基于自适应特征与多级反馈模型的中英文混排文档分割[1)]

夏 勇 王春恒 戴汝为
(中国科学院自动化研究所 北京 100080)
(E-mail: yong.xia@ia.ac.cn)

摘 要 提出了一种基于自适应特征与多级反馈模型的新颖的字符分割方法，对文字图像质量与中英文混排格式有较好的自适应能力. 该方法的主要思想就是将一个分割过程分成很多层，每层都会由一个主要特征来指导字符分割与中英文预分类，然后将分割层的结果反馈至当前分割层或前面的分割层，并指导下一层的分割. 该方法将字符分割、中英文预分类和字符识别这三者进行了很好的融合，大大提高了字符分割与识别的正确率.

关键词 中英文混排文档分割，中英文预分类，自适应特征与多级反馈模型，对文档图像的自适应特性，OCR
中图分类号 TP391

Segmentation of Mixed Chinese/English Document Based on AFMPF Model

XIA Yong WANG Chun-Heng DAI Ru-Wei
(*Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)
(E-mail: yong.xia@ia.ac.cn)

Abstract This paper proposes a novel method to segment document image based on adaptive feature and multiple phase feedback (AFMPF) model, which is adaptive to the blurring of document image and various patterns of mixed Chinese/English document. First, the whole process of segmentation is divided into several phases and each phase is allocated a primary feature to segment document image. Second, the segmentation result of each phase is fed back to the current or previous phase, and directs the next phase segmentation. This method causes good and effective combination among character segmentation, pre-classification of Chinese/English and character recognition, which improves greatly the accuracy of character segmentation and recognition.

Key words Segmentation of mixed Chinese/English document, pre-classification of Chinese/English, adaptive feature and multiple phase feedback model (AFMPF model), adaptation to various document image, OCR

1 引言

目前，有关字符分割的文献大多讨论的是纯英文、纯数字或纯汉字的分割 [1~3]，而对于中英文混排文档的字符分割较少有论文涉及，目前中英文混排文档的分割依然还是字符

1) 国家自然科学基金项目 (60472066) 资助
Supported by National Natural Science Foundation of P. R. China (60472066)
收稿日期 2005-9-1 收修改稿日期 2005-11-23
Received September 1, 2005; in revised form November 23, 2005

分割领域的一个难点. [4] 中首先利用连通域分析的方法得到连通域单元, 并对有上下依赖关系的单元进行合并. 然后, 利用笔画特征, 对每一个单元进行分类, 即西文和汉字 (包括汉字部首). 然后利用识别的结果来对分类的结果进行校验. [5,6] 中将字符的中心点间距作为一个字符预分类的特征. 一般而言, 相邻两个汉字的中心点间的距离比较一致, 而英文字符除了具有固定宽度的字体外, 相邻两个字符中心间的距离是不一致的, 变化会比较大. [7] 中对造成中英文分割困难的原因进行了分析, 然后采用了两步法来进行分割. 即首先利用投影法得到一个初始的分割块, 然后对英文字符串进行检测, 若是英文字符串就调用英文的分割引擎, 否则就调用汉字的分割引擎. [8] 中通过投影法来获得初始的分割块, 然后根据各块的宽高比信息对相邻块进行合并, 接着利用了基于不同特征得到的两个分类器的识别结果, 以识别结果的可信度为引导来进行分割. [9] 提出了一个适用于多语种混排分割的通用的框架, 首先利用连通域分析得到初始分割块, 然后利用字符的识别结果及字符几何特征的校验的方法来进行进一步分割, 最后利用上下文语义上的关系来进行校验. [10] 与 [11] 讨论了有关韩语与英文混排时的分割方法, 由于韩语与中文在结构上有一定的相似性, 所以这些方法对中英文混排的分割也有一定的借鉴意义. [10] 中利用了多种几何与结构特征, 如字高、笔画特征等, 来将韩语字符与英文字符分开. [11] 中利用了韩语与英文字混排时的基线特征来指导文字的分割.

