刷品被通过扫描等光学输入方式实现纸质文字到图像信息的转化。在获得图像信息之后，需要通过字符识别来实现图像信息到可使用的计算机文字的转换。

技术内容

本申请提供了一种字符识别中识别结果的校验方法和装置。

一种字符识别中识别结果的校验方法，包括：

通过字符识别过程输出的识别结果构建待识别字符串的预测状态统计模型，所述识别结果包含所述待识别字符串中各待识别字符的候选字符；每个待识别字符对应至少一个候选字符；

根据所述预测状态统计模型和预训练的状态转移矩阵求解形成候选字符串的最优路径；所述候选字符串包含分别对应各待识别字符的至少一个所述候选字符；

根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符；

根据所述待识别字符串中的非中文字符进行所述识别结果校验，并将校验结果反馈至所述字符识别过程。

一种字符识别中识别结果的校验装置，所述装置包括：

一个或一个以上存储器；

一个或一个以上处理器；其中，

所述一个或一个以上存储器存储有一个或者一个以上指令模块，经配置由所述一个或者一个以上处理器执行；其中，

所述一个或者一个以上指令模块包括：

构建模块，通过字符识别过程输出的识别结果构建待识别字符串的预测状态统计模型，所述识别结果包含待识别字符串中各待识别字符的候选字符；每个待识别字符对应至少一个候选字符；

最优路径求解模块，根据所述预测状态统计模型和预训练的状态转移矩阵求解形成候选字符串的最优路径；所述候选字符串包含分别对应各待识别字符的至少一个所述候选字符；

非中文字符识别模块，根据所述路径中的状态转移概率识别所述待识别字符串中的非中文字符；

反馈模块，根据所述待识别字符串中的非中文字符进行所述识别结果校验，并将校验结果反馈至所述字符识别过程。

一种非易失性计算机可读存储介质，其上存储有计算机可读指令，可以使至少一个处理器执行上述方法。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性的，并不能限制本申请。

附图简要说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实例，并于说明书一起用于解释本发明的原理。

图 1 是根据一示例性实例示出的一种装置的框图；

图 2 是根据一示例性实例示出的一种字符识别中识别结果的校验方

法的流程图；

图 3 是一待识别字符串的实例图；

图 4 是图 2 对应实例的根据最优路径中的状态转移概率识别待识别字符串中的非中文字符步骤的流程图；

图 5 是根据另一示例性实例示出的一种字符识别中识别结果的校验方法的流程图；

图 6 是一种场景下字符识别中识别结果校验的示意图；

图 7 是根据一示例性实例示出的一种字符识别中识别结果的校验装置的框图；

图 8 是图 7 对应实例的非中文字符识别模块的框图；

图 9 是根据另一示例性实例示出的一种字符识别中识别结果的校验装置的框图。

实施方式

这里将详细地对示例性实例执行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

图 1 是根据一示例性实例示出的一种装置 100 的框图。例如，装置 100 用于为本申请提供所需要的实施环境，因此，装置 100 可以是服务器。

参见图 1，该服务器 100 可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器（central processing units，CPU）122（例如，一个或一个以上处理器）和存储器 132，一个或一个以上存储应用程序 142 或数据 144 的存储介质 130（例如一个或一个以上海量存储设备）。其中，存储器 132 和存储介质 130 可以是短暂存储或持久存

储。存储在存储介质 130 的程序可以包括一个或一个以上模块（图示未示出），每个模块可以包括对服务器中的一系列指令操作。更进一步地，中央处理器 122 可以设置为与存储介质 130 通信，在服务器 100 上执行存储介质 130 中的一系列指令操作。服务器 100 还可以包括一个或一个以上电源 126，一个或一个以上有线或无线网络接口 150，一个或一个以上输入输出接口 158，和/或，一个或一个以上操作系统 141，例如 Windows ServerTM，Mac OS XTM，UnixTM, LinuxTM，FreeBSDTM 等等。下述图 2、图 3 和图 4 所示实例中所述的由服务器所执行的步骤可以基于该图 1 所示的服务器结构。

图 2 是根据一示例性实例示出的一种字符识别中识别结果的校验方法的流程图。该字符识别中识别结果的校验方法用于图 1 所示的装置 100。如图 2 所示，该字符识别中识别结果的校验方法，可以由服务器执行，可以包括以下步骤。

在步骤 210 中，通过字符识别（或称字符识别过程）输出的识别结果构建待识别字符串的隐马尔科夫链，识别结果包含待识别字符串中各字符的候选字符。

可理解的是，隐马尔科夫链仅是预测状态统计模型中的一种，还可以通过字符识别（或称字符识别过程）输出的识别结果构建其他类型的预测状态统计模型，因此步骤 S201 也可以描述成“通过字符识别过程输出的识别结果构建待识别字符串的预测状态统计模型，所述识别结果包含所述待识别字符串中各待识别字符的候选字符；每个待识别字符对应至少一个候选字符”。

隐马尔科夫链，可以理解为隐马尔科夫模型，是一种统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。在简单的马尔可夫模型（如马尔可夫链），所述状态是直接可见的观察者，因此状态转移概率是唯一的参数。在隐马尔可夫模型中，状态是不直接可见的，但输出依赖于该状态下，是可见的。每个状态通过可能的输出记号有了可能的概率分布。因此，通过隐马尔科夫模型产生标记序列提供了有关状态的

一些序列的信息。注意，“隐藏”指的是，该模型经其传递的状态序列，而不是模型的参数；即使这些参数是精确已知的，我们仍把该模型称为一个“隐藏”的马尔可夫模型。

其中，需要说明的是，对待识别字符串进行字符识别所输出的识别结果为一个或者两个以上。在具体实现中，字符识别所输出的识别结果数量大都是两个以上的，并且每一识别结果都包含了对应于待识别字符串中各字符（即待识别字符）的候选字符，以及此候选字符即为待识别字符串中相应字符的分数。分数越高，则候选字符与待识别字符串中相应字符相一致的可能性越高。

