

一种多字体印刷藏文字符的归一化方法^{*}

王 华, 丁晓青

(清华大学 电子工程系, 北京 100084)

摘要: 消除输入字符在位置和大小上的差异的归一化操作是字符识别系统中一个重要环节。在详细分析藏文字符字形特征的基础上, 提出了一种多字体印刷藏文字符归一化方法: 综合运用字符重心和外边框信息实现位置归一化, 然后采用三次 B 样条函数将字符归一化到 48 × 96 的目标点阵。根据所提方法进行的实验证明了其有效性。

关键词: 藏文字符识别; 归一化; 基线; 三次 B 样条

中图法分类号: TP391.43

文献标识码: A

文章编号: 1001-3695(2004)06-0041-03

A Normalization Method of Multi-font Printed Tibetan Characters

WANG Hua, DING Xiao-qing

(Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: In an OCR (Optical Character Recognition) system, character normalization is a crucial step to eliminate variations in character size or position. In this paper, based on the detailed analysis of the characteristics of shape and stroke distribution of multi-font printed tibetan characters, a new normalization algorithm for tibetan OCR is proposed. Firstly character position is normalized combing profile information with the centroid of input character images. Then the 48 × 96 block is introduced to perform the size normalization by cubic B-spline. The effectiveness of proposed algorithm is demonstrated by experimental results.

Key words: Tibetan Character Recognition; Normalization; Baseline; Cubic B-spline

藏文字符识别技术是中文多文种信息处理系统的一个重要组成部分, 开展藏文字符识别的研究具有很高的理论价值和广阔的应用前景。在印刷体藏文中, 由于排版、字号、字体变化等因素的影响, 所以同一藏文字符的不同样本的图像点阵在位置和大小上存在很大的差异。对藏文字符的识别主要是在它的图形结构的基础上进行的, 若经过归一化处理消除字符在位置和大小上的差异, 就无法正确比较藏文字符点阵之间的相似性。

目前国内外对藏文字符识别的研究非常有限, 据对现有文献的检索, 只有文献[3]介绍了一种基于藏文字符基线的位置归一化和基于字符左右外边框的大小归一化方法。该方法在文本质量优良、字符受干扰小和排版规则的情况下, 对单字体藏文字符归一化能取得很好的结果。它最大的缺点在于无法适应多字体藏文字符归一化的要求, 而且它对基线非常敏感, 受输入字符噪声的影响性能下降很快。本文在对藏文字符的字形特点统计分析的基础上, 提出了一种多字体藏文字符的归一化方法, 实验结果证明了其有效性。

1 藏文字符特点简述^[4]

藏文是一种以辅音字母为主要组成部件的特殊拼音文字, 左右拼写、上下叠加, 既不同于西文, 也有别于汉字。藏文以音

节为构词单位, 音节拼写的每一个横向单元称为一个字丁。现代藏文共有 592 个字丁。每个音节中基字所在的字丁是纵向叠加的组合体(叠加层数为 1 ~ 4 层), 该字丁前面可以有前加字, 后面可以有后加字和又后加字。藏文的书写方向从左到右, 起笔在同一水平线上, 该水平线就是基线。藏文字符识别以字丁为基本的识别单位。图 1 为一个 4 字丁藏文音节。



图 1 实际藏文音节示意

2 藏文字符归一化算法

2.1 位置归一化

常用的有重心归一化和外框归一化两种主要的位置归一化方法。外框归一化时各边框的搜索是局部性的, 容易受污点或笔画缺损等干扰的影响; 而重心的计算则是全局性的, 因此重心归一化的抗局部干扰能力强, 能够得到比较稳定的结果。

大多数汉字的笔画在上、下、左、右四个方向的分布比较均匀, 其重心与字形的中心基本重合, 所以采用重心归一化基本上不会造成字形的失真^[1]。而在藏文字符中, 除了以某些本身笔画分布均匀的字母为主要部件构成的字丁的重心与字形中心较接近(图 2(a))外, 大多数字丁的笔画在各方向上分布很不均匀(图 2(b))。若仅采用重心归一化, 相当比例的字丁将产生极大的字形畸变; 而由于边框对噪声和变形的敏感性,

