

## 古代汉字文献切分研究

倪恩志<sup>1</sup>, 蒋旻隽<sup>2</sup>, 周昌乐<sup>1</sup>

NI Enzhi<sup>1</sup>, JIANG Minjun<sup>2</sup>, ZHOU Changle<sup>1</sup>

1. 厦门大学 信息科学与技术学院, 艺术认知与计算实验室, 福建 厦门 361005

2. 上海应用技术学院 计算机科学与信息工程学院, 上海 201418

1. Mind, Art and Computation Lab, School of Information Science and Technology, Xiamen University, Xiamen, Fujian 361005, China

2. School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai 201418, China

NI Enzhi, JIANG Minjun, ZHOU Changle. Research on segmentation of historical Chinese books. *Computer Engineering and Applications*, 2013, 49(2): 29-33.

**Abstract:** In this paper, the methods of text line segmentation and character segmentation are proposed according to the characteristics of historical Chinese documents. The method of line segmentation analyzes stroke projection, and adopts a recursive segmentation algorithm based on various project thresholds and gap thresholds. This algorithm is robust in the cases of text line adhesion and skew, especially short text lines. The method of character segmentation has two steps. A rough segmentation is applied to get the approximate positions of segmentation. A fine segmentation based on the analysis of connected components and the judgment of adhesion points is carried out. This algorithm can extract the characters even though they overlap and connect each other. The experimental results show the methods have good performance and are suitable for the segmentation of historical Chinese documents.

**Key words:** document image processing; Chinese character segmentation; ancient books digitalization

**摘 要:** 针对古代汉字文档的特点, 提出了适合于古文档的列切分方法和字切分方法。提出的列切分方法直接对文档的笔画投影进行分析, 采用一种基于分层投影过滤和变长间隙阈值的递归切分算法。该算法在列间隔较小、列与格线存在粘连、文档具有一定程度的倾斜的情况下, 也能准确地抽取出列, 尤其对短列的切分达到了较好的效果。提出的字切分方法分为两步, 进行粗切分确定大致的切分位置, 采用基于连通域分析与粘连点判断的方法做进一步的细切分。该算法对具有较多粘连和重叠汉字的列, 也能较好地切分出完整的单字。实验结果表明, 提出的方法用于古代汉字文档切分能够获得较好的效果。

**关键词:** 文档图像处理; 文档切分; 古籍数字化

**文献标志码:** A **中图分类号:** TP311 **doi:** 10.3778/j.issn.1002-8331.1209-0246

我国历史悠久, 文化丰富, 现今留存了大量的珍贵古籍文献。随着互联网和计算机技术的发展, 古籍数字化渐渐引起人们的关注, 文档切分是其中一项重要内容。到目前为止, 对汉字文档切分已有一些研究(行列切分<sup>[1-4]</sup>和字切分<sup>[5-16]</sup>)。行列切分方法较为关注行列歪曲倾斜情况下的切分<sup>[4]</sup>; 字切分方法可分为三类: 统计方法<sup>[5-7]</sup>, 结构方法<sup>[8-9]</sup>和基于识别的方法<sup>[6, 10-13]</sup>。现有的切分方法主要针对现代汉字文档, 如银行票据<sup>[14]</sup>、信封地址<sup>[6, 8, 10-11]</sup>、视频字幕<sup>[15]</sup>等。然而古代文档与现代文档有较大区别: 两者的书写方向不同, 古代文档采用毛笔书写, 笔画较粗, 间隙较小, 且多数文档中画有竖直格线; 由于这些文档年代久远, 受到不同

程度的破坏, 图像质量较差, 经过二值化后, 相邻字之间存在较多的粘连, 而格线更是成为准确切分的障碍(图1)。现有的文档切分方法较少关注这些情况, 目前为止, 尚无专门针对古代文档列切分和字切分的研究。

根据古代文档的结构和特点, 本文提出了适用于古代文档的切分方法。首先将文档进行列切分, 得到每个列; 在此基础上, 再对每列进行字切分得到每个单字。列切分方法采用了一种基于分层投影过滤和变长间隙阈值的递归切分算法。字切分算法则分为两步, 首先进行粗切分确定大致的切分位置, 然后采用基于连通域分析与粘连点判断的方法做进一步的细切分。实验分析对比表明, 本文算

