基于 PCA 学习子空间算法的有限汉字识别

蒋伟峰 刘济林
(浙江大学信电系,杭州 310027)

摘 要 采用 PCA 学习子空间方法来进行灰度图象上字符的识别,不仅克服了传统的基于二值化字符特征提取和识别所带来的主要困难,还尽量多地保存了字符特征.该算法在 PCA 子空间的基础上,通过反馈监督学习的方法使子空间作旋转调整,从而获得了更好的分类效果.特别当字符类别数不是很大时,子空间的训练时间也将在可接受的范围之内.应用效果也表明,采用 PCA 学习子空间算法对车牌汉字这一有限汉字集进行识别,取得了较好的效果,实用价值较高.

关键词 灰度图象的 OCR 识别 PCA 学习子空间算法 字符特征信息
中图法分类号:TP391.43 **文献标识码**:A **文章编号**:1006-8961(2001)01-0186-05

Recognition of a Limited Chinese Character Set Based on PCA Learning Subspace Algorithm

JIANG Wei-feng, LIU Ji-lin
(*Department of Information & Electronic Engineering, Zhejiang University, Hangzhou* 310027)

Abstract This paper is to realize the optical character recognition on grey-scale level by adopting learning subspace method of principal component analysis(PCALSM). Compared with Arabic number images, the resolution of Chinese character images is small, which creates great difficulty in extracting the character features. And it will get worse especially when the quality of image is low. PCALSM can overcome the main shortages of classification on binary images, and keeps integrity features of character information dramatically. On the basis of PCA subspaces, training of each subspace is rotated in different ways of the supervised-feedback learning algorithm; and better classification is therefore obtained. The time-consuming subspace training can be accepted especially when the number of character classes is not large. Our experimental results have proved that recognition of car license plate characters (a limited Chinese character set) has been improved by PCALSM, which makes it highly worth applying this optical character recognition (OCR) method.

Keywords Gray-scale character recognition, PCALSM, Image information features

0 引 言

统计模式识别的理论基础是 Bayes 决策理论,它主要依靠统计手段来实现模式分类,但这种方法要获得好的识别率,通常需采用 Parzen 窗对模式作出准确的概率分布,虽然从理论上讲,该方法对规则和不规则的、单峰或多峰的分布都可以得到准确的概率分布,但其所要求的训练样本及相应的计算量是十分大的.虽然也可以直接通过训练样本来获得判决函数,进而实现良好的分类,但若要采用这种方法对高维数据进行特征分类,往往会陷入“维数灾难”这一困境,所以在实际应用中很少采用该方法.

大家知道,子空间分类器的设计往往并不需要准确知道每个模式的先验概率,如由 Wantable 所提出的 CLAFIC (Class-Featuring Information Compression)[1]算法就是由训练样本相关矩阵的特征向量构成的,但该算法中,各模式的子空间建立彼此独立,相互之间没有任何联系,并且一旦子空间建立以后,就再不能根据分类效果作修正;针对这一缺点,Kohonen 提出了学习子空间算法(LSM),它是一种反馈监督学习算法,即通过训练样本的测试结果来对子

收稿日期:2000-01-05 改回日期:2000-05-08

空间作适当的旋转,以提高分类能力;而后,Oja 在分析 LSM 的基础上,又提出了平均学习子空间算法(ALSM)[2],由于这种算法每次学习要更新 1 ~ 2 类相关矩阵,因此所需的计算量和存储量均比较大.最近几年,人们对子空间的研究又重新重视起来,如 Parkash 和 Marky 提出了基于 Hebb 规则学习的子空间算法(HLSM)[3]和生长子空间算法(GSM)[4]等等.同传统的统计模式识别方法相比,这种子空间算法有其自身的特点:

(1) 子空间作为模式 x 的分类标准,每个子空间可代表一类模式,而且它将类定义为某些矢量的线性组合,其类间边界一般是非线性的,可直接同决策分类相挂钩;

(2) 由于采用子空间的方法可以降低特征空间的维数,因此缩短了处理时间.子空间的判决函数一般仅涉及到一些向量运算,若采用专用的向量处理机,则可达到很高的处理速度;

(3) 子空间学习算法是一种多特征的分类寻优过程,即可根据样本的训练结果适当地对子空间进行调整,因此它也是一种监督学习方法.通过这种学习反馈机制能提高子空间的识别效果.