由此可知，对于待识别字符串中的一独立字符而言，由于存在着一个或者两个以上的识别结果，因此，此独立字符存在着一个或者两个以上的候选字符。换而言之，为此独立字符输出候选字符的识别结果为一个或者多个，需通过识别结果来得到为待识别字符串输出候选字符的隐马尔科夫链。

为待识别字符串输出候选字符的隐马尔科夫链中，标示了为待识别字符串所分别输出候选字符的识别结果，并且其所进行的识别结果标示是与待识别字符串中字符的顺序相符的。也就是说，隐马尔科夫链标示了待识别字符串输出候选字符的路径，即待识别字符串中字符所对应候选字符所来自的识别结果。

在步骤230中，根据隐马尔科夫链和预训练的状态转移矩阵求解识别结果输出候选字符的最优路径，也可称为候选字符串的最优路径。

在每一个待识别字符的所有候选字符中选择一个候选字符，然后针对待识别字符串中各个待识别字符所选择出来的候选字符按顺序形成一个候选字符串，这个候选字符串可以称为一条路径，在候选字符中选择出不同的字符会形成不同的路径，最优路径是指最接近待识别字符串的路径。在步骤 230 中，“据隐马尔科夫链和预训练的状态转移矩阵求解识别结果输出候选字符的最优路径”的过程也可以理解为是“根据所述隐马尔科夫链和预训练的状态转移矩阵求解形成与所述待识别字符串相对应的候选字符串的最优路径”的过程。

举例来说，参见图 3，待识别字符串“北京大学深圳医院临检组检验报告单”中，待识别字符“北”的候选字符有“北”、“比”、“韭”，待识别字符“京”的候选字符有“京”、“宗”、“家”，待识别字符“大”的候选字符有“大”、“太”、“犬”，等等。其中一条路径为“北宗大孝深洲医阮临楂姐检险恨吉单”，一条路径为“比宗大孝深洲医院临楂姐检险恨吉单”，还有很多其他的路径。最优路径当然是“北京大学深圳医院临检组检验报告单”。

其中，状态转移矩阵是预先进行训练所获得的，其用于表征两个字符之间转换的概率，换而言之，通过状态转移矩阵，可以获知一字符与另一字符构成一整体，即连接在一起的可能性。

如前所述的，通过隐马尔科夫链即可获得输出候选字符的路径。在此，需要利用预训练的状态转移矩阵进行最优路径的求解，以便于能够得到识别结果输出候选字符的最优路径。

最优路径与隐马尔科夫链相类似的，最优路径标示了输出候选字符的识别结果，以及识别结果之间的状态转移概率。

通过最优路径的求解使得对待识别字符串中字符所分别对应输出的候选字符不再仅仅考虑独立的字符，而是在整体上进行待识别字符串所对应候选字符的选取，提高了字符识别的整体性，进而也将得以提高字符识别的最终准确性。

在一个示例性实例中，最优路径的求解可以通过 viterbi 算法实现。

通过预训练的状态转移矩阵来实现识别效果的整体判定。通过字符识别所获得的识别结果为一个或者多个，每一识别结果都包含了针对待识别字符串中的字符而输出的相应候选字符，并且此候选字符具备一定的分数。

针对待识别字符串中字符而输出的候选字符相互连接，即构成了待识别字符串的识别结果。

状态转移矩阵是基于互联网文本预先进行训练而得到的，通过此状态转移矩阵，即可获知两个候选字符之间的状态转移概率，即两个候选字符连接在一起的可能性。

因此，通过预训练的状态转移矩阵来获得识别结果中候选字符之间的状态转移概率，进而从整体考虑所获得的若干个识别结果中，哪些候选字符构成最终的识别结果，此即为最优路径的获得。

在步骤250中，根据最优路径中的状态转移概率识别待识别字符串中的非中文字符。

其中，可以理解的，待识别字符串中，字符是通过一定的顺序连接在一起，进而形成待识别字符串的。相对应的，针对待识别字符串所输出的候选字符也是存在着此顺序，并且两两之间也是相互连接的，进而对于最优路径中输出候选字符的识别结果而言，相互之间也存在着关联性，此关联性是指在待识别字符串中字符的顺序上，最优路径中输出字符的一识别结果变换为另一识别结果的可能性，即状态转移概率。

也就是说，最优路径中，构成路径的识别结果分别作为隐藏状态而相互连接，以分别输出待识别字符串中字符的候选结果。

基于此，可根据最优路径中的状态转移概率进行待识别字符串中非中文字符的识别，进而在整体上实现待识别字符串中非文字符的识别，通过整体识别来有效分割中文字符和非中文字符，提高了准确率。

在步骤270中，根据待识别字符串中的非中文字符进行识别结果的校验反馈。

其中，根据最终所识别出的非中文字符进行识别结果的校验反馈，将有助于能够灵活调整待识别字符串中不同类型字符的识别算法，提高识别效果和矫正准确率。

通过如上所述的校验环节的实现，得以形成了有效的反馈机制，使得所进行的字符识别中能够自动判定整体和各字符的识别效果，进而灵活调整算法来提高识别效果，最大限度地避免了字符识别中低级错误的发生。

在一示例性实例示出的对步骤210的细节进行的描述。该步骤210，可以包括以下步骤。

根据待识别字符串中字符的顺序以输出相应候选字符的识别结果

为隐藏状态，隐藏状态相互连接构建得到待识别字符串的隐马尔科夫链。

其中，由于候选字符是与待识别字符串中的字符存在着对应关系的，因此，根据待识别字符串中字符的顺序以将输出相应候选字符的识别结果作为隐藏状态，且按照其对应于待识别字符串中字符的顺序将此隐藏状态相互连接，由此即获得了隐马尔科夫链。

通过隐马尔科夫链的构建，一方面标示了为待识别字符串输出候选字符的路径；另一方面，也是将有利于寻找得到最优路径，进而评估当前对待识别字符串所进行的识别是否正常，以便于对此进行反馈。

