基于可训练机制的联机维吾尔手写字母识别技术研究

木塔力甫·沙塔尔[1] 李春庚[1] 艾斯卡尔·艾木都拉[2] 安居白[1]

[1](大连海事大学信息科学技术学院 辽宁 大连 116026)

[2](新疆大学信息科学与工程学院 新疆 乌鲁木齐 830046)

摘 要 在深入研究英文和汉字手写识别的基础上,结合维吾尔文字母的特点,提出一种基于支持向量机机器学习算法的维吾尔文联机手写字母识别方法,系统研究了样本采集、预处理、特征提取和分类等模块。在预处理中,为了消除干扰和噪声及比较中的相似性,采用了平滑滤波和线性归一化处理;考虑到维吾尔文相似字母较多,为了有效提取特征,将结构特征和统计特征相结合,提取了字符的梯度方向特征;分类器采用支持向量机。实验表明,随着训练样本的增加,识别率可以从 90.62% 提高到 96.09%。

关键词 支持向量机 联机手写识别 维吾尔文字母 归一化 特征提取

中图分类号 TP391.43 **文献标识码** A

ON TECHNOLOGY OF TRAINABLE MECHANISM BASED ONLINE UYGHUR HANDWRITTEN LETTER RECOGNITION

Mutallip Sattar[1] Li Chungeng[1] Askar Hamdulla[2] An Jubai[1]

[1](*School of Information Science and Technology, Dalian Maritime University, Dalian 116026, Liaoning, China*)

[2](*College of Information Science and Engineering, Xinjiang University, Urumqi 830046, Xinjiang, China*)

Abstract In this paper, through in-depth study on the state-of-the-art of on-line Chinese and English handwriting recognition technology, and in combination with the unique shape of Uyghur letters, the authors proposed a method for online Uyghur letter recognition based on support vector machines, and researched systematically its modules including data sampling, pre-processing, feature extraction and classifier, etc. In the pre-processing, in order to eliminate the noise and the similarity in comparison, the authors use smoothing filtration and linear normalisation. Taking into account that there are more similar letters in Uyghur alphabet, the authors use the method of gradient directional feature of the character for effective feature extraction, which combines the structural features and statistical features together. The classifier uses support vector machines, a branch of statistical learning theory. Experimental results show, with the increase of the training data, the recognition rate increases from 90.62% to 96.09%.

Keywords Support vector machines Online handwriting recognition Uyghur characters Normalization Feature extraction

0 引 言

随着智能手机、掌上电脑、个人数字助理(PDA)等便携式移动计算设备的普及,由于磁性笔简洁,输入舒适,联机手写识别技术成为模式识别领域中一个"热"点研究分支。联机手写识别技术能给用户提供自然、方便的人机交互方法。联机手写识别是通过手写板等轨迹捕获设备获得手写者的书写信息,并对它进行实时地识别操作。手写者也能够很容易地发现和纠正识别错的字符。相对于脱机识别而言,联机识别的优势是在笔尖运动过程中可获取动态信息。已经有很多种中文和英文的联机手写识别产品问世,但联机维吾尔文手写识别技术还处在初步研究状态。维吾尔文是新疆维吾尔自治区广泛使用的官方语言文字,约有 900 多万人使用(2009 年总人口达 987 万),研究它的手写识别方法对促进该地区的信息交流与科技发展具有重要意义。

迄今为止,为解决联机手写字符识别问题,各国研究者提出了很多种处理方法[1,2]。一般,联机或实时字符识别包括以下几个主要步骤:手写数据的预处理,特征提取和分类等。

为了设计一个实时、可靠的识别器,我们需要速度快的特征提取和分类算法。为了使用尽可能简单的分类器来得到较理想的识别率,抽取的特征应该具有高的判别性能。对联机手写识别而言,实时捕获运动轨迹是比较容易的。Nouboud 等人通过各种措施来尝试从笔尖运动过程的动态信息中提取特征[3]。