若仅采用外框归一化, 结果欠缺稳定性。因此, 将两种方法结合起来使用, 取长补短, 是一种好的选择。

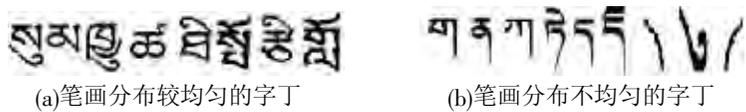


图2 具有不同笔画分布的藏文字丁

设二值化字符图像点阵为 $f(i, j)$, $i = 1, 2, \dots, m, j = 1, 2, \dots, n$, 且 f 在黑像素位置取值为 1, 背景处取值为 0。令其重心和外边框中心分别为 $G(G_I, G_J)$ 和 $C(C_I, C_J)$, 则有:

$$\begin{cases} C_I = m/2 \\ C_J = n/2 \end{cases} \quad (1)$$

$$\begin{cases} G_I = \left(\sum_{j=1}^n \sum_{i=1}^m (i \cdot f(i, j)) \right) / \left(\sum_{j=1}^n \sum_{i=1}^m f(i, j) \right) \\ G_J = \left(\sum_{i=1}^m \sum_{j=1}^n (j \cdot f(i, j)) \right) / \left(\sum_{i=1}^m \sum_{j=1}^n f(i, j) \right) \end{cases} \quad (2)$$

令 $M(M_I, M_J)$ 为介于 $G(G_I, G_J)$ 与 $C(C_I, C_J)$ 之间的一点, 即:

$$\begin{cases} M_I = G_I + (1 - \alpha) C_I \\ M_J = G_J + (1 - \alpha) C_J \end{cases} \quad \text{其中 } \alpha \text{ 为常数且 } 0 < \alpha < 1 \quad (3)$$

移动字符点阵使 M 位于归一化后新的字符点阵的中心, 从而完成输入字符的位置归一化。

2.2 大小归一化

由于字号不同而引起的藏文字符的尺寸相差最大可达十倍, 为使同一识别字典适应多字号字符识别的要求, 必须对识别字符实行有效的大小归一化。方法是先确定输入字符的外接边框, 再将其放大或缩小到规定大小的目标点阵。

藏文字符与汉字的一个显著差别是它并非方块字, 仅字符宽度具有相对稳定性, 而各字符间高度差异很大, 所以不能像汉字那样把藏文字符归一化为诸如 48×48 或 64×64 的方形点阵。考虑到在藏汉混排的文本中, 藏文字丁与汉字的宽度基本相等, 少数情况下比汉字略窄, 我们将归一化后藏文点阵的宽度定为 48。

对收集到的 1 050 套藏文字符样本中共 621 600 个 (6 种字体、7 种字号, 每套样本 592 个字丁) 字丁的高宽比特性做了统计, 得到图 3。藏文字丁高宽比分布具有如下特点: 不同字体间字丁的高宽比特性差异显著; 同一字体字丁的高宽比分布范围非常大; 各字体的高宽比均有一个聚集了 50% 以上字丁的相对集中的分布区间。这些特点决定了归一化目标点阵大小的选择必须考虑各种字体, 兼顾大多数的情况, 同时又要方便处理。据此, 取归一化之后的藏文字符的高宽比为 2 较合理, 这不失为差别各异的各字体字丁高宽比的一个折中。

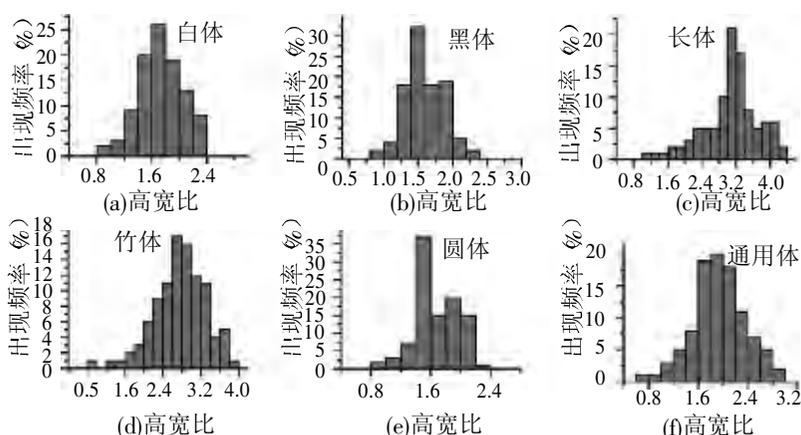


图3 不同字体的藏文字符的宽高比分布直方图

设输入 $m \times n$ 字符图像为 $f(i, j)$, $i = 1, 2, \dots, m, j = 1, 2, \dots, n$, 归一化之后的 $p \times q$ 字符点阵为 $g(i, j)$, $i = 1, 2, \dots, p, j = 1, 2, \dots, q$, 则有:

$$g(i, j) = f(i/r_i, j/r_j) \quad (4)$$

其中 r_i 和 r_j 分别为 i 和 j 方向的尺度变换因子: $r_i = p/m, r_j = q/n$ 。根据上式, 输出图像点阵中的点 (i, j) 对应于输入字符中的点 $(i/r_i, j/r_j)$ 。 $f(i, j)$ 为离散函数, 而 $i/r_i, j/r_j$ 的取值一般不为整数, 故需要根据 f 中已知的离散点处的值来估计其在 $(i/r_i, j/r_j)$ 处的取值。本文采用三次 B 样条函数来进行插值运算, 以减少归一化后字符点阵出现诸如阶梯状边缘等畸变。对于给定 (i, j) , 令:

$$\begin{cases} x = i/r_i = x_0 + \alpha \\ y = j/r_j = y_0 + \beta \end{cases} \quad 0 \leq \alpha, \beta < 1 \quad (5)$$

