

文章编号: 2095-2163(2020)05-0206-05

中图分类号: TP391.4

文献标志码: A

基于连体段的印刷维吾尔文特征提取

贾钰峰, 章蓬伟, 贾园园, 邵小青, 刘茂霞

(新疆科技学院 工商管理系, 新疆 库尔勒 841300)

摘要: 典型的印刷识别系统由预处理, 特征提取, 训练模型, 识别器组成。本文讨论了基于连体段的印刷维吾尔文特征提取方法。结合维吾尔文特点提取了: 孔洞数, 尾点, 交叉点, 方向码, 笔划位置特征, 前后景比值等。并利用以上特征进行了印刷维吾尔文的模型训练和识别。

关键词: 印刷维吾尔文; 连体段; 特征提取

Printed Uighur feature extraction methods based on Word-part

JIA Yufeng, ZHANG Pengwei, JIA Yuanyuan, SHAO Xiaoqing, LIU Maoxia

(Xinjiang University of science and technology Department of business administration, Korla Xinjiang 841300, China)

【Abstract】 A typical printing recognition system consists of preprocessing, feature extraction, training model and recognizer. This article discusses the printed Uighur feature extraction methods based on word-part. Combining features of Uighur, feature extraction are as follows: hole number, end point, cross-point, orientation codes, relative position features of gestures, black and white pixel ratio etc. The above features are used to train and recognize the printed Uighur model.

【Key words】 printed Uighur; word-part; feature extraction

0 引言

关于维吾尔文印刷识别方面, 相关的研究文献资料较少。但维吾尔文与阿拉伯文很相似, 参考了阿拉伯文及相关印刷识别方法^[1-2]; 典型的识别系统模块是由预处理、特征提取、训练模型、识别器组成的, 如图1。由维吾尔文的特点得知: 印刷的文字切分不论以笔划, 字母还是词, 切分都是相当困难的^[3-5]。同时还有图像文本噪点等因素影响, 如: 粘连, 断裂, 伪字母切分等。基于连体段(WordPart)^[3]的段切分是一个很好的解决方案。它能够保留出整体的完备信息, 从而为提取良好的特征做准备。本文在已经预处理的基础上, 对连体段提取各类特征。

1 维吾尔文的特点

维吾尔族的语言属于阿尔泰语系突厥语族。1938年形成现行文字, 有32个字母拼音自右向左横写, 且有120多个字符形式, 以词的形式表达^[6]。其部分特点如下:

(1) 维吾尔文字母包括主体部分和附加部分, 其中20个字母有附加部分, 它的形式为一“◆”、两“”、多“”、以及“”、“”等。附加部分在主体笔划的内部、下面或上面。如: 一、二、三、四、五、六、七、八、九、十、十一、十二、十三、十四、十五、十六、十七、十八、十九、二十、二十一、二十二、二十三、二十四、二十五、二十六、二十七、二十八、二十九、三十、三十一、三十二、三十三、三十四、三十五、三十六、三十七、三十八、三十九、四十、四十一、四十二、四十三、四十四、四十五、四十六、四十七、四十八、四十九、五十、五十一、五十二、五十三、五十四、五十五、五十六、五十七、五十八、五十九、六十、六十一、六十二、六十三、六十四、六十五、六十六、六十七、六十八、六十九、七十、七十一、七十二、七十三、七十四、七十五、七十六、七十七、七十八、七十九、八十、八十一、八十二、八十三、八十四、八十五、八十六、八十七、八十八、八十九、九十、九十一、九十二、九十三、九十四、九十五、九十六、九十七、九十八、九十九、一百、一百零一、一百零二、一百零三、一百零四、一百零五、一百零六、一百零七、一百零八、一百零九、一百一十、一百一十一、一百一十二、一百一十三、一百一十四、一百一十五、一百一十六、一百一十七、一百一十八、一百一十九、一百二十、一百二十一、一百二十二、一百二十三、一百二十四、一百二十五、一百二十六、一百二十七、一百二十八、一百二十九、一百三十、一百三十一、一百三十二、一百三十三、一百三十四、一百三十五、一百三十六、一百三十七、一百三十八、一百三十九、一百四十、一百四十一、一百四十二、一百四十三、一百四十四、一百四十五、一百四十六、一百四十七、一百四十八、一百四十九、一百五十、一百五十一、一百五十二、一百五十三、一百五十四、一百五十五、一百五十六、一百五十七、一百五十八、一百五十九、一百六十、一百六十一、一百六十二、一百六十三、一百六十四、一百六十五、一百六十六、一百六十七、一百六十八、一百六十九、一百七十、一百七十一、一百七十二、一百七十三、一百七十四、一百七十五、一百七十六、一百七十七、一百七十八、一百七十九、一百八十、一百八十一、一百八十二、一百八十三、一百八十四、一百八十五、一百八十六、一百八十七、一百八十八、一百八十九、一百九十、一百九十一、一百九十二、一百九十三、一百九十四、一百九十五、一百九十六、一百九十七、一百九十八、一百九十九、二百、二百零一、二百零二、二百零三、二百零四、二百零五、二百零六、二百零七、二百零八、二百零九、二百一十、二百一十一、二百一十二、二百一十三、二百一十四、二百一十五、二百一十六、二百一十七、二百一十八、二百一十九、二百二十、二百二十一、二百二十二、二百二十三、二百二十四、二百二十五、二百二十六、二百二十七、二百二十八、二百二十九、二百三十、二百三十一、二百三十二、二百三十三、二百三十四、二百三十五、二百三十六、二百三十七、二百三十八、二百三十九、二百四十、二百四十一、二百四十二、二百四十三、二百四十四、二百四十五、二百四十六、二百四十七、二百四十八、二百四十九、二百五十、二百五十一、二百五十二、二百五十三、二百五十四、二百五十五、二百五十六、二百五十七、二百五十八、二百五十九、二百六十、二百六十一、二百六十二、二百六十三、二百六十四、二百六十五、二百六十六、二百六十七、二百六十八、二百六十九、二百七十、二百七十一、二百七十二、二百七十三、二百七十四、二百七十五、二百七十六、二百七十七、二百七十八、二百七十九、二百八十、二百八十一、二百八十二、二百八十三、二百八十四、二百八十五、二百八十六、二百八十七、二百八十八、二百八十九、二百九十、二百九十一、二百九十二、二百九十三、二百九十四、二百九十五、二百九十六、二百九十七、二百九十八、二百九十九、三百、三百零一、三百零二、三百零三、三百零四、三百零五、三百零六、三百零七、三百零八、三百零九、三百一十、三百一十一、三百一十二、三百一十三、三百一十四、三百一十五、三百一十六、三百一十七、三百一十八、三百一十九、三百二十、三百二十一、三百二十二、三百二十三、三百二十四、三百二十五、三百二十六、三百二十七、三百二十八、三百二十九、三百三十、三百三十一、三百三十二、三百三十三、三百三十四、三百三十五、三百三十六、三百三十七、三百三十八、三百三十九、三百四十、三百四十一、三百四十二、三百四十三、三百四十四、三百四十五、三百四十六、三百四十七、三百四十八、三百四十九、三百五十、三百五十一、三百五十二、三百五十三、三百五十四、三百五十五、三百五十六、三百五十七、三百五十八、三百五十九、三百六十、三百六十一、三百六十二、三百六十三、三百六十四、三百六十五、三百六十六、三百六十七、三百六十八、三百六十九、三百七十、三百七十一、三百七十二、三百七十三、三百七十四、三百七十五、三百七十六、三百七十七、三百七十八、三百七十九、三百八十、三百八十一、三百八十二、三百八十三、三百八十四、三百八十五、三百八十六、三百八十七、三百八十八、三百八十九、三百九十、三百九十一、三百九十二、三百九十三、三百九十四、三百九十五、三百九十六、三百九十七、三百九十八、三百九十九、四百、四百零一、四百零二、四百零三、四百零四、四百零五、四百零六、四百零七、四百零八、四百零九、四百一十、四百一十一、四百一十二、四百一十三、四百一十四、四百一十五、四百一十六、四百一十七、四百一十八、四百一十九、四百二十、四百二十一、四百二十二、四百二十三、四百二十四、四百二十五、四百二十六、四百二十七、四百二十八、四百二十九、四百三十、四百三十一、四百三十二、四百三十三、四百三十四、四百三十五、四百三十六、四百三十七、四百三十八、四百三十九、四百四十、四百四十一、四百四十二、四百四十三、四百四十四、四百四十五、四百四十六、四百四十七、四百四十八、四百四十九、四百五十、四百五十一、四百五十二、四百五十三、四百五十四、四百五十五、四百五十六、四百五十七、四百五十八、四百五十九、四百六十、四百六十一、四百六十二、四百六十三、四百六十四、四百六十五、四百六十六、四百六十七、四百六十八、四百六十九、四百七十、四百七十一、四百七十二、四百七十三、四百七十四、四百七十五、四百七十六、四百七十七、四百七十八、四百七十九、四百八十、四百八十一、四百八十二、四百八十三、四百八十四、四百八十五、四百八十六、四百八十七、四百八十八、四百八十九、四百九十、四百九十一、四百九十