**基金项目:** 国家自然科学基金(No.6097507)。

**作者简介:** 倪恩志(1982—), 男, 博士生, 主要研究方向: 汉字识别、人工智能; 蒋旻隽(1981—), 女, 博士, 讲师, 主要研究方向: 计算机作曲; 周昌乐(1959—), 男, 博士, 教授, 主要研究方向: 人工智能。E-mail: nienzhi@gmail.com

**收稿日期:** 2012-09-24 **修回日期:** 2012-11-23 **文章编号:** 1002-8331(2013)02-0029-05

法对古文档切分具有较好的效果。



图1 古代文档样本

## 1 列切分方法

文档的投影数组中与列所在位置相对应的部分称为列段,而与间隙和格线位置相对应的部分称为间段。由于相邻列之间存在格线与粘连,这使得投影数组中与相邻列间隙对应的投影值通常不会为0,因此不能简单地认为投影值为0的段就是间段。此外,每个列的长度可以不相等,而且可能存在一些只包括一两个字的短列。采用统一的投影值对列段和间段进行划分,则很有可能将实际上的间段当成列段,而将短列当成间段。本文的列切分方法是:使用不同的投影阈值和间隙阈值对投影数组进行切分,对每个切分方案进行评估,选择其中最好的切分方案,再对这个切分方案中不符合要求的间段或列段按同样方法进行递归切分。

### 1.1 列宽估计

为了评价列切分方案的好坏以及每个切分段是否符合要求,算法需要事先估计列宽。列与列之间存在格线。如果对文档图像的像素点进行投影,则在格线位置上会有较大的投影值,为了减少格线的干扰,采用的是笔画投影,它的投影值等于在竖直方向上通过的笔画数目。令  $I$  为文档图像,  $I(x, y)$  为文档图像  $I$  在第  $x$  列第  $y$  行的像素值,当该处为背景时  $I(x, y)$  值为0,存在点时  $I(x, y)$  值为1。笔画投影值用下列公式计算:

$$p_x = \frac{1}{2} \sum_{1 \leq y \leq H} |I(x, y) - I(x, y-1)|$$

其中,  $p_x$  表示在  $x$  处的笔画投影,  $H$  为图像  $I$  的高度。在计算格线处的投影时,如果格线垂直并且中间没有断开,则得到的投影值仅为1。列宽用以下方法估算。

(1)取最大投影值的一半作为阈值,即

$$t = \frac{1}{2} \max(p)$$

将小于阈值的投影滤除:

$$p_x = \begin{cases} p_x, & p_x \geq t \\ 0, & p_x < t \end{cases}$$

(2)计算所有列段的平均宽度:

$$\hat{w} = \frac{1}{n} \sum_{i=1}^n w_i$$

其中  $n$  为列段数目。

(3)则列宽估计值可用以下公式计算:

$$\hat{w} = \frac{1}{m} \sum_{w_i > \bar{w}} w_i$$

其中  $m$  为列宽大于平均宽度的列段数目。

### 1.2 列切分算法

首先,计算文档图像在横轴上的笔画投影。扫描投影数组,记录最小的投影值  $x_{\min}$  与最大投影值  $x_{\max}$ 。分别采用  $x_{\min}$  到  $\frac{x_{\min} + x_{\max}}{2}$  之间的连续整数作为投影阈值,对投影数组进行过滤,将所有小于等于投影阈值的位置设为0,

得到一系列新的投影数组(总共有  $\frac{x_{\max} - x_{\min}}{2}$  个投影数组产生)。然后继续采用同样范围内的整数作为间隙阈值,在每一个新产生的投影数组中寻找长度超过间隙阈值且连续为0的段,将这些段作为间段,而在两个间段之间的部分就是列段。于是可以初步得到针对某个投影阈值和间隙阈值的列切分方案,该步骤会产生  $\frac{x_{\max} - x_{\min}}{2} \times \frac{x_{\max} - x_{\min}}{2}$

个切分方案,要在其中选择最优的切分方案。根据估计列宽,确定列宽上下限,计算所有列段宽度的方差,但是不包括宽度超过列宽上下限的列段,具有最小方差的切分方案即作为最优方案。