本文利用主分量分析(Principal Component Analysis-PCA)学习子空间算法来实现汽车牌照的汉字识别,且在整个识别过程中,汉字特征的提取和分类是直接在灰度图象上进行的,它同传统的基于二值化图象上的光学字符识别(Optical Character Recognition-OCR)技术不同,它不仅避免了对一些质量比较差的图象进行二值化时,所造成的字符特征丢失,而且采用该方法,还可尽量多地保存字符的信息特征.PCA 学习子空间算法是首先通过主分量分析算法来建立各类别的子空间,然后通过学习子空间算法,根据训练样本的分类结果对 PCA 子空间作适当的调整,从而使得子空间的分类效果得到进一步的提高.

1　PCA 子空间分类器的设计

子空间和统计模式识别一样,通常需将原始样本表示成特征矢量,而子空间分类器就是从这些原始特征中提取有用的特征信息,来实现模式的分类.对于一个未标定的 n 维特征向量 x,并且 x 必定属于 K 类子空间 $\{C_s\}_{s=1}^{K}$ 中的一类(不考虑拒识情况),此时,可对子空间定义一个分类判决函数,再按下式进行模式分类:

$$\text{当 } i = \arg\min_s g_s(x, L^i) \text{ 时, } C(x) = C_i \quad (1)$$

式中,$g_s(x, L^i)(x \geq 0)$ 是子空间的判决函数,它代表向量 x 同子空间(L^i)的接近程度.很显然,分类即希望向量 x 在其所对应子空间的距离是最小的,从而获得小的分类错误.

在 PCA 子空间中,分类函数可按下式通过计算 x 在相应子空间 L^i 的投影残差距离来获得.

$$g_s(x, L^i) = \|x - M_i\|^2 - \|P_i(x - M_i)\|^2 \quad (2)$$

$$M_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j^i \quad (3)$$

$$L^i = L(u_1^{(i)}, u_2^{(i)}, \cdots, u_m^{(i)}) \quad (4)$$

$$U^{(i)} = (u_1^{(i)}, u_2^{(i)}, \cdots, u_m^{(i)}) \quad (5)$$

$$u_j^{(i)} = [u_{j_1}^{(i)}, u_{j_2}^{(i)}, \cdots, u_{j_n}^{(i)}]^{\mathrm{T}} \quad j = 1, 2, \cdots, m \quad (6)$$

$$P_i = U^{(i)}U^{(i)\mathrm{T}} \quad \text{当 } u_j^{(i)} \text{ 为正交基时} \quad (7)$$

$$P_i = U^{(i)}(U^{(i)\mathrm{T}}U^{(i)})^{-1}U^{(i)\mathrm{T}} \quad \text{当 } u_j^{(i)} \text{ 互不相交时} \quad (8)$$

式中,$\|\cdot\|$ 表示向量的 2-泛数;L^i 代表第 i 类子空间,它由 m 个不相关的向量组成;P_i 表示第 i 类子空间的投影矩阵,它是 $n \times n$ 维的正交矩阵,当 $u_j^{(i)}$ 相互不相关时,该矩阵可通过式(8)来获得,容易证明投影矩阵 P_i 是一个等幂矩阵,有 $P_i = P_i^2$ 符号 x_j^i 表示第 i 类训练样本集的第 j 个训练向量,它是将字符图象按行(列)排列而获得的;N_i 代表第 i 类的训练样本总数;M_i 是其均值向量,在子空间 L 中的向量 x 可由子空间基向量 U_k 的线性组合表示,即 $L = L(u_1, u_2, \cdots, u_m) = \{x \mid x = \sum_{k=1}^{m} y_k u_k, y_k \in R\}$;$m$ 为子空间的维数.相应地,向量 x 可通过其在子空间 L^i 的投影分量来重构,即有

$$\bar{x} = \sum_{k=1}^{m} y_k u_k = \sum_{k=1}^{m} (x^{\mathrm{T}}u_k^{(i)})u_k^{(i)} = P_i x \quad (9)$$

$$\|\bar{x}\|^2 = \sum_{k=1}^{m} \langle x, u_k^{(i)} \rangle^2 = \sum_{k=1}^{m} (x^{\mathrm{T}}u_k^{(i)})^2 = x^{\mathrm{T}}P_i x \quad (10)$$

其正交残差距离可定义为 $Dist(x, L^i) = \|x\|^2 - \|\bar{x}\|^2$,但应注意在采用 PCA 时,通常是在假定 $E(x) = 0$ 的条件下进行的,因此 x 减去类均值 M_i,即有式(2)成立.