二、四百九十三、四百九十四、四百九十五、四百九十六、四百九十七、四百九十八、四百九十九、五百、五百零一、五百零二、五百零三、五百零四、五百零五、五百零六、五百零七、五百零八、五百零九、五百一十、五百一十一、五百一十二、五百一十三、五百一十四、五百一十五、五百一十六、五百一十七、五百一十八、五百一十九、五百二十、五百二十一、五百二十二、五百二十三、五百二十四、五百二十五、五百二十六、五百二十七、五百二十八、五百二十九、五百三十、五百三十一、五百三十二、五百三十三、五百三十四、五百三十五、五百三十六、五百三十七、五百三十八、五百三十九、五百四十、五百四十一、五百四十二、五百四十三、五百四十四、五百四十五、五百四十六、五百四十七、五百四十八、五百四十九、五百五十、五百五十一、五百五十二、五百五十三、五百五十四、五百五十五、五百五十六、五百五十七、五百五十八、五百五十九、五百六十、五百六十一、五百六十二、五百六十三、五百六十四、五百六十五、五百六十六、五百六十七、五百六十八、五百六十九、五百七十、五百七十一、五百七十二、五百七十三、五百七十四、五百七十五、五百七十六、五百七十七、五百七十八、五百七十九、五百八十、五百八十一、五百八十二、五百八十三、五百八十四、五百八十五、五百八十六、五百八十七、五百八十八、五百八十九、五百九十、五百九十一、五百九十二、五百九十三、五百九十四、五百九十五、五百九十六、五百九十七、五百九十八、五百九十九、六百、六百零一、六百零二、六百零三、六百零四、六百零五、六百零六、六百零七、六百零八、六百零九、六百一十、六百一十一、六百一十二、六百一十三、六百一十四、六百一十五、六百一十六、六百一十七、六百一十八、六百一十九、六百二十、六百二十一、六百二十二、六百二十三、六百二十四、六百二十五、六百二十六、六百二十七、六百二十八、六百二十九、六百三十、六百三十一、六百三十二、六百三十三、六百三十四、六百三十五、六百三十六、六百三十七、六百三十八、六百三十九、六百四十、六百四十一、六百四十二、六百四十三、六百四十四、六百四十五、六百四十六、六百四十七、六百四十八、六百四十九、六百五十、六百五十一、六百五十二、六百五十三、六百五十四、六百五十五、六百五十六、六百五十七、六百五十八、六百五十九、六百六十、六百六十一、六百六十二、六百六十三、六百六十四、六百六十五、六百六十六、六百六十七、六百六十八、六百六十九、六百七十、六百七十一、六百七十二、六百七十三、六百七十四、六百七十五、六百七十六、六百七十七、六百七十八、六百七十九、六百八十、六百八十一、六百八十二、六百八十三、六百八十四、六百八十五、六百八十六、六百八十七、六百八十八、六百八十九、六百九十、六百九十一、六百九十二、六百九十三、六百九十四、六百九十五、六百九十六、六百九十七、六百九十八、六百九十九、七百、七百零一、七百零二、七百零三、七百零四、七百零五、七百零六、七百零七、七百零八、七百零九、七百一十、七百一十一、七百一十二、七百一十三、七百一十四、七百一十五、七百一十六、七百一十七、七百一十八、七百一十九、七百二十、七百二十一、七百二十二、七百二十三、七百二十四、七百二十五、七百二十六、七百二十七、七百二十八、七百二十九、七百三十、七百三十一、七百三十二、七百三十三、七百三十四、七百三十五、七百三十六、七百三十七、七百三十八、七百三十九、七百四十、七百四十一、七百四十二、七百四十三、七百四十四、七百四十五、七百四十六、七百四十七、七百四十八、七百四十九、七百五十、七百五十一、七百五十二、七百五十三、七百五十四、七百五十五、七百五十六、七百五十七、七百五十八、七百五十九、七百六十、七百六十一、七百六十二、七百六十三、七百六十四、七百六十五、七百六十六、七百六十七、七百六十八、七百六十九、七百七十、七百七十一、七百七十二、七百七十三、七百七十四、七百七十五、七百七十六、七百七十七、七百七十八、七百七十九、七百八十、七百八十一、七百八十二、七百八十三、七百八十四、七百八十五、七百八十六、七百八十七、七百八十八、七百八十九、七百九十、七百九十一、七百九十二、七百九十三、七百九十四、七百九十五、七百九十六、七百九十七、七百九十八、七百九十九、八百、八百零一、八百零二、八百零三、八百零四、八百零五、八百零六、八百零七、八百零八、八百零九、八百一十、八百一十一、八百一十二、八百一十三、八百一十四、八百一十五、八百一十六、八百一十七、八百一十八、八百一十九、八百二十、八百二十一、八百二十二、八百二十三、八百二十四、八百二十五、八百二十六、八百二十七、八百二十八、八百二十九、八百三十、八百三十一、八百三十二、八百三十三、八百三十四、八百三十五、八百三十六、八百三十七、八百三十八、八百三十九、八百四十、八百四十一、八百四十二、八百四十三、八百四十四、八百四十五、八百四十六、八百四十七、八百四十八、八百四十九、八百五十、八百五十一、八百五十二、八百五十三、八百五十四、八百五十五、八百五十六、八百五十七、八百五十八、八百五十九、八百六十、八百六十一、八百六十二、八百六十三、八百六十四、八百六十五、八百六十六、八百六十七、八百六十八、八百六十九、八百七十、八百七十一、八百七十二、八百七十三、八百七十四、八百七十五、八百七十六、八百七十七、八百七十八、八百七十九、八百八十、八百八十一、八百八十二、八百八十三、八百八十四、八百八十五、八百八十六、八百八十七、八百八十八、八百八十九、八百九十、八百九十一、八百九十二、八百九十三、八百九十四、八百九十五、八百九十六、八百九十七、八百九十八、八百九十九、九百、九百零一、九百零二、九百零三、九百零四、九百零五、九百零六、九百零七、九百零八、九百零九、九百一十、九百一十一、九百一十二、九百一十三、九百一十四、九百一十五、九百一十六、九百一十七、九百一十八、九百一十九、九百二十、九百二十一、九百二十二、九百二十三、九百二十四、九百二十五、九百二十六、九百二十七、九百二十八、九百二十九、九百三十、九百三十一、九百三十二、九百三十三、九百三十四、九百三十五、九百三十六、九百三十七、九百三十八、九百三十九、九百四十、九百四十一、九百四十二、九百四十三、九百四十四、九百四十五、九百四十六、九百四十七、九百四十八、九百四十九、九百五十、九百五十一、九百五十二、九百五十三、九百五十四、九百五十五、九百五十六、九百五十七、九百五十八、九百五十九、九百六十、九百六十一、九百六