

基于方向线素特征的手写体维文字符识别

姜 文, 卢朝阳, 李 静

(西安电子科技大学 通信工程学院, 陕西 西安 710071)

摘 要: 手写体维吾尔文字符识别过程, 其中一个重要的环节是特征提取, 对此提出了一种基于方向线素特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别算法. 实验表明, 此算法是一种简单有效且识别率比较高的识别方法.

关键词: 手写体维吾尔文字符; 方向线素; KNN 识别分类器; 训练样本; 测试样本

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2013)10-0097-04

Recognition of Handwritten Uyghur Character Based on Directional Element Feature

JIANG Wen, LU Zhao-yang, LI Jing

(School of Telecommunication Engineering, Xidian University, Xi'an 710071, China)

Abstract: In recognition of hand-written Uyghur character, one of important steps is feature extraction. This paper gives an algorithm of directional element feature extraction for hand-written Uyghur character, which uses KNN recognition as classifier. The experimental result shows, the algorithm is simple, effective with remarkable recognition rate.

Key words: Hand-written Uyghur Character; directional element; KNN Recognition Classifier; training sample; testing sample

1 引言

手写体维吾尔文字符的书写形式^[1]具有明显的连笔草体书写特征; 书写随意性大; 字符结构简单、笔划较少对字符图像进行特征提取时难以准确地将字符特征提取出来. 基于以上这三个原因, 手写体维吾尔文字符识别技术发展比较缓慢. 但是由于维吾尔文字符是从阿拉伯文字符发展而来的, 近些年来阿拉伯文字符识别技术的发展, 可以根据维吾尔文字符的特点, 借鉴一些阿拉伯文字符识别的方法.

手写体维吾尔文单字符的笔划含有丰富的方向信息. 从时间的角度上而言, 笔划上的方向信息表明了此笔划从前一时刻到后一时刻的走向与运动轨迹. 因此, 本文提出一种基于方向线素特征的手写体维吾尔文字符识别算法. 本算法对经过归一化为 64

$\times 64$ 大小的字符图像之后, 将字符图像划分成 64 个 8×8 的大小相同的网格, 提取每一个网格内字符笔划的方向向量. 由于, 字符样本数较多, 需要将提取出来的方向向量经过降维之后得到的向量作为特征向量, 使用 KNN 分类器对手写体维吾尔文字符进行识别. 该算法已经取得较好的识别结果.

2 基于方向线素特征提取算法的基本原理

维吾尔文单字符进行预处理后, 对 64×64 大小的字符图像进行均匀网格划分, 将每一个网格中字符图像方向线性的特征^[1-4]提取出来, 使用分类器对字符图像进行识别处理.

基于方向线素特征提取算法步骤如下:

(1) 提取每一个笔划点的方向线素特征

计算每个维吾尔文单字符笔划点的方向线素. 假设每个字符的笔划点个数为 N , 总共可以提取到

N 个 8 维的方向特征向量 $f_1, f_2, \dots, f_N$, 如图 1 所示. 其中, D_1 至 D_8 分别代表划分的 8 个方向, 方向像素 $f = (D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8)$ 是将一个字符的笔划方向映射到这八个方向上来描述它的方向特征. 映射方法如图 1(b) 所示, 其中 V 为笔划中相邻两点间的方向向量, 方向向量 V 投影到方向 D_1 上的 a^1 和 D_8 上的 a^2 , 由于方向向量 V 在其余的方向上没有投影, 所以 V 在 8 方向坐标轴上的投影, 两点间的方向向量为 $f = \{a^1, 0, 0, 0, 0, 0, 0, a^2\}$.