分类器方面,从曲线匹配[4]到马尔可夫模型[5]等多种统计学方法都被应用于分类器的设计。具有计算简单、自适应、可训练等特点的人工神经网络也在脱机和联机手写识别中也得到了广泛应用[6]。这些算法的提出虽然不是针对维吾尔文,但与维吾尔文的识别有一致性。

本文介绍基于 SVM(Support Vector Machine)的维吾尔文手写字母识别系统的原理和设计。在 SVM 分类器训练(学习)时,本文选用序列最小最优化算法 SMO (Sequential Minimal Optimization Algorithm)[7]。由于基本的支持向量机只适用于两类分类,而维吾尔文手写字母识别属于多类(128 类)分类问题,需

收稿日期:2010-05-10。国家自然科学基金项目(60772118)。**木塔力甫·沙塔尔**,硕士生,主研领域:信号及图形的处理和识别。

要将 SVM 推广到解决多类分类问题。为此,本文用一对一分类法[8]来解决这个问题。

1 维吾尔文字母的特点

维吾尔语属于阿尔泰语系突厥语族西匈语支。维吾尔文以及在新疆维吾尔自治区使用的哈萨克、柯尔克孜等文种都借用了阿拉伯文和部分波斯文字母。维吾尔文由 32 个字母组成,其中有 24 个辅音字母和 8 个元音字母。32 个字母共有 126 种字符形式,另外还有一个复合字符ﻻ和一个前缀符ئ,他们又各有两种形式,这样信息处理中需要处理的维吾尔文字母的形式共有 130 种。但是,在手写字母识别中,字母“ە”的四种形式里只识别两种就可以了,因为它的独立形式和首写形式、中间形式和尾写形式一般人眼分不清楚。因此,在维吾尔文手写字母识别中实际需要处理的字符为 128 个。

维吾尔文字母与由统一大小方形字组成的汉字和由拉丁字母拼写而成的英文有着明显差异,无法直接应用汉字和英文的识别技术来识别维吾尔文。下面简单介绍维吾尔文字符的基本特点。

(1) 与拉丁文等文字相比,维吾尔文字的手写方向不是从左到右,而是从右到左。

(2) 每个字母根据在字里面的位置不同而有不同的字符形式:首尾与相邻字符都不相连的独立形式ں、尾部与下一个字符连接的首写形式ﻧ、首尾与相邻字符连接的中间形式ﻨ和首部与上一个字符连接的尾写形式ﻦ等。

(3) 不同位置的字符形式也有 0 到 2 不同,这样不同字母总的字符形式有 2、4、8 等三种。其中,5 个字母有两种形式,25 个字母有四种形式,2 个字母有八种形式。如表 1 所示。

表 1 部分维吾尔文字母

中间形式	中间形式	首写形式	独立形式
ﯘ			ﯗ
ب	ﺒ	ﺑ	ب
ﻰ ى	ﺌ ﯩ	ﯨ ﺋ	ى ﻯ

(4) 从自动识别的角度来看,很多维吾尔字母包含点和其他附加的符号,点数和点位置的不同或其他附加符号的差异也是区别主体部分相同字母的重要依据。如字母و、ۆ、ۇ、ۈ、ڎ、د可以出现如下具有附加符号的形式ٴ、ۅ、٧、١。这些附加符号也增加了字母识别的难度。

2 手写数据的预处理

联机手写字母识别的第一步工作是手写字母轨迹的采集。字符通过鼠标或磁性笔输入。手写的轨迹数据是根据笔尖在手写板上的运动轨迹按时间顺序获取,这样我们不但能获得每个点的坐标信息,而且能得到它们的时间序列、笔画数等信息。在手写维吾尔文过程中,笔尖运动轨迹由它在手写板上的 x、y 坐标和“落笔”、“抬笔”的状态来描述,将笔尖在手写板上的运动轨迹分隔为笔划序列。这样,第 i 个字符类 C_i 的第 μ 个样本的坐标序列通过下面的公式来描述:

$$C_i^{\mu} = \{(x_{i0}^{\mu},y_{i0}^{\mu}),\cdots,(x_{ik}^{\mu},y_{ik}^{\mu}),\cdots,(x_{in}^{\mu},y_{in}^{\mu})\} \quad (1)$$

