第 3 卷 第 3 期
1998 年 3 月

中国图象图形学报
Journal of Image and Graphics

Vol. 3, No. 3
Mar. 1998

基于特征分布的图象信息抽取*

叶 衍 张 凌 曹明明 何永保

(复旦大学计算机科学系,上海 200433)

摘 要 为抽取实时图象中的有效信息,有别于传统的将整个特征域看做一个信源的特征筛选方法,将具有统计意义的特征区间看作一个信息源,计算其熵值,取熵值较小、且类间离散度较大、类内离散度较小的特征区间为有效特征域,每个模式都拥有自己的一组贡献值不等的有效特征域构成其专用特征空间。此算法的有效性在工业流水线上的工件识别系统中得到了较满意的验证。

关键词 模式识别,特征筛选,专用特征空间,熵值

1 引 言

对一般用于模式识别的实时图象,在众多的普遍特征中,抽取怎样的有效特征用于识别是一个重要的问题。这不仅关系到识别的精度,也直接影响识别的速度。在高维特征筛选问题上,A. K. Jain 和 W. G. Waller[1]早在 1978 年就已经证明了样本个数与最优特征维数之间的关系。当样本个数固定时,减少分类错误率的一个显而易见的办法就是增加新的独立特征,然而新增特征超过了一定的限度会导致分类器性能变坏。产生这个问题的根源在于用于设计分类器的样本数目是有限的。

一般的特征空间压缩方法,如 Fisher 方法[2],认为有效的特征表现为样本在该特征轴上投影的类间离散度较大,而类内离散度较小。用 M_i 表示 d 维 i 类样品的离散度,即 $M_i = \frac{1}{n_i}\sum_{i=1}^{n_i} X_i$,各类样本均值之差可用来度量投影点之间的类间离散度,而类内离散度则可用类内标准方差度量:$S_i^2 = \sum_{X\in\chi}(X - M_i)^2$。

一般的特征筛选方法(参见文献[3])描述了一种理想的特征分布(图 1)。但是,这样的衡量标准常会忽略一些事实上有效的特征。图 2 所示的是一种具有普遍性的特征分布,由于其总体类间离散度较小,从一般的标准看来它是无用的,然而,就模式 3 而言,可以依据此特征将模式 3 与其它特征区分开。从这个意义上说,该特征对于区分模式 3 有突出贡

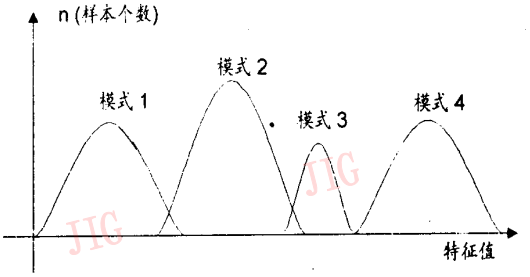

图 1 一般特征标准描述的特征分布

Fig. 1 Standard feature distribution

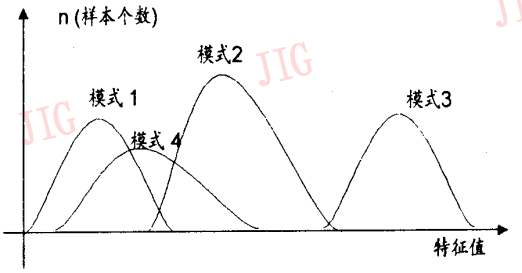

图 2 一般特征标准忽略的特征分布

Fig. 2 Common feature distribution

* 本文得到国家攀登计划和国家自然科学基金的资助
收稿日期:1996-07-03;收到修改稿日期:1997-12-09

献，不应被忽略。基于这种思想，我们提出了相对于一般通用特征空间 CFS(Common-used Feature Space)我们提出了专用特征空间 SFS(Special-used Feature Space)[4]。

通用特征空间 CFS 认为，每个可用的特征对判别每一类模式都有贡献，这样生成的分类器(判别函数或判别规则决策)对特征的适用域不加区分。图1为这些算法适用的特征域分布，在这样的分布中，各类模式的分布区域比较集中，且重叠的区域不多。

专用特征空间 SFS 认为，有必要对不同模式选用的特征集加以区分，并不要求一个可用的特征能区分所有的模式，只要它对某一类的区分特别有贡献，就将它归入该类的专用特征空间。每一类模式都享有自己的 SFS，这些 SFS 可以是相互交叉的。这样选取的特征也符合人的识别规律，即人脑识别不同的事物往往是根据各事物独特特征的认识轨迹。