以上介绍的这些方法能解决一些情况下中英文混排文字的分割, 但都不是很全面、很系统, 且大多数对于混排文档分割的方法过多的依赖字符识别的结果. 由于中英文混排的形式很多, 对于一些较为复杂的情况, 用现有的方法很难做到既准确又快速. 我们提出了一种基于自适应特征与多级反馈模型 (Adaptive feature and multiple phase feedback model) 的分割方法来解决中英文混排文档的分割问题, 该方法提高了对文字图像质量及字符混排形式的自适应能力, 大大提高了字符分割的正确率, 并且速度相对于完全基于识别的分割的方法也有明显的提高. 该方法最为显著的特点就是充分利用字符的几何与结构特点, 加快中英文预分类的过程, 同时充分利用每一个分割阶段的信息, 及时进行反馈与计算, 提高了对中英文混排文档的自适应能力.

2 基于自适应特征与多级反馈模型的分割

2.1 自适应特征与多级反馈模型

多级反馈模型是指将一行文字的分割过程按字符特征类别分成一种层状结构. 每一层由一种特征为主来指导分割, 层与层之间引入反馈, 层内部也有反馈. 这样就形成了一个多特征融合的反馈分割结构. 所谓特征的自适应是指根据当前层的分割结果进行分析计算, 确定下一层应该采用什么特征来指导分割, 即在实际分割时, 分割层不一定是依照一个固定的顺序来选择的, 有可能会发生跳跃的现象. 模型框图如图 1 所示.

图 1 中的控制模块, 主要负责整个分割过程的分析与调度. 图中的分割层根据字符特征的复杂度来进行排列. 控制模块对每一层的分割结果进行评估, 并对该层特征值进行修正, 如果符合要求, 则进入下一层继续分割. 注意, 特征值的选择遵循由简单到复杂的顺序, 但并不一定是连续的, 也就是说如果当前分割层是第 i 层, 则下一个可能选择的分割层为第 j 层, j 一定大于 i, 但 j 并不一定等于 $i+1$.

如图 2 所示, 控制模块由调度模块、评估模块和特征选择模块组成. 调度模块主要是负责分割层之间的转换. 评估模块主要是对当前分割层的结果进行评价, 就是判断当前层是否需要内部自反馈, 或者是需要将当前层的分割结果反馈至前面的分割层, 或者是进入后续的分割层. 特征选择模块以当前层的评估结果为输入, 经过分析得出下一个将要采用的最合适的特征. 这是一个非常关键的模块, 对系统的分割性能有重大的影响. 特征的自适应选择是一件很困难的事情. 首先, 字符的特征是很多的, 如何选取有效的特征就是一

个问题. 其次，在什么情况下，采用什么样的特征来指导分割，这无疑也是很难确定的. 这里，我们采用了分层分割的方法，利用每层的反馈信息来进行特征的自适应分析，取得了较好的效果.

