矢量量化在 OCR 特征库压缩中的应用

陈光磊，罗林开
（厦门大学信息科学与技术学院模式识别与智能系统研究所　福建 厦门　361005）

【摘　要】：OCR(Optical Character Recognition)光学字符识别技术已被广泛应用于企业与个人的信息化处理,而随着嵌入式系统的发展,特别是中文手写识别技术的成熟,对系统容量与识别速度提出了新的要求.为了便于在资源有限的嵌入式硬件设备上实现 OCR 系统,寻求一种能保持识别率基本不变,又有较好压缩比的 OCR 特征库压缩方法是很有理论意义与商业应用价值的.本文通过对矢量量化算法作相应修改,用 C++语言实现 OCR 特征库的压缩,并在实验中取得了良好的性能.

【关键词】：矢量量化,OCR,模式识别,特征库压缩

1. OCR 的应用现状与出现的问题

OCR(Optical Character Recognition),是指通过光学技术对文字进行识别,即光学字符识别,一般指文字识别.随着计算机的处理能力的提高与嵌入式系统的发展,经过不断发展改进,特别是手写体的各种 OCR 技术的研究取得了令人瞩目的成果,人们对 OCR 产品的功能要求也从原来的单纯注重识别率,发展到对整个 OCR 系统的识别速度、用户界面的友好性、操作的简便性、产品的稳定性、体积的便携性等各方面提出更高的要求.

除了传统的以 PC 机为平台的扫描识别，现在的 OCR 技术越来越广泛地应用在了体积越来越小的嵌入式系统设备上.为了不降低字符识别率,较少去压缩用于模板匹配的字符特征库,而汉字字符集大与嵌入式系统资源的有限性在一定程序上限制了其发展的广度与速度[1].而这方面的应用还比较少,促使我们迫切需要一种能够保证识别率，又能够减少字符特征库的数据压缩技术,从而缩短汉字识别过程中的模板匹配时间,提高识别速度.矢量量化是一种压缩率较高的算法,其编、解码算法简单,易于硬件实现，已在数字视频与音频压缩及其图像压缩方面得到了广泛的应用.但还未在 OCR 领域得到广泛应用.在此本文先从 OCR 与 VQ 的工作原理中说明其对 OCR 特征库压缩的适用性.接着对传统 VQ 算法进行相应的修改，并通过实际程序与实验结果说明其性能.最后对后续工作做进一步的探讨.

2. OCR 基本原理与工作过程

OCR 的工作原理为通过扫描仪或数码相机等光学输入设备获取纸张上的文字图片信息，利用各种模式识别算法分析文字形态特征,判断出汉字的标准编码,并按通用格式存储在文本文件中。

一个 OCR 识别系统,从影像到结果输出,须经过影像输入、影像前处理、文字特征抽取、单字识别、最后经人工校正将认错的文字更正,将结果输出。如印刷体汉字识别就是将印刷在纸张上的文字通过扫描仪扫描或数码相机拍摄等光学方式输入后得到灰度图像或者二值图像，然后利用模式识别算法对文字图像进行分析,提取文字的特征,与标准文字进行匹配判别,从而达到识别文字的目的[2]。当输入文字运算完特征后,不管是用统计或结构的特征，都须有一比对数据库或特征数据库来进行模式匹配。即根据不同的特征特性,选用不同的数学距离函数,找出最相似的匹配,通常是不断试验同一个字符的不同模板来比较,识别出对应的字符。

在实际的工作中,我们通常称这个"模式"为"模板",相应的识别方法叫模板匹配.而上面的比对数据库或特征数据库简称特征库.如果把整个识别过程比做查字典,那么这个特征库就是一本字典.而对于我们的汉字来说,要形成这样一本"字典",就需要比较大的空间了.由于上面提到的汉字字符集大与嵌入式系统资源的有限性的矛盾,我们引入矢量量化的压缩方法.

3. 矢量量化和矢量量化器

3.1 矢量量化的定义与基本原理

矢量量化(VQ - Vector Quantization)是 70 年代后期发展起来的一种数据压缩技术.其理论基础[3]是香农的速率-失真理论.矢量量化可以定义为从 k 维欧几里德空间 R^k 到一个包含 N 个输出(重构)点的有限集合 C 的映射,即 $Q: R^k \rightarrow C$,其中 $C=\{Y_1, Y_2,\cdots,Y_N | Y_i \in R^k, i\in\{1,2,\cdots,N\}\}$,集合 C 称为码书,N 为码书长度.输入矢量空间 R^k 通过大小为 N 的量化器 Q 后，被分割成 N 个互不重叠的区域,称为胞腔.该映射应满足:$Q(X|X\in R^k)=Y_i^*$,其中 $X=(x_1,x_2,\cdots,x_n)$为 R^k 中的 k 维矢量，$Y_i^*\in\{Y_1,Y_2,\cdots,Y_n\}$为码书 C 中的码字并满足:

$$d(X,Y_i^*)=\min_{1\le j\le n}(d(X,Y_j)) \quad (1)$$

其中 $d(X,Y_j)$为矢量 X 与码字 Y_i 这间的失真测试,本文则采用均方误差(MSE),即欧氏距离的平方:

$$d(X,Y_j)=\sum_{j=1}^{k}(X_j-Y_{ij})^2 \quad (2)$$

对于给定训练矢量划分，最优码书的各码字为各胞腔的中心,对应于中心矢量.