在这里 $i = 1,2,\cdots,128$,也就是本文研究的 128 个字符,$(x_{i0}^{\mu},y_{i0}^{\mu})$ 和 $(x_{in}^{\mu},y_{in}^{\mu})$ 分别表示某个字符 C_i^{μ} 的第一个和最后一个点的坐标值。

人们使用手写板书写时,由于书写的随意性,人手抖动,书写的速度变化等,导致各种干扰和噪声,又因为手写过程中会产生冗余点,不能直接用于识别,为了消除这些影响,必须对采集到的信号进行平滑滤波处理。

设采样得到的数值化坐标为 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,则平滑处理计算公式为:

$$x_{pt} = ax_{p(t-1)} + (1-a)x_t \quad (2)$$

$$y_{pt} = ay_{p(t-1)} + (1-a)y_t \quad (3)$$

其中 (x_t,y_t) 是笔划平滑处理前在 t 时刻的坐标,(x_{pt},y_{pt}) 和 $(x_{p(t-1)},y_{p(t-1)})$ 分别表示在 t 时刻和 $(t-1)$ 时刻经过平滑后的坐标数据。平滑系数为 $0 \leqslant a \leqslant 1$。由平滑公式可知,平滑后各点坐标值与该点平滑前的坐标值和前一点平滑后的坐标值有关。选择不同的 a 值,可以得到不同的平滑效果。

维吾尔文字母点阵的归一化是十分重要的,因为维吾尔文字母识别主要是基于字母的图形结构而识别的,如果不能将字母点阵在位置和大小上经归一化处理一致起来,维吾尔文字母点阵的相似性比较就无法正确进行。

本文采用线性归一化方法[9,10]对维吾尔文字母进行归一化。

采集得到的原始数据坐标点经线性归一化得到新坐标点的线性变换可以表示成:

$$\begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} a_1 b_1 \\ a_2 b_2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad (4)$$

其中,系数矩阵 $\begin{pmatrix} a_1 b_1 \\ a_2 b_2 \end{pmatrix}$ 可以反映字母原来图像与线性归一化后的结果之间的变换,$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ 是变换向量,x、y 是原始图像的坐标,m、n 是归一化后图像的坐标。

线性归一化以后得到的坐标点为:

$$m = \left[x \times \frac{M}{X}\right] \quad n = \left[y \times \frac{M}{Y}\right] \quad (5)$$

其中 $X \times Y$ 表示原始图像的大小,$M \times N$ 表示归一化以后的图像大小。

可见,进行这一变换后,所有图像将具有同样的大小,这种方法的优点是简单易行,是字符识别中应用很广的一种归一化方法。

下一步对归一化后的字母进行重采样。假设对于所有字符的所有样本而言,采样率是固定的。这样字符的总点数取决于它在手写板上写的时间,所以每一个字符的点数 n 是不一样的。由于书写速度不同而不同字符实际轨迹的点数 n 大不一样,为了处理的简单,本文选用了一个比一般点数小的,并且均匀的分散在时间序列上的固定数 p。

从总点数为 n 的字符数据中选出 m 个点的公式为:

$$m = j \times (n \text{ 点之间的总距离} / p) \quad (6)$$

其中,j 的取值范围为 $0,1,2,\cdots,p$。

任何一个字符第 μ 个样本的时间序列信号的进一步处理通过一个简单的坐标序列公式来表示:

$$C^{\mu} = \{(x_0^{\mu},y_0^{\mu}),\cdots,(x_k^{\mu},y_k^{\mu}),\cdots,(x_p^{\mu},y_p^{\mu})\} \quad (7)$$

考虑到公式表示的复杂性,在式(7)中没有使用式(1)中表示某个字符 C_i 的下标 i,所以现在开始把 x_{ik}^{μ} 和 y_{ik}^{μ} 分别替换为 x_k^{μ} 和 y_k^{μ}。