十二、九百六十三、九百六十四、九百六十五、九百六十六、九百六十七、九百六十八、九百六十九、九百七十、九百七十一、九百七十二、九百七十三、九百七十四、九百七十五、九百七十六、九百七十七、九百七十八、九百七十九、九百八十、九百八十一、九百八十二、九百八十三、九百八十四、九百八十五、九百八十六、九百八十七、九百八十八、九百八十九、九百九十、九百九十一、九百九十二、九百九十三、九百九十四、九百九十五、九百九十六、九百九十七、九百九十八、九百九十九、一千、一千零一、一千零二、一千零三、一千零四、一千零五、一千零六、一千零七、一千零八、一千零九、一千一十、一千一十一、一千一十二、一千一十三、一千一十四、一千一十五、一千一十六、一千一十七、一千一十八、一千一十九、一千二十、一千二十一、一千二十二、一千二十三、一千二十四、一千二十五、一千二十六、一千二十七、一千二十八、一千二十九、一千三十、一千三十一、一千三十二、一千三十三、一千三十四、一千三十五、一千三十六、一千三十七、一千三十八、一千三十九、一千四十、一千四十一、一千四十二、一千四十三、一千四十四、一千四十五、一千四十六、一千四十七、一千四十八、一千四十九、一千五十、一千五十一、一千五十二、一千五十三、一千五十四、一千五十五、一千五十六、一千五十七、一千五十八、一千五十九、一千六十、一千六十一、一千六十二、一千六十三、一千六十四、一千六十五、一千六十六、一千六十七、一千六十八、一千六十九、一千七十、一千七十一、一千七十二、一千七十三、一千七十四、一千七十五、一千七十六、一千七十七、一千七十八、一千七十九、一千八十、一千八十一、一千八十二、一千八十三、一千八十四、一千八十五、一千八十六、一千八十七、一千八十八、一千八十九、一千九十、一千九十一、一千九十二、一千九十三、一千九十四、一千九十五、一千九十六、一千九十七、一千九十八、一千九十九、二千、二千零一、二千零二、二千零三、二千零四、二千零五、二千零六、二千零七、二千零八、二千零九、二千一十、二千一十一、二千一十二、二千一十三、二千一十四、二千一十五、二千一十六、二千一十七、二千一十八、二千一十九、二千二十、二千二十一、二千二十二、二千二十三、二千二十四、二千二十五、二千二十六、二千二十七、二千二十八、二千二十九、二千三十、二千三十一、二千三十二、二千三十三、二千三十四、二千三十五、二千三十六、二千三十七、二千三十八、二千三十九、二千四十、二千四十一、二千四十二、二千四十三、二千四十四、二千四十五、二千四十六、二千四十七、二千四十八、二千四十九、二千五十、二千五十一、二千五十二、二千五十三、二千五十四、二千五十五、二千五十六、二千五十七、二千五十八、二千五十九、二千六十、二千六十一、二千六十二、二千六十三、二千六十四、二千六十五、二千六十六、二千六十七、二千六十八、二千六十九、二千七十、二千七十一、二千七十二、二千七十三、二千七十四、二千七十五、二千七十六、二千七十七、二千七十八、二千七十九、二千八十、二千八十一、二千八十二、二千八十三、二千八十四、二千八十五、二千八十六、二千八十七、二千八十八、二千八十九、二千九十、二千九十一、二千九十二、二千九十三、二千九十四、二千九十五、二千九十六、二千九十七、二千九十八、二千九十九、三千、三千零一、三千零二、三千零三、三千零四、三千零五、三千零六、三千零七、三千零八、三千零九、三千一十、三千一十一、三千一十二、三千一十三、三千一十四、三千一十五、三千一十六、三千一十七、三千一十八、三千一十九、三千二十、三千二十一、三千二十二、三千二十三、三千二十四、三千二十五、三千二十六、三千二十七、三千二十八、三千二十九、三千三十、三千三十一、三千三十二、三千三十三、三千三十四、三千三十五、三千三十六、三千三十七、三千三十八、三千三十九、三千四十、三千四十一、三千四十二、三千四十三、三千四十四、三千四十五、三千四十六、三千四十七、三千四十八、三千四十九、三千五十、三千五十一、三千五十二、三千五十三、三千五十四、三千五十五、三千五十六、三千五十七、三千五十八、三千五十九、三千六十、三千六十一、三千六十二、三千六十三、三千六十四、三千六十五、三千六十六、三千六十七、三千六十八、三千六十九、三千七十、三千七十一、三千七十二、三千七十三、三千七十四、三千七十五、三千七十六、三千七十七、三千七十八、三千七十九、三千八十、三千八十一、三千八十二、三千八十三、三千八十四、三千八十五、三千八十六、三千八十七、三千八十八、三千八十九、三千九十、三千九十一、三千九十二、三千九十三、三千九十四、三千九十五、三千九十六、三千九十七、三千九十八、三千九十九、四千、四千零一、四千零二、四千零三、四千零四、四千零五、四千零六、四千零七、四千零八、四千零九、四千一十、四千一十一、四千一十二、四千一十三、四千一十四、四千一十五、四千一十六、四千一十七、四千一十八、四千一十九、四千二十、四千二十一、四千二十二、四千二十三、四千二十四、四千二十五、四千二十六、四千二十七、四千二十八、四千二十九、四千三十、四千三十一、四千三十二、四千三十三、四千三十四、四千三十五、四千三十六、四千三十七、四千三十八、四千三十九、四千四十、四千四十一、四千四十二、四千四十三、四千四十四、四千四十五、四千四十六、四千四十七、四千四十八、四千四十九、四千五十、四千五十一、四千五十二、四千五十三、四千五十四、四千五十五、四千五十六、四千五十七、四千五十八、四千五十九、四千六十、四千六十一、四千六十二、四千六十三、四千六十四、四千六十五、四千六十六、四千六十七、四千六十八、四千六十九、四千七十、四千七十一、四千七十二、四千七十三、四千七十四、四千七十五、四千七十六、四千七十七、四千七十八、四千七十九、四千八十、四千八十一、四千八十二、四千八十三、四千八十四、四千八十五、四千八十六、四千八十七、四千八十八、四千八十九、四千九十、四千九十一、四千九十二、四千九十三、四千九十四、四千九十五、四千九十六、四千九十七、四千九十八、四千九十九、五千、五千零一、五千零二、五千零三、五千零四、五千零五、五千零六、五千零七、五千零八、五千零九、五千一十、五千一十一、五千一十二、五千一十三、五千一十四、五千一十五、五千一十六、五千一十七、五千一十八、五千一十九、五千二十、五千二十一、五千二十二、五千二十三、五千二十四、五千二十五、五千二十六、五千二十七、五千二十八、五千二十九、五千三十、五千三十一、五千三