[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06K 9/00
G06K 7/00

[12] 发明专利说明书

[21] ZL 专利号 99123532.0

[45] 授权公告日 2003 年 10 月 15 日

[11] 授权公告号 CN 1124562C

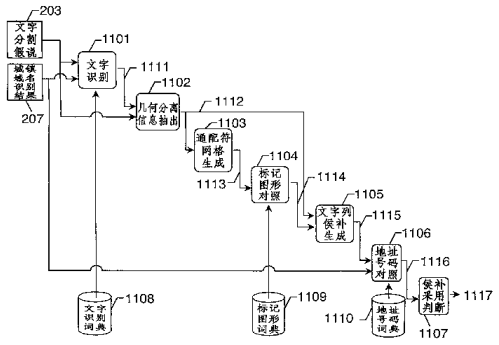
[22] 申请日 1999.11.10 [21] 申请号 99123532.0
[30] 优先权
[32] 1998.11.10 [33] JP [31] 318693/1998
[71] 专利权人 株式会社日立制作所
地址 日本东京
[72] 发明人 绪方日佐男 藤泽浩道 古川直广 洒匂裕 寺本正人 渡边成
审查员 哈雅坤
[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 王以平

权利要求书 4 页 说明书 17 页 附图 23 页

[54] 发明名称 文书识别装置和信件分检机

[57] 摘要

本发明提供在地址名标记中存在意义的解释有不确定性的文字间的空白信息或换行信息时高精度地识别解释地址信息的装置。经文字识别处理输出各文字图形的候补文字网格。几何分离信息抽出单元抽出地址文字串中的分离信息假说，并与候补文字网格关联地输出。通过通配符网格生成处理生成用表示任意的英文数字的通配符符号表现的网格。标记图形对照单元标记图形词典与通配符网格进行对照。



ISSN1008-4274

1. 一种文书识别装置，其特征在于包括：

将记载在纸页上的地址信息作为数字图像数据进行读取的图像读取单元；

从上述数字图像数据中分割并输出包含上述地址信息的文字行的文字行分割单元；

具有文字识别词典、输入所述分割出的上述文字行并分割出文字信息、参照上述文字识别词典进行文字识别并输出多个文字识别候补的文字识别单元；

具有城镇域名对照词典的输入上述多个文字识别候补、参照上述城镇域名对照词典对上述多个文字识别候补对照城镇域名并输出该城镇域名候补的城镇域名对照单元；

输入上述城镇域名候补并判断是否将上述城镇域名候补作为城镇域名识别结果采用，在判定采用时将上述城镇域名候补作为城镇域名识别结果而输出的第 1 候补采用判断单元；

号码对照单元，具有保持包含用符号或数字表示的街道号码、门牌号码、房间号码中的某一个的号码的标记图形的标记图形词典，以及保持包含城镇域名与街道号码、门牌号码、房间号码中的某一个的号码的对应关系的号码词典，用于根据上述多个文字识别候补输出包含街道号码、门牌号码、房间号码中的某一个的号码候补；以及

判断是否将上述号码候补作为号码识别结果采用在判定采用时就将上述号码候补作为号码识别结果输出的第 2 候补采用判断单元，

其中，上述号码对照单元包含几何分离信息抽出单元、通配符变换单元、标记图形对照单元和文字串候补生成单元，所述几何分离信息抽出单元检测在上述多个文字识别候补中包含的一个或多个文字的几何配置关系、把分离上述多个文字识别候补中包含的文字串的分离信息的属性加到所述文字识别候补中；所述通配符变换单元把该附加了分离信息的上述文字识别候补中的文字图形用通配符变换规则

进行置换；所述标记图形对照单元把上述变换后的文字识别候补中所含的一个或多个文字与上述标记图形词典相对照，输出用所述分离信息得到的标记图形候补；以及所述文字串候补生成单元把上述标记图形候补中由通配符变换单元变换后的文字复原，然后输出文字串候补；

所述号码对照单元把上述文字串候补和上述号码词典相对照，输出号码候补。

2. 如权利要求 1 所述的文书识别装置，其特征在于：上述号码对照单元进行对照号码的上述多个文字识别候补是从上述文字识别单元输出的上述多个文字识别候补中排除了由上述城镇域名对照单元对照过的上述多个文字识别候补的多个文字识别候补。

3. 如权利要求 1 所述的文书识别装置，其特征在于：上述几何配置关系是在上述多个文字识别候补中包含的 2 个文字间的空白或在上述多个文字识别候补中包含的换行。

4. 如权利要求 1 所述的文书识别装置，其特征在于：上述几何分离信息抽出单元通过求在上述多个文字识别候补中包含的文字的文字图像的重心、测定相邻的文字间的距离来检测上述 2 个文字间的空白。

5. 如权利要求 1 所述的文书识别装置，其特征在于：上述号码对照单元将在上述多个文字识别候补中包含的 1 个或多个文字置换为符号或数字，与上述标记图形词典进行对照。

6. 一种信件分捡机，包括：

提供纸页的供给机构；

将记载在被供给的纸页上的姓名、地址信息作为数字图像数据进行读取的扫描器；

从上述数字图像数据中分割并输出包含上述姓名、地址信息的文字行的文字行分割单元；

具有文字识别词典、输入该分割出的上述文字行并分割出文字信息、参照上述文字识别词典进行文字识别并输出多个文字识别候补的

文字识别单元;

具有城镇域名对照词典的输入上述多个文字识别候补、参照上述城镇域名对照词典对上述多个文字识别候补对照城镇域名并输出该城镇域名候补的城镇域名对照单元;

传送由所述扫描器读取的纸页的传送机构，以及

根据由上述第2候补采用判断单元所采用的地址号码进行纸页的分检的分检机构，

输入上述城镇域名候补并判断是否将上述城镇域名候补作为城镇域名识别结果采用，在判定采用时将上述城镇域名候补作为城镇域名识别结果而输出的第1候补采用判断单元;

地址号码对照单元；具有保持包含用符号或数字表示的街道号码、门牌号码、房间号码中的某一个的地址号码的标记图形的标记图形词典和保持包含城镇域名与街道号码、门牌号码、房间号码中的某一个的地址号码的对应关系的地址号码词典，用于输入上述多个文字识别候补和上述城镇域名识别结果，根据上述多个文字识别候补输出包含街道号码、门牌号码、房间号码中的某一个的地址号码候补;

输入上述地址号码候补、判断是否将上述地址号码候补作为地址号码识别结果采用，在判定采用时就将上述地址号码候补作为地址号码识别结果输出的第2候补采用判断单元;

其特征在于：上述地址号码对照单元包含几何分离信息抽出单元、通配符变换单元、标记图形对照单元和文字串候补生成单元，所述几何分离信息抽出单元检测在上述多个文字识别候补中包含的一个或多个文字的几何配置关系、把分离上述多个文字识别候补中包含的文字串的分离信息的属性加到所述文字识别候补中；所述通配符变换单元把该附加了分离信息的上述文字识别候补中的文字图形用通配符变换规则进行置换；所述标记图形对照单元把上述变换后的文字识别候补中所含的一个或多个文字与上述标记图形词典相对照，输出用所述分离信息得到的标记图形候补；以及所述文字串候补生成单元把上述标记图形候补中由通配符变换单元变换后的文字复原，然后输

出文字串候补;

所述地址号码对照单元把上述文字串候补和上述地址号码词典相对照，输出地址号码候补。

7. 如权利要求 6 所述的信件分检机，其特征在于：上述地址号码对照单元进行地址号码对照的上述多个文字识别候补是从上述文字识别单元输出的上述多个文字识别候补中排除了由上述城镇域名对照单元对照过的上述多个文字识别候补的多个文字识别候补。

8. 如权利要求 6 所述的信件分检机，其特征在于：上述几何配置关系是在上述多个文字识别候补中包含的 2 个文字间的空白或在上述多个文字识别候补中包含的换行。

9. 如权利要求 6 所述的信件分检机，其特征在于：上述几何分离信息抽出单元通过求在上述多个文字识别候补中包含的文字的文字图像的重心、测定相邻的文字间的距离来检测上述 2 个文字间的空白。

10. 如权利要求 6 所述的信件分检机，其特征在于：上述地址号码对照单元将在上述多个文字识别候补中包含的 1 个或多个文字置换为符号或数字，与上述标记图形词典进行对照。

文书识别装置和信件分检机

本发明涉及使用文字识别技术读取并解释记载在纸面上的文字信息的文书识别装置和使用该文书识别装置的信件分检机。

以往，在使用文字识别技术读取用文字串表示的纸面上的文字信息时，是按照从左到右（从上到下）的顺序一个字一个字地读取该文字串的。并且，根据该读取结果，在比方说表示地址信息时就依次给出（日本的场合）县名、市镇村名、街道号码、门牌号码、房间号码等解释。其中，为了对街道号码、门牌号码、房间号码给出解释，使用与它们相关的标记的知识是很有效的，如日本专利公报“住所读取装置和方法”（特开平 8－234505）所公开的那样。这里，简单地描述一下它的内容。作为街道号码的例子，虽然“1－2”、“1番2号”、“一番二号”是不同的标记，但是，却全部表示相同的住所。只要能判别纸面上的地址信息与哪种标记对应，就可以提高读取精度，这是很容易想象的。因此，为了使用这些标记知识，预先将表示标记图形的文字串保持在词典中，将该标记图形与一个字一个字地读取的结果进行对照，从而给出对街道号码和房间号码的解释。与上述“1－2”、“1番2号”、“一番二号”对应的标记图形为“N－N”、“N番N号”、“K番K号”。这里，“N”是表示数字的通配符符号，“K”是表示汉字数字的通配符符号。“－”、“番”、“号”是将英文数字串分开的分离符号。