其中: $\begin{cases} x_0 = \lfloor x \rfloor, & x = x_0 + \alpha \\ y_0 = \lfloor y \rfloor, & y = y_0 + \beta \end{cases}$, $\lfloor \cdot \rfloor$ 为取整函数。插值过程可表示为:

$$g(i, j) = f(x_0 + \alpha, y_0 + \beta) = \sum_{k=-1}^2 \sum_{l=-1}^2 f(x_0 + k, y_0 + l) R_B(k - \alpha) R_B(l - \beta) \quad (6)$$

式中的 $R_B(z)$ 为三次 B 样条函数:

$$R_B(z) = \frac{1}{6} [(z+2)^3 U(z+2) - 4(z+1)^3 U(z+1) + 6z^3 U(z) - 4(z-1)^3 U(z-1)] \quad (7)$$

其中 $U(z)$ 为阶跃函数, $U(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}$

3 实验结果

我们设计了一个藏文字符识别的实验系统, 抽取归一化后字符的方向线素特征^[2] 采用欧氏距离分类器进行识别。从 1 050 套样本中随机抽取 700 套用于训练, 其余留作测试。每次实验将待识别字符归一化成宽度为 48、高度各不相同的点阵, 选取典型的宽高比, 实验共进行 7 次, 分别记录系统的识别率, 得到表 1。从表 1 中可知, 在相同的实验条件下, 当归一化后字符点阵为 48×96 时, 系统的识别性能达到最佳, 从而从实验的角度证明了将藏文字符归一化为 48×96 的合理性。

表1 不同归一化点阵大小下的系统识别率

点阵大小	训练集 (%)	测试集 (%)	点阵大小	训练集 (%)	测试集 (%)
48 × 48	98.73	97.60	48 × 112	99.58	99.18
48 × 64	99.37	98.74	48 × 128	99.63	99.26
48 × 80	99.59	99.15	48 × 144	99.64	99.25
48 × 96*	99.67	99.27			

表 2 给出了固定归一化点阵大小为 48×96 时, 采用不同的位置归一化方法时系统的识别率。可见, 本文的重心与边框相综合的方法得到的识别性能相对于单独采取重心或边框信息均有相当程度的改善。

表2 不同的位置归一化下的系统识别率

位置归一化方法	训练集 (%)	测试集 (%)
重心	98.92	98.05
边框	99.12	98.84
重心 + 边框	99.67	99.27

已有的方法^[3] (简称方法 A) 先在垂直方向上将字符基线移到指定位置 (字符点阵高度的 1/4 处); 然后在水平方向移动字符

左、右外边框的中心到指定位置;最后利用字丁的宽度信息,根据左、右外边框,按比例将图像线性放大或缩小到指定大小的点阵。图 4 列出了一些字符用方法 A 与本文的方法(简称方法 N)进行归一化的结果,而表 3 总结了这两种方法的异同点。



图 4 两种方法对若干白体字符的归一化结果

表 3 方法 A 与方法 N 的比较

归一化阶段	比较项目	方法 A	方法 N
位置归一化	受输入字符外边框的影响	大	小
	对基线位置的敏感性	大	无
大小归一化	用到的外边框	左、右	上、下、左、右
	是否适用于多字体问题	否	是

需要指出的是,方法 A 的最大缺点在于无法适应多字体的情况。图 5 中 2 个白体字符,2 个长体字符,字号均相同。用方法 A 归一化后,字符在宽度和高度两个方向的差异全部集中到高度上,不仅有可能造成部分图像超出点阵范围而丢失(图 5(c) 中的长体字符虚线框内部分),而且使得不同字体的同一字符归一化后的点阵差异太大,无法进行相似性比较;而方法 N 不存在这个问题,能够较好地适应多字体的需要。



图 5 两种方法对于两个白体、长体字符的归一化结果

(上接第 25 页)

$$\hat{F}(R | \langle w_h, sc_h \rangle, \langle w_c, sc_c \rangle) = \frac{C(R, \langle W_h, SC_h \rangle, \langle W_c, SC_c \rangle)}{C(\langle W_h, SC_h \rangle, \langle W_c, SC_c \rangle)} \quad (6)$$

最后,我们可以得到:

$$P(\text{LSF} | w_1, \dots, w_j) = \prod_{k=i, w_k}^j P(\text{SR}(k) | h, w_k) = \prod_{k=i, w_k}^j \hat{F}(R | \langle w_h, sc_h \rangle, \langle w_c, sc_c \rangle) \quad (7)$$