一个完整的 PCA 子空间算法可描述为:

(1)由训练样本 $x_j^i(j = 1, 2, \cdots, N_i)$ 来计算子空间 L^i 的协方差矩阵 $R_i(i = 1, 2, \cdots, K)$

$$\boldsymbol{R}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}(\boldsymbol{x}_j^i - \boldsymbol{M}_i)(\boldsymbol{x}_j^i - \boldsymbol{M}_i)^{\mathrm{T}} \tag{11}$$

(2)采用奇异值分解(SVD)算法求 $\boldsymbol{R}_i$ 的特征向量 $\boldsymbol{u}_j^{(i)}(j=1,2,\cdots,p)$ 和特征值 $\lambda_j^{(i)}$;

(3)对特征值以降序排列($\lambda_1^{(i)} \geq \lambda_2^{(i)} \geq \cdots \geq \lambda_m^{(i)}$),按下式确定子空间的维数 m,并以相应的基来构成子空间 $L^i = L(\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}, \cdots, \boldsymbol{u}_m^{(i)})$;

$$\frac{\sum_{j=1}^{m}\lambda_j^{(i)}}{\sum_{j=1}^{p}\lambda_j^{(i)}} \geq r_1 \quad \text{且} \frac{\lambda_m^{(i)}}{\lambda_1^{(i)}} \leq r_2 \tag{12}$$

其中,r_1、r_2 均为小于 1 大于零的正数,m 为使上式成立的最小正整数.

(4)根据 $U^{(i)}$ 求各子空间的投影矩阵 $\boldsymbol{P}_i$,$\boldsymbol{P}_i = \boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)\mathrm{T}}$,其中,$\boldsymbol{U}^{(i)} = (\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}, \cdots, \boldsymbol{u}_m^{(i)})$;

(5)未标定向量 x 按式(2)来计算其在各子空间 L^i 的投影残差距离;

(6)按式(1)实现向量 x 的模式分类,即以最短投影残差距离的子空间作为其所属类别.

2 基于 PCA 学习子空间的字符识别算法

由于 PCA 学习子空间算法是在所获得的各类字符特征子空间的基础上,根据训练样本的测试结果,对两类子空间作适当的旋转,以使子空间与所属字符向量 x 在空间上更加接近,而同原先最相近的、非所属类的子空间作适当的分离,因而采用该方法有利于提高字符的识别率. 整个学习和识别算法见图 1.

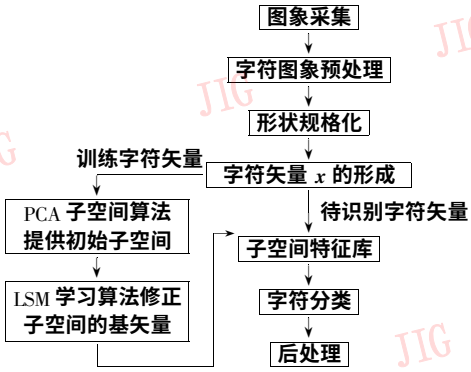

图 1 基于 PCA 学习子空间的字符识别算法流程

根据矩阵变换理论分析,学习子空间(LSM)算法所进行的子空间变换是一种线性变换. 通常在子空间模式识别中,采用以下一类特殊的矩阵迭代公式[1],即

$$\begin{cases} L_{k+1}^i = (I + \mu\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})L_k^i \\ L_k^i = L[\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}, \cdots, \boldsymbol{u}_m^{(i)}] \end{cases} \tag{13}$$

其中,L_k^i 代表通过 LSM 的 k 次迭代处理后所形成的第 i 类子空间,它由子空间的基矢量 $\boldsymbol{u}$ 形成,其中 p 代表第 i 类子空间的维数,μ 是学习系数,其值的正负取决于训练样本是否属于子空间 L_k^i.