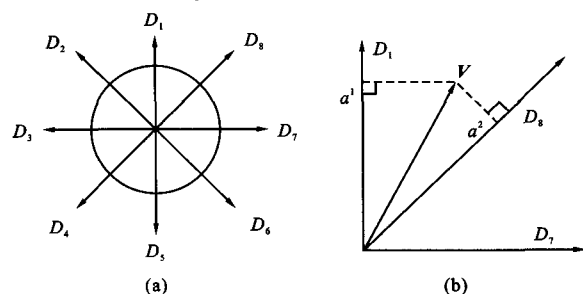


图 1 八方向的方向像素

按照文献[2]中所描述的方法对特征向量进行计算.

(2) 网格划分

字符图像经过归一化处理之后的大小为 64×64 , 实验中对字符图像采用 8×8 均匀网格划分. 然后将每一个网格内笔划点特征向量的 8 方向特征分别相加, 即可得到该网格中的方向特征 $F = (a^1, a^2, a^3, a^4, a^5, a^6, a^7, a^8)$. N 个特征向量的 8 方向特征的表达式形式如式(1)所示.

$$\begin{cases} F \cdot a^1 = \sum_{i=1}^N f_i \cdot a^1 \\ F \cdot a^2 = \sum_{i=1}^N f_i \cdot a^2 \\ \dots\dots\dots \\ F \cdot a^8 = \sum_{i=1}^N f_i \cdot a^8 \end{cases} \quad (1)$$

得到了 64 个 8 维的向量, 将它们依次串联起来, 最后得到 512 维向量, 即字符图像的网格方向特征.

(3) 降维处理

由于字符图像的特征维数较高, 可采用加权采样的方法对提取出来的特征进行了降维处理, 降维过程如下:

①网格特征扩展. 由于实验中将字符图像划分为 8×8 均匀网格, 需要对提取出来的特征向量进行隔点加权采样, 完成降维处理. 为了方便采样需将原有的 8×8 网格补充一行一列扩展为 9×9 的网格. 新增

行列的特征与原网格最后一行一列的特征相同.

②采样权值. 实验中采用的权值如表达式(2)所示, 权值采样分配见图 2(a).

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \quad (2)$$

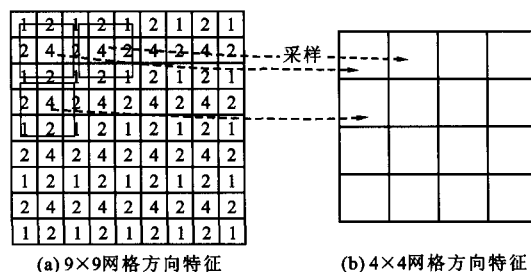


图 2 采样降维

③经过采样之后, 使用加权求和的方法得到的 4×4 网格的特征向量的特征值, 经过采样降维处理过程后, 降维后的新的特征向量维数为 128 维, 如图 2(b)所示.

3 KNN 识别分类器

KNN 算法(K-Nearest Neighbor)^[3] 是一种懒惰的(即直至给定一个测试元组才开始构造泛化模型, 也称为基于实例的学习法)、有监督的、基于机器实例的学习方法, 也是空间向量模型下最好的分类方法之一.

3.1 KNN 算法原理

KNN 算法的定义是取未知样本 x 的 k 个近邻, 看这 k 个近邻中多数属于哪一类, 就把 x 归为那一类. 基本规则是, 在 N 个已知样本中, 找出 x 的 k 个近邻. 设这 k 个样本中, 来自 ω_1 类中的样本有 k_1 个, 来自 ω_2 类中的样本有 k_2 个, $\dots\dots\dots$, 来自 ω_c 类中的样本有 k_c 个, 将判别函数定义为

$$g_i(x) = k_i, i = 1, 2, \dots, c \quad (3)$$

决策规则的表达式形式如下:

$$g_i(x) = \max_i k_i \quad (4)$$

则决策 $x \in \omega_j$. 通常, 为了避免出现 $k_1 = k_2$, k 的取值一般为奇数.

在实际应用中, 确定了 k 的取值之后, 从训练样本中找出 k 个与其最相近的样本, 然后看这 k 个样本中哪个类别的样本多, 则待判定的值(或说抽样)就属于这个类别.