图 4 是根据一示例性实例示出的对步骤 250 的细节进行的描述。如图 4 所示，步骤 250，可以包括以下步骤。

在步骤 251 中，根据最优路径中等于预设阈值的状态转移概率，获得状态转移概率所连接的隐藏状态。

其中，如前所述的，通过对隐马尔科夫链所进行的最优路径求解将得到为待识别字符串输出候选结果的最优路径。由于最优路径是由隐马尔科夫链中标示的路径而选取是到的，因此，其与隐马尔科夫链可类似的，也是由相互连接的隐藏状态构成的，通过预训练的状态转移矩阵使得连接具备一定的概率，即状态转移概率。

也就是说，最优路径中，也将通过状态转移概率来评估作为隐藏状态的识别结果输出候选字符的可靠性。

所谓的状态转移概率是指相邻两个字符，前一个字符的状态转移到后一个字符的状态的概率，例如，图 3 中的两个相邻字符“大”和“学”，“大”为中文字符，“学”为中文字符，这两个相邻字符对应的状态转移概率为中文状态转移到中文状态的概率。再例如，相邻的两个字符为“唱”和“K”，字符“唱”为中文字符，字符“K”为英文字符，这两个字符对应的状态转移概率为中文字符转移到英文字符的概率。也就是说，一个状态转移概率对应两个相邻的字符。

本申请实例中，一个隐藏状态对应一个候选字符，每一个候选字符

都是待识别字符的隐藏状态。因此上述步骤中“根据最优路径中等于预设阈值的状态转移概率，获得状态转移概率所连接的隐藏状态”也可以描述成“获取所述最优路径中等于预设阈值的状态转移概率所对应的隐藏状态”。

预设阈值一方面用于最优路径中状态转移概率的评估，另一方面用于在预训练的状态转移矩阵中替代非中文字符的状态转移概率。

状态转移矩阵在由中文语言模型训练并进行平滑处理得到之后，将使用预设阈值替换其中非中文字符对应的状态转移概率。由此便使得状态转移概率中非中文字符对应的状态转移概率即为此预设阈值。

因此，在最优路径中，等于预设阈值的状态转移概率在待识别字符串中对应的字符即为非中文字符。

基于此，将由最优路径中具备等于预设阈值的状态转移概率的连接获得隐藏状态，此隐藏状态即为具备等于预设阈值的状态转移概率的连接所连接的隐藏状态。

在步骤253中，将待识别字符串中对应于隐藏状态的字符作为非中文字符从待识别字符串中分割出来。即，将所述待识别字符串中步骤251所获取的隐藏状态对应的字符识别为非中文字符。

其中，作为隐藏状态的识别结果所输出的候选字符是对应于待识别字符串中的字符的，因此隐藏状态将是对应于待识别字符串中的字符的。由于此隐藏状态与其所连接的其它隐藏状态之间状态转移矩阵即为预设阈值，因此，可以获知此隐藏状态所对应的待识别字符串中的字符是非中文字符的可能性非常高，进而将被从待识别字符串中分割出来。

通过如上所述的过程，得以对待识别字符串中存在的非中文字符，例如，英文字符和/或符号字符进行有效准确地识别。

可以理解的，待识别字符串所进行的字符识别大都是通过一种识别算法进行的，进而输出相应的识别结果，单一的识别算法并无法适配于各种类型的字符，并且在具体实现中，此识别算法大都是实现中文识别的，在此，通过识别结果的校验，来实现中文字符和非中文字符的有效区分，进而方能够提高识别效果。

在一些实例中，步骤250中，在执行步骤251之前，还可以包括以下步骤。

根据最优路径中的状态转移概率进行路径评分得到最优路径的路径分数。

如果路径分数等于预设阈值，则执行步骤251。

在一些实例中，还可以包括以下步骤：

如果路径分数大于预设阈值，则根据最优路径中大于预设阈值的状态转移概率从待识别字符串中分割出中文字符，待识别字符串中余下的字符即为分割得到的非中文字符。即，确定所述最优路径中大于预设阈值的状态转移概率对应的隐藏状态，将所述待识别字符串中所确定的隐藏状态对应的字符识别为中文字符，将所述待识别字符串中余下的字符作为所述非中文字符

其中，最优路径的路径分数可以是最优路径中最大状态转移概率，也可以是最优路径中状态转移概率的均值。

通过所进行的路径评分，将得以在获得识别结果的基础上对待识别字符串进行评估，进而判断当前所进行的字符识别是否真正适用于待识别字符串，由此方能够便于进行非中文字符的识别错误的纠正，指导系统在必要时重新进行识别。

通过此方式，也是相当于实现了全局统筹，进而优化全局的识别效果，避免实际中低级错误的存在。

在一示例性实例示出的一种字符识别中识别结果的校验方法中，还可以包括以下步骤。

将待识别字符串分割得到的非中文字符与预置的英文词表进行匹配，根据获得的匹配进行分割得到待识别字符串中的英文字符和符号字符。

其中，通过前述过程，获得了待识别字符串中的中文字符和非中文字符，而对于非中文字符，还将识别所存在的英文字符和/或符号字符。

具体的，预置了英文词表，将待识别字符串分割得到的非中文字符与预置的英文词表相匹配，进而与英文词表相匹配的字符即为英文字

符，与英文词表不相匹配的字符即为非英文字符。

在一示例性实例中，所进行的非中文字符和英文词表的匹配可以通过最小编辑距离的运算实现，也是可以通过其它距离运算实现，在此不进行限定。

图 5 是根据一示例性实例示出的一种字符识别中识别结果的校验方法。如图 5 所示，该字符识别中识别结果的校验方法，可以包括以下步骤。

在步骤 310 中，以互联网文本为语料预先进行中文语言模型的训练，并进行平滑处理得到预训练的状态转移矩阵。

其中，首先利用海量的互联网文本为作语料。在一示例性实例中，采用的中文语言模型 N-gram 中的 By-gram，也可以是其它的 N-gram。

此外，所采用的平滑处理算法可以是 good-turing 方法，也可以是其它平滑处理模型，例如，add-one 模型。

在步骤 330 中，通过预设阈值在预训练的状态转移矩阵进行阈值更新（即用该预设阈值去更新状态转移概率），使预训练的状态转移矩阵中非中文字符对应的状态转移概率被预设阈值替代。