作为应用这样的文书识别装置的领域之一，有采用信件分检机进行邮件分拣的业务。在该业务中，除了地址中的镇域名外，为了确定街道号码、门牌号码、房间号码等住所，必须全部读取所需要的文字信息，重新排列为邮件的投递人进行投递的顺序。以往，该业务全部是由人手进行的，随着使用信件分检机的机械化，可以压缩人工费和解决人手不足的问题。

另一方面，由于分检机不能读取的邮件还得由人手进行分检，所以，这部分工作将增加人工费用。因此，提高信件分检机的地址信息的读取率，对降低投递成本是非常重要的。特别是10万件邮件的1%就是1000件，提高读取精度所带来的降低成本的效果是非常高的。因此，街道号码、门牌号码、房间号码的高识别率在信件分检机的领域具有非常重要的意义。

但是，在信件的地址名中，存在下列影响高精度的识别率的标记上的问题。

（1）文字图形间的空白

在信件地址中，如图1（1）所示的那样，为了表现街道号码与房间号码的分界，使文字图形间的间隔比周围的文字分得开，存在在该处存在空白的标记。在打字时，文字多数是按等间距书写的，所以，该空白信息的检测很容易。但是，如（2）所示的那样用手写时，文字图形的间距及大小的变化很大，所以，在文字图形间是否存在空白的信息就不一定明确。因此，在检测到存在空白时，判断数字之间是否分离并读取正确的地址信息是非常困难的。

（2）换行

在信件地址名中，如图1（3）所示的那样，为了表现街道号码与房间号码的分界，有时也使用换行标记。另一方面，如（4）的例子所示的那样，在无意分界的地方同样也存在被强制地换行的情形。这样的例子，在直接邮寄的信件中是常见的标记，在1行的文字数超过一定值时，就强制地换行来打印住所。因此，在地址标记中存在换行的信息时，根据换行来判断数字之间是否分离并读取正确的地址信息是非常困难的。

因此，本发明的目的旨在高精度地读取街道号码、门牌号码、房间号码，解决以下2个问题:

（1）在识别装置根据文字图形间的空白信息的解释和使用该信息的地址的高精度识别文字图形间的间距及文字大小检测到在该处存在空白的信息时，必须判断该信息是否可以解释为就是用于分离

街道号码、门牌号码、房间号码的分离信息。

本发明的目的在于，在解释中存在不确定性时，就建立是分离信息和不是分离信息两方面的假说，并在以后的处理中进行验证，从而实现高精度的识别方式。

（2）换行信息的解释和使用该信息对地址的高精度识别

识别装置根据文字图形间的相对位置关系检测到在该处存在行信息时，必须判断该信息是否可以解释为就是用于分离街道号码、门牌号码、房间号码的分离信息。本发明的目的在于，在解释中存在不确定性时，就建立是分离信息和不是分离信息两方面的假说，并在以后的处理阶段进行验证，从而实现高精度的识别的方式。

为了解决上述问题，本发明提供了一种文书识别装置，其特征在于包括：将记载在纸页上的地址信息作为数字图像数据进行读取的图像读取单元；从上述数字图像数据中分割并输出包含上述地址信息的文字行的文字行分割单元；具有文字识别词典、输入所述分割出的上述文字行并分割出文字信息、参照上述文字识别词典进行文字识别并输出多个文字识别候补的文字识别单元；具有城镇域名对照词典的输入上述多个文字识别候补、参照上述城镇域名对照词典对上述多个文字识别候补对照城镇域名并输出该城镇域名候补的城镇域名对照单元；输入上述城镇域名候补并判断是否将上述城镇域名候补作为城镇域名识别结果采用，在判定采用时将上述城镇域名候补作为城镇域名识别结果而输出的第 1 候补采用判断单元；号码对照单元，具有保持包含用符号或数字表示的街道号码、门牌号码、房间号码中的某一个的号码的标记图形的标记图形词典，以及保持包含城镇域名与街道号码、门牌号码、房间号码中的某一个的号码的对应关系的号码词典，用于根据上述多个文字识别候补输出包含街道号码、门牌号码、房间号码中的某一个的号码候补；以及判断是否将上述号码候补作为号码识别结果采用在判定采用时就将上述号码候补作为号码识别结果输出的第 2 候补采用判断单元，其中，上述号码对照单元包含几何分离信息抽出单元、通配符变换单元、标记图形对照单元和文字串候补生

成单元，所述几何分离信息抽出单元检测在上述多个文字识别候补中包含的一个或多个文字的几何配置关系、把分离上述多个文字识别候补中包含的文字串的分离信息的属性加到所述文字识别候补中；所述通配符变换单元把该附加了分离信息的上述文字识别候补中的文字图形用通配符变换规则进行置换；所述标记图形对照单元把上述变换后的文字识别候补中所含的一个或多个文字与上述标记图形词典相对照，输出用所述分离信息得到的标记图形候补；以及所述文字串候补生成单元把上述标记图形候补中由通配符变换单元变换后的文字复原，然后输出文字串候补；把上述文字串候补和上述号码词典相对照，输出号码候补。

本发明还提供了一种信件分捡机，包括：提供纸页的供给机构；将记载在被供给的纸页上的姓名、地址信息作为数字图像数据进行读取的扫描器；从上述数字图像数据中分割并输出包含上述姓名、地址信息的文字行的文字行分割单元；具有文字识别词典、输入该分割出的上述文字行并分割出文字信息、参照上述文字识别词典进行文字识别并输出多个文字识别候补的文字识别单元；具有城镇域名对照词典的输入上述多个文字识别候补、参照上述城镇域名对照词典对上述多个文字识别候补对照城镇域名并输出该城镇域名候补的城镇域名对照单元；传送由所述扫描器读取的纸页的传送机构，以及根据由上述第 2 候补采用判断单元所采用的地址号码进行纸页的分检的分检机构，输入上述城镇域名候补并判断是否将上述城镇域名候补作为城镇域名识别结果采用，在判定采用时将上述城镇域名候补作为城镇域名识别结果而输出的第 1 候补采用判断单元；地址号码对照单元；具有保持包含用符号或数字表示的街道号码、门牌号码、房间号码中的某一个的地址号码的标记图形的标记图形词典和保持包含城镇域名与街道号码、门牌号码、房间号码中的某一个的地址号码的对应关系的地址号码词典，用于输入上述多个文字识别候补和上述城镇域名识别结果，根据上述多个文字识别候补输出包含街道号码、门牌号码、房间号码中的某一个的地址号码候补；输入上述地址号码候补、判断是

否将上述地址号码候补作为地址号码识别结果采用，在判定采用时就将上述地址号码候补作为地址号码识别结果输出的第2候补采用判断单元；其特征在于：上述地址号码对照单元包含几何分离信息抽出单元、通配符变换单元、标记图形对照单元和文字串候补生成单元，所述几何分离信息抽出单元检测在上述多个文字识别候补中包含的一个或多个文字的几何配置关系、把分离上述多个文字识别候补中包含的文字串的分离信息的属性加到所述文字识别候补中；所述通配符变换单元把该附加了分离信息的上述文字识别候补中的文字图形用通配符变换规则进行置换；所述标记图形对照单元把上述变换后的文字识别候补中所含的一个或多个文字与上述标记图形词典相对照，输出用所述分离信息得到的标记图形候补；以及所述文字串候补生成单元把上述标记图形候补中由通配符变换单元变换后的文字复原，然后输出文字串候补；把上述文字串候补和上述地址号码词典相对照，输出地址号码候补。

图1是表示街道号码、门牌号码、房间号码标记的示例图。

图2是表示文书识别装置的基本功能结构图。

图3是表示地址名文字信息的示例图。

图4是表示文字行分割结果的示例图。

图5是表示文字分割假说网络的示意图。

图6是表示城镇域名识别数据流的示意图。

图7是表示与所选择的路径对应的文字候补网格图。

图8是表示城镇域名知识的示例图。

图9是表示使用有限自动装置的单词对照的示意图。

图 10 是表示单词候补网格的形式的例子的示意图。

图 11 是表示街道号码、门牌号码、房间号码识别处理流程图。

图 12 是表示几何分离信息抽出结果的示例图。

图 13 是表示带几何分离信息属性的网格的示例图。

图 14 是表示通配符变换规则的示意图。

图 15 是表示通配符网格的示意图。

图 16 是表示街道号码、门牌号码、房间号码的标记图形知识的示例图。

图 17 是表示使用有限自动装置的表示图形的单词对照的示意图。

图 18 是表示单词候补网格的形式的示例图。

图 19 是表示标记图形候补生成结果的示意图。

图 20 是表示英文数字复原候补的示意图。

图 21 是表示地址号码词典的结构例的示意图。

图 22 表示了一个分类符号串候补树的例子。

图 23 是表示包括文书识别装置的信件分检机的结构图。

首先，简单地说明一下本发明的原理。

作为地址名标记的例子，先考虑图 1（3）的情况。图中，在“1－2”与“1－201”之间存在换行信息。但是，仅根据图像信息不能判断该信息是表示将文字串分开的分离信息还是表示非分离信息。因此，使用换行这样的文字图形间的几何信息、文字图形的识别结果以及街道号码的标记的知识生成假定换行不表示分离信息的街道号码假说“1－21－201”和假定换行表示分离信息的假说“1－2－1－201”这样 2 个假说。然后，将存储实际存在的住所代码的地址号码词典与上面各个街道号码假说进行对照，并认为登录在词典上的一方是正确的解释，并输出该假说。

这样，在换行及空白这样的解释中有存在不确定的信息时，本发明的基本出发点就是建立多个假说并在后级的处理中验证哪个假说是正确的多重假说检验方式。

下面，详细说明本实施例。

图 2 是文书识别装置的基本功能结构。读取对象的纸张 301 上有（比方说）图 3 所示的记载。该纸张经过图像读入处理 201，通过扫描器输入图像，然后变换成数字图像数据。文书识别装置对该图像数据进行文字区域的分割 202、文字分割假说的生成 203、城镇域名识别 204 和街道号码识别 205 的处理，并输出识别结果 208。