通过上述过程,可以得到一个初步的列切分方案。由于某些列与格线间隙较小,因此当前设置的阈值可能无法准确将其切分,而得到较长的列段,它们的宽度大于列宽上限;另一面,格线和太短的列可能产生一些较短的列段,它们的宽度小于列宽的下限;因此需要进行递归切分。具体方法如图2所示。在图2(a)中,已经获得一个初步的切分方案,但其中五个列没有被正确切分,而是被当成两个较长的列段,同时也有三处格线被当成较短的列段。在图2(b)中,将这三处较短的列段合并到相邻的间段,然后再次使用上述算法,对所有间段和超过列宽上限的两个列段进行递归切分。最后将每个片段的切分结果合并就可以得到最终的切分方案,如图2(c)所示。整个算法的描述如下所示。

输入:

$P$ : 文档图像的投影数组,  $P = p_1 p_2 \cdots p_l$ , 其中  $l$  为文档图像的宽度。

$T^P$ : 投影阈值集合,  $T^P = \{t_1^P, t_2^P, \cdots, t_m^P\}$ , 其中  $m$  为包括的投影阈值的个数。

$T^S$ : 间隙阈值集合,  $T^S = \{t_1^S, t_2^S, \cdots, t_n^S\}$ , 其中  $n$  为包括的间隙阈值的个数。

$\hat{w}$ : 估计列宽。

$\lambda_{\max} \hat{w}, \lambda_{\min} \hat{w}$ : 设置的列宽上下限。

输出:

$S$ : 切分方案,  $S = s_1 s_2 \cdots s_l$ , 其中  $s_i = 0$  或 1, 连续为0的段为间段,连续为1的段为列段。

算法:

Function Line\_Segmentation( $P, T^P, T^S, \hat{w}, \lambda_{\max} \hat{w}, \lambda_{\min} \hat{w}, S$ )

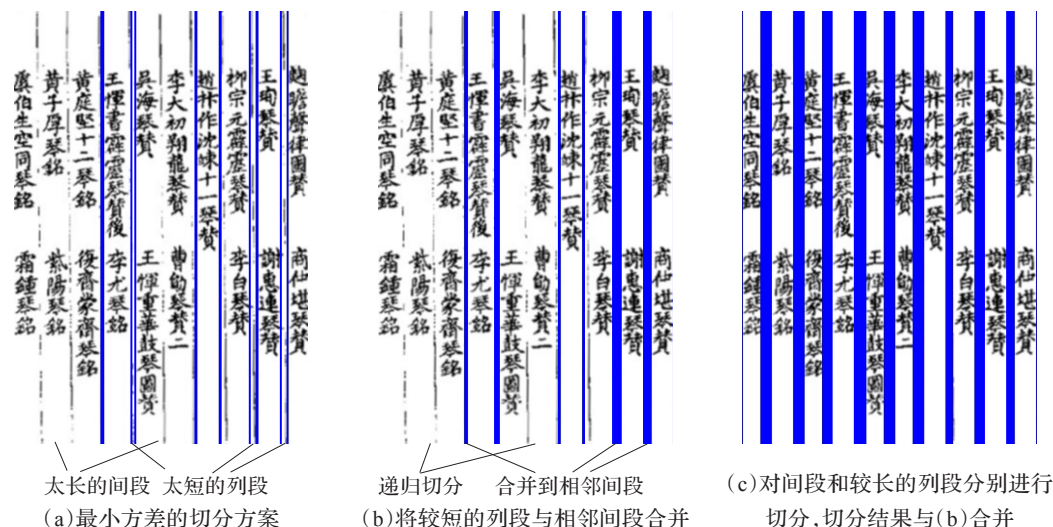


图2 列切分过程

For  $i=1$  to  $m$

计算经过投影阈值过滤后的投影数组  $P^i$ ,  $P^i = p_1^i p_2^i \cdots p_l^i$ ,

$$\text{其中 } p_k^i = \begin{cases} p_k^i, & p_k^i > t_i^p \\ 0, & p_k^i \leq t_i^p \end{cases}.$$

For  $j=1$  to  $n$

计算初步切分方案  $S^{ij}$ ,  $S^{ij} = s_1^{ij} s_2^{ij} \cdots s_l^{ij}$ , 当  $P^i$  中存在连续为0的段, 且段长大于或等于  $t_j^s$  时, 相应位置上的  $s_j^{ij}$  设为0, 其他  $s_j^{ij}$  设为1;