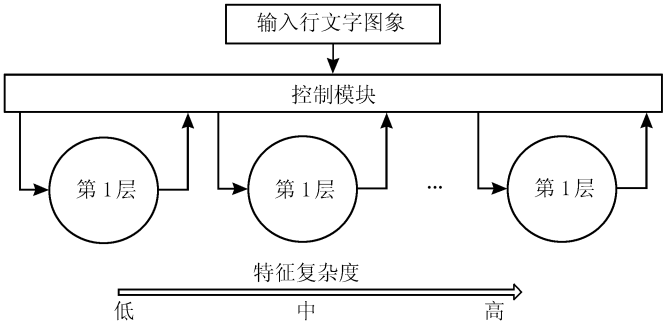

图 1　AFMPF 模型结构
Fig. 1　Structure of AFMPF model

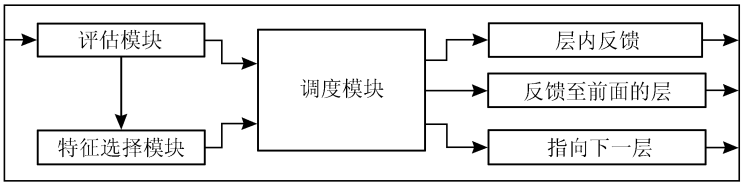

图 2　控制模块结构
Fig. 2　Structure of control module

AFMPF 模型是一个多层反馈分割模型，对每一层分割的结果都会进行评价，并根据评价结果自适应的动态调用一些新的字符特征来指导分割. 从特征的选择上而言，我们既选用了几何与结构特征，同时也选用了统计特征. 一般的识别引擎实质上就是利用了统计特征，而其它的特征如字宽字高等均属于几何特征. 很好的几何与结构特征能够进行字符识别，但由于结构特征的提取比较困难，且鲁棒性不好，所以在字符识别中较少直接采用. 但如果仅仅只是对字符集进行一个粗略的划分，如中英文的预分类，利用字符的结构特征将是一个很好的方法，并且有很好的鲁棒性.

一个典型的按分割层顺序连续分割的过程如图 3 所示. 该过程只考虑了层内的分割反馈，没有考虑层间反馈的情况. 在一般情况下，仅考虑层内分割的反馈就已经可以达到很好的效果了，当然如果引进层间的反馈，则效果会更好，但同时处理时间会随着分割层数的增加而急剧增长，这主要针对那些质量较差的难以分割的图像会有理想的效果.

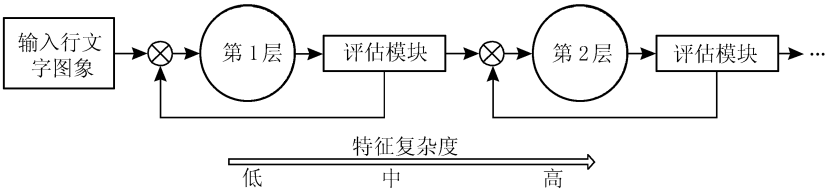

图 3　基于 AFMPF 模型分割的典型结构
Fig. 3　A typical frame of segmentation based on AFMPF model

2.2 特征的自适应分析与评估

中英文混排文档分割的一个关键问题就是如何确定各分割单元的语言属性, 即中英文的预分类. 下面将着重对采用 AFMPF 模型的方法对中英文进行预分类的过程进行详细的阐述.

2.2.1 基本特征的选取

基本特征包括字宽、字高、字间隙等. 这些特征经常被用来进行字符的分割. 这些特征不仅要从单个分割单元中提取, 而且还要对整行文本进行统计. 另外还有一些特征对字符的预分类与分割有很好的效果, 如字符的中心点间距、上下穿刺特征、基线特征、笔画特征、连通域特征等. 下面对上面提到的几个特征做一个简单描述.

字符的中心点间距特征[5,6]. 在一行文本中, 一般汉字为同一字号, 则相邻汉字的间距一般为一个恒定值. 而相邻英文字符的间距一般为一个变化值, 且小于汉字的间距.

字符的上下穿刺特征[10]. 在英文字符中, 每一次上下穿刺的交点数不会超过 4 个, 而汉字中存在上下穿刺的交点数大于 4 个的情况.

字符的基线特征[11]. 如图 4 所示, 中英文混排的字符一般存在 6 种基线. 基线特征对中英文的预分类有很好的效果, 对粘连字符的分割也会有很大的帮助. 在没有任何先验知识的情况下, 直接提取中英文混排文档的基线是很困难的. 在实际文档的测试中, 我们发现一般对于一个文本行, 可以利用已有的文字预分类信息, 根据已经确定类别信息的文字来估算文字基线, 然后根据基线信息指导还未确定类别属性的文字的预分类.

图 4 中英文混排文档字符基线特征图

Fig. 4 Baseline of character in mixed Chinese/English text line

字符的笔画特征. 笔画特征无论是对中英文的预分类还是对字符的识别, 都会有很好的效果, 但笔画特征的提取一直没有得到很好的解决, 尤其是曲线笔画的提取, 难度很大, 鲁棒性也不是很好. 这里我们仅提取两种简单的笔画, 即水平笔画和垂直笔画. 由于大部分汉字中这两种笔画的数量较多, 而英文字符中这种笔画较少, 所以笔画特征对中英文的预分类会有很大的帮助.