十二、五千三十三、五千三十四、五千三十五、五千三十六、五千三十七、五千三十八、五千三十九、五千四十、五千四十一、五千四十二、五千四十三、五千四十四、五千四十五、五千四十六、五千四十七、五千四十八、五千四十九、五千五十、五千五十一、五千五十二、五千五十三、五千五十四、五千五十五、五千五十六、五千五十七、五千五十八、五千五十九、五千六十、五千六十一、五千六十二、五千六十三、五千六十四、五千六十五、五千六十六、五千六十七、五千六十八、五千六十九、五千七十、五千七十一、五千七十二、五千七十三、五千七十四、五千七十五、五千七十六、五千七十七、五千七十八、五千七十九、五千八十、五千八十一、五千八十二、五千八十三、五千八十四、五千八十五、五千八十六、五千八十七、五千八十八、五千八十九、五千九十、五千九十一、五千九十二、五千九十三、五千九十四、五千九十五、五千九十六、五千九十七、五千九十八、五千九十九、六千、六千零一、六千零二、六千零三、六千零四、六千零五、六千零六、六千零七、六千零八、六千零九、六千一十、六千一十一、六千一十二、六千一十三、六千一十四、六千一十五、六千一十六、六千一十七、六千一十八、六千一十九、六千二十、六千二十一、六千二十二、六千二十三、六千二十四、六千二十五、六千二十六、六千二十七、六千二十八、六千二十九、六千三十、六千三十一、六千三十二、六千三十三、六千三十四、六千三十五、六千三十六、六千三十七、六千三十八、六千三十九、六千四十、六千四十一、六千四十二、六千四十三、六千四十四、六千四十五、六千四十六、六千四十七、六千四十八、六千四十九、六千五十、六千五十一、六千五十二、六千五十三、六千五十四、六千五十五、六千五十六、六千五十七、六千五十八、六千五十九、六千六十、六千六十一、六千六十二、六千六十三、六千六十四、六千六十五、六千六十六、六千六十七、六千六十八、六千六十九、六千七十、六千七十一、六千七十二、六千七十三、六千七十四、六千七十五、六千七十六、六千七十七、六千七十八、六千七十九、六千八十、六千八十一、六千八十二、六千八十三、六千八十四、六千八十五、六千八十六、六千八十七、六千八十八、六千八十九、六千九十、六千九十一、六千九十二、六千九十三、六千九十四、六千九十五、六千九十六、六千九十七、六千九十八、六千九十九、七千、七千零一、七千零二、七千零三、七千零四、七千零五、七千零六、七千零七、七千零八、七千零九、七千一十、七千一十一、七千一十二、七千一十三、七千一十四、七千一十五、七千一十六、七千一十七、七千一十八、七千一十九、七千二十、七千二十一、七千二十二、七千二十三、七千二十四、七千二十五、七千二十六、七千二十七、七千二十八、七千二十九、七千三十、七千三十一、七千三十二、七千三十三、七千三十四、七千三十五、七千三十六、七千三十七、七千三十八、七千三十九、七千四十、七千四十一、七千四十二、七千四十三、七千四十四、七千四十五、七千四十六、七千四十七、七千四十八、七千四十九、七千五十、七千五十一、七千五十二、七千五十三、七千五十四、七千五十五、七千五十六、七千五十七、七千五十八、七千五十九、七千六十、七千六十一、七千六十二、七千六十三、七千六十四、七千六十五、七千六十六、七千六十七、七千六十八、七千六十九、七千七十、七千七十一、七千七十二、七千七十三、七千七十四、七千七十五、七千七十六、七千七十七、七千七十八、七千七十九、七千八十、七千八十一、七千八十二、七千八十三、七千八十四、七千八十五、七千八十六、七千八十七、七千八十八、七千八十九、七千九十、七千九十一、七千九十二、七千九十三、七千九十四、七千九十五、七千九十六、七千九十七、七千九十八、七千九十九、八千、八千零一、八千零二、八千零三、八千零四、八千零五、八千零六、八千零七、八千零八、八千零九、八千一十、八千一十一、八千一十二、八千一十三、八千一十四、八千一十五、八千一十六、八千一十七、八千一十八、八千一十九、八千二十、八千二十一、八千二十二、八千二十三、八千二十四、八千二十五、八千二十六、八千二十七、八千二十八、八千二十九、八千三十、八千三十一、八千三十二、八千三十三、八千三十四、八千三十五、八千三十六、八千三十七、八千三十八、八千三十九、八千四十、八千四十一、八千四十二、八千四十三、八千四十四、八千四十五、八千四十六、八千四十七、八千四十八、八千四十九、八千五十、八千五十一、八千五十二、八千五十三、八千五十四、八千五十五、八千五十六、八千五十七、八千五十八、八千五十九、八千六十、八千六十一、八千六十二、八千六十三、八千六十四、八千六十五、八千六十六、八千六十七、八千六十八、八千六十九、八千七十、八千七十一、八千七十二、八千七十三、八千七十四、八千七十五、八千七十六、八千七十七、八千七十八、八千七十九、八千八十、八千八十一、八千八十二、八千八十三、八千八十四、八千八十五、八千八十六、八千八十七、八千八十八、八千八十九、八千九十、八千九十一、八千九十二、八千九十三、八千九十四、八千九十五、八千九十六、八千九十七、八千九十八、八千九十九、九千、九千零一、九千零二、九千零三、九千零四、九千零五、九千零六、九千零七、九千零八、九千零九、九千一十、九千一十一、九千一十二、九千一十三、九千一十四、九千一十五、九千一十六、九千一十七、九千一十八、九千一十九、九千二十、九千二十一、九千二十二、九千二十三、九千二十四、九千二十五、九千二十六、九千二十七、九千二十八、九千二十九、九千三十、九千三十一、九千三十二、九千三十三、九千三十四、九千三十五、九千三十六、九千三十七、九千三十八、九千三十九、九千四十、九千四十一、九千四十二、九千四十三、九千四十四、九千四十五、九千四十六、九千四十七、九千四十八、九千四十九、九千五十、九千五十一、九千五十二、九千五十三、九千五十四、九千五十五、九千五十六、九千五十七、九千五十八、九千五十九、九千六十、九千六十一、九千六十二、九千六十三、九千六十四、九千六十五、九千六十六、九千六十七、九千六十八、九千六十九、九千七十、九千七十一、九千七十二、九千七十三、九千七十四、九千七十五、九千七十六、九千七十七、九千七十八、九千七十九、九千八十、九千八十一、九千八十二、九千八十三、九千八十四、九千八十五、九千八十六、九千八十七、九千八十八、九千八十九、九千九十、九千九十一、九千九十二、九千九十三、九千九十四、九千九十五、九千九十六、九千九十七、九千九十八、九千九十九、