计算  $S^{ij}$  的列段宽度方差  $\text{Var}(S^{ij})$ : 不考虑宽度大于  $\lambda_{\max} \hat{w}$  或小于  $\lambda_{\min} \hat{w}$  的列段, 令剩余列段数目为  $k$ , 则

$$\text{Var}(S^{ij}) = \begin{cases} \frac{\sum_{i=1}^k (w_i - \bar{w})^2}{k}, & k > 1 \\ (\hat{w} - \bar{w})^2, & k = 1 \\ \hat{w}^2, & k = 0 \end{cases}$$

End For

End For

$S = \arg \min(\text{Var}(S^{ij}))$ , 对  $S$  做以下操作:

将列宽小于  $\lambda_{\min} \hat{w}$  的列段删除, 并合并到相邻的间段中;

如果  $S$  只有一个列段或一个间段, 且段长等于  $l$ , 则返回  $S$ ;

对每个间段和宽度大于  $\lambda_{\max} \hat{w}$  的列段进行列切分:

假设该段为  $\{s_i \cdots s_j\}$ , 调用  $\text{Line\_Segmentation}(\{p \cdots p_j\})$ ,

$T^p, T^s, \hat{w}, \lambda_{\max} \hat{w}, \lambda_{\min} \hat{w}, S'$ , 用  $S'$  替换  $\{s_i \cdots s_j\}$ ;

返回  $S$ 。

## 2 字切分方法

在列切分的基础上, 要进一步进行字切分。提出的字切分方法包括粗切分和细切分两个步骤。粗切分是为了得到单字切分的大致位置, 同样采用基于分层投影过滤和变长间隙阈值的算法。由于相邻的字之间可能有重叠(如图3(a))或粘连(如图3(b)), 如果在粗切分位置上用直线进行切分, 就无法得到完整的单字。因此, 使用细切分解决这个问题。细切分通过对连通域的分析 and 粘连位置的判断实现单字的准确切分。



图3(a) 重叠



图3(b) 粘连

### 2.1 粗切分

首先对列图像在纵轴上进行投影, 得到它投影数组。然后采用与列切分类似的分层投影过滤和变长间隙阈值方法, 进行第一步的粗切分。粗切分算法与列切分算法相似, 后者针对整个文档图像在横向做切分, 粗切分则针对单列在纵向上做切分, 得到单字在某列上的大致位置。具体算法不再赘述, 整个粗切分的步骤如下:

(1) 使用不同的投影阈值, 对投影数组进行过滤, 得到与投影阈值对应的一组新的投影数组。

(2) 使用不同的间隙阈值, 在这些投影数组中寻找长度超过间隙阈值的连续为0的段, 将这些段作为间段, 而在两个间段之间的部分就是切分出来的单字, 称之为字段。

(3) 于是得到了多个初步的切分方案。由于相邻字之间的重叠与粘连, 一些相邻字无法得到切分。因此需要对较大的字段进行进一步切分, 对较小的字段进行合并。

得到的切分方案中可能存在两种极端的情况: 一是包含了许多太短的切分字段, 二是仅得到少数太长的切分字段。使用列宽作为估计的单字高度, 如果切分方案的平均字段高度与列宽相差较多, 则直接排除这个方案。最后采用所有字段与平均字段高度的方差作为评价切分方案优劣的标准, 输出具有最小方差的切分方案。

### 2.2 细切分

通过粗切分, 得到了切分的大致位置。由于相邻字之间的重叠和粘连, 使用直线切分会使切分得到的单字不完整或者在单字中包含了相邻字的笔画, 这会导致单字的识别率降低。在这里使用基于连通域和粘连点判断的方法进行细切分。首先找出列图像中所有的连通域, 根据连通域与切分字段交叠的情况分别做如下处理:



(1) 连通域不与任何字段交叠, 则将该连通域合并到最靠近的字段中。

(2) 连通域与一个字段交叠, 则将该连通域合并到交叠字段中。

(3) 连通域与超过一个字段交叠, 则接着判断该连通域是否大部分(三分之二)属于一个字段, 如果是, 则将其直接合并到该字段, 否则对其做粘连切分, 然后对将切分后的部分合并至最靠近的字段中。

粘连切分在连通域交叠的间段的中线位置进行。首先细化连通域图像, 在细化图像中找出所有端点、折点、叉点和笔段。称与间段中线相交的笔段为粘连笔段, 粘连切分是通过对粘连笔段的切分实现的。这需要通过以下几个步骤完成:

确定粘连笔段上的切分位置。由于汉字中的粘连点通常表现为折点和叉点, 因此当粘连笔段靠近间段中线的那一端是折点或叉点时, 则判定这个点是粘连点。确定粘连点后要将此处的笔段分割开来, 具体切分示例如图4所示。

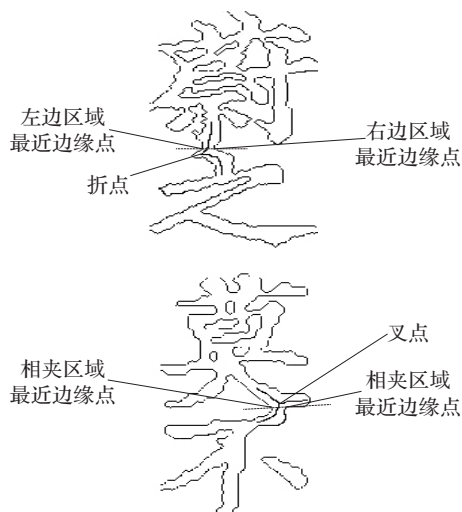


图4 粘连笔段分割

先考虑折点的情况, 粘连点连接的两个笔段将图像分成两个区域。在这两个区域内分别寻找分割点  $(x', y')$ , 满足:

$$D(x', y')(x^*, y^*) = \min_{(x, y) \in Area \cap Contour} D(x, y)(x^*, y^*)$$

其中  $D$  为两点距离,  $Area$  为该区域,  $Contour$  为字的边缘轮廓,  $(x^*, y^*)$  为粘连点, 用连接这两个分割点的直线分割笔段。

再考虑叉点的情况, 粘连点连接三个以上的笔段。从中选取笔段  $l'$ , 满足  $\theta_{l', l''} = \min(\theta_{l, l''})$ , 和笔段  $l''$ , 满足  $\theta_{l', l''} = \max(\theta_{l, l''})$ 。在粘连笔段与笔段  $l'$ 、 $l''$  分别相夹的两个区域内寻找满足以上公式条件的分割点, 用连接这两个分割点的直线分割笔段。从图4中可以看出用这种方法分割笔画不会形成毛刺, 具有较好的效果。

### 3 实验设计与结果

#### 3.1 列切分

从《四书章句集注》, 明代《琴书大全》和《太音大全集》中挑选了共85页文档作为列切分的实验样本。评价列切

分结果的优劣, 不但要考虑实际的列被正确切分出来的比例, 也要考虑算法正确切分出来的列占总切分列数的比例, 这两个标准可以分别用召回率和准确率来表示, 列切分召回率和列切分准确率分别用如下公式计算:

$$\text{列切分召回率} = \frac{\text{正确切分的列数}}{\text{文档中的实际列数}}$$

$$\text{列切分准确率} = \frac{\text{正确切分的列数}}{\text{切分列数}}$$

采用基于投影的列切分方法作为比较算法。列切分结果如表1所示。本文的列切分算法获得了较高的列切分召回率和准备率, 无论在召回率或准确率上, 本文算法都比投影法要高, 并且召回率高出了将近10个百分点。对具体的切分结果进行分析, 发现投影法很难切分出文档中的短列, 而本文的算法对短列的切分效果较好。除此之外, 本文算法切分得到的列比投影法切分得到的列更为完整。本算法的实际切分效果以及投影法的切分效果可见图5。

表1 列切分结果

样本	算法	实际列数	切分列数	正确切分列数	召回率 / (%)	准确率 / (%)
四书章句集注	投影法	226	211	195	86.28	92.42
琴书	本文算法	226	224	224	99.12	100.00
大全	投影法	728	680	673	92.45	98.97
太音	本文算法	728	715	712	97.80	99.58
大全集	投影法	475	417	382	80.42	91.61
大全集	本文算法	475	472	456	96.00	96.61
合计	投影法	1 429	1 322	1 254	87.75	94.86
	本文算法	1 429	1 411	1 392	97.41	98.65