3　特征提取

提取速度快、稳定性好、分类能力强的特征是模式识别的关键。特征提取算法有基于统计模式识别的特征提取法和基于结构模式识别的特征提取法。统计法适合识别有噪声的字符,抗干扰性强,但是,可以用来区分结构的敏感部位的差异也随之被淹没了,所以它不能充分地利用字符结构信息;而结构法可以利用字形的结构规律来识别,对字符变体、变形适应性好,能较好反映事物的结构特性,但是,基元的提取很不容易,对结构特征的敏感性,导致它的不稳定和抗干扰能力低。所以,把统计法和结构法两者结合起来,存优去劣,在统计法中,字符特征的选择和抽取充分考虑字形结构信息,在结构法中应用统计方法的模式分布性质,这是当前字符识别方法的主要发展方向。梯度方向特征就是这两种方法结合的产物。

所以,结合维吾尔文字母的特征,本文选用梯度方向特征[11]算法对其进行特征提取。

把每个预处理后的字母点阵图像分成 16 网格。独立计算每个网格扇区的特征。

(1)把预处理后得到的大小相同的归一化字符分成 4×4 的 16 块。

(2)对每个块中的笔迹点进行方向链表特征提取。定义东、西、南、北、东南、东北、西南、西北八个方向。如图 1 所示。

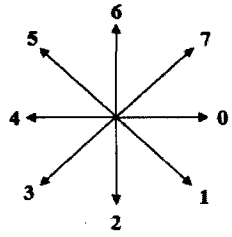

图 1　笔画走向

(3) 计算出每个块中对应每个方向的笔迹点的总数并把这个总数作为特征。这样可以得到 4×4×8 =128 维特征向量。

4　分类器

本文采用支持向量机(SVM)作为实验系统的分类器。考虑如图 2 所示的二维两类线性可分的情形,图中实心点和空心点分别表示两类训练样本,H 为把两类没有错误地分开的分类线,$H1$、$H2$ 分别为过各类样本中离分类线最近的点且平行于分类线的直线,$H1$ 和 $H2$ 之间的距离叫做两类的分类空隙或分类间隔。所谓最优分类线就是要求分类线不但能将两类无错误地分开,而且要使两类的分类空隙最大。前者是保证经验风险最小(为 0),而后者是置信风险最小,从而使真实风险最小。推广到高维空间,最优分类线就成为最优分类面。

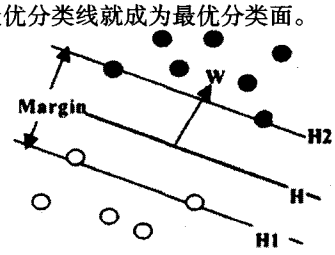

图 2　SVM 示意图

进行分类之前,SVM 需要学习(训练)。为学习选用序列最小最优化算法(SMO)[7]。SMO 将工作集的规模减到最小,一个直接的后果就是迭代次数的增加。然而序列最小最优化算法的优点在于:两个变量的最优化问题可以解析求解,因而在算法中不需要迭代地求解二次规划问题。它在每一个迭代步只选择两个变量 C_i 和 C_j 进行调整,同时固定其他变量,通过求解最优化问题,得到关于这两个变量的最优值,然后用它们来改进相应的 C 的分量。

基本的支持向量机用于两类分类,而本文研究的维吾尔文手写字母识别属于多类(128 类)分类问题,需要将 SVM 推广来解决多类分类问题。为此,本文用一对一分类法[8]来解决这个问题。

一对一分类中需要许多两类问题的分类机。对 M 类问题,就有(M-1)M/2 个两类分类机。这个数目比一对多分类法得到的分类机数目大得多。但是每一个分类问题的规模却小了许多,实际上,这里的每一个两类分类机,都比一对多中的两类分类机的规模小(有较少的支持向量),这是因为训练集规模小,并且由于类别有较小的重合,要学习的问题也比较简单。因此,该算法优点就是分类速度相比一对多分类法要快。

5　实验结果与分析

5.1　数据的采集

维吾尔文联机手写字母识别的第一步工作是手写字母图像的采集,在本文中利用手写板采集到不同手写者的 128 个字母样本图像,对手写者的手写字母顺序是特殊规定的。将字母图像信息保存为二进制格式文件中,样本文件数据结构为:(1)字符的总长度;(2)字符的内码;(3)字符的总笔画数;(4)每一个笔画的每一个点的 x、y 坐标值;(5)每一个笔画的结束标志;(6)字符的结束标志。