其基本思想是:将若干个标量数据组构成一个矢量,然后在矢量空间给以整体量化，在量化时用输出组集合中最匹配的一组输出值(矢量)来代替一组输入采样值.从而压缩了数据而不损失太多的信息.

3.2 传统矢量量化器的算法与适用性

具体的矢量量化需要一个具体的矢量量化器算法去实现.而许多矢量量化器的码书设计主要以穷尽搜索矢量量化器的 GLA 设计算法为基础进行改进与优化.在数据压缩领域称为 LBG 算法[4].穷尽搜索矢量量化器,也称最近邻矢量量化器(nearest neighbor VQ),其原理是:计算输入矢量与所有码字之间的失真,通过比较找到失真最小的码字作为输入矢量的重构矢量.此设计问题也可以用两个优化准则[5]来描述:(1)最近邻条件(最佳划分);(2)质心条件(最佳码书,即中心矢量条件).传统 LBG 算法在每次迭代时都要对矢量库中的矢量作一次穷尽搜索.如对矢量数为 n,矢量维数为 k 的穷尽搜索矢量量化器,每输入一个矢量的算法复杂度为 O(n·k),这样随 k 和 n 的增大,算法的复杂度和计算时间都将显著增加.

在传统的预测和变换编码中，首先将信号经某种映射变换变成一个数的序列,然后对其一个一个地进行标量量化编码.而在矢量量化编码中,则是把输入数据几个一组地分成许多组,成组地量化编码,即将这些数看成一个 K 维矢量,然后以矢量为单位逐个矢量进行量化.

而对比我们的 OCR 特征库,单字对应 N 个模板,而每个模板则由 K 个特征组成，我们可以把这 K 个特征看成 VQ 中的 K 维矢量,再以此特征组成的矢量为单位,对单字的不同模板进行量化,最终达到压缩特征库的目的.

4. VQ 压缩 OCR 特征库工作原理与流程

4.1 适用于 OCR 的工作准备

VQ 压缩 OCR 特征库工作原理与上述矢量量化原理相同,只是针对 OCR 汉字识别中汉字字符集大,不但字数多,单个字的模板多且数目相关较大.故为了保证压缩后的模板库与测试

模板库的识别率基本不变,在此先采用按字分组的方法,对属同一个字的模板进行分组,再对同一个字中的各个模板特征做为输入矢量进行量化,最后将各个字的量化后的特征再加上前两个字节的字输出,实现最后的压缩特征库输出.

4.2 对矢量量化算法的相应修改

注意到本例应用的特殊性与前面的工作准备,在具体的矢量量化过程中,笔者对矢量量化器中传统的 LBG 算法做了些修改:第一,传统的 LBG 算法必须事先规定码书的大小 N,而本算法未确定码书大小,为探讨矢量量化压缩 OCR 特征库在不同量化距离下的压缩与识别效果,故不必事先规定码书的大小,码书的大小决定于最大量化距离和输入矢量的具体情况;第二,由于 LBG 算法规定的初始码书是针对待量化的同类矢量,对比本例,即同一汉字的模板集。而本算法应用的是 OCR 特征库中的全部汉字字符集.故本算法采用全局的矢量量化距离。第三,传统的 LBG 算法是在把所有与 Yj 最近的输入矢量归类,归完全部的 n 类后,然后再计算各输入的特征矢量与现有中心矢量集中最相近的中心矢量的距离平方之和,得到本次迭代的总失真;如果本次迭代的总失真与上次迭代的总失真之间的相对误差满足阀值要求时停止迭代. 由于本文的应用在于数千个汉字的数万个字符模板集,为保持后续文字识别过程得到较高的识别率,本文的矢量量化采用全局的最大量化距离,但不是对全局汉字特征库进行,而是在按字分组的基础上细化于每一个字的对应特征库中.进而利用文字特征库中同一汉字对应的特征的相似性,应用自适应的方法在输入矢量的同时不断与已有训练矢量集类进行重构中心矢量,逐步更新太到最优的中心矢量.