#### 3.2 字切分

从古籍中挑选了共54列作为实验样本。评价切分结果的优劣, 不但要考虑实际的字被正确切分出来的比例, 也要考虑算法正确切分出来的字占总切分字数的比例。这两个标准可以分别用召回率和准确率来表示:

$$\text{字切分召回率} = \frac{\text{正确切分的字数}}{\text{实际字数}}$$

$$\text{字切分准确率} = \frac{\text{正确切分的字数}}{\text{切分字数}}$$

在实验中, 选择了文献[8]中的算法作为比较算法。该算法首先根据汉字结构特征, 对连通域的合并, 然后根据规则对可能存在粘连的切分块做进一步的切分。由于这些规则是针对横写汉字归纳的, 而且文献[8]没有给出具体的规则, 因此为了应用到本实验中, 对粘连切分这个步骤做了必要的修改。另外, 也用文献[5]的算法进行了实验, 但由于实验样本粘连和重叠较多, 并且字之间的间隙较小, 该算法的效果不大理想, 所以在此处就不做进一步的比较。

字切分结果如表2所示。从表2可以看出, 无论是召回率还是准确率, 本文的算法都高出了将近10个百分点。图6中给出了一些字切分的结果。从图6中可以看出, 通过细切分, 即使相邻两个字存在重叠或粘连, 也能被正确

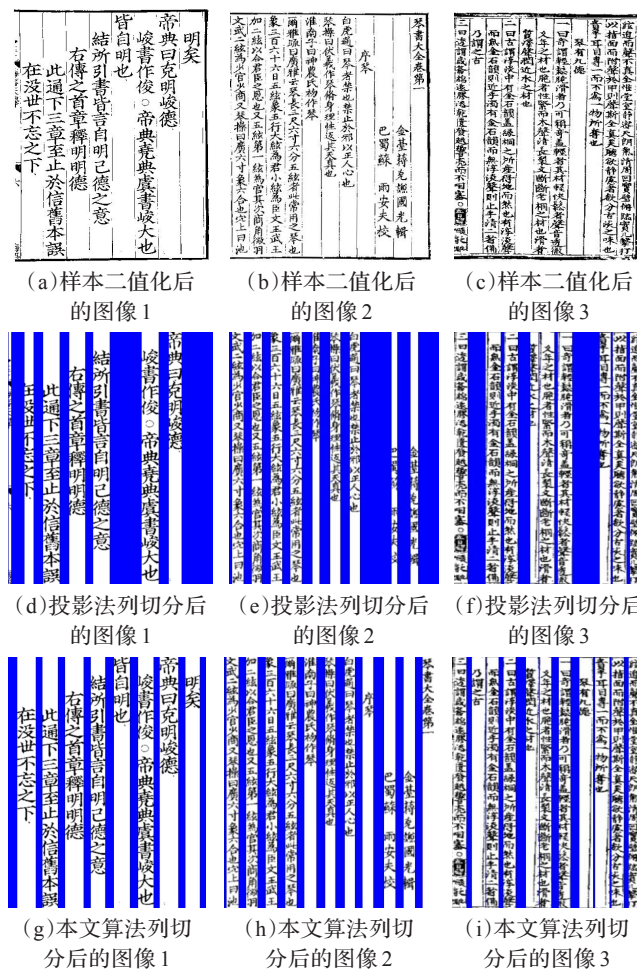


图5 列切分结果示例

(每列为一组, (a)取自《四书章句集注》, (b)取自《琴书大全》, (c)取自《太音大全集》)

完整地切分出来。这些实验结果表明本文提出的字切分算法可以用于字间存在粘连和重叠的情况, 因此它较为适合应用于古文档的字切分。

#### 4 结束语

本文根据古代汉字文档的特点, 提出了适合于古文档切分的方法。实验结果表明本文的方法有如下优点: 本文的列切分算法不受格线的影响, 可以直接对古文档进行列切分。在列间隔较小、列与格线存在粘连、文档具有一定程度的倾斜的情况下, 仍能准确抽取出列。尤其对短列的切分达到了较好的效果。如果文档倾斜较小, 则在列切分的同时能准确地去除格线。对于间隙较小且具有较多粘连和重叠汉字的列, 本文的字切分算法也能较好地切分得到完整单字。

#### 参考文献:

- [1] 周昌乐. 手写汉字的机器识别[M]. 北京: 科学出版社, 1997.
- [2] Tseng Y H, Lee H J. Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm[J]. Pattern Recognition Letters, 1999, 20: 791-806.