字符的连通域特征. 这里是指一个分割单元中连通域的数量、密度、位置等特性. 一般英文字符为单连通域, 除了 i, j 为两个连通域外. 而汉字中有很多包含多个连通域. 所以此特征对中英文的预分类会有一定的帮助, 但此特征对字符的断笔极其敏感, 所以仅适用于那些图像质量较好, 没有断笔的情况下使用.

2.2.2 特征的分析与评估

对于中英文混排文档的分割, 我们的目标不仅是要求有较高的分割正确率, 而且也要保证速度. 简单的特征一般计算复杂度低, 但预分类的效果不一定很好; 同样, 一个复杂的特征, 在预分类时的效果不一定就比简单特征好. 对 2.2.1 中提到的各种特征, 我们对每一种特征分别在实际文档样本库中进行了中英文预分类实验. 在实验中, 我们对每一行的文本提供一个行高信息, 然后分别以不同的特征来进行预分类, 统计分类的结果. 根据实验结果, 我们发现字高、字中心点间距的分类效果和鲁棒性较好, 其次为统计特征和基线特征, 再次为字间隙、字笔画、穿刺特征. 考虑到特征的计算复杂度与预分类效果的整体性能, 我们最终选定的各分割层的主导特征分别为: 字高、字中心间距、字间隙、基线特征、上下穿刺特征、字笔画、统计特征.

我们用 $f(x)$ 表示基于单个特征的分类的可信度函数，其中 x 为各种分类特征．每种特征都有两个可信度域值，即高域值 TH_i 和低域值 TL_i, 其中 i 的取值范围为 $1,\cdots,7$.

单个特征的可信度函数 $f(x_i)=\begin{cases}1, & x_i > TH_i\\ 0.5, & TL_i < x_i \leqslant TH_i,\ i=1,\cdots,7\\ 0, & x_i \leqslant TL_i\end{cases}$.

多特征融合的累积可信度函数 $g(x)=\dfrac{1}{n}\sum_{j=1}^{n} f(x_j)$, j 为分割层代号, n 为分割层数量.

在分割的过程中，我们采用了以行为单位来进行分割．即一个分割层的作用对象是一行文字，而不是单个字符．对于文字行的分割，采用了行信息与局部信息结合的方法．行信息主要是行文字的平均高度、平均间隙等多种信息，随着分割层的不断增多，会有一些新的行信息增加进来．对于每层的分类可信度，如果不是 1, 则转入下一层的分割；凡是可信度为 1 者，则直接跳至最后一个分割层，即利用统计特征进行识别，若可信度为 1, 则分类结束，给出结果．否则，此字符的类别暂时设定为不确定状态，等待下一次循环中进行处理．对一行文字预分类完毕，若有新的字符被确定类别，则需重新统计行信息．若行信息有明显的变化，则基于新的行信息，对该行文字中没有确定类别属性的文字再进行一次预分类；若没有新的字符被确定类别，而又存在没有被确定类别的字符，则根据累积可信度的结果来判定类别，以 0.5 为域值，高于此值为一个类别，低于此值为另一个类别．

当一行文本的预分类过程结束后，每个分割单元的类别属性就确定了，需要根据各类别的特点来进行字块的融合与粘连字符的分拆．对于印刷体的中文字符而言，由于字符的粘连情况不是很常见，所以主要需要解决的问题就是分割单元间的融合，因为汉字是多部首结构，一个汉字可能由几个分割单元组成．英文字符分割主要是解决英文字符的粘连问题，这里主要采用了基于识别的分割方法[12]．

3　实验结果

图 5 给出了利用 AFMPF 模型对一个文本行进行分割与识别的全过程.