文书识别的第 1 个处理，就是文字区域的分割 202。从纸张 301 的图像中检测记载区域 302，然后进而对各文字行如图 4 所示的那样分割出图像区域 401、402。

其次的处理就是文字分割假说的生成处理 203。这里，所述的“假说”和“候补” 意义相同。

然后，处理逐行分割出的文字行图像 401、402，如图 5 所示的那样，生成文字分割假说，作为网络进行表示（图中只表示出了图 4 的文字行的一部分）。

通过称为“标记”的处理，抽出构成图像的像素的值为 1（表示黑）的连续的区域。将抽出的连结的黑像素的块（图形）称为黑连结成分。将相邻的黑连结成分（1 个或多个）视为构成 1 个文字的最小部分（基本成分），由它们构成图 5 所示的网络的基干部。其次，汇总有可能构成 1 个文字和不能构成 1 个文字的几个相邻的基本成分，附加在图案的旁边。通过对对全体进行该处理，便可作成图 5 所示的文字分割假说网络。

文字的分割很难，在文字识别之前不能唯一地决定。因此，如图 5 所示，在假说的形式中保留有可能存在的分割方式，并以这些多个假说为对象进行以下各个识别处理。

这里，图 5 只表示出了图 4 的文字行 401 的前半部分。即，与“1－2”以后的文字串对应的部分虽然也有数据存在，但是，在图 5 中为了简化图示而没有示出。

该网络将分割出的文字（或该部分）的图像和它们的组合（也有 1 个的情况）图像附属于弧线（图中线上面的梯形），将各个不

同的分割方式表现在1个图案（网络）中。图5是用图示表示它们，但是，也可以作为计算机内部的数据来表现。图中，弧线3表示与“都”的左偏旁部分对应的部分图像，弧线4表示与“都”的右偏旁部分对应的部分图像。另外，弧线16是将这2部分图像组合的图像，表示另一种分割假说。

城镇域名识别204是在输入上述分割假说网络206后使用事前知识进行文字识别和知识处理，并输出在城镇域名候补和分割假说网络中到何处为止进行了识别对照的指示信息207。

下面，使用图6说明城镇域名识别处理的详细情况。文字识别601是在参照文字识别词典604对附属于上述文字分割假说网络的弧线部分图像进行处理后根据该信息决定弧线的组合，并输出图7所示的文字候补网格。

作为文字识别方式，也可以使用先有技术。作为识别印刷汉字和手写汉字的方式，已知的有抽出文字笔划的方向性的“方向性图形整合法”。这里，虽然省略了详细说明，但是，用图形整合法可以将未知图形的图像（或特征矢量）与另外有限存储在文字识别词典中的参照图形（图像或特征矢量）的类似性作为称为类似度的0到1的数值（实数）进行计量。

图7所示的识别结果，是按类似度大的顺序将文字分类及其类似度进行排列的结果，称为文字候补网格。图中，第2行的标识符就是在图6中选择的弧线号码，是将文字识别结果的类似度高的图形的组合罗列的结果。

这里的意图在于：在图5中，如弧线1、2那样，如果部分图像或它们的组合图像是实际可以存在的图形，它就与登录在图形词典中的参照图形类似，所以，可以推测其类似度大。相反，如弧线4、7那样，如果部分图像是实际不存在的图形，则在图形词典中也不存在与其对应的参照图形，所以，其类似度就小。这样，通过使用文字识别作成类似度大的部分图像的组合，来执行文字识别和分割。

实际上，有时最佳路径也是错误的，所以，不仅要选择类似度

的总和最大的路径而且还要选择第 2 位、第 3 位等多个路径，向后级的处理传送。

因此，只作成该数量的文字候补网格。以后的处理，就以这些文字候补网格为对象。在图 5 和图 7 中，只表示出了一部分，但是，这些处理是对图 4 所示的全部文字行进行的。

其次，城镇域名对照处理部分 602 是对根据所有的文字行作成的多个文字候补网格进行的。该处理部分参照以表格数据的形式保持图 8 所示的城镇域名知识的城镇域名对照词典进行处理。该词典具有出现在都道府县名、市镇村名、街区名中的单词知识(图 8(ａ)、(ｂ)、(ｃ))。此外，在该单词知识中也包括哪个单词可以与哪个单词接续的信息。

例如，在图 8(ｂ)中，如果市标识符为 C 0517 的“千代田区”的主标识符为 K 31，就进行存储，看一下图 8(ａ)的都道府县名表，它存储的是“东京都”。即，表示在单词“东京都”之后可以接续单词“千代田区”。同样，对于“千代田区”，根据图 8(ｃ)的街区名表“千代田区”，存储的是接续“神田骏河台”等。

城镇域名对照处理，是使用这些单词知识和它们的组合知识从文字候补网格中识别单词串(句)例如“东京都·千代田区·神田骏河台”的处理，这有已知的方式。下面，为了帮助理解本发明，说明该方式。

为此，先根据文字候补网格构成图 9 所示的有限自动装置，通过将单词(文字串)输入该自动装置，可以判断该单词是否为可以根据该文字候补集合再构成的单词即是否可以将该单词作为识别结果。该处理就是单词对照的处理。

具体而言，作成比网格的行数多 1 的状态(图 9 中的圆圈符号)，用从某一状态向下一状态的迁移链(箭头)表示 1 个文字候补。例如，作成第 1 文字的第 1 位候补“东”为从状态 0 至状态 1 的第 1 迁移链、第 1 文字的第 2 位候补“果”为从状态 0 至状态 1 的迁移链、第 2 文字的第 1 位候补“宗”为从状态 1 至状态 2 的迁移链…

这样的自动装置。利用众所周知的方法，在软件方面使用状态迁移表可以很容易地作成有限自动装置。

某一单词是否可以根据文字候补集合再构成，可以根据构成该单词的文字串的各文字通过该有限自动装置的哪个迁移链进行判断。

例如，为了看出单词“东京都”是否位于最初的 3 个文字，先看文字“东”从状态 0 通过哪个链迁移到状态 1。使用状态迁移表的自动装置，就可以高速地实现该检索。图 9 的情况，可知第 1 个链就是“东”。其次，同样检索文字“京”从状态 1 通过哪个链向状态 2 迁移。这时，可知是第 2 个链。

同样，可知文字“都”是通过第 2 个链从状态 2 向状态 3 迁移的。这时，单词“东京都”的存在的确信度可以用各个文字候补的类似度总和表示。

或者，也有使用通过哪个链来计算恶化度的方法。这时，通过第 1 链时恶化度为 0、通过第 2 链时恶化度为 1、通过第 3 链时恶化度为 2。本例的情况，“东京都”的恶化度为 0 + 1 + 1 = 2。

在该检索处理中，如果对某一文字不存在对应的迁移链时，就假定通过图 9 的“其他”的迁移链。与“其他”的链对应的类似度或恶化度，事前补充作为一个参量的常数。

该处理在逻辑上对具有可能性的所有的单词都从所有的状态开始进行。其结果可以用图 10 的单词候补网格的形式进行表示。例如，图 10 表示从状态 0（第 1 文字）中将单词“东京都”和“ＸＸＸ”作为候补、从状态 1（第 2 文字）中将单词“京都”作为候补、从状态 2（第 3 文字）中将将单词“ＹＹＹ”作为候补。最后的候补是从状态 6（第 7 文字）开始的单词“ＳＳＳ”和“神田骏河台”。这样，单词候补网格就是以在哪个位置可以有什么样的单词这样的确信度或恶化度进行表示的。到作成该单词候补网格为止，是城镇域名对照处理部分 602 的前半部分的处理。

城镇域名对照处理部分 602 的后半部分的处理，是从单词候补

中选出只满足组合规则的单词。这就是从单词候补网格的末端（右端）的单词候补开始，参照图 8 所示的城镇域名知识中的单词接续信息顺序检验可以与左边接续的单词在单词候补网格中是否存在。

结果如图 10 所示，“神田骏河台”可以接续“千代田区”，但是，由图可知，可以与单词“S S S”接续的单词候补不存在。

最后，可以将“东京都·千代田区·神田骏河台”作为句子的候补。

根据情况不同，有时存在多个句子成为候补，这时，可以使用综合确信度或综合恶化度选择顺位好的一个句子。

或者，也可以将多个候补并列地传送给后级的处理，在后级进行最终的判断。

在候补采用判断部分 603 中，进行从城镇域名对照处理部 602 输出的城镇域名候补是否为妥当的对照结果的判断。例如，参考城镇域名候补的确信度的值进行阈值处理。

按照上述方式进行城镇域名识别处理 204，结果，就是识别出了例如文字串“东京都千代田区神田骏河台”，并将该文字串和表示该文字串的街区标识符“T 2121”作为数据 207 向后级的处理部分传送。

但是，该数据中不仅包含所谓的识别结果“东京都千代田区神田骏河台”，而且也包含表示与文字候补网格中的哪一部分对应的数据。这样，后级的处理例如街道号码识别处理 205 就可以知道自己下一个要处理的部分是何处。

更具体而言，就是包含表示间从图 4 的文字行、第 1 行的“东京都...”开始到“...骏河台”的识别对照处理的指示信息。这样，就可以知道下一个应处理的部分是“1－2”以后的文字串。这时，指针信息可以用图 7 的文字候补网格中的标识符号码的形式表示。