可以证明,当 $\mu \neq (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}$ 时,通过式(13)进行子空间旋转,能保持子空间的维数不变. Oja 曾证明了当 $\mu > 0$ 时,矢量 x 其在旋转后的空间投影将获得更大的投影分量[1],即矢量 x 在空间上,同变换后的子空间更加接近,而当 $-\|\boldsymbol{x}\|^{-2} < \mu < 0$ 时,按式(13)作空间旋转所起的作用正好相反. 而当 $\mu = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}$ 时,由式(13)所形成的子空间将朝着字符矢量 x 的垂直方向旋转,但这样会造成整个子空间的变换过分依赖于当前的字符矢量 x,实践已证明这不是一个好的方法,因为这样容易使子空间的旋转过于快速,并易震荡. 比较好的方法是,对该值乘上一个小的修正系数 η,使得子空间的旋转变得缓慢一点. LSM 算法,其学习前后子空间投影矩阵 $\boldsymbol{P}_k$ 的关系可由下式表示[1]

$$\begin{aligned} \boldsymbol{P}_{k+1}^{(i)} = \boldsymbol{P}_k^{(i)} + & \frac{\mu_k^{(i)}}{1 + (\mu_k^{(i)})^2\|\boldsymbol{x}_k^{(i)}\|^2 + 2\mu_k^{(i)})\boldsymbol{x}_k^{(i)\mathrm{T}}\boldsymbol{P}_k^{(i)}\boldsymbol{x}_k^{(i)}} \\ & \times \{\boldsymbol{P}_k^{(i)}\boldsymbol{x}_k^{(i)}\boldsymbol{x}_k^{(i)\mathrm{T}} + \boldsymbol{x}_k^{(i)}\boldsymbol{x}_k^{(i)\mathrm{T}}\boldsymbol{P}_k^{(i)} - \\ & 2\boldsymbol{P}_k^{(i)}\boldsymbol{x}_k^{(i)}\boldsymbol{x}_k^{(i)\mathrm{T}}\boldsymbol{P}_k^{(i)} + \\ & \mu_k^{(i)}[(\boldsymbol{x}_k^{(i)\mathrm{T}}\boldsymbol{P}_k^{(i)}\boldsymbol{x}_k^{(i)})\boldsymbol{x}_k^{(i)}\boldsymbol{x}_k^{(i)\mathrm{T}} - \\ & \|\boldsymbol{x}_k^{(i)}\|^2\boldsymbol{P}_k^{(i)}\boldsymbol{x}_k^{(i)}\boldsymbol{x}_k^{(i)\mathrm{T}}\boldsymbol{P}_k^{(i)}]\} \end{aligned} \tag{14}$$

其中,下标 k 表示当前训练序号,上标 i 表示所对应的子空间类别.

综上所述,采用子空间学习算法,对任一训练样本 x 来讲,在其最易误判的子空间上的投影总是递减的,而在其对应的类别子空间上的投影总是增加的. 在一个训练循环中,该算法通过对所有类的样本进行训练,其中一些旋转的效果,可能会相互抵消,为了得到最佳效果,可将训练循环重复几次.

下面就 PCA 学习子空间算法作一个小结,其训练共分为如下 5 个步骤:

(1)初始化. 设样本的训练次数为 N,并令 $k=1$;

(2)按均匀分布概率的方式来输入训练样本 x,根据式(2)来计算其在各子空间 $L^i(i=1,2,\cdots,K)$ 的投影残差距离 $g_s(x, L^i)$;

(3)标记两类子空间. 以 L^o 表示 x 的所属子空间,以 L^r 表示匹敌子空间,该子空间即为除 L^o 外的在各子空间中投影残差最短的一类,即有

$r = \arg\min_s g_s(\boldsymbol{x}, L^i)$;

(4)按下式更新 L_k^o 和 L_k^r 子空间,形成新的子空间;

$$L_{k+1}^o = (I + \mu_0 \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) L_k^o, L_{k+1}^r = (I - \mu_r \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) L_k^r \quad (15)$$

参数 $\mu_0 = \eta_1/(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})$, $\mu_r = \eta_2/(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})$,其中 η_1、η_2 是大于零的修正系数.

(5) $k = k + 1$,若 $k \leq N$,则重复步骤(2)~(5);否则,结束训练.