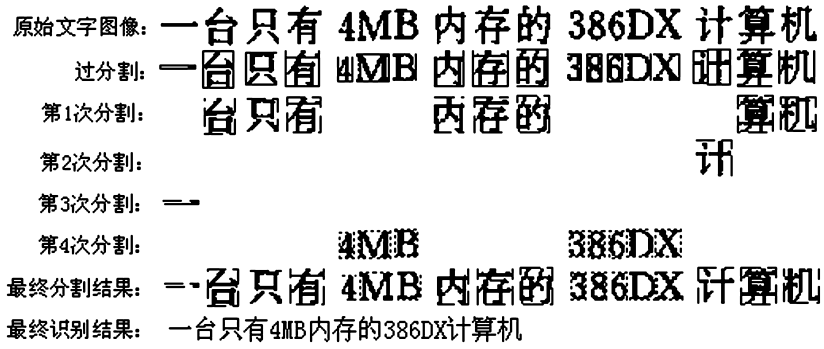

图 5　基于 AFMPF 模型分割的一个例子
各次基于行信息的分割采用的特征：第 1 次分割：高度特征、统计特征；第 2 次分割间隙特征、中心点间距特征、统计特征第 3 次分割：笔画特征、中心点间距特征、统计特征；第 4 次分割统计特征

Fig. 5　Structure of AFMPF model
Features adopted by segmentation based on row infomation: First segmentation: Height, statistics; Second segmentation: Gap, distance of middle statistics; Third segmentation: Stroke, distance of middle, statistics; Fourth segmentation: Statistics

图 5 中的过分割步骤就是指先对文本行进行连通域分析，然后基于规则对具有包含关系或上下交错的连通域进行合并，以合并后的连通域作为一个基本分割单元，即为图中的

实线框内的部分. 从图中可以看出, 经过 4 次分割, 一行文本就已经成功的完成了分割过程, 并且字符的类别属性也被标定了, 其中长虚线框表示为中文属性, 短虚线框表示为非中文属性.

对于中英文混排文档而言, 一般中文的数量要远大于英文, 且有许多汉字的特点比较突出, 很容易和英文区分开, 所以, 我们在做预分类的时候是先找可信度最高的汉字, 然后利用这些已经确定类别属性的分割单元来指导其它分割单元的分割与预分类. 对于实验中的识别器的设计, 我们选用了 256 维的轮廓方向线素特征与 64 维的网格灰度特征, 并经 LDA 压缩为 128 维, 分类器采用基于欧氏距离的最近邻法.

我们对 150 篇中英文混排的杂志和书籍文档进行了测试. 字符总数为 65872, 其中中文字符个数为 45005, 英文字符个数为 20867. 我们组建了两个测试系统, 系统 1 采用了基于识别反馈的分割方法; 系统 2 采用了基于 AFMPF 模型的方法. 测试在 Pentium IV-3.2GHz CPU 的 PC 机上进行. 测试结果如表 1 所示.

表 1 测试结果

Table 1 The result of test

测试系统	预分类错误数	预分类正确率 (%)	分割错误数	分割正确率 (%)	识别错误数	识别正确率 (%)	速度 (字/ 秒)
1	811	98.77	985	98.50	1103	98.33	1000
2	503	99.24	583	99.11	632	99.04	1300

从以上数据可见, 对中英文混排文档仅仅利用识别的反馈来指导分割是不够的, 应该融合多种特征来提高分类与分割的精度, 同时还可以适当提高速度.

4 结束语

目前, 基于反馈的分割思想主要是指将识别的结果反馈来指导分割. 这种思想已经有很多文章作了较为详细的阐述. 很早就已经有学者指出, 在进行字符的分割时不引入识别是荒谬的, 这一点已经被广大学者所认同, 并且基于识别的分割的思想也被应用到了实际的文档分割中, 取得了较好的分割效果. 但是, 在中英文混排的文档中, 如果仅仅只利用识别的结果来指导分割是不够的, 因为字符识别的结果存在一个可信度的问题, 英文字符与汉字或汉字部首存在一些相似字, 仅仅通过一些统计特征是很难将它们区分开的.