在上面的公式中,我们可以使用基于 Back-Off 方法的参数平滑技术^[5]。分析过程中的动态规划剪枝过程、概率计算过程与规则概率模型是类似的。如果一个单元格中两个部分分析具有相同的属性结构,则具有较低概率的部分分析结果被废弃,不参与后面的分析组合过程。

假设自动应答系统输入的句子为“*She eats pizza without anchovies*”,则现在有:

$$P(T_1) = P(\text{LSF}_1) = P(\text{AGT} | \text{eat, she}) P(\text{OBJ} | \text{eat, pizza}) P(\text{MOD} | \text{pizza, anchovies}) \quad (8)$$

$$P(T_2) = P(\text{LSF}_2) = P(\text{AGT} | \text{eat, she}) P(\text{OBJ} | \text{eat, pizza}) P(\text{MOD} | \text{eat, anchovies}) \quad (9)$$

假设通过语料统计得到的相关模型参数如表 1 所示,则:

$$P(T_1) = 0.0025 \times 0.002 \times 0.003 = 1.5 \times 10^{-6}$$

$$P(T_2) = 0.0025 \times 0.002 \times 0.0001 = 5 \times 10^{-8}$$

表 1 相关模型参数

P(AGT eat, she)	0.0025
P(OBJ eat, pizza)	0.002
P(MOD pizza, anchovies)	0.003
P(MOD eat, anchovies)	0.0001
P(MOD pizza, hesitation)	0.0001
P(MOD eat, hesitation)	0.0008

我们可以根据此结果选择 T_1 作为正确的分析结果。如果将 anchovies 转换为 hesitation,则 $P(T_1) = 5 \times 10^{-8}$, $P(T_2) = 4 \times 10^{-7}$ 。可以看出,随着句子中词语的变化,语言模型仍然可以

4 结束语

在对藏文字符特点进行详细分析的基础上,本文完成了多字体印刷体现代藏文字符识别中的一个重要环节——归一化操作:根据藏文字符的笔画分布特性,综合运用重心和外边框信息实现位置归一化;全面考虑各字体字符的高宽比分布特性,确定了目标点阵大小为 48×96 ,采用三次 B 样条函数完成字符尺度变换。实验证明了本文所提方法的有效性和合理性。最后,将本文的方法与已有方法做了比较分析,表明前者是适合多字体藏文字符的稳定的、有效的归一化方法。

参考文献:

- [1] 吴佑寿,丁晓青. 汉字识别——原理、方法与实现(第一版) [M]. 北京:高等教育出版社,1992.
- [2] ato N, Suzuki M, Omachi S, et al. A Handwritten Character Recognition System Using Directional Element Feature and Asymmetric Mahalanobis Distance [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(3): 258-262.
- [3] 王浩军,赵南元,邓钢铁. 藏文识别的预处理 [J]. 计算机工程, 2001, 27(9): 93-96.
- [4] 王维兰. 藏文基本字符识别算法研究 [J]. 西北民族学院学报(自然科学版), 1999, 22(3): 20-23, 51.
- [5] ou H, Andrews H. Cubic Splines for Image Interpolation and Digital Filtering [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(6): 508-517.

为我们选择合理的分析结果,这正是它的优点所在。

4 结束语

总之,自动应答系统的核心技术是自然语言理解技术。自然语言理解技术涉及到知识库和语料库的建设、文本的切分和标注、句子的语法分析和语义分析等。本文重点论述了自动应答系统中知识信息的语义网络表示和 LSF 随机化句法分析模型,并对模型进行了参数训练,实践证明这些技术是可行的。而且,通过我们与太原威廉公司合作开发的服务于银行的“受限领域自动应答系统”,这些技术被证明是高效的、可推广的。

参考文献:

- [1] ugene Charniak. Statistical Parsing a Context-free Grammar and Word Statistics [C]. Proceedings of the 14th National Conference on Artificial Intelligence, 1997. 78.
- [2] obert Kass. Modeling User's Interests in Information Filtering [C]. Proceedings 5th Text Retrieval Conference, 2001. 101-103.
- [3] 田盛丰. 人工智能与知识工程 [M]. 北京:中国铁道出版社,2001.
- [4] 姚天顺. 自然语言理解 [M]. 北京:清华大学出版社,2002.
- [5] ollins M. A New Statistical Parser Based on Bigram Lexical Dependencies [C]. Proceedings of the 34th Annual Meeting of the ACL, 1996. 184-191.

作者简介:

郭庆琳(1973-),男,副教授,博士研究生,主要研究领域为自然语言处理与理解、智能人机接口技术;樊孝忠(1948-),男,教授,博士生导师,主要研究领域为自然语言处理、自然语言理解、模式识别、人工智能、智能人机接口、多媒体数据压缩。