(2) 维吾尔文的单词是由多个或者单个字母组合而成。根据书写规则这些字母的组合形成一个或几个前后相连接的音节或称连体段。连体段中的字母, 在印刷的时候总是沿着水平线的, 这条水平线被称为基线。维吾尔文的结构详情见图2。

2 特征提取

特征提取的过程就是将图像文本映射到文字独有的特征空间, 以便压缩信息量, 方便后续的分类、训练和识别。本文根据维吾尔文的特点, 基于连体段提取出一系列常用特征, 并最终建立一个特征库

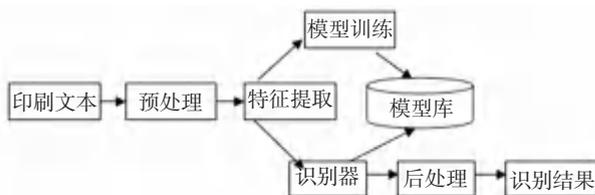


图1 印刷维吾尔文识别系统框架

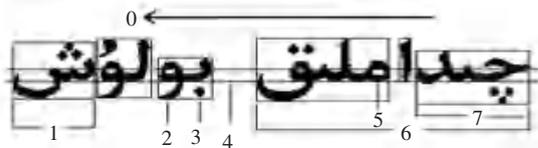
Fig. 1 Printed Uighur recognition system framework

基金项目: 自治州科学技术基金项目(2019018)。

作者简介: 贾钰峰(1986-), 男, 硕士, 助教, 主要研究方向: 图像处理与模式识别、大数据; 章蓬伟(1985-), 男, 硕士, 助教, 主要研究方向: 人工智能与模式识别、大数据; 邵小青(1988-), 女, 硕士, 助教, 主要研究方向: 模式识别、自动化; 贾园园(1988-), 女, 硕士, 助教, 主要研究方向: 电子商务、人工智能; 刘茂霞(1991-), 女, 本科生, 主要研究方向: 电子商务。

收稿日期: 2020-01-07

以备使用,现将方法介绍如下:



- 0.六个连体段组成两个词,方向自右向左; 1.一个字母组成连体段; 2.主笔划; 3.主笔划下方的副笔划; 4.基线和基线的上下限; 5.环; 6.三个连体段组成的词; 7.三个字母组成的连体段

图2 维吾尔文结构特点说明

Fig. 2 Structural features of Uighur

2.1 笔划位置特征

维吾尔文中连体段由主体部分和附加部分组成。由图2可知附加笔划在主笔划的下面、上面或内部。笔划位置特征就是求出主笔划,副笔划个数,副笔划与主笔划的相对位置关系。副笔划的位置特征可以通过与基线的相对位置来识别。但有些特殊情况无法完美判断,如: پ 与 پ 。故介绍一种基于联通域的更细致的判别方法:

选定一点作为一个种子,由图像的区域生长法^[7-8]可以求得一个相同像素相互联通的区域(即笔划)。因此根据种子点就可以知道联通域,根据联通域就知道连体段笔划的数量。根据图像像素点读取的顺序,位图行内是由左往右,行间由下往上。将二值图像扫描时第一个碰到的黑色像素作为种子,再由区域生长法生长蔓延出一个笔划区域,得到一个笔划。为了防止重复计算笔划,将已经蔓延过的笔划反色(使笔划和背景同色)后继续循环寻找下一个种子点,继续生长蔓延找到下一个笔划,直至图循环完毕。该算法主要步骤:

(1)为了不破坏原图,复制初始化为连体段图像的缓冲区两个(设为A和B)。

(2)对A图像域按由左往右,由下往上的方式依次扫描像素点。碰到黑色像素点就作为种子点记录下来。

(3)由区域生长法得出此联通域。在区域生长时同时统计黑点的个数并记录黑点的坐标(黑点总个数为连体段的面积;坐标的平均值可以用来判断笔划的位置关系)。

(4)根据记录的黑点坐标,把B图像域相同坐标处像素值反色(白色)。此时B域就是除去联通域剩下的部分。

(5)将B域的像素值覆盖掉A域。

(6)重复步骤(2),(3),(4),直至循环完毕。

由此算法可统计出连体段的联通域个数,即笔

划数;其中面积最大的为主笔划。其它笔划的平均纵坐标与主笔划比较,就可以判断出副笔划的上下位置特征。同理副笔划之间的平均横坐标可以判断出副笔划之间的左右位置特征,而这些特征都是全局精确特征,可以显著提高识别分类效果。连体段笔划位置特征提取步骤如图3所示。

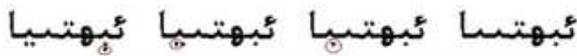


图3 连体段笔划位置特征提取步骤

Fig. 3 Stroke position feature extraction steps of wordpart

其特征结果如下:

(1)笔划数为8个。面积分别为:15,16,16,815,14,82,14,16。最大面积即主笔划面积为:815。

(2)位置特征由下往上依次为:-16,-11,-11,0,-11,13,13,15。其中0代表主笔划;小于0的共4个,代表在主笔划下方;大于0的共3个,代表在主笔划上方。

(3)同时还可以知道副笔画之间的位置关系。

2.2 孔洞数

连体段孔洞数也是基于联通域个数的^[7]。为保证孔洞数目准确性,将图像黑白像素调换后继续递归循环算法对不是笔划的区域进行联通域个数计算,如图4所示,可得到总的不是笔划的连通区域个数,减去和边界相交的背景区域,就可以得到连体段孔洞数数量这一特征。这也是全局精确特征。此特征在后续使用中发现非常高效。

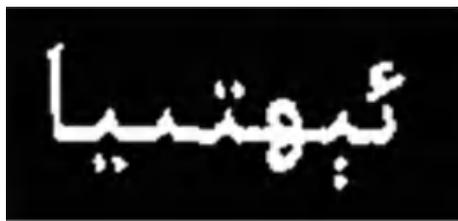


图4 黑色联通域共有3块 孔洞数共有2个

Fig. 4 There are 3 holes in the black connecting area and 2 holes in total figure

2.3 方向码

方向码特征是字符识别中非常经典有效的特征,方向码就是把平面分成八个方向^[9-10],如图5所示。取出主笔划的轮廓,在轮廓上选取一个开始点和一个结束点作为一个线段,并判断这个线段归属于那个方向,然后用方向序号标识。继续重复直至循环轮廓一周后得到一组方向特征向量,这组向量就是主笔划的方向码。它具有较高稳的定性和抗干扰能力,代表主笔划的形状特征。方向码特征提取过程为:求连体段轮廓,点聚类,直线逼近,得出方向码。

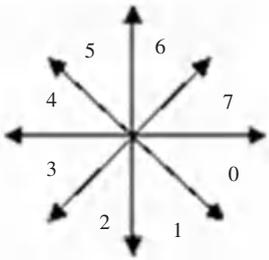


图5 八方向示意图

Fig. 5 Eight direction schematic diagram

(1)取轮廓。取轮廓使用边界跟踪法,从图6得到主笔划,从主笔划选取种子点,顺时针环绕主笔划边界一周后得到连体段的轮廓^[7-11]。



图6 连体段原始图像

Fig. 6 Original image of wordpart

主笔划轮廓选取算法如下:

定义初始搜索方向为左上方,搜索到的第一个黑色像素作为轮廓种子点,找到下一个黑色像素,记录边界点坐标和个数。找不到就顺时针旋转 45° 继续寻找直至找到下一个点。重复以上方法直到返回最初起始种子点为止。轮廓像素点个数就是周长特征,坐标点就是主笔划的离散形式。用此方法取得的轮廓如图7所示。



图7 取出轮廓的连体段

Fig. 7 Take out the wordpart of the contour

(2)点聚类。如隔几个像素点取一个点,使轮廓点稀疏。但点聚类依然不够简化。

(3)直线逼近。在尽量保留拐点、关键点的情况下,仍对轮廓进行进一步的简化采样,将轮廓变为保持原有形状的折线,如图8所示。



图8 直线逼近后的连体段

Fig. 8 The wordpart after the straight line approximation

在这里我选用了经典的 Douglas-Peucker 算

法^[12]实现直线逼近功能,其算法描述如如图9所示。

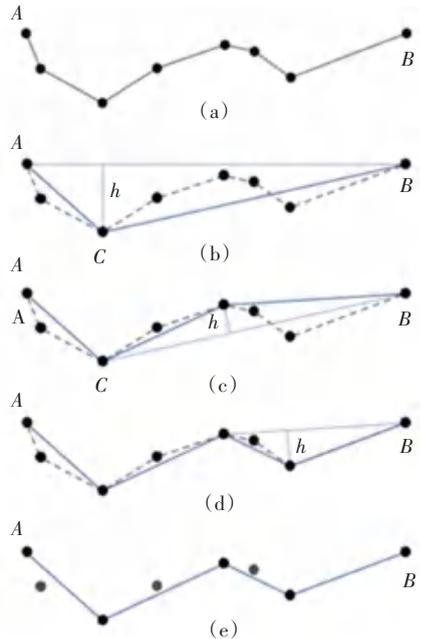


图9 Douglas-Peucker 算法描述过程

Fig. 9 Douglas Peucker algorithm description process

①在曲线 AB 之间连接一条线段 AB ,线段 AB 为曲线的弦,如图9(b)所示。

②求出曲线上离线段 AB 距离最远的点 C ,计算其与线段 AB 的距离 h 。

③若 h 小于预定阈值,则该直线段 AB 作为曲线的近似,该段曲线 AB 处理完毕。

④若 h 大于预定阈值则,利用距离最远的 C 点将曲线 AB 分为两段 AC 和 BC ,线段 AC 和 BC 重复步骤1到3,直至处理完毕,如图9(c)和图9(d)。