我们采用基于 AFMPF 模型的方法, 解决了字符的多特征融合问题, 并将其应用于字符的预分类与分割中, 取得了很好的效果. 由于特征的选取对解决中英文字符的预分类与分割是很重要的, 如何选取更有效的特征来指导分割是一个值得研究的问题, 因为一个好的特征不仅可以提高字符分割的准确度, 而且还可以大大提高速度. 基于特征进行预分类的可信度评价方法是 AFMPF 模型能否充分发挥其反馈特性来提高分割与识别精度的关键所在. 目前, 对于严重退化的文档, 效果还不是很理想, 以后将对基于特征进行预分类的可信度评价方法进行更深入的研究与实验, 同时对于反馈的调度策略也将做进一步深入的研究.

References

1 Casey R, Lecolinet E. A survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996, **18**(7): 690∼706

2 Morita M, Sabourin R, Bortolozzi F, Suen C Y. Segmentation and recognition of handwritten dates. In: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition, Niagara-on-the-Lake, Ontario, Canada: IEEE Press, 2002. 105∼110

3 Lin Y T, Chen R C. A new method for segmenting handwritten Chinese characters. In: Proceedings of the Fourth International Conference on Document Analysis and Recognition, Ulm, Germany: IEEE Press, 1997. **2**: 568∼571

4 Kuo H H, Wang J F. A new method for the segmentation of mixed handprinted Chinese/English characters. In: Proceedings of the Second International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan: IEEE Press, 1993. 810∼813
5 Wang K, Jin J M, Pan W M, Shi G S, Wang Q R. Mixed Chinese/English document auto-processing based on the periodicity. In: Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, China: IEEE Press, 2004. **6**: 3616∼3619
6 Wang K, Wang Q R. Research on Chinese/English mixed document recognition. *Journal of Software*, 2005, **16**(5): 786∼798
7 Guo H, Ding X Q, Zhang Z, Guo F X, Wu You-Shou. Realization of a high-performance bilingual Chinese-English OCR system. In: Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, Canada: IEEE Press, 1995. **2**: 978∼981
8 Feng Z D, Huo Q. Confidence guided progressive search and fast match techniques for high performance Chinese/English OCR. In: Proceedings of the Sixteenth International Conference on Pattern Recognition, Quebec, Canada: IEEE Press, 2002. **3**: 89∼92
9 Wen D, Ding X Q. A general framework for multi-character segmentation and its application in recognizing multilingual Asian documents. Smith E.H. Barney, Hu J, Allan J. (Eds.), In: Document Recognition and Retrieval XI, SPIE: 2004. **5296**: 147∼154
10 Hwang Y S, Moon K A, Chi S Y, Jang D G, Oh W G. Segmentation of a text printed in Korean and English using structure information and character recognizers. In: Proceedings of 2000 IEEE International Conference on Systems, Man, and Cybernetics, Nashville, USA: IEEE Press, 2000. **3**: 1586∼1591
11 Kim J H, Kim K K, Chien S I, Choi H M. Segmentation of touching characters in printed Korean/English document recognition. In: Proceedings of 1996 IEEE International Conference on Systems, Man, and Cybernetics, Beijing, China: IEEE Press, 1996. **1**: 438∼443
12 Su L, Ahmadi M, Shridhard M. Segmentation of touching characters in printed document recognition. In: Proceedings of the Second International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan: IEEE Press, 1993. 569∼572

夏　勇　博士研究生，研究方向包括模式识别、图像处理、文字识别、文档检索.

(**XIA　Yong**　Ph. D. candidate at Institute of Automation, Chinese Academy of Sciences. His research interests include pattern recognition, image processing, character recognition, and document retrieval.)

王春恒　研究员，博士生导师，研究方向包括模式识别和智能系统理论与方法、文字识别、综合集成理论、复杂系统理论.

(**WANG Chun-Heng**　Professor. His research interests include pattern recognition, intelligent system, character recognition, meta-synthetic theory, and complexity theory.)

戴汝为　研究员，博士生导师，中国科学院院士，研究方向包括模式识别和智能系统理论与方法、文字识别、综合集成理论、复杂系统理论.

(**DAI Ru-Wei**　Professor, academician of Chinese Academy of Sciences. His research interests include pattern recognition, intelligent system, character recognition, meta-synthetic theory, and complexity theory.)