下面，使用图 11 说明街道号码识别 205. 在街道号码识别中，进行尚未进行处理的处理对象“1－2”以后的文字串的识别对照处理。

在文字识别处理1101中输入文字分割假说203和城镇域名识别结果207后，使用文字识别词典1108利用和城镇域名识别204中的文字识别601相同的方式进行文字识别，并输出候补文字网格1111。

几何信息抽出处理1102是在输入文字分割假说203和候补文字网格1111后，输出与文字间的空白或换行这样的几何分离信息假说属性相关联的带几何分离信息属性的网格1112。该处理可以使用已知的方式实现。下面，使用图12说明本处理的详细情况。

为了抽出空白信息，首先根据文字分割假说附带的几何信息求出与行方向对应的各个分割的文字图像的重心，并测定相邻的重心间的距离。1201表示其结果。1202的部分具有相邻的重心间的距离的2倍的距离。并且，如果该重心间的距离相对于行的高度大于某一一定的阈值，就认为该处存在空白，使之与分割出的图像相关联，赋予该属性。1204表示空白信息的候补存在于该文字图形之后。对于换行属性，也判定文字图形间的相对的位置关系，并赋予同样的属性。1203是它们的处理结果，几何分离信息的假说以符号串的形式来表现。最后，几何分离信息假说如图13那样对于文字候补网格与对应的位置关联地作为带几何分离信息属性的网格1112而输出。

通配符网格生成处理1103接收带几何分离信息属性的网格1112，参照图14所示的通配符变换规则，将网格中的候补文字变换为通配符符号。例如，如图13中的记载文字号码13那样，对于记载文字“1”，考虑候补“1”、“7”上升到网格中的情况。这些候补文字分别变换为符号“N”、“N”，这时，由于成为相同的符号，所以，合并为1个。对于记载文字号码14的记载文字“－”，候补“丿（之）”、“－”上升到网格中，所以，分别变换为符号“丿”、“－”。以后，反复进行同样的处理。变换后的通配符网格示于图15。在该网格中，几何分离信息属性与通配符文字网格相关联地存储。并且，为了后级的处理而输出该通配符网格1113。

标记图形对照处理1104接收通配符网格1113，查找标记图形词

典 1109 进行对照处理。该词典在内部以表形式的数据保持图 16 所示的街道号码、门牌号码、房间号码以及它们的标记知识。具体而言，对于街道号码标记、门牌号码标记、房间号码标记（例如，“1－2”的部分的标记），是存储可以具有什么样的形式的表。图中，符号“N”表示阿拉伯数字、符号“NN”表示 2 位的阿拉伯数字、符号“NNN”表示 3 位的阿拉伯数字、“A”表示字母、“の”、“ノ”、“－”表示分离符号（定界符）。图中虽然未示出，但是，在街区标记图形表中，预先也登录了与上述“1 番 2 号”或竖写时的“一の二”等标记方法对应的图形。

首先，根据通配符网格 1113 作成图 17 所示的有限自动装置，向该处输入与 P 001、P 002、P 003、P 004 对应的标记图形（“N－N”等），在各状态间迁移，累计与通过的路径对应的成本。

该处理使用和城镇域名对照处理 602 相同的方法执行。通过控制自动装置中对照的开始位置，在任意的位置进行单词对照。但是，在该自动装置中，街道号码位于开头，所以，只从状态 1 开始进行单词对照。单词对照的结果作为图 18 所示的单词网格而输出。

进行单词对照后，将街道号码、门牌号码、房间号码以及它们的标记图形单词连接，生成将街道号码、门牌号码、房间号码并列的一连串的标记。这里，根据文字候补网格附带的空白属性或换行属性控制各单词的连接。即，连接对象标记图形单词的连接部分为数字时，在连接的部分存在空白或换行的属性时，就将分离信息的属性附加到该处进行连接。相反，在与空白或换行不一致时，如果连接，将生成重复的标记图形假说，所以，不进行连接。

例如，在图 18 中，用箭头所示的单词间连接的情况下，在连接位置存在空白或换行的属性，所以，是判定为可以连接的标记图形候补。假定不存在空白属性 1804 时，若着眼于单词 1801、1802、1803，将 1801 与 1803 连接时，将重复生成与 1802 相同的假说。这样，将数字之间连接生成标记假说时，如果在该处不存在分离信息的属性，就生成与其他的标记图形单词相同的假说。因此，不进行标记

图形单词的连接。

按照上述规则连接的结果，在被视为重复的标记时，就删除重复的标记。例如，将“N”与“N－”连接的标记和“N－N”相同，所以，将其中之一删除。

此外，多数情况是街道号码、门牌号码、房间号码的标记的末尾与空白或行末一致。因此，末尾与空白或换行不一致的候补，就视为不适当的假说的可能性高，从而增加恶化度。

这些标记图形对照的结果是输出图 19 所示的标记图形的候补 1114。图中，“＝”表示几何分离信息插入到该处。

在文字串候补生成处理 1105 中，输入上升到高位的标记图形候补 1114 和带几何分离信息属性的网格 1112 后，将标记图形中的英文数字部分与对应的网格的候补文字进行比较，输出复原为原来的文字的文字串候补 1115。图 20（1）～（4）分别是与图 19 的（1）～（4）对应的候补。

在地址号码对照 1106 中，参照地址号码词典 1110 对照在文字串候补生成处理 1105 中得到的候补实际上是否存在。图 21 表示地址号码词典的结构例。如图所示，将实际存在的住所按层次保存在词典中，判断在登录到该词典中的住所中是否存在在英文数字复原处理 1105 中得到的候补。如果判定存在，就向后级输出候补，如果判定不存在，就排除该候补。在图 20 的候补中，选择（1）作为地址号码对照候补 1116 而输出。

在候补采用判断 1107 中，和城镇域名识别 204 的候补采用判断 603 一样进行在前级的对照中得到的候补是否妥当的判断。

在本实施例中，作为几何分离信息，采用的是文字间的空白或换行，但是，并不限于此，也可以使用颜色信息、有无下划线、字体形状、文字的大小或行宽的不同等其他信息。

此外，使用几何分离信息的标记图形假说的生成方式不限于本实施例。在本实施例中，是通过在标记图形单词连接的控制中使用分离信息来实现插入几何分离信息的标记图形假说和不插入的标记

图形假说的各假说生成的。例如，如以下所述，也可以通过其他方式来实现。

首先，作为第 1 方式，和本实施例一样，输入几何分离信息属性网格，与标记图形单词进行对照。然后，在标记图形单词连接的库中不使用分离信息而将单词之间连接。结果，对于上升到高位的标记图形候补，执行以下的处理。着眼于存在确定的几何分离信息的位置，假定在该处存在分离信息，追加生成将分离信息插入到标记图形单词中的标记图形假说。例如，在图 19(2)的例中，根据“N－N N－N N N”的标记图形单词生成“N－N＝N－N N N”及“N－N＝N－N N＝N”。这里，“＝”表示根据几何信息生成的分离信息。

作为第 2 方式，将空白插入到标记图形单词中的单词和不插入的单词都登录到词典中后，着眼于存在确定的几何分离信息的位置时，分别生成将例如空白的文字代码插入到候补文字网格的相应的位置的网格和不插入的网格，通过与上述各候补文字网格进行对照便可生成。即，第 2 方式是生成以文字代码的形式表示几何分离信息的多个候补文字网格并对它们分别进行标记图形对照的方式。

作为第 3 方式，标记图形的信息不是作为存储在词典中的文字串的信息来表现的，例如，对于在日本专利公报特开平 6－124366 中所示的方式，也可以生成使用同样的空白或换行的信息的标记图形假说。在本发明中，考虑位置关系的制约，将文字分割的候补以树结构的形式来表现，此外，将与其节点对应的识别结果根据字种附加上标号，检验该标号的排列，生成标记图形的假说。这里，在检验排列的规则中预先追加将空白及换行作为分离信息使用的排列的规则。此外，在将空白或换行这样的几何分离信息的属性赋予各节点、检验排列并生成标记图形假说时，分别生成插入空白或换行的分离信息的标记图形假说和不插入的标记图形假说。

作为第 4 方式，如特开平 8－180142 中所示的那样，对于将满足预先决定的标记规则的分类符号串候补用树结构表现标记图形的方

式，也可以生成使用同样的空白或换行的信息的标记图形假说。在本发明中，对与文字分割的候补对应的识别结果根据字种进行分类。然后，根据文字分割候补的位置关系的制约和分类符号间的连接关系的制约，用树结构的形式表现分类符号串候补。根据该树结构再次展开作为各分类符号的发生源的文字，并作成文字串候补后，应用街道号码范围规则来选择文字串候补。这里，在预先将在文字分割的候补间可以存在的空白或换行这样的几何分离信息的属性赋予文字分割的候补、检验分类的排列并生成树结构时，分别生成插入空白或换行的分离信息的标记假说和不插入的标记假说。图22表示所生成的候补树的例子。用虚线包围的部分是由本发明追加的。图中，“N”是表示数字的分类符号、“Ｓｈ”是表示“－”的分类符号、“Ｓｉ”是表示几何分离信息的分类符号。图23是表示搭载文书识别装置的信件分检机的概略结构图。

2000是信件分检机的总体体系（机构部分），由将多张纸页顺序供给扫描器的供给机构2001、扫描器2002、将扫描过的纸页向分检机构传送的传送机构2003、从中央控制计算机2006接收决定纸页的传送前往地的分检信息进行纸页的分检的分检机构2004构成。此外，机构部2000具有多个控制用的计算机，这些计算机在中央控制计算机2006的指示下控制机构部分2000全体。

2007是文书识别装置，由输入从扫描器2002读取了记载在纸张上的地址信息的图像数据并向中央处理装置（ＣＰＵ）输出的输入接口2008、输入图像数据并进行地址信息识别的中央处理装置（ＣＰＵ）2009及2010、存储用于进行地址信息处理的程序及供识别对照使用的事前知识的存储器2013、存储装载在存储器2013中的程序及事前知识的2次存储装置2014和将由中央处理装置（ＣＰＵ）识别的识别结果数据向中央控制计算机2006输出的通信接口2016构成。