2 专用特征空间的生成

2.1 符号与约定

设原始特征空间共 K 维，$F_0=\{f_1,f_2,\cdots,f_K\}$，模式类别共 M 类，记为 $C_1,C_2,\cdots,C_M$，试验样本空间记为 $S_0=\{s_1,s_2,\cdots,s_N\}$，每类模式的样本数为 n_i，则总样本数为 $N=\sum_{i=1}^{m}n_i$ 对于每类样本，生成 k_i 维的专用特征空间 T_i，$k_i\ll K$。

我们假定模式在有效的特征域中的投影总是聚合在一段特征域 $\alpha(C_i,f_l)=[\min(C_i,f_l),\max(c_i,f_l)]$ 中，式中 $\min(C_i,f_l)$，$\max(C_i,f_l)$ 分别表示模式 C_i 在特征 f_l 的分布图中投影的最小值和最大值，它们由样本数据统计得出。

2.2 特征有效性的度量

若试验样本 x 的特征值 $f_l(x)$ 落在模式 C_i 占据的特征域 α 中，则样本属于模式 C_i 的后验概率用 $P(C_i|f_l\in\alpha)$ 表示；先验概率 $P(C_i)$ 和条件概率 $P(f_l\in\alpha|C_i)$ 由试验样本统计得出；那么，根据 Bayes 法则，

先验概率：

$$P(C_i)=C_i\text{的样本总数}\ n_i/\text{总样本数}\ N \tag{1}$$

条件概率：

$$P(f_l\in\alpha|C_i)=\#\{x|x\in C_i\ \text{AND}\min(C_i,f_l)\leqslant f_l(x)\leqslant\max(C_i,f_l)\}/\#\{x|x\in C_i\} \tag{2}$$

后验概率：

$$P(C_i|f_l\in\alpha)=\frac{P(f_l\in\alpha|C_i)P(C_i)}{\sum_j P(f_l\in\alpha|C_j)P(C_j)} \tag{3}$$

式中 $\#\{\}$ 表示集合{}的元素个数。

从 N 个可能的事件中，选出其中一个事件 x 所需要的信息度量，称作信息量，用 $I(x)$ 表示：

$$I(x)=-\log P(x) \tag{4}$$

每个消息或符号的自信息量统计均值称为熵，即熵就是平均信息量，用 $H(x)$ 表示

$$H(x)=\sum_{i=1}^{n}P(x_i)I[P(x_i)]=-\sum P(x_i)\log P(x_i) \tag{5}$$

熵 $H(x)$ 有如下性质：(1)$H(x)>0$；(2)当各个 x_i 出现概率相同时，最大熵为 $H(x)=\log M$；(3)当 $P(x)=0$ 或 $P(x)=1$ 时，$H(x)=0$。

这里，定义信息函数 $I(C_i|f_l\in\alpha_i)$ 表示在 $\alpha(C_i,f_l)$ 区间内，从 M 个可能的模式中确定一个模式 C_i 所需要的信息量：

$$I(C_i|f_l\in\alpha)=-\log(P(C_i|f_l\in\alpha)) \tag{6}$$

定义区间 $\alpha(C_i,f_l)$ 的熵为：

$$H[\alpha(C_i,f_l)]=\sum_{i=1}^{M}\{P(C_i|f_l\in\alpha_l)*I[P(C_i|f_l\in\alpha_l)]\}=-\sum_{i=1}^{M}[P(C_i|f_l\in\alpha_l)*\log P(C_i|f_l\in\alpha_l)] \tag{7}$$

由熵的性质可知，$0<H[\alpha(C_i,f_l)]<\log M$。当区间 $\alpha(C_i,f_l)$ 被模式 C_i 样本独占时，$H[\alpha(C_i,f_l)]=0$，最坏情况下，在区间 $\alpha(C_i,f_l)$ 中的各模式的样本平均分布，$H[\alpha(C_i,f_l)]=\log M$。故可用 $H[\alpha(C_i,f_l)]$ 来度量区间 $\alpha(C_i,f_l)$ 对判别模式 C_i 的可靠性，$H[\alpha(C_i,f_l)]$ 越小，区间 $\alpha(C_i,f_l)]$ 越可靠。