由此可见,PCA 学习子空间算法是采用监督学习的方法对原有子空间作适当的旋转,以得到一个分类效果更好的子空间.由于 LSM 与样本输入顺序有关,若采用常规的方法训练,则其效果不好.本文采用随机选取训练样本的方法来对各子空间训练,然后根据测试结果来决定是否要对修正系数作调整,最后通过训练样本的多次学习来达到识别性能的提高.

3　实验论证

本文将 PCA 学习子空间算法应用于汽车牌照字符的识别,与其他的字符识别系统相比较,该牌照识别系统有如下特点:首先,识别的字符集比较小,即在牌照上出现的汉字字符只包括全国 30 个省、直辖市和部队、警车的简称,总计不到 60 个字符,其中最常用的约 30 个,因此这对手写字所要求的汉字集来讲,是一个非常‘有限的汉字集”,它使得采用该算法所花费的训练和识别时间处在一个可以接受的范围内,这里所指的‘有限汉字集的识别’就是针对这种情况;其次,通过摄像机所获得的字符点阵的分辨率一般不高;再次,由于汽车牌照识别系统通常在室外进行,受环境的干扰较大,因此在实际采集过程中应采取措施,尽量将这种影响减少到最小.这些特点给汽车牌照的识别既带来了方便,但也造成了识别上的困难.

从现有的汽车牌照字符库中取警、沪、吉、粤、鲁、浙等 30 类字符,共计30 000个字符作模拟测试.每一类字符各取1 000个,并分成两类,其中一类为训练类,另一类为测试类,其中,训练样本每类为 100 个,其它均为测试样本.同时采用交叉证实(Cross Validation)的方法对该算法进行论证.由于这些样本是在各汽车收费站或停车场不同时段实拍的,因此这些图象具有代表性.在图象样本中,有近 1/2 灰度图象质量是比较好的,而有 1/4 左右的图象质量是比较差的,如有在光照不均条件下获得的字符灰度图象,有的字符图象的信噪比较低,还有相当数量字符图象有稍许的几何失真(图 2).

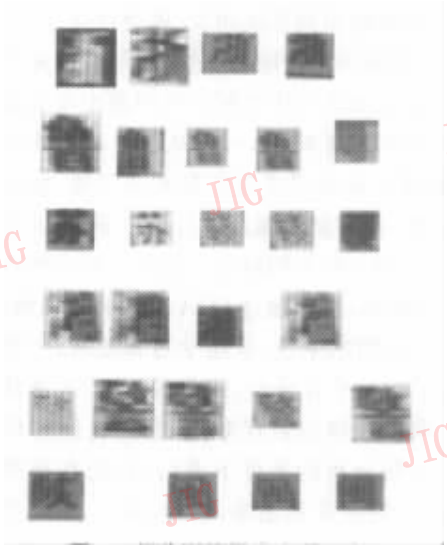

图 2　部分训练样本字符图象

下面就 PCA 学习子空间算法,对以上30 000个样本进行测试实验.在实验中,对原始样本处理后,可认为以下条件成立.

(1)图象大小经二次采样后得到一致;

(2)对字符图象作线性灰度拉伸,并将灰度限制在[0　1]之间;

(3)子空间的维数由 r_1 和 r_2 决定(见式(12)),本实验中,取 $r_1 = 0.65$, $r_2 = 0.17$.

其结果见表 1.

表 1　基于子空间算法的字符识别率

类型	30 类测试汉字的平均识别率
PCA 子空间算法	93.5%
PCA 学习子空间算法	95.32%

在每类训练样本中首先选择 50 个字符,按 PCA 子空间算法构成初始观测子空间,然后采用 LSM 算法对所有训练样本对各类子空间进行训练调整,以提高子空间分类器的泛化能力,并使得子空间在相对低维的空间中仍能获得好的分类效果.在实验过程中,需特别留意一些从传统的二值化图象上提取特征后,在识别时经常要相互误判的字,如‘浙’和“湘”;“鲁”和‘警’等字,这些字在分辨率不高的情况下,采用传统方法其误判概率是比较高的,可称这类字为‘相似字”.在实验过程中,还发现采用 PCA 学习子空间算法,这种误判情况得到了一定程度的改善,平均识别率从原来的 89.6%上升到 91%.