基于上述文件结构,在新疆大学采集了 600 多人的手写样本。其参加采集的人包括文、理、工科专业,本科生到博士生,甚至教师,可以说是千变万化的笔画都有。然后对样本数据进行挑选,得到了 530 个质量达标的样本。

5.2　实验结果

本文用的实验系统的运行测试环境为 Windows XP,主要的开发工具为 Visual C++ 6.0,部分辅助性的操作使用了 Matlab 2007。

本文实验所用的手写样本共 400 份,每份样本包含 128 个字符的数据,并用梯度方向特征提取方法得到样本的特征数据,利用 SMO 算法,通过不同数量的样本对 SVM 分类器进行学习,相应的不同数量的样本对 SVM 分类器进行测试,识别效果如表 2 所示。

表 2　识别效率比较

训练样本	测试样本 1	识别率(%)	测试样本 1	识别率(%)
128 * 50	128 * 50	89.92	128 * 1	90.62
128 * 100	128 * 100	90.80	128 * 1	92.86
128 * 200	128 * 200	91.23	128 * 1	94.53
128 * 400	训练样本自己	94.19	128 * 1	96.09

从表 2 可以看出,随着训练样本数量的增加,分类器的识别效率也提高,同时识别过程需要的时间也增加,当然训练需要的时间更多。最好的时候识别率达到 96%,这是个很理想的识别

结果,需要的识别时间为3秒左右,因为实际的识别系统是一个字符、一个字符的进行识别,如果输入的是1个字符,而不是128个字符,这样识别所用的时间会更短。

本文为了进行识别结果比较,又利用模板匹配法做了识别率测试。模板匹配是较早提出的一种分类方法。该方法的基本思想是:首先需要为每一类模式建立一个对应的模板,然后对待识别对象进行分类判决。分类判决时用待识别模式与已有的模板进行比较。

本文采用均值法来建立模板,采用最近邻分类器[12]作为分类器。均值法的基本思想是:首先利用某种特征提取算法对不同手写者的样本数据进行特征提取;然后利用所得到的特征数据,对每一类进行特征值的平均值计算。比如,第 i 个字符类 C_i 的 μ 个特征样本的平均值计算公式为:

$$T_i = \frac{1}{\mu}\sum C_i^{\mu} \tag{8}$$

在这里 $i = 1,2,\cdots,128$,也就是本文研究的128个字符,T_i 表示某个字符 C_i 的模板。

表3是利用最近邻分类器进行的分类结果,此方法用的特征数据是用SVM分类器分类时使用的特征数据,只不过是分类器不一样而已。

表3 识别效率比较

训练样本	测试样本1	识别率(%)	测试样本2	识别率(%)
128*50	128*1	45.31	128*1	57.81
128*100	128*1	42.97	128*1	50.00
128*200	128*1	44.53	128*1	57.81
128*400	128*1	40.63	128*1	60.94

表3中的测试样本1是参加过训练过程的同一组样本,而测试样本2是没参加过训练过程的一组样本。对于同样数量的训练样本而言,测试样本2的识别率总是高于测试样本1的识别率。但是随着训练样本数的增多,分类器的识别率没有表现出规律性的变化,所以很难得出一个规律性的结论。跟表2的识别率比较,最近邻分类器的识别率远远不如SVM分类器的识别率。最近邻分类器的最好识别率达到60.94%,而SVM分类器的识别率达到96.09%。

6 结 语

本文研究了基于SVM算法的联机维吾尔文手写识别方法,并给出了识别系统的整体框架。该方法在数据采集阶段中,采用自定义的数据结构和相应的文件格式来保存手写样本数据;预处理阶段中,首先对原始数据进行平滑滤波,然后利用线性归一化方法进行归一化,最后通过重采样方法压缩信息量,这样可以提高下一步的计算速度;特征提取中,选用了对于字符的扭曲和变形具有较好稳定性的、结合结构特征和统计特征的梯度方向的特征提取方法;分类过程采用支持向量机进行分类。本文通过用实际采样获取的手写样本数据对该系统进行验证,实验结果表明采用此方法能够获得较理想的结果,最高分类精度达到96.09%,最差不低于90%。最后跟模板匹配法进行比较,从比较结果可以看出,SVM分类器的识别率明显高于最近邻分类器的识别率。这些研究对于新疆维吾尔自治区的哈萨克文、柯尔克孜文等相似的文字研究也有一定的参考价值。