在图22中，中央处理装置（ＣＰＵ）示出了2台，但是，在处理能力不足时，可以实际装配2台以上。或者，将文书识别装置2007

并列地设置所需要的台数。这些文书识别装置分担处理高速读取的图像数据2015。

中央控制计算机2006从通信接口2011解释在通信数据2016中包含的识别结果数据，按照预先编程的逻辑作成分检信息并向分检机构输出。

在本实施例中，说明了用扫描器作为图像信息读取写在纸面上的文字信息从而进行文字识别的方法，但是，最近，也有很多情况是电子的图像信息已存储到了计算机系统中，不言而喻，本发明也可以适用于识别和解释这样的图像信息。

如上所述，按照本发明，在着眼于存在用于将文字串分开的几何分离信息假说的确定的位置时，生成假定该处存在分离信息的标记和假定不存在的标记的各标记图形假说。

然后，根据这些标记图形假说和文字识别结果生成住所代码，通过将实际的住所与存储的地址名词典进行对照，来确定正确地解释的标记从而识别住所。通过这些处理，便可正确地识别使用空白或换行等几何分离信息标记的地址信息。

另外，如图1的（3）、（4）所示，开始可以由本发明正确地解释换行的信息并读取地址的标记，大多都是大量地作成邮购广告单那样的相同种类的邮件时的地址。在邮电局，这种相同种类的邮件多数是集中地放到信件分检机上进行分检的。这里，信件分检机具有读取邮件上的地址信息并按各投递地址分检邮件的功能。因此，在上述那样的信件分检机的运用方法中，提高读取邮件的地址进行分检的分检效率的效果特别大。

图1

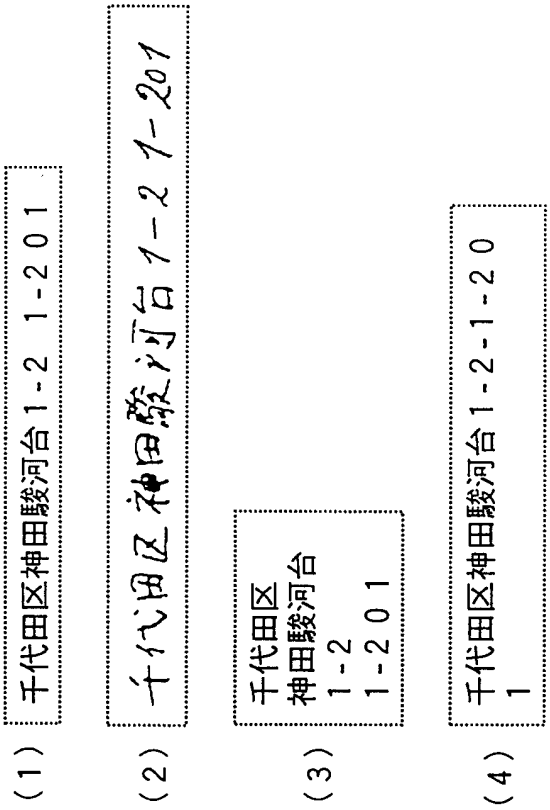

（1）
千代田区神田駿河台1-2 1-201
（2）
千代田区神田駿河台1-2 1-201
（3）
千代田区
神田駿河台
1-2
1-201
（4）
千代田区神田駿河台1-2-1-20
1

图2

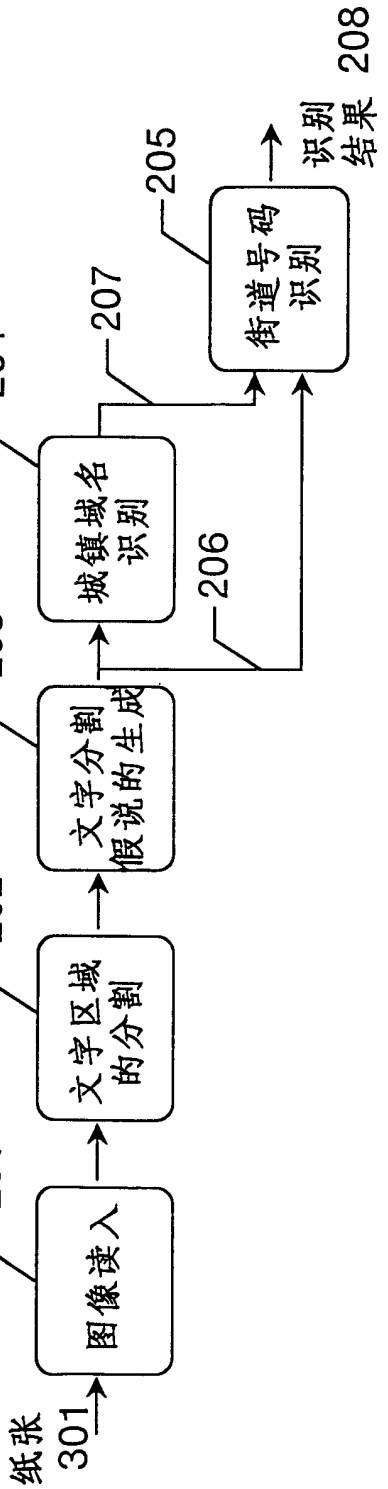

纸张
301
图像读入
201
文字区域的分割
202
文字分割假说的生成
203
城镇域名识别
204
街道号码识别
205
206
207
识别结果
208

图3

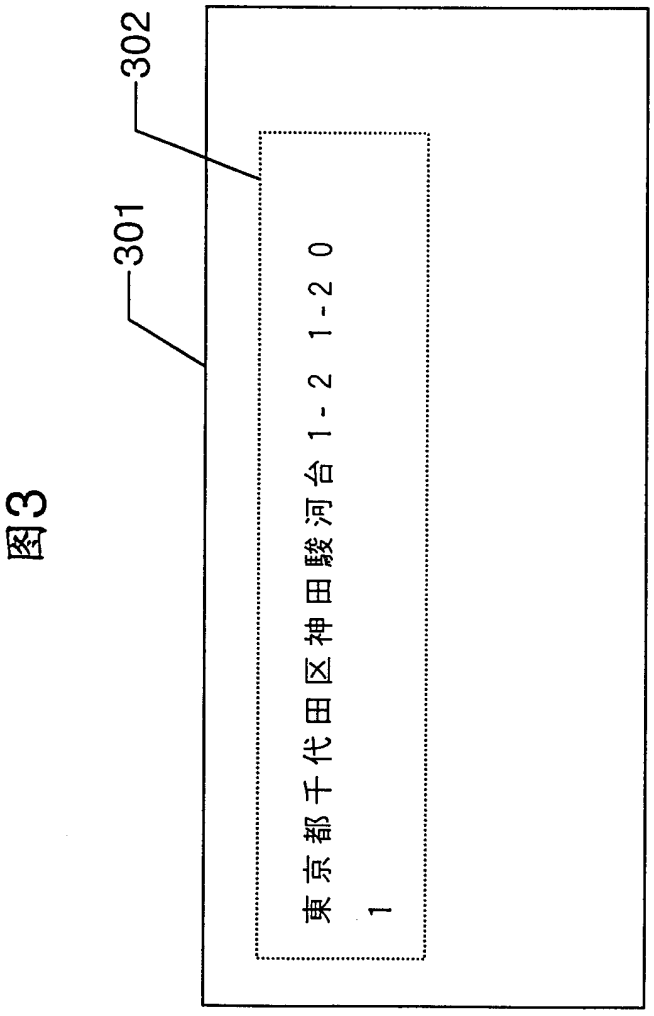


图4

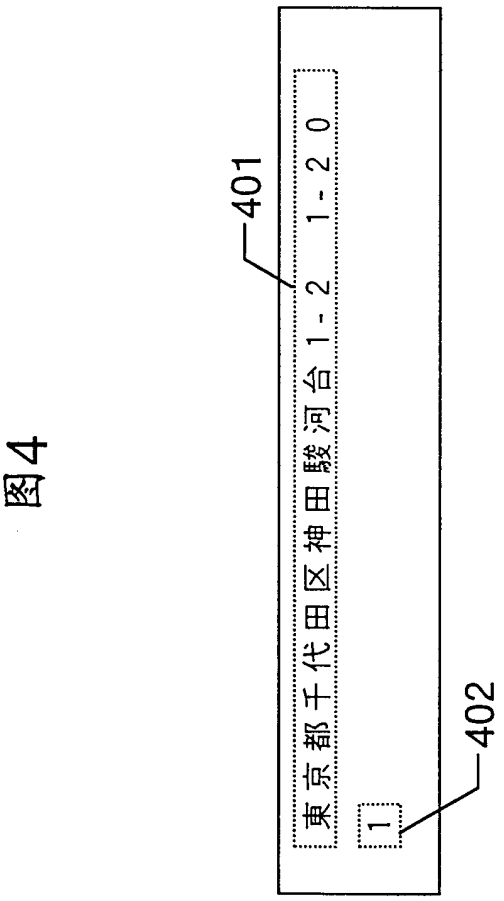


图5

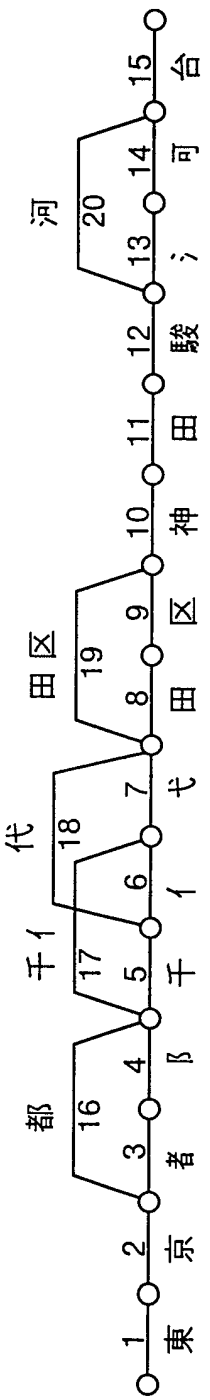

1 東
2 京
3 者
4 ß
5 千
6 亻
7 弋
8 田
9 区
10 神
11 田
12 駿
13 氵
14 可
15 台
16 都
17 千亻
18 代
19 田区
20 河