表2 字切分结果

算法	实际 字数	切分 字数	正确切 分字数	召回率 /(%)	准确率 /(%)
文献[8]的算法	1 009	957	771	76.41	80.56
本文算法	1 009	951	858	85.03	90.22

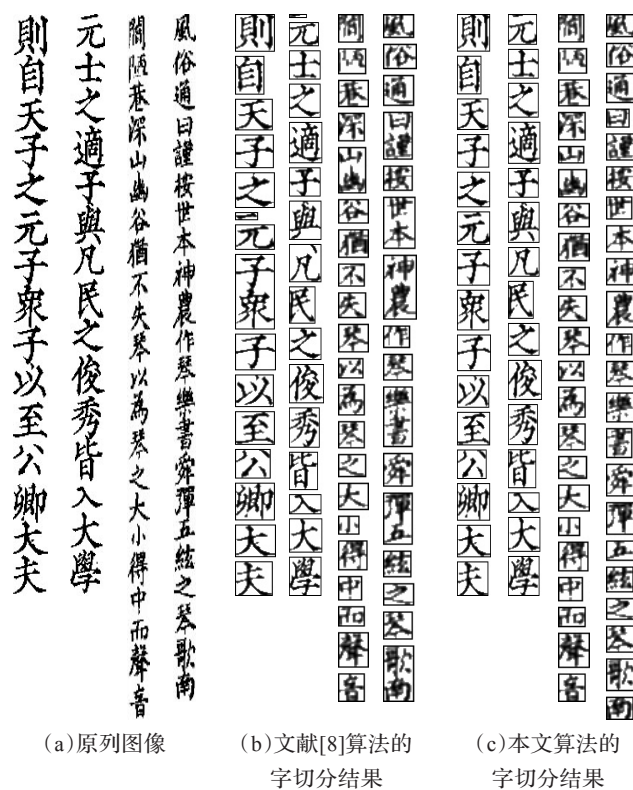


图6 字切分结果示例

- [3] Yin F, Liu C L. Handwritten text line segmentation by clustering with distance metric learning[C]//Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition. Quebec, Canada: [s.n.], 2008.
- [4] Yin F, Liu C L. Handwritten text line segmentation by clustering with distance metric learning[C]//International Conference Frontiers in Handwriting Recognition. Montreal, Canada: [s.n.], 2008: 229-234.
- [5] Hong C, Loudon G, Wu Y, et al. Segmentation and recognition of continuous handwriting Chinese text[J]. International Journal of Pattern Recognition and Artificial Intelligence, 1998, 12(2): 223-232.
- [6] 韩智, 刘昌平, 殷绪成. 手写中文信封的地址行字符切分算法[J]. 中文信息学报, 2006, 20(1): 85-90.
- [7] 马洋涛, 陶志穗, 张金焕, 等. 一种用于手写汉字切分的优化模型[J]. 计算机工程与应用, 2008, 44(2): 227-229.
- [8] 吕岳, 施鹏飞, 张克华. 基于汉字结构特征的自由格式手写体切分[J]. 电子学报, 2000, 28(5): 102-104.
- [9] 李小园, 杨芳, 张望博. 基于结构聚类和笔画分析的粘连手写汉字切分[J]. 计算机工程与应用, 2008, 44(34): 163-165.

(下转 38 页)

```

<Efficiency rdf:ID="Efficiency_lowpressureturbine_1">
  <Max rdf:datatype="http://www.w3.org/2001/
    XMLSchema#float"
    >-1.0</Max>
</Efficiency>
</Detail>
</Fault_Phenomenon>
</Relation_RP>
...

```

对于某故障现象——“高压涡轮效率降低1.2%”的故障诊断过程为:首先,通过查询Feature类各子类相应实例中的Max属性,判断出该故障现象从属于故障现象实例8中;然后,装载入规则进行推理,在推理结果中查找出与该故障现象实例有关的推理结果为:(Fault\_Phenomenon\_8 Relation\_PR Fault\_Reason\_3),即高压涡轮效率降低1.2%的故障原因是高压涡轮故障引起的;最后,根据Fault\_Phenomenon\_8与Fault\_Reason\_3之间的关系路径,根据定理1计算出该故障与原因间的关联度为 $0.8 \times 0.8 = 0.64$ 。

## 6 结束语

针对航空维修行业内不同角色企业故障诊断知识、系统、资源的分散、封闭、异构等原因导致的故障知识不易共享、重用的问题,采用本体来构建故障知识库;鉴于故障现象与故障结论间存在的不确定性关系,从知识的表示及本体的推理等方面考虑,给出了故障现象间、故障现象与故障原因间的复杂关系在故障本体模型中的设计。进一步,编写相应的推理规则,设计了基于本体的不确定性推理故障诊断系统。最后,以飞机发动机故障诊断为例进行验证,体现了该方法支持知识重用及基于本体的不确定性推理的优越性。