3.2 基于 KNN 分类器的手写体维吾尔文单字符识别算法

使用 KNN 分类器对手写体维吾尔文单字符进

行识别时,字符将被转换成特征空间中的特征向量. 需要根据特征提取步骤中所提取出来的字符图像的 Gabor 能量特征以及网格特征分别使用 KNN 识别分类器来完成的. 使用 KNN 分类器对手写体维吾尔文字符进行识别的步骤如下所示:

(1)将提取出来的特征转换为向量. 定义一个结构体 vector:

```
typedef struct
{
    double attributes[NATTRS];
    int classlabel;
} vector;
```

其中,attributes 表示字符图像的特征,NATTRS 表示特征的个数,classlabel 表示分类号(即手写体维吾尔文单字符的序号).

(2)确定 k 的取值,根据 KNN 算法的原理,必须要确定出未知样本的 k 个近邻,才能对未知样本进行分类识别. 近邻的个数是影响分类准确与否的关键因素. 通常,使用 KNN 算法时, k 的取值为 1、3、5……等奇数^[4-6].

(3)设定训练样本的数目 N , $N = \text{手写体字符数} \times \text{手写体字符组数}$ (本数据库中共有 115 组,通常选择其中的 70 组进行训练),组成的训练集. 训练样本的结构为 vector.

(4)选定一个测试样本与训练集中的所有元素进行相似度比较,进行相似度比较时,计算相似度使用的时欧氏距离法,表达式如下:

$$d(X,Y) = \sqrt{\sum_{i=0}^{N-1} (X. \text{ attributes}[i] - Y. \text{ attributes}[i])^2}$$

式中, N 为样本向量的特征数.

(5)定义一个结构体数组 knn[k]用来存放经过相似度计算与测试样本相似度最高的的 k 个向量的相似度值和字符编号. 结构体 item 被定义为

```
typedef struct
{
    double distance;
    int classlabel;
} item ;
```

其中,distance 表示测试样本与训练集中元素的欧氏距离,即二者的相似度;classlabel 表示分类号.

测试样本与训练样本逐一进行比较,找到 k 个最相似的样本,将这 k 个训练样本与测试样本的距离以及训练样本的字符编号存放在 knn[k]数组中,

knn[k]数组存放按照 distance 取值从小到大的顺序进行排列.

(6)通过数组 knn[k]对测试样本进行分类识别. 对数组 knn[k]中的字符编号进行比较分析:若哪一个字符编号最多,则判定测试样本的字符编号与这个出现次数最多的训练样本的字符编号相同;若数组 knn[k]中的字符编号均不相同,则测试样本的字符编号判定为与排在数组 knn[k]中最前面的那个字符编号.

通过对基于 KNN 识别分类算法步骤,可以看出使用基于 KNN 识别分类算法对手写体维吾尔文单字符进行识别能够充分的利用所提取的特征,是一种快速有效的分类识别算法.

4 实验结果及分析

本文采用基于方向线索特征提取算法和 KNN 识别分类器对手写体维吾尔文单字符进行识别实验. 实验使用的是实验室自建的手写体维吾尔文单字符数据库,本数据库维吾尔文单字符图像共计 115 套,每套包含 128 个维吾尔文单字符. 这 115 套维吾尔文单字符由 23 人书写,每人写了 5 套.

基于方向线索特征识别手写体维吾尔文单字符时. 本文选择这 115 套中的前 70 套作为训练样本由其中的 14 人书写,后 45 套作为为测试样本由其余的 9 人书写. 训练样本中不包含被测试样本书写者的笔迹样本.

基于方向线索特征的手写体维吾尔文单字符笔迹特征时,将这 115 套样本数据按书写人分为 23 组,每组 5 套,选择每 5 套中的 3 套作为训练样本,剩余 2 套为测试样本. 训练样本共 69 套,测试样本共 46 套. 训练样本中包含被测试样本书写者的笔迹样本.