其中，在通过中文语言模型的训练和平滑处理将得到非稀疏的状态转移矩阵。在此非稀疏的状态转移矩阵中，根据状态转移概率所对应数值的大小，可以获知存在着小概率事件，例如，英文字符的存在由于受到语料中组成成分的影响，而存在着被识别为中文字符的概率，此即为小概率事件的发生，此时将对存在的小概率事件都使用预设阈值替代，以方便后续对此进行识别，并且也是将减小状态转移矩阵所占用的存储空间，有利于其在内存中的存储和后续的计算。

在此需要说明的是，预训练的状态转移矩阵的获得以及所进行的阈值更新是可以在离线下训练完成的，进而在所进行的字符识别中直接使用即可。

结合具体应用场景，描述该字符识别中识别结果的校验方法。例如，

图 6 是一种场景下字符识别中识别结果校验的示意图。

首先将进行离线的语言模型训练，即步骤 510 所示的过程。在此过程中，将利用文本和 N-gram 语言模型训练、平滑处理过程来获得状态转移矩阵，并在状态转移矩阵中设置阈值，以此调整状态转移矩阵中的数值，为字符识别的反馈提供预训练的状态转移矩阵。

如图 6 所示，在完成 OCR 字符识别，即步骤 520 之后，便输出了相应的识别结果，由此识别结果进行隐马尔科夫链的构建，以形成隐马尔科夫链。

在步骤 550 中，将分别以状态转移矩阵和隐马尔科夫链作为输入，利用 Viterbi 算法求解最优路径。

对最优路径进行评分得到路径分数，以根据路径分数来判定待识别字符串是否正常，例如，是否存在着非中文字符，如果存在，则需要进行返回，以通知所进行的 OCR 字符识别过程，即步骤 520 重新采用与当前待识别字符串相适应的识别算法进行识别，如果正常，则直接输出结果即可，即如步骤 580 所示的。

下述为本申请装置实例，可以用于执行本申请上述字符识别中识别结果的校验方法。对于本申请装置实例中未披露的细节，请参照本申请字符识别中识别结果的校验方法实例。

图 7 是根据一示例性实例示出的一种字符识别中识别结果的校验装置的框图。如图 7 所示，该字符识别中识别结果的校验装置包括但不限于：

一个或一个以上存储器；

一个或一个以上处理器；其中，

所述一个或一个以上存储器存储有一个或者一个以上指令模块，经配置由所述一个或者一个以上处理器执行；其中，

所述一个或者一个以上指令模块包括：

隐马尔科夫链构建模块 710、最优路径求解模块 730、非中文字符

识别模块 750 和反馈模块 770。

隐马尔科夫链构建模块 710，用于通过字符识别输出的识别结果构建待识别字符串的隐马尔科夫链，识别结果包含待识别字符串中各字符的候选字符。其中，隐马尔科夫链仅仅是预测状态统计模型中的一种，当采用其他的预测状态统计模型时，隐马尔科夫链构建模块 710 可以称为创建模块。

在一示例性实例中，隐马尔科夫链构建模块 710 进一步用于根据待识别字符串中字符的顺序以输出相应候选字符的识别结果为隐藏状态，隐藏状态相互连接构建得到待识别字符串的隐马尔科夫链。

最优路径求解模块 730，用于根据隐马尔科夫链和预训练的状态转移矩阵求解识别结果输出候选字符的最优路径。

非中文字符识别模块 750，用于根据路径中的状态转移概率识别待识别字符串中的非中文字符。

反馈模块 770，用于根据待识别字符串中的非中文字符进行识别结果校验，并将校验结果反馈至字符识别过程。

图 8 是根据一示例性实例示出的对非中文字符识别模块的细节进行的描述。如图 8 所示，非中文字符识别模块 730，可以包括但不限于：隐藏状态获得单元 731 和分割单元 733。

隐藏状态获得单元 731，用于根据最优路径中等于预设阈值的状态转移概率，获得状态转移概率所连接的隐藏状态；

分割单元 733，用于将待识别字符串中对应于隐藏状态的字符作为非中文字符从待识别字符串中分割出来。

在一些实例中，非中文字符识别模块 730 还包括路径评分单元。该路径评分单元用于根据最优路径中的状态转移概率进行路径评分得到最优路径的路径分数。

如果路径分数等于预设阈值，则通知隐藏状态获得单元 731。

在一些实例中，非中文字符识别模块 730 还包括中文字符分割单元。

该中文字符分割单元用于如果路径分数大于预设阈值，则根据最优

路径中大于预设阈值的状态转移概率从待识别字符串中分割出中文字符，待识别字符串中余下的字符即为分割得到的非中文字符。

在一些实例中，非中文字符识别模块730还包括匹配单元。该匹配单元用于将待识别字符串分割得到的非中文字符与预置的英文词表进行匹配，根据获得的匹配结果分割得到待识别字符串中的英文字符和符号字符。

图9是根据另一示例性实例示出的一种字符识别中识别结果的校验装置。如图9所示，该字符识别中识别结果的校验装置还包括但不限于：状态转移矩阵预训练模块810和阈值更新模块830。

状态转移矩阵预训练模块810，用于以互联网文本为语料预先进行中文语言模型的训练，并进行平滑处理得到预训练的状态转移矩阵。

阈值更新模块830，用于通过预设阈值在预训练的状态转移矩阵进行阈值更新，使预训练的状态转移矩阵中非中文字符对应的状态转移概率被预设阈值替代。

在一些实例中，本申请还提供一种字符识别中识别结果的校验装置，该字符识别中识别结果的校验装置，执行图2、图4和4任一所示的字符识别中识别结果的校验方法的全部或者部分步骤。所述装置包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为执行：