考察特征 f_l 对判别模式 C_i 所作的贡献，还应考虑 C_i 在 f_l 域上投影值的类内离散度和类间离散度，类似 Fisher[2] 的定义 $J(W)$：

$$J(C_i,f_l)=\frac{\min\limits_{j\neq i}\{|f_l(x|x\in C_i)-f_l(x|x\in C_j)|\}}{S_i^2} \tag{8}$$

$$S_i^2=\sum_{j=1}^{n_i}(f_l(x_j)-M_i)^2 \tag{9}$$

$$M_i=\frac{1}{n_i}\sum_{j=1}^{n_i}f_l(x_j) \tag{10}$$

综上所述，定义特征 f_l 对模式 C_i 的贡献度 $F(C_i,f_l)$ 如下：

$$F(C_i,f_l)=1-\frac{H[\alpha(C_i,f_l)]}{\log M}+J(C_i,f_l) \tag{11}$$

显然有两种情况,即:当区间 $\alpha(C_i,f_l)$ 内无其它模式的样本点时,$H[\alpha(C_i,f_l)]=0$,相当可靠。

$$F(C_i,f_l)=1+J(C_i,f_l),\text{大于 1;}$$

当区间 $\alpha(C_i,f_l)$ 内有其它模式的样本点时,$J(C_i,f_l)=0$,可靠性差,

$$F(C_i,f_l)=1-\frac{H[\alpha(C_i,f_l)]}{\log M},\text{小于 1。}$$

这样,对每个模式,每个特征计算 $F(C_i,f_l)$ 的结果,生成了一个 $M\times K$ 维的矩阵,为了模糊决策的需要,对每个模式的 K 个 $F(C_i,f_l)$ 进行了规一化处理。

若用可信度规则表示则为:

IF $f_l(x)\in\alpha(C_i,f_l)$ THEN $x\in C_i$ with $cf=F(C_i,f_l)$

2.3 生成模式 C_i 的专用特征空间

首先,对 d 维原始特征空间中的每一个特征 f_l $(l\leqslant K)$ 求 $F(C_i,f_l)$;然后对第一步求得的一组 $F(C_i,f_l)(1\leqslant l\leqslant K)$ 排序,设定阈值 TH,选取对判别 C_i 贡献最大的几个特征形成模式 C_i 的专用特征空间 T_i。阀值约束为:

$$\sum_l F(C_i,f_l)\approx TH \tag{12}$$

3 本算法在工件识别系统中的实验结果

工件实时自动识别是为提高工业流水线上的自动化程度而提出的一个重要课题,此课题要解决的问题是,依据流水线上的摄像机所提供的图象信息,实时地判别流水线上当前位置零件的品种、稳态(可稳定摆放的状态)及方位,然后以识别结果确定机械手或机器人应采取的相应动作。

实验系统的构成如图 3 所示,模拟了工业流水线的情况。在实验中,我们使用 2 台摄像机分别拍摄试验工件(4 个不同的继电器零件)的俯视图和侧视图;使用一块 CA540 图象采集卡,在微机的控制下,对由摄像机获得的图象数据进行处理,并进行识别。系统运行过程中,原始图象、经过处理的图象都可以显示在图象监视器上。系统运行过程可分为:

(1) 专用特征空间生成阶段:首先分别对拍摄得到的两幅图象各抽取 56 个常规特征,如面积、欧拉数、整体凹度、方位投影宽度、中心矩函数等等。这 112 个特征构成原始的特征空间。本实验共有 11 个工件稳态,对每个稳态拍摄 8 个方位的俯视图、侧视图,得到 88 个测试样本。按照我们得出的专用特征空间生成步骤,对样本逐个进行学习,即可获得专用特征空间。实验结果,在原有的 112 个特征中,我们得到了 31 个有效特征,使特征空间维数得以压缩。

(2) 测试样本识别阶段:对单个工件随意放置,我们用由此得到的 88 个样本对专用特征空间进行测试。在筛选后的特征空间中应用最小近邻分类法,并考虑特征的贡献度,对试验样本进行识别。

实验结果:对试验样本的识别正确率为97.7%,对测试样本的识别正确率为 90.9%。如果不加选择地将 112 个常规特征直接用于最小邻近分类法,一方面维数的增加将导致处理过程中的组合爆炸,112 维特征空间使实时识别变得不现实;另一方面,维数增加超过一定限度时,将使识别性能变差。