⑤依次连接各个最终确定的逼近点作为连体段的近似轮廓,完成简化采样,如图9(e)。

(4)得出方向码。根据笔划骨架中每一段线段归属的方向得出方向码特征向量。

图8的方向码特征向量为(假设阈值为3,具体根据字体大小判断):4,5,5,6,1,0,0,0,7,5,5,6,0,1,0,6,1,7,5,6,0,1,2,3,3,4,2,3,3。若将每个方向的个数求和,则方向码向量降维为:6,4,1,3,5,4,2。

2.4 尾点、交叉点

尾点与交叉点作为细化后字符识别提取的常用特征已被广泛使用^[13]。前人已经研究了各种经典的细化算法,经过论证,本文采用细化算法提取此特征,得到了较好的效果^[7]。在细化后的连体段图像中,交叉点一般都是在当前点的8邻域模板中,即以当前点 p 为中心选择一个 3×3 的模板来判别其尾点、交叉特征属性。计算 P 点的交点数的公式(1):

$$NUM = \sum_{i=0}^6 |P_{i+1} - P_i| + |P_0 - P_7|, SUM = \sum_{i=0}^7 |P_i| \quad (1)$$

模板如图 10 所示。

P_3	P_2	P_1
P_4	P	P_0
P_5	P_6	P_7

图 10 8 邻域模板

Fig. 10 8 Neighborhood templates

其中, SUM 表示当前点 8 领域中像素点的个数。 NUM 表示当前点 P 的 8 邻域模板中像素值的 0,1 变化次数。由算法可以判定出:

- (1) 当 $NUM = 2, SUM = 1$ 时, P 为尾点;
- (2) 当 $SUM > 3$ 时为交叉点。其中, $NUM = 6, SUM = 3$ 时, P 为三叉点; $NUM = 8, SUM = 4$ 时, P 为四叉点。

图 11 为图 3 的主笔划细化后的情况, 共有尾点、三叉以上点各 6 个。

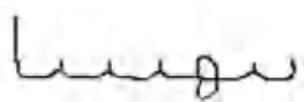


图 11 细化后的主笔划

Fig. 11 Main stroke after thinning

2.5 前后景比值

前后景比值特征是基于统计, 计算出连体段矩

形框中黑像素和与白像素的比值^[14]。这样提取特征的优点在于计算简单, 不受字符大小的影响, 只要字形固定, 对于印刷体就可以作为特征, 如图 12 的前后景比值为 0.25。另外, 印刷体维吾尔文连体段具有一定的高度和宽度, 计算连体段宽高比特征值, 如图 12 宽为 60, 高为 56, 宽高比: 1.071。前后景色比值结合连体段宽高比, 可以作为联合特征, 针对部分连体段可获得较高的稳定性和抗干扰性。



图 12 连体段的像素模型

Fig. 12 Pixel model of wordpart

3 结束语

本文重点表述了维吾尔文连体段常用特征提取方法。在分析维吾尔文书写特点的基础上提取了如下特征: 宽高比, 前后景比值, 孔洞数, 尾点, 交叉点, 方向码, 笔划位置特征。基于以上方法通过对四十多对样张进行批处理, 建立起了基于连体段的特征库图 14 为对图 13 从右往左提取连体段相关特征的截图。

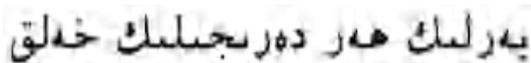


图 13 原始样张部分截取

Fig. 13 Part of the original sample

Unicode	HorRatio	EndPtStr	LoopNum	EndPoint	CrossPoint	DirectionCode	AreaNum	SubGesture	SubGestureDir	LetterStr	EndPtStr
87	1.26	45	1	4	2	11120220	2	0	1	2	38:51
88	1.57	26	0	2	0	01111121	1	0	0	1	28
89	0.68	57	0	6	2	44143141	2	1	0	3	33:54:31:28
90	0.56	47	3	1	5	11121121	1	0	0	2	50:44
91	1.66	27	0	2	0	02101012	1	0	0	1	27
92	1.73	71	0	2	0	02011221	1	0	0	1	50:91
93	3.69	63	1	1	1	01101110	1	0	0	1	63
94	1.63	27	0	2	0	01200221	1	0	0	1	27
95	0.29	39	0	12	0	77384375	3	1	1	0	57:28:50:25:53:29:29
96	0.75	34	1	6	4	21133122	2	1	0	2	35:43
97	1.50	34	1	6	1	33322141	3	2	0	2	39:29

图 14 特征库截图

Fig. 14 Screenshot of feature library

图 14 为特征库的截图: 从左往右每一列的特征分别为: 连体段 Unicode 码, 高宽比, 孔洞数, 尾点个数, 交叉点数, 压缩后的 8 方向码, 笔划数, 基线上方副笔划数, 基线下方副笔划数, 连体段字母个数, 前后景比值等。

将特征应用于维吾尔文字印刷识别系统, 由识别器识别后证明所采用的算法思想和特征维度是有

效的。但还需要进一步完备连体段的有效特征, 以便抽取那些对不同类别最为重要的特征, 组合成良好且优秀的特征组合。后续还需要研究局部特征和全局特征合并训练的方法, 完备字符特征集合, 改进特征选取准则函数, 结合分类器的改进, 尽可能的提高识别率^[15]。