通过字符识别输出的识别结果构建待识别字符串的隐马尔科夫链，识别结果包含待识别字符串中各字符的候选字符；

根据隐马尔科夫链和预训练的状态转移矩阵求解识别结果输出候选字符的最优路径；

根据最优路径中的状态转移概率识别待识别字符串中的非中文字符；

根据待识别字符串中的非中文字符进行识别结果校验，并将校验结果反馈至字符识别过程。

该实例中的装置的处理器执行操作的具体方式已经在有关该字符识别中识别结果的校验方法的实例中执行了详细描述，此处将不做详细阐述说明。

本申请另一实例还提供一种非易失性计算机可读存储介质，其上存储有计算机可读指令，可以使至少一个处理器执行上述方法，例如：

通过字符识别输出的识别结果构建待识别字符串的隐马尔科夫链，所述识别结果包含待识别字符串中各字符的候选字符；

根据所述隐马尔科夫链和预训练的状态转移矩阵求解所述识别结果输出所述候选字符的最优路径；

根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符；

根据所述待识别字符串中的非中文字符进行所述识别结果校验，并将校验结果反馈至字符识别过程。

本领域普通技术人员可以理解实现上述实例的全部或部分步骤可以通过硬件来完成，也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，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

应当理解的是，本发明并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围执行各种修改和改变。本发明的范围仅由所附的权利要求来限制。

权利要求书

1、一种字符识别中识别结果的校验方法，包括：

通过字符识别过程输出的识别结果构建待识别字符串的预测状态统计模型，所述识别结果包含所述待识别字符串中各待识别字符的候选字符；每个待识别字符对应至少一个候选字符；

根据所述预测状态统计模型和预训练的状态转移矩阵求解形成候选字符串的最优路径；所述候选字符串包含分别对应各待识别字符的至少一个所述候选字符；

根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符；

根据所述待识别字符串中的非中文字符进行所述识别结果校验，并将校验结果反馈至所述字符识别过程。

2、根据权利要求 1 所述的方法，其中，所述预测状态统计模型为隐马尔科夫链。

3、根据权利要求 2 所述的方法，其中，所述通过字符识别过程输出的识别结果构建待识别字符串的预测状态统计模型的步骤包括：

根据所述待识别字符串中字符的顺序以输出所述相应候选字符的识别结果为隐藏状态，所述隐藏状态相互连接构建得到待识别字符串的隐马尔科夫链。

4、根据权利要求 2 所述的方法，其中，所述根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符的步骤包括：

获取所述最优路径中等于预设阈值的状态转移概率所对应的隐藏状态，所述预设阈值用于在所述预训练的状态转移矩阵中替代非中文字符的状态转移概率；

将所述待识别字符串中所获取的隐藏状态对应的字符识别为非中文字符。

5、根据权利要求 4 所述的方法，其中，所述根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符的步骤还包括：

根据所述最优路径中的状态转移概率，进行路径评分，得到所述最优路径的路径分数；

如果所述路径分数等于所述预设阈值，则执行所述获取所述最优路径中等于预设阈值的状态转移概率所对应的隐藏状态的步骤。

6、根据权利要求 5 所述的方法，其中，所述根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符的步骤还包括：

如果所述路径分数大于所述预设阈值，则确定所述最优路径中大于预设阈值的状态转移概率对应的隐藏状态，将所述待识别字符串中所确定的隐藏状态对应的字符识别为中文字符，将所述待识别字符串中余下的字符作为所述非中文字符。

7、根据权利要求 4 所述的方法，其中，所述根据所述最优路径中的状态转移概率识别所述待识别字符串中的非中文字符的步骤还包括：

将所述待识别字符串中的所述非中文字符与预置的英文词表进行匹配，根据获得的匹配结果确定所述待识别字符串中的英文字符和符号字符。

8、根据权利要求 1 或 2 所述的方法，其中，所述方法还包括：

以互联网文本为语料预先进行中文语言模型的训练，并进行平滑处理得到所述预训练的状态转移矩阵；

通过预设阈值对所述预训练的状态转移矩阵进行状态转移概率更新，使所述预训练的状态转移矩阵中非中文字符对应的状态转移概率被所述预设阈值替代。

9、一种字符识别中识别结果的校验装置，所述装置包括：

一个或一个以上存储器；

一个或一个以上处理器；其中，

所述一个或一个以上存储器存储有一个或者一个以上指令模块，经配置由所述一个或者一个以上处理器执行；其中，

所述一个或者一个以上指令模块包括：

构建模块，通过字符识别过程输出的识别结果构建待识别字符串的预测状态统计模型，所述识别结果包含待识别字符串中各待识别字符的候选字符；每个待识别字符对应至少一个候选字符；

最优路径求解模块，根据所述预测状态统计模型和预训练的状态转移矩阵求解形成候选字符串的最优路径；所述候选字符串包含分别对应各待识别字符的至少一个所述候选字符；；

非中文字符识别模块，根据所述路径中的状态转移概率识别所述待识别字符串中的非中文字符；

反馈模块，根据所述待识别字符串中的非中文字符进行所述识别结果校验，并将校验结果反馈至所述字符识别过程。

10 根据权利要求 9 所述的装置，所述预测状态统计模型为隐马尔科夫链。

11、根据权利要求 10 所述的装置，其中，所述构建模块进一步根据所述待识别字符串中字符的顺序以输出所述相应候选字符的识别结果为隐藏状态，所述隐藏状态相互连接构建得到待识别字符串的隐马尔科夫链。

12、根据权利要求 10 所述的装置，其中，所述非中文字符识别模块包括：

隐藏状态获得单元，获取所述最优路径中等于预设阈值的所述状态转移概率对应的隐藏状态；

分割单元，将所述待识别字符串中所获取的隐藏状态对应的字符识

别为非中文字符。

13、根据权利要求 12 所述的装置，其中，所述非中文字符识别模块还包括：

路径评分单元，根据所述最优路径中的状态转移概率，进行路径评分，得到所述最优路径的路径分数；

如果所述路径分数等于所述预设阈值，则通知所述隐藏状态获得单元执行所述获取所述最优路径中等于预设阈值的所述状态转移概率对应的隐藏状态的步骤。

14、根据权利要求 13 所述的装置，其中，所述非中文字符识别模块还包括：

中文字符分割单元，如果所述路径分数大于所述预设阈值，则确定所述最优路径中大于所述预设阈值的状态转移概率对应的隐藏状态，将所述待识别字符串中所确定的隐藏状态对应的字符识别为中文字符，将所述待识别字符串中余下的字符识别为所述非中文字符。