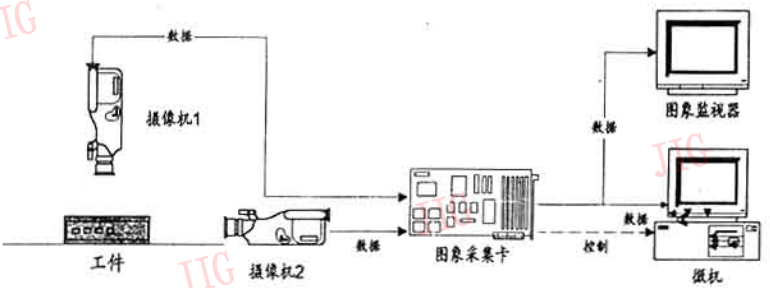

图 3　实验系统

Fig. 3　Experimental System

表1 实验结果,专用特征空间

Tabel1 Experiment result,SMFS

	特征1	特征2	特征3	特征4	特征5	特征6
稳态1	H-面积	H-M(0,0)	H-m(0,1)	H-m(0,0)	V-面积	
稳态2	H-m(0,1)	H-m(1,0)	H-270投影	V-面积	H-凹度	H-180直径
稳态3	H-180投影	H-M1	H-周长	H-M(2,0)	H-0投影	H-M2
稳态4	H-m(0,1)	H-M(0,2)	H-面积	H-M(0,0)		
稳态5	H-面积	V-M3	V-M1	H-M(2,0)	V-周长	
稳态6	V-面积	H270投影	V-M3	V-m(1,0)	V-M(0,1)	
稳态7	H-270投影	V-面积	V-M3	V-m(0,1)	H-凹度	H-圆度
稳态8	H-周长	V-270投影	V-周长	V-面积		
稳态9	H-270投影	H-M(0,1)	H-m(1,0)	H-M(0,2)	H-m(1,1)	
稳态10	V-圆度	H-周长	H-315投影	H-270投影	H-135投影	
稳态11	H-315投影	V-315周长	V-135投影	H-180投影		

说明:"H-面积"表示水平图象的区域面积;"V-周长"表示垂直图象的区域周长;"H-圆度"表示水平图象的整体圆度;"V-凹度"表示垂直图象的区域凹度;"H-M(0,1)"表示水平图象的一阶中心矩;"V-m(0,1)"表示垂直图象的一阶统计矩;"H-M1"表示水平图象的中心矩函数M1;"H-270投影"表示水平图象的270度方向投影宽度;"H-135"直径表示水平图象的135度广义方向直径;依此类推。

4 结 论

本文用信息论和概率统计的方法对图象信息的特征分布进行了剖析,认为有效的物征筛选方法应该适应于不同的模式,提出了生成专用特征空间的特征筛选算法。此算法将每个特征分布中某模式占有的一段特征域看作一个信源,计算其熵值,并结合类间离散度和类内离散度,确定该特征对判别此模式的贡献度。选取贡献值最大的数个特征构成此模式的专用特征空间。这样特征判据不仅有效地降低了维数,而且更符合模式特征的分布特性。此方法在工件识别系统中的实验结果较令人满意。

参考文献

1 Jain A K. On the Optimal Number of Features in the classification of Multivariate Gaussian Data. Pattern Recognition 1978, 10:365~374.

2 沈清,汤霖. 模式识别导论. 长沙:国防科技大学出版社. 1992:99~104.

3 Kittler J. Feature Set Search Algorithm. Ed. by Chen C H. Patern Recognition & Signal Processing. the Netherlands:SIJTHOFF & NOORDHOFF Internatinal Publishers B. V. ,Alphen aan den Rijn,Th. 1978,41~59.

4 叶衍,吴卫中,何永保. 基于多维专用特征空间的视觉模糊识别方法. 模式识别与人工智能,1996,9(3):258~263.

叶衍,女,24岁,福州市人。复旦大学计算机系计算机应用专业96届研究生。主要研究方向包括人工智能,模式识别,神经网络,模糊技术等。现已赴美国留学。

张凌,复旦大学计算机系计算机应用专业研究生。主要研究方向包括人工智能,模式识别,图象处理等。

曹明明,复旦大学计算机科学系硕士研究生,主要研究方向包括人工智能与模式识别。

何永保,复旦大学计算机系教授,主要研究方向包括人工智能,神经网络,专家系统,模糊技术,非线性预测预报。

A Hierarchical Fuzzy Recognition Algorithm Based on feature Distribution Analysis

Ye Yan, Zhang Ling, Chao Mingming, He Yongbao

(*Computer Science Department, Fudan University, Shanghai* 200433)

Abstract In this paper, a new feature selection method is presented based on analysis of the feature distribution of the images using information theory. In order to extract more useful information of the feature distributions, we calculate the entropy of the feature intervals instead of the whole fearue space, then Special-used Multi-dimension Feature Space (SMFS) is constructed, which was comprise of effective feature intervals with small entropy, small inter-pattern dispersion and large intra-pattern dispersion. Different pattern has its distinctive SMFS. The experiment of the component recognition on the assembly lines using this method shows quite satisfing results.

Keywords Pattern recognition, Feature selection, Special-used multi-dimension feeature space, Entropy value