本文的识别方法存在的问题是,在训练样本不够多的情况下,主题部分相似的字符进行识别时出现混淆。这个问题可以从两个方面解决。(1)增加训练样本;(2)设计一个代表性更强的特征提取算法。对于训练样本的增加已经做了实验,并有了一定的效果。但使训练效率降低。这样,下一步工作将集中在针对维吾尔文的更有效的特征提取算法。

参 考 文 献

[1] C C Tappert, et al. The state of the art in on-line handwriting recognition[J]. IEEE Trans, PAMI, 1990,12(8):787-808.

[2] R Plamondon, et al. On-line handwriting recognition[J]. Encycl, Electr, Electron, Eng, 1999, 15:123-146.

[3] Nouboud F,Plamondon R. On-line recognition of handprinted characters: survey and beta tests[J]. Pattern Recognition,1990,25(9): 1031-1044.

[4] Li X,Yeung D Y. On-line handwritten alphanumeric character recognition using dominant points in strokes[J]. Pattern Recognition,1997, 30(1):31-44.

[5] Anquetil E,Lorette G. On-line cursive handwritten character recognition using hidden Markovmodels[J]. Trait. Signal,1995,12(6):575-583.

[6] Le Cun Y, et al. Handwritten digit recognition with a back-propagation network[J]. Advances in Neural Information Processing Systems, 1992,2:575-583.

[7] Platt J C. Fast training of support vector machines using sequential minimal optimization[C]. Massachusetts:MIT Press,1998,185-208.

[8] Knerr S, Personnaz L, Dreyfus G. Single-layer learning revisited: a stepwise procedure for building and training a neural network[M]// Fogelman J, et al. Neurocomputing: Algorithms, Architectures and Applications. NATO ASI. Springer-Verlag,1990.

[9] Casey R G. Moment normalization of handprinted characters[J]. IBM J. Res. Develop,1970,14:548-557.

[10] Nagy G,Tuong N,et al. Normalization techniques for handprinted numerals[J]. CACM,1970,13(8):465-481.

[11] Qing Wang, Zheru Chi, David D Feng, et al. Match Between Normalization Schemes and Feature Sets for Handwritten Chinese Character Recognition[C]//Sixth International Conference on Document Analysis and Recognition, Seattle, September 10 - 13, 2001. Washington: IEEE, C2001.

[12] 边肇祺. 模式识别[M]. 北京:清华大学出版社,2000.

(上接第17页)

[2] Zhong Hua,Zhang Xiaohua,Jiao Licheng. Digital watermarking and image authentication:algorithm and applications[M]. Xi'an:Xidian University Press,2006.

[3] 钱振兴,程义民,谢春辉,等. 基于嵌入矩阵的二值图像隐藏方法[J]. 电路与系统学报,2008,13(6):128-131.

[4] 谢建全,阳春华,谢勋,等. 一种大容量的二值图像信息隐藏算法[J]. 小型微型计算机系统,2008,29(10):1874-1877.

[5] 王国新,平西建,张涛,等. 空域LSB信息伪装及其隐写分析[J]. 计算机工程,2008,34(1):173-174,189.

[6] 杨全周,蔡晓霞,陈红. 一种基于游程编码的二值图像隐藏方案[J]. 舰船电子对抗,2008,31(2):86-88.

[7] 侯整风,高汉军. 一个新的基于多重秘密共享的图像隐藏方案[J]. 计算机应用,2008,28(4):902-905.

[8] LIBAI. Areliable(k,n) miage secretsharing scheme[C]//Dependable,Autonomic and SecureComputing,2nd IEEE.