4.3 单个字模板集中的矢量量化算法

根据前面已对同一汉字进行了分组与相关算法修改,所以下面的工作则把同一个汉字的所有模板集看成是由各个模板顺序连接成.把单个模板中的 k 个特征(如以一个字节存储一个特征)看成一个 k 维的矢量作为输入,称为特征矢量.算法步骤如下:

Step1 预置最大量化距离 MAXDIST;

Step2 以初始输入特征矢量为初始矢量 X_1;并以此作为初始中心矢量 Y1,构造一初始类,类号为 n=1;

Step3 当所有模板的特征集都以矢量输入完毕时,算法停止;否则输入下一个特征矢量 X_i,找出其与已有类中心矢量 Y_k (k=1,2,…,n)距离最近的类号 j 和距离 MinDist,并作如下判断:

Step3.1 如果 MinDist<=MAXDIST,则把此特征矢量归入类 j,重新计算类 j 的中心矢量 Y_j,转到 Step3;

Step3.2 如果 MinDist>MAXDIST,则构造一新的类,类号为 n=n+1,其中心矢量 Y_{n+1} 为特征矢量 X_i,转到 Step3;

以下应用实验证明在本例压缩后的 OCR 特征库在汉字识别中具有较好的性能.

5. 实验仿真与结果分析

本实验使用的 OCR 模板特征库已实际应用于 OCR 名片识别系统中的简体中文汉字字符识别. 是一个由 54113 个模板组成的特征文件. 特征文件由多个模板构成,每个模板占 126 字节,依次顺序存放.单个模板格式如下:

第 0 和 1 个字节存放单个中文字的编码,3~16 字节为预留的保留位(在实际产品中如无特殊用途,则删除本段字节位),17~125 存放此模板中文字的特征,共有 110 个特征,每个特征占用一个字节.在 OCR 字符识别程序中,我们以一个测试文件与在不同失真测试下 VQ 压缩后的特征库文件进行识别,计算识别率. 识别过程采用基于最近距离的码字全搜索算法进行模板匹配. 为便于说明算法的推广能力,测试样本采用一小一大两个特征库,分别由 4592(655KB)和 8985(1106KB)个模板组成.

通过调整最大失真测度 max(d(X,Yj)),即取实际程序中的最大量化距离阀值 MAXDIST,得到不同的 VQ 压缩特征库,再通过与 OCR 识别程序,测出 VQ 压缩后的识别率,选取识别率在有效范围内的数据,根据在不同矢量量化距离下压缩 OCR 特征库后对大小两测试库的压缩比与识别率,得到如下实验结果:

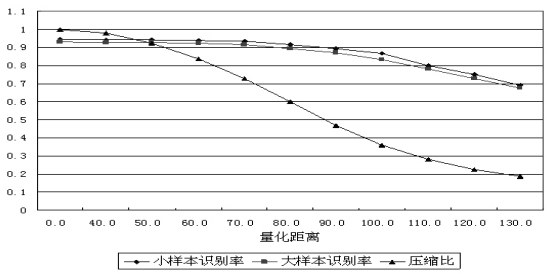

图 1 压缩的 OCR 特征库后对大小两测试库的压缩比与识别率折线图

通过以上实验及其对应图表数据可以看出,采用本矢量量化算法对 OCR 特征库压缩可以达到较好的压缩效果.特别是压缩比在 28.05%~97.84%之间时,利用压缩后的特征库对测试样本的识别率也变化不大,基本在 80.05%~94.53%之间.而通过图 2,更明显地体现了在压缩比快速变小的同时,识别率却变化不大,并基本保持在 90%以上平滑过渡.同时由两个一小一大的测试样本识别率来看,变化性质稳定,基本重合,说明本算法具有较好的推广能力.相对未压缩的特征库,在实际嵌入式应用当中,我们只要适当地选取识别率在允许范围内,特征库文件的大小又能适应特定硬件设备的压缩特征库,就可以使 OCR 识别产品得到良好性能,并在同类商品中获得市场竞争力.

6. 结论与展望

通过改进矢量量化算法并应用于适应 OCR 特征库的压缩中,实验证明能保持识别率基本不变,又有较好的压缩比.由于矢量量化算法也容易在硬件上实现. 这样压缩后的特征库有利于 OCR 系统在嵌入式硬件设备上运行,使得 OCR 识别产品得到良好性能.

参考文献:

1. 蔡韦, 徐国治. 面向嵌入式系统的 OCR 算法[J]. 计算机应用与软件, 2006,(01)

2. 岳晓峰, 焦圣喜, 韩立强, 李洪洲. 模式识别中的光字符识别技术及应用综述[J]. 河北工业科技, 2006,(05)

3. (美)托马斯,林奇(编著).吴家安,杜淑玲(译).数据压缩技术及其应用[M].北京:人民邮电出版社,1985

4. Y.Linde,A.Buzo,R.M.Gray.An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications,1980,28(1):84- 95

5. 孙圣和,陆哲明.矢量量化技术与应用[M].北京:科学出版社,2002