采用 PCA 算法的核心是 KL 变换,虽从变换本身而言,其所获得的特征并不具备有光照和位移不变性,但针对牌照识别系统而言,可通过适当的预处理,使车牌字符获得比较准确的切分,尽管这并不是

十分困难的,但要彻底解决光照识别问题也是有相当难度的.在实际识别系统中,可以从硬件和软件两方面着手,来减缓环境光照变化所带来的问题,如对图象采集系统加一些辅助光源来进行补偿等等.就识别系统而言,可选用一些不同时段字符和一些有特点的字符作为训练样本,然后通过 PCA 算法来提取特征.

从实验的结果看,采用学习算法后,字符的识别率提高了近 2 个百分点,这是采用 PCA 算法来进行特征抽取所获得的最优效果,它是在假定信号或数据处于高斯分布的条件下进行的,而在很多实际情况中,这一假定并不正确,因为,第一,数据并不符合高斯分布,而是存在着非高斯噪声干扰;第二,训练样本通常十分有限,二阶统计特征并不能完全反映数据的分布特性,在这种情况下,PCA 子空间并非是基于均方误差最优的.本文采用 PCA 学习子空间算法是对训练样本进行监督学习,并对构成的子空间作了调整,从而提高了处理效果.

关于子空间维数的确定,除了采用常用的方差贡献率 r_1 外[4],为了防止重要特征的丢失,还引入参数 r_2,以减少 r_1 确定的盲目性,从而获得了好的分类效果.

4 小 结

由于在一些 OCR 的字符处理系统中,所要处理的类别是相当有限的,如信件邮政编码的识别、机动车牌照识别等应用领域所面对的分类是相当有限的,通常只有几十类,因此这时采用子空间法来处理是很有效的,而且子空间的建立所需的计算量也不很大,因而具有很高的实用价值.

本文采用 PCA 学习子空间算法,直接在字符的灰度图象上实现了车牌汉字识别.采用该方法有以下 4 个方面的优点:

(1)整个识别对象是直接在字符的灰度图象上进行的,这样既克服了传统 OCR 识别技术在字符图象质量比较差的情况下,难以二值化所带来的技术困难,也避免了有些汉字识别技术在字符识别过程中,对字符进一步细化时,对二值化提出的更高要求.

(2)识别速度快.由于在字符识别时,只需比较字符矢量 x 在各个子空间的投影距离的大小,而字符矢量 x 对子空间作的投影运算可转化为矢量的内积运算,这样不但加快了处理速度,而且也便于硬件实现.

(3)学习子空间模式识别是一种多特征寻优模式识别技术,通常在相同方差贡献率的条件下,其选取的子空间维数会随样本数的增加而增长,例如,在方差贡献率在 0.7 的条件下,若‘浙’字的训练样本数只有 14 个,其子空间的维数通常在 5 左右,而当训练的样本数达 120 时,则其子空间的维数可达 35.当然其识别率也大为提高,它比只采用单一特征来进行模式识别的效果要好得多.

(4)从识别效果看,学习子空间算法比非学习子空间算法好,而且其模式判决面通常是非线性的.

相对而言,该算法在字符训练阶段的计算量是比较大的,而且修正系数 η_1、η_2 的大小对训练效果也有影响,通常在开始时可以选用一个比较小的正数,然后再根据识别效果作适当的调整.但是,在实际应用时,各类样本子空间还可以事前训练好,这个要求在许多应用场合并不过分.实践也表明,这种集成技术对提高 OCR 识别率是十分有效的.该技术已应用到我们开发的汽车牌照识别系统中,并取得了满意的效果.

参 考 文 献

1 Oja E. Subspace methods of pattern recognition. England: Research Studies Press, 1983.

2 Oja E. The ALSM algorithm——An improved subspace method of classification. Pattern Recognition, 1983, 16(4): 421427.

3 Prakash M, Murky M. Hebbian learning subspace method: A new approach. Pattern Recognition, 1997, 30(1): 141149.

4 Prakash M, Murky M. Growing subspace pattern recognition methods and their neural-network models. IEEE Trans. On Neural Networks, 1997, 8(1): 161168.

5 焦李成. 神经网络的应用与实现. 西安: 西安大学出版社, 1993.

蒋伟峰 1967 年生,浙江大学信电系博士生.主要从事模式识别、信号处理、神经网络等方面的研究.

刘济林 1947 年生,教授,博士生导师.主要从事计算机视觉、模式识别、并行处理等方面的研究.