15、根据权利要求 12 所述的装置，其中，所述非中文字符识别模块还包括：

匹配单元，将所述待识别字符串中的所述非中文字符与预置的英文词表进行匹配，根据获得的匹配结果确定所述待识别字符串中的英文字符和符号字符。

16、根据权利要求 9 或 10 所述的装置，其中，所述装置还包括：

状态转移矩阵预训练模块，用于以互联网文本为语料预先进行中文语言模型的训练，并进行平滑处理得到所述预训练的状态转移矩阵；

阈值更新模块，用于通过预设阈值对所述预训练的状态转移矩阵进行状态转移概率更新，使所述预训练的状态转移矩阵中非中文字符对应的状态转移概率被所述预设阈值替代。

17、一种非易失性计算机可读存储介质，其上存储有计算机可读指令，可以使至少一个处理器执行权利要求1~8任一项所述的方法。

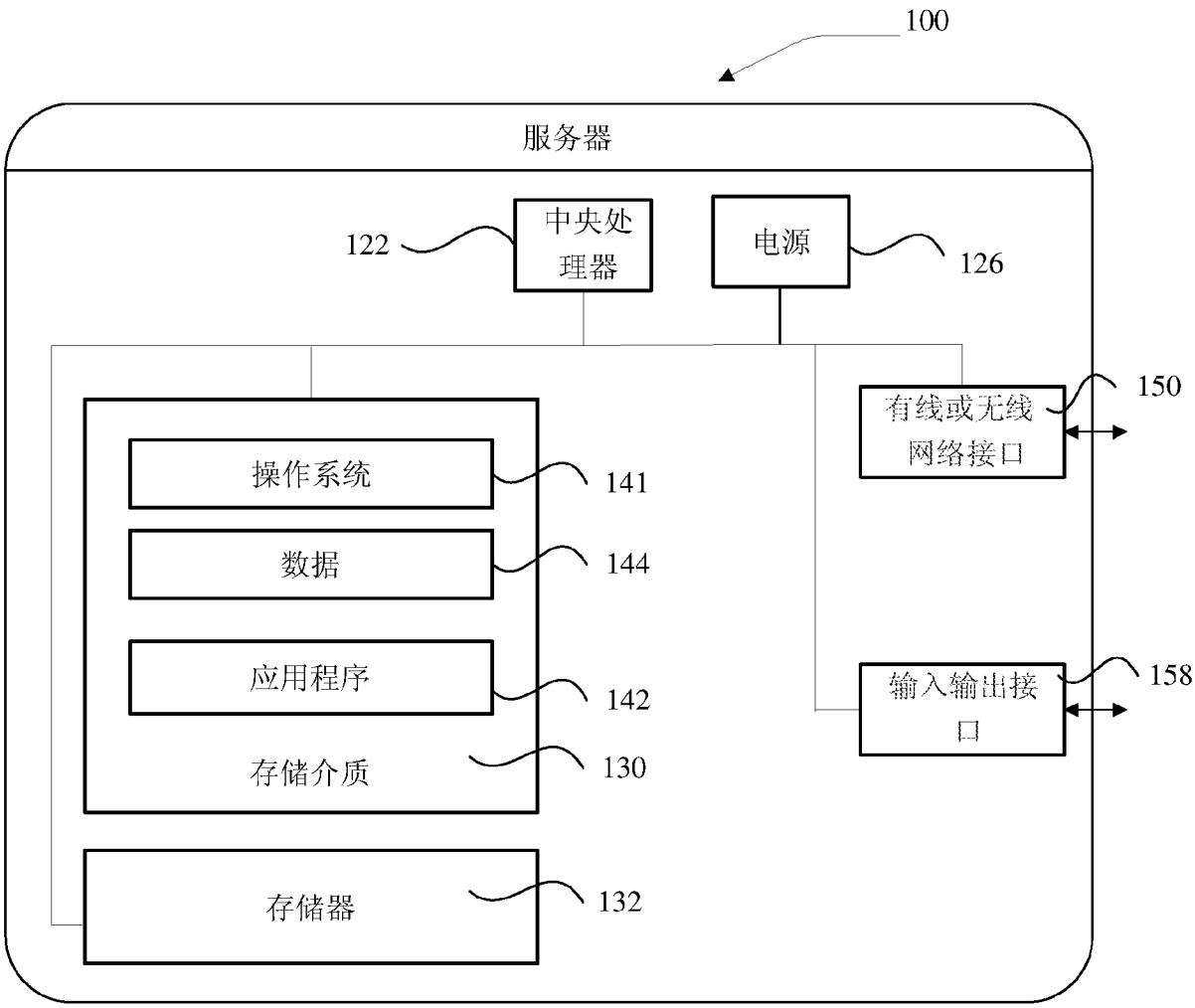
100
服务器
122
中央处理器
电源
126
150
有线或无线网络接口
操作系统
141
数据
144
应用程序
142
存储介质
130
158
输入输出接口
存储器
132