图6

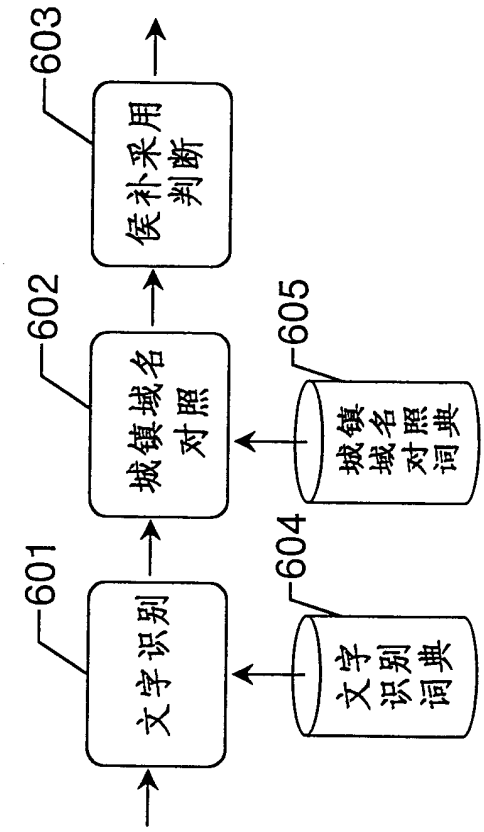


图7

记载文字序号		1	2	3	4	5	6	7	8	9	10	11	12
分割标识		1	2	16	5	18	8	9	10	11	12	20	15
文字分割图像		東	京	都	千	代	田	区	神	田	駿	河	台
候补文字顺序	第1位	東	宗	部	千	代	田	区	神	田	駿	河	台
	第2位	菓	京	都	干	伐	囲	巨	禅	由	駿	珂	旨
	第3位	巣	未	郁	十	付	由	匡	伸	図	駐	何	白
	第4位	束	亨	郭	チ				枠	国	鮫	洞	垢
	第5位	栗		却	斗								

图8

(a) 都道府县名列表(第1层)

县标识符	都道符县名
K01	北海道
K02	青森県
K03	秋田県
. .	. . .
K31	東京都
. .	. . .

(b) 市街区村名列表(第2层)

市标识符	市街村名	主标识符
C0001	. . .	
. .	. . .	
C0501	足立区	K31
C0502	荒川区	K31
. .	. . .	K31
C0517	千代田区	K31
. .	. . .	
. .	. . .	

(c) 街道名列表(第3层)

街道名标识符	街道名	主标识符
T0001	. . .	
. .	. . .	
T2101	飯田橋	C0517
T2102	一番町	C0517
. .	. . .	C0517
T2121	神田駿河台	C0517
. .	. . .	C0517
T2137	一ツ橋１丁目	C0517
T2138	一ツ橋２丁目	C0517
. .	. . .	
. .	. . .	

图9

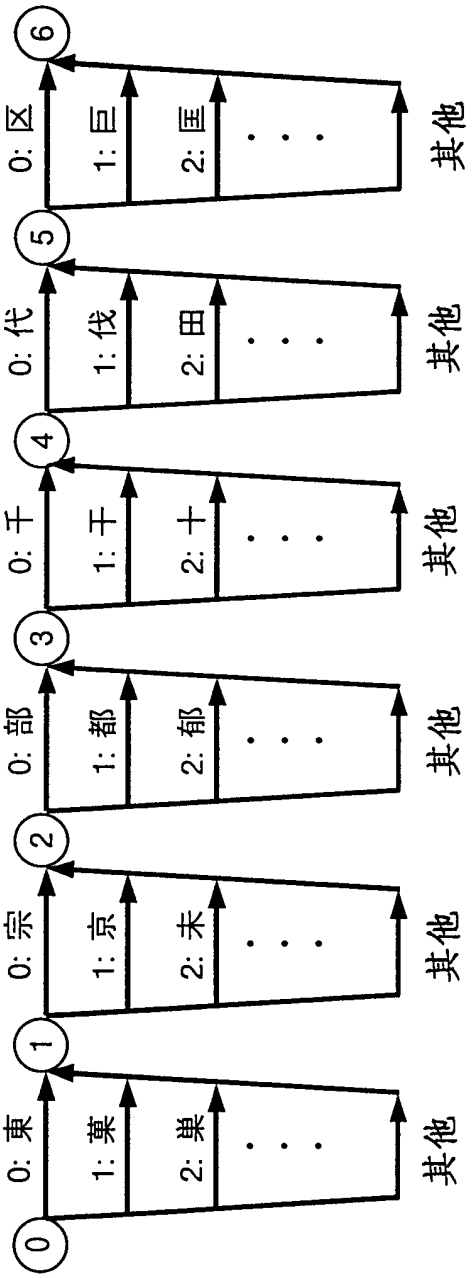

0
0: 東
1: 菓
2: 巢
其他
1
0: 宗
1: 京
2: 未
其他
2
0: 部
1: 都
2: 郁
其他
3
0: 千
1: 干
2: 十
其他
4
0: 代
1: 伐
2: 田
其他
5
0: 区
1: 巨
2: 匡
其他
6