测试环境: CPU Intel Core (TM) i3-2310M, 2 GB内存,Windows 7 操作系统.

将基于方向线索特征的 KNN 分类识别结果与基于方向线索特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别结果进行对比. 识别结果如表 1 所示.

表 1 特征提取方法分别使用 KNN 识别分类器的平均识别率 %

	方向线索特征	方向线索笔迹特征
K=1	65. 69	78. 69
K=3	66. 18	75. 19
K=5	67. 34	74. 73

应用于手机的手写体维吾尔文单字符笔迹识别通常在屏幕上显示多个候选字符供用户选择. 基于方向线素特征的 KNN 识别分类法以及基于方向线素特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别法在前 1 候选、前 3 候选以及前 5 候选的条件下的平均候选识别率如表 2 所示.

表 2 方向线素特征使用 KNN 识别分类器的候选平均识别率 %

	方向线素特征	方向线素笔迹特征
前 1 候选	65. 69	78. 69
前 3 候选	83. 23	90. 91
前 5 候选	89. 11	94. 94

根据对手写体维吾尔文字符的基于方向线素特征,使用 KNN 分类器进行识别的平均识别率和平均候选识别率的数据分析可以得到以下结果:

(1)基于方向线素特征的 KNN 识别分类算法. 识别率随着 k 取值不断增大而增大.

(2)基于方向线素特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别法,在 k 分别取 1、3、5 的时候,识别率随着 k 的增大而减小;由表 1 可知当 $k=1$ 时,识别准确率最高. 当 k 值相同时,基于方向线素特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别法的识别率要远好于基于方向线素能量特征的 KNN 识别分类法,当 $k=1$ 时,识别准确率最高达到 78. 69%.

(3)对于平均候选识别率而言,基于方向线素特征的 KNN 识别分类法与基于方向线素特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别法的识别率均随着 k 值的增大而增大. 由表 2 可知,当 k 值相同时,基于方向线素特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别法的识别率要远好于基于方向线素特征的 KNN 识别分类法,当 $k=5$ 时,识别准确率最高达到 94. 94%.

5 结束语

本文所提出的基于方向线素特征的手写体维吾尔文单字符笔迹特征的 KNN 分类识别法与基于方

向线素的 KNN 识别分类算法相比,是一种更为简单有效且识别率大幅提高的识别方法. 若要进一步提高识别率,需要在进行分类识别之后,对识别结果进行后处理.

参考文献:

[1] 任宏宇,袁保社,田园. 基于 BP#神经网络的维吾尔联机手写字符识别[J]. 微电子学与计算机,2010,27(8): 238-241.

[2] 金连文,徐秉铮. 手写体汉字识别中的一种新的特征提取方法——弹性网格方向分解特征[J]. 电路与系统学报,1997,2(3):7-12.

[3] Sanjeev R Kulkarni,Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling[J]. IEEE Transaction on Information Theory,1995,41(4):1028-1039.

[4] 陈友斌,丁晓青,吴佑寿. 一种手写汉字特征提取的新方法[J]. 信号处理,1998,14(2):1177-122.

[5] 张睿,丁晓青,方弛. 脱机手写汉字识别的最优采样特征新方法[J]. 中国图形图像学报,2002,7(2): 176-180.

[6] Nei Kato, Masato Suzuki, Shin ichiro Omachi,et al. A handwritten character recognition system using directional element features and asymmetric mahalanobis distance [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1999,21(3):258-262.

作者简介:



姜文 女,(1986-),硕士研究生. 研究方向为图像处理与分析,文字信息分析处理.

卢朝阳 男,(1963-),博士,教授,博士生导师. 研究方向为图像传输与处理、文字信息分析处理、图像分析与图像识别等.

李静 女,(1979-),博士,副教授,硕士生导师. 研究方向为文字信息分析处理、图像分析与图像识别等.