图 1

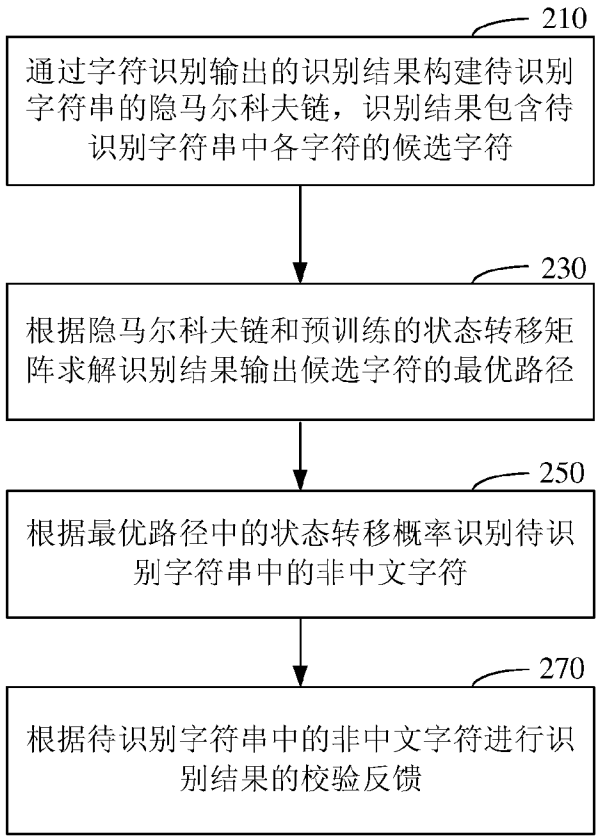
210
通过字符识别输出的识别结果构建待识别字符串的隐马尔科夫链，识别结果包含待识别字符串中各字符的候选字符
230
根据隐马尔科夫链和预训练的状态转移矩阵求解识别结果输出候选字符的最优路径
250
根据最优路径中的状态转移概率识别待识别字符串中的非中文字符
270
根据待识别字符串中的非中文字符进行识别结果的校验反馈

图 2


北 京 大 学 深 圳 医 院 临 检 组 检 验 报 告 单
北 京 大 学 深 圳 医 院 临 检 组 检 验 报 告 单
94.9791 94.3612 95.0932 94.6586 95.1069 95.0739 94.611 93.3442 95.0921 93.6531 96.2801 93.8983 95.7302 93.3324 96.9318 95.6215
比 京 太 学 深 圳 医 院 检 检 检 险 报 告 单
89.0444 91.9996 91.8511 92.3726 91.7893 91.4242 92.4556 93.2962 90.1383 92.8246 95.9401 92.6399 93.2811 93.0594 93.6755 91.608
基 深 大 学 深 圳 医 院 检 检 检 验 报 告 单
88.6287 91.222 89.6745 91.0885 91.7135 90.4902 92.4313 91.0937 89.8689 92.6651 93.0306 92.2574 92.8618 92.5848 89.7083 91.5862

图 3

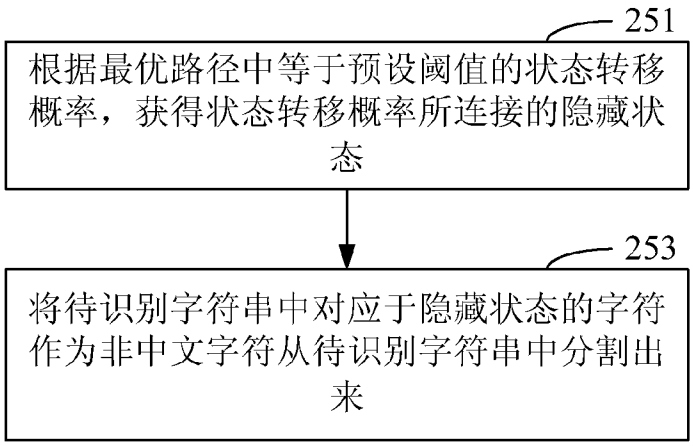
251
根据最优路径中等于预设阈值的状态转移概率，获得状态转移概率所连接的隐藏状态
253
将待识别字符串中对应于隐藏状态的字符作为非中文字符从待识别字符串中分割出来

图 4

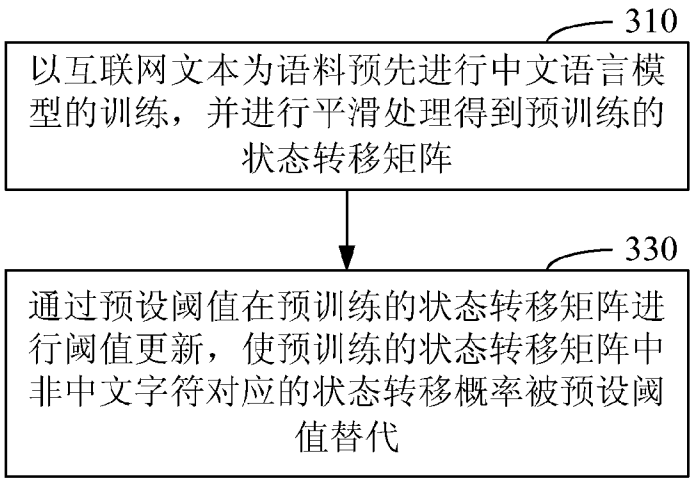
310
以互联网文本为语料预先进行中文语言模型的训练，并进行平滑处理得到预训练的状态转移矩阵
330
通过预设阈值在预训练的状态转移矩阵进行阈值更新，使预训练的状态转移矩阵中非中文字符对应的状态转移概率被预设阈值替代