图10

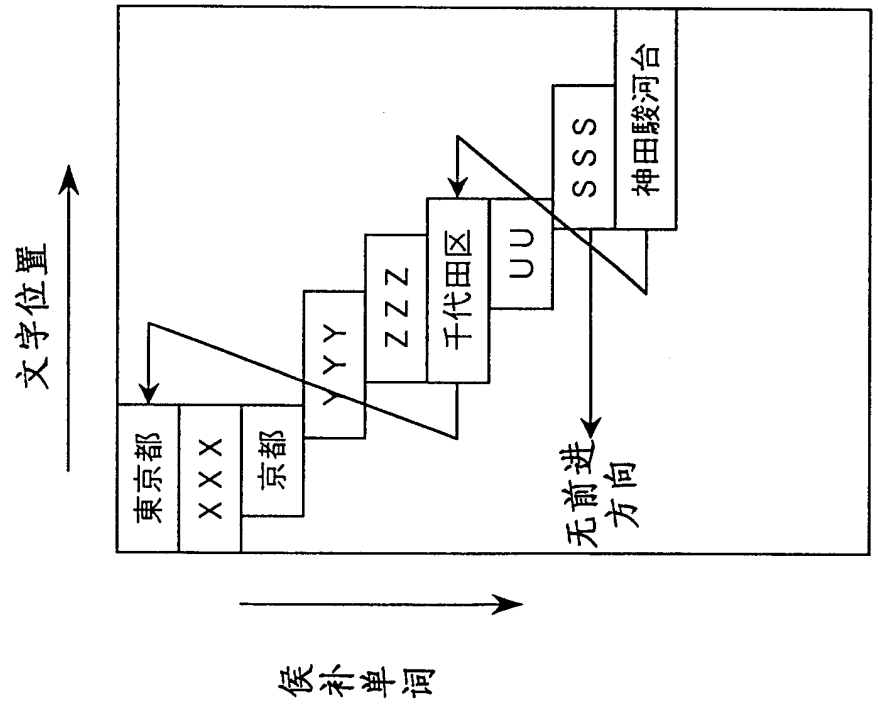


图11

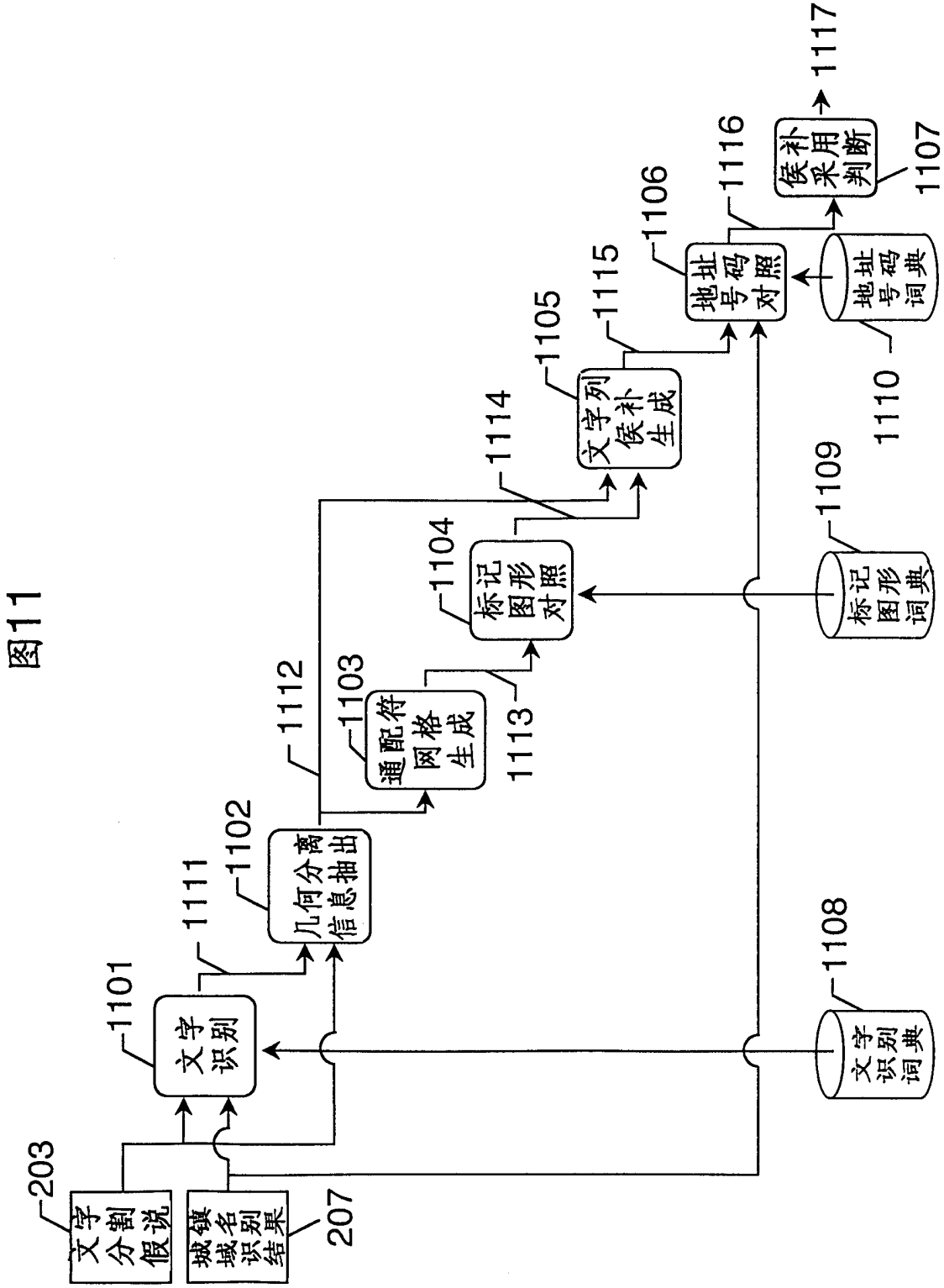

203
文字分割假说
207
城镇域名识别结果
1101
文字识别
1111
1102
几何分离信息抽出
1112
1103
通配符网格生成
1113
1104
标记图形对照
1114
1105
文字列侯补生成
1115
1106
地址号码对照
1116
1117
侯补采用判断
1107
1108
文字识别词典
1109
标记图形词典
1110
地址号码词典

图12

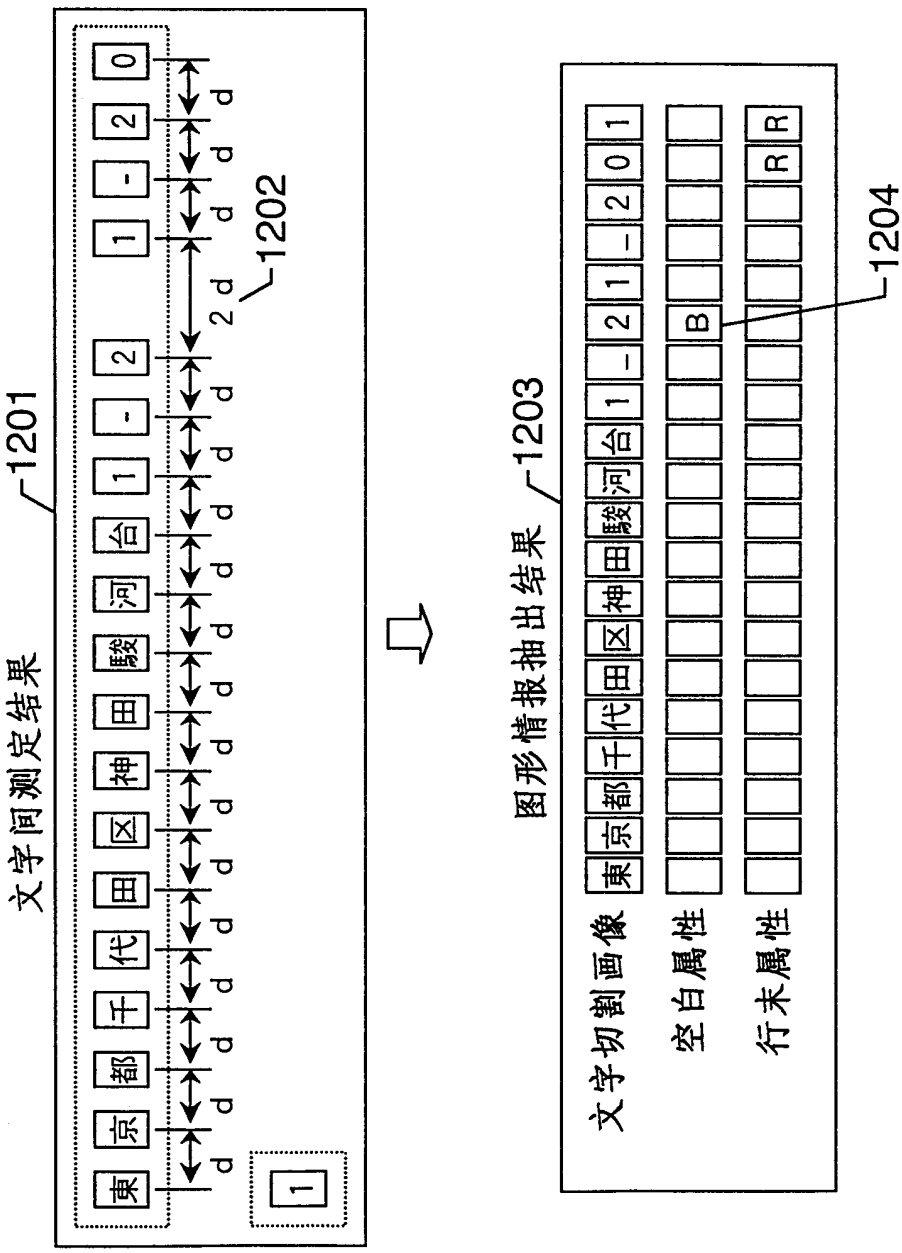


图13



记载文字序号		13	14	15	16	17	18	19	20
记载文字		1	一	2	1	一	2	0	1
候补文字顺序	第1位	1	ノ	Z	丨	一	日	口	1
	第2位	7	一	2	1	一	Z	0	丨
	第3位	丨	一	日	丨	ノ	2	O	7
	第4位	丨	~		J	.	8	6	丨
	第5位	J	一				7	9	
	第6位		一				z		
空白属性				B					
行末属性								R	R

图14

(a)街道号码表示模式列表

形式识别符	表示形式
P0001	N-N
P0002	NのN
P0003	N／N
P0004	N-N N-
：	：

(b)楼号表示模式列表

形式识别符	表示形式
Q0001	N-
Q0002	N
Q0003	N
Q0004	A
：	：

(c)房号表示模式列表

形式识别符	表示形式
R0001	N
R0002	N N
R0003	N N N
R0004	N
：	：

图15

输入文字种类	通配符
0，1，2，3，4，5，6，7，8，9	N
〇，一，二，三，四，五，六，七，八，九，十，百，千	K
A，B，C，： a，b，c，：	A
丁	丁
目	目
番	番
地	地
号	号
の	の
ノ	ノ
一，～，／	一
—	—
上述以外的文字种类	E

图16


记载文字序号 13 14 15 16 17 18 19 20
记载文字 1 — 2 1 — 2 0 1
候补文字顺序
第1位 N ノ A A — E E N
第2位 | — N N K A N |
第3位 A K E | ノ N A A
第4位
第5位
第6位
空白属性 B
行末属性 R R

图17

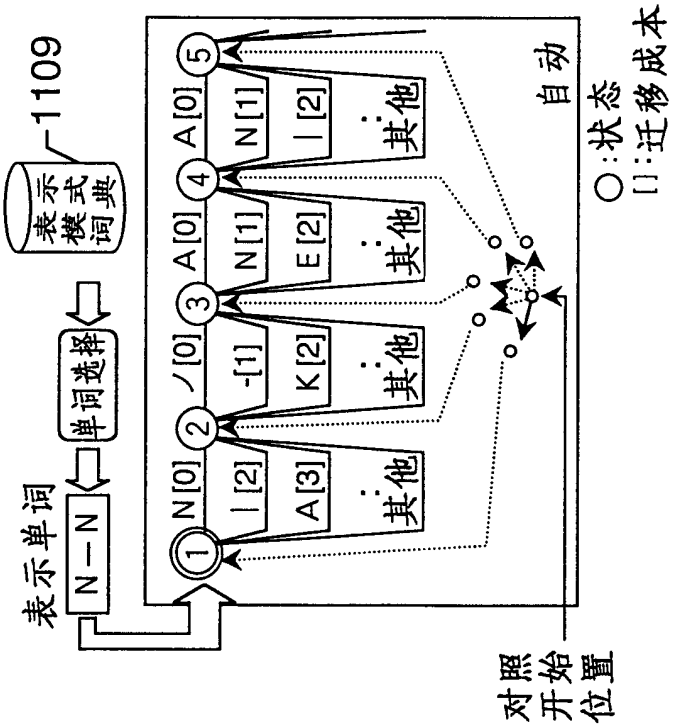

表示单词
N－N
单词选择
1109
表示模式词典
N [0]
| [2]
A [3]
其他
ノ [0]
-[1]
K [2]
其他
A [0]
N [1]
E [2]
其他
A [0]
N [1]
| [2]
其他
自动
○:状态
[]:迁移成本
对照开始位置