## 参考文献:

- [1] Aretakis N, Mathioudakis K, Stamatis A. Nonlinear engine component fault diagnosis from a limited number of measurements using a combinatorial approach[J]. Transactions of the ASME, 2003, 125: 642-650.
- [2] Zedda M, Singh R. Gas turbine engine and sensor fault diagnosis using optimization techniques[J]. Journal of Propulsion & Power, 2002, 5.
- [3] Borguet S, Dewallef P. On-line transient engine diagnostics in a Kalman filtering framework[J]. ASME Turbo Expo, 2005, 1.
- [4] Borguet S, Dewallef P, Leonard O. A way to deal with model-plant mismatch for a reliable diagnosis in transient operation[J]. Journal of Engineering for Gas Turbines & Power, 2008, 130(3).
- [5] 陈恬, 孙健国, 杨蔚华, 等. 自组织神经网络航空发动机气路故障诊断[J]. 航空学报, 2003, 24(1): 46-48.
- [6] 徐启华, 师军. 应用SVM的发动机故障诊断若干问题研究[J]. 航空学报, 2005, 26(6): 686-690.
- [7] Jia Q S, Zhao Q C. A SVM-based method for engine maintenance strategy optimization[C]//Proceedings of the 2006 IEEE International Conference on Robotics and Automation. [S.l.]: IEEE Press, 2006: 1066-1071.
- [8] Lee S M, Roh W J, Choi D W. Defect diagnostics of gas turbine engine using hybrid SVM-artificial neural network method[C]//43rd AIAA/ASME/SAE/ASEE Joint Propulsion Conference. Cincinnati: [s.n.], 2007: 1107-1119.
- [9] 李端, 艾永乐. 神经网络信息融合用于电梯故障诊断的研究[J]. 计算机工程与应用, 2010, 46(14): 231-234.
- [10] Volponi A J. Data fusion for enhanced aircraft engine prognostics and health management, NASA/CR-2005-214055[R]. Houston: NASA, 2005.
- [11] 包健, 刘晓蕾. 基于信息融合技术的电梯控制系统故障诊断[J]. 计算机应用研究, 2011, 28(3): 869-872.
- [12] 杜宁, 黄道. 基于本体论的化工过程故障诊断研究[J]. 计算机集成制造系统, 2005, 11(4): 587-591.
- [13] 赵静, 鄢萍, 陈国荣, 等. 基于本体的机电设备故障诊断知识表示方法研究[J]. 制造业自动化, 2011, 33(8): 79-84.
- [14] 郭建文, 于德介, 刘坚, 等. 网格环境下基于本体的设备故障诊断服务发现方法[J]. 中国机械工程, 2009, 20(20): 2440-2447.
- [15] Dickinson I. The Jena ontology API[EB/OL]. [2012-07-30]. <http://jena.sourceforge.net/ontology/index.html>.
- [16] Su T H, Zhang T W, Guan D J, et al. Off-line recognition of realistic Chinese handwriting using segmentation-free strategy[J]. Pattern Recognition, 2009, 42: 167-182.
- [17] 马瑞, 杨静宇. 一种有效的手写汉字多步分割方法[J]. 中国图形图像学报, 2007, 12(11): 2062-2067.
- [18] 杨武夷, 张树武. 一种视频中字符的集成型切分与识别算法[J]. 自动化学报, 2010, 36(10): 1468-1476.
- [19] Zhou P, Ma S P, Jiang Z. Modification of the Chinese character segmentation method based on units amalgamation[J]. Journal of Chinese Information Processing, 1998, 13(2): 33-39.
- [20] Liu C L, Koga M, Fujisawa H. Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(11): 1425-1437.
- [21] 蒋焰, 丁晓青, 任征. 基于地址结构匹配的手写中文地址的切分与识别[J]. 清华大学学报, 2006, 46(7): 1235-1238.
- [22] Wang Q F, Yin F, Liu C L. Integrating language model in handwritten Chinese text recognition[C]//The 10th International Conference on Document Analysis and Recognition. Barcelona, Spain: [s.n.], 2009: 1036-1040.

(上接33页)