图 5

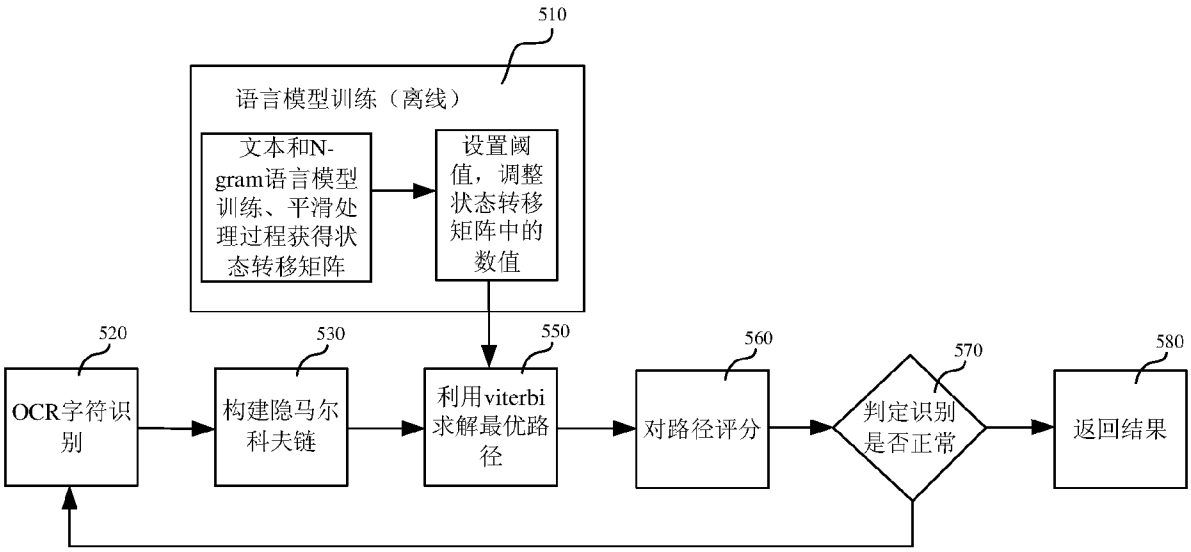
510
语言模型训练（离线）
文本和N-gram语言模型训练、平滑处理过程获得状态转移矩阵
设置阈值，调整状态转移矩阵中的数值
520
OCR字符识别
530
构建隐马尔科夫链
550
利用viterbi求解最优路径
560
对路径评分
570
判定识别是否正常
580
返回结果

图 6

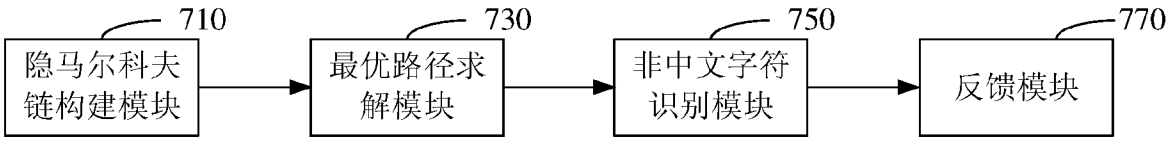
710
隐马尔科夫链构建模块
730
最优路径求解模块
750
非中文字符识别模块
770
反馈模块

图 7

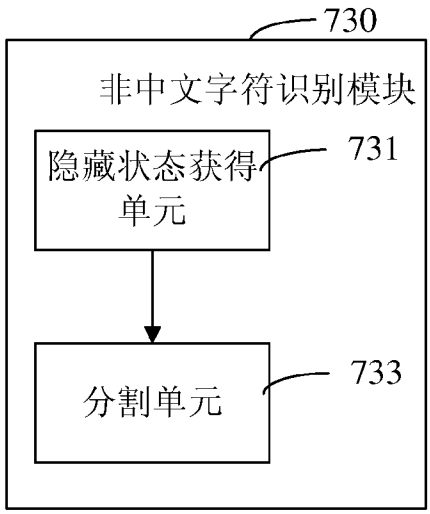
730
非中文字符识别模块
隐藏状态获得单元
731
分割单元
733

图 8

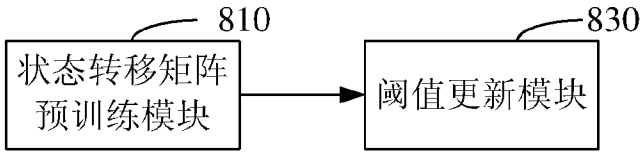
810
状态转移矩阵预训练模块
830
阈值更新模块

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/095992

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/20 (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)

CNPAT, EPODOC, WPI, GOOGLE, CNKI: checkout, recognition, statistics, correct, feedback, predict+, probability, matrix, transfer, shift, candidate, static, path, character

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104951779 A (CHINA UNIONPAY CO., LTD.), 30 September 2015 (30.09.2015), description, paragraphs 18-37	1-17
A	CN 102024139 A (FUJITSU LIMITED), 20 April 2011 (20.04.2011), the whole document	1-17
A	CN 102982330 A (SINA.COM TECHNOLOGY (CHINA) CO., LTD.), 20 March 2013 (20.03.2013), the whole document	1-17
A	US 2016125275 A1 (KABUSHIKI KAISHA TOSHIBA et al.), 05 May 2016 (05.05.2016), the whole document	1-17

☐ 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	Date of mailing of the international search report
24 October 2017 (24.10.2017)	**03 November 2017 (03.11.2017)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **WANG, Jing** Telephone No.: (86-10) **62413685**

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/095992

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104951779 A	30 September 2015	None	
CN 102024139 A	20 April 2011	JP 2011065646 A	31 March 2011
CN 102982330 A	20 March 2013	None	
US 2016125275 A1	05 May 2016	CN 105574523 A	11 May 2016
		JP 2016091200 A	23 May 2016

国际检索报告

国际申请号
PCT/CN2017/095992

A. 主题的分类

G06K 9/20(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）)

CNPAT, EPODOC, WPI, GOOGLE, CNKI：校验，反馈，预测，概率，识别，字符，转移，矩阵，候选，统计，路径，correct, feedback, predict+, probability, matrix, transfer, shift, candidate, static, path, character

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
A	CN 104951779 A (中国银联股份有限公司) 2015年 9月 30日 (2015-09-30) 说明书第18-37段	1-17
A	CN 102024139 A (富士通株式会社) 2011年 4月 20日 (2011-04-20) 全文	1-17
A	CN 102982330 A (新浪网技术中国有限公司) 2013年 3月 20日 (2013-03-20) 全文	1-17
A	US 2016125275 A1 (KABUSHIKI KAISHA TOSHIBA 等) 2016年 5月 5日 (2016-05-05) 全文	1-17

☐ 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
2017年 10月 24日	2017年 11月 3日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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	受权官员 王晶 电话号码 (86-10)62413685

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/095992

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	104951779	A	2015年 9月 30日		无		
CN	102024139	A	2011年 4月 20日	JP	2011065646	A	2011年 3月 31日
CN	102982330	A	2013年 3月 20日		无		
US	2016125275	A1	2016年 5月 5日	CN	105574523	A	2016年 5月 11日
				JP	2016091200	A	2016年 5月 23日