图18

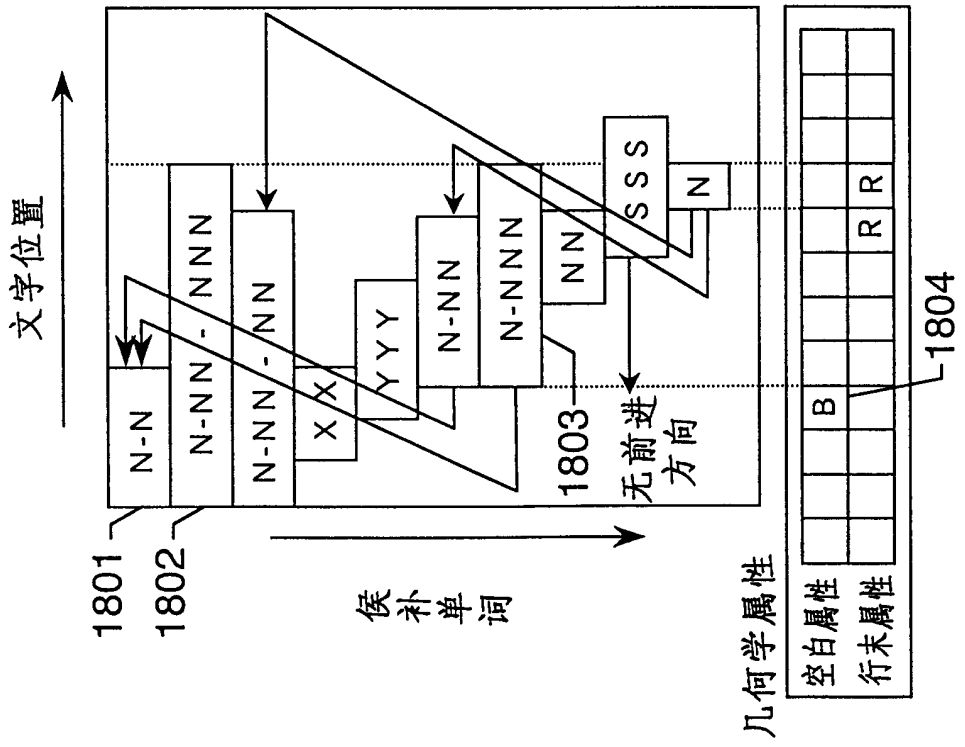

文字位置
候补单词
1801
1802
1803
1804
N-N
N-NN-NNN
N-NN-NN
XX
YYY
N-NN
N-NNN
NN
SSS
N
无前进方向
几何学属性
空白属性
行末属性
B
R R

图19

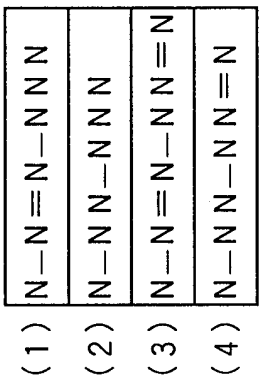


图20

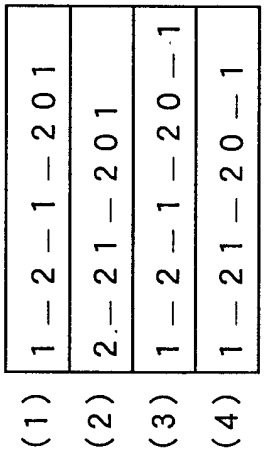


图21

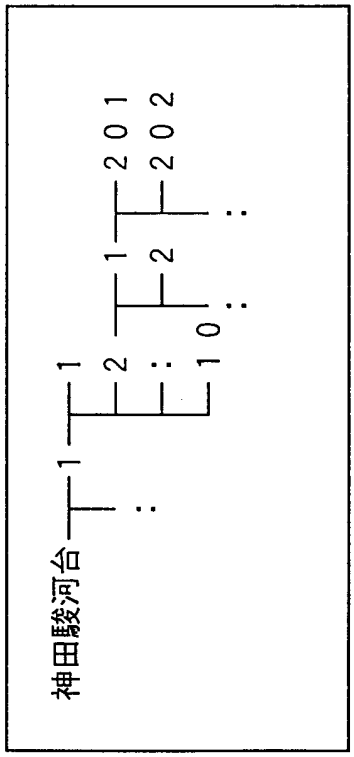


图22

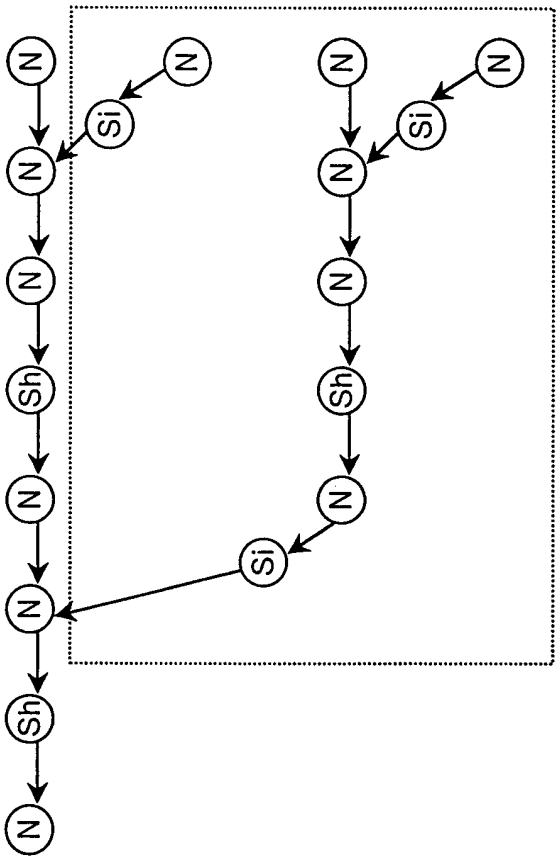
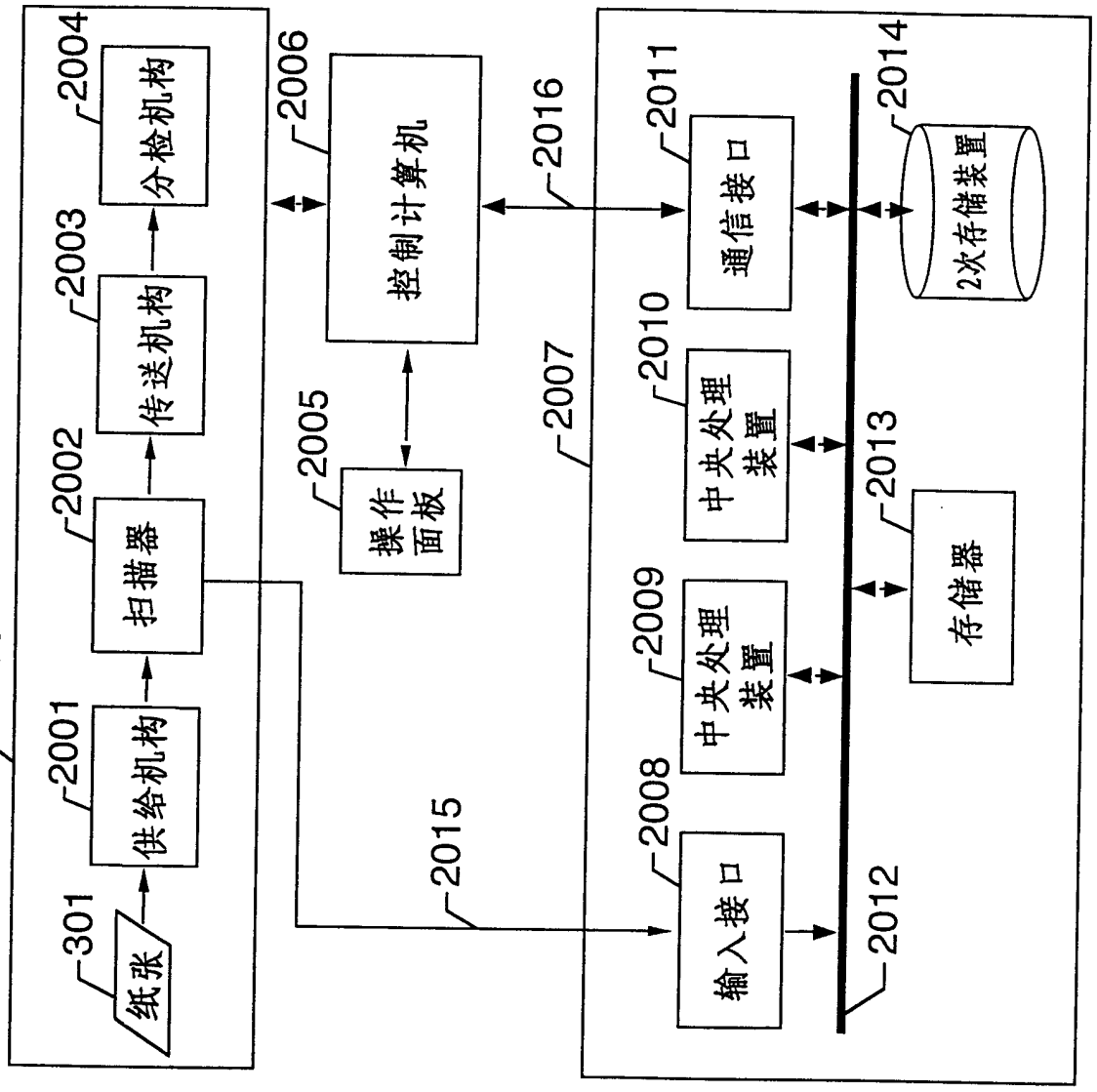


图23


2000
2001
供给机构
2002
扫描器
2003
传送机构
2004
分检机构
301
纸张
2005
操作面板
2006
控制计算机
2007
2008
输入接口
2009
中央处理装置
2010
中央处理装置
2011
通信接口
2012
2013
存储器
2014
2次存储装置
2015
2016