第5卷 第5期 光 学 学 报 Vol. 5, No. 5
1985年5月 ACTA OPTICA SINICA May, 1985

用光学功率谱识别癌细胞工作中的特征选择

张少颖 郭履容
(四川大学物理系)

提 要

本文研究了利用光学功率谱识别宫颈细胞工作中的特征选择问题。从正常细胞与癌细胞的形态特征以及相应的频谱分布的规律性出发，找出了识别分类的最佳特征。当采用此单个特征时，正确识别率可达94.5%，比以前类似工作中用6个特征所达到的识别率92%[1]有所提高。若采用3个特征，则正确识别率可达95.5%。此外，本文还给出了分类正确率与特征维数的关系曲线。结果表明，当采用更多数目的特征时，反而会使分类正确率下降。

一、引 言

近年来为了发挥光学方法快速、两维并行处理的优点，以及计算机的灵活性及高精度，利用光-电混合处理的频谱分析系统对癌细胞进行识别分类的工作，已收到了明显的效果[1~3]。本文采用了类似的系统，但着重研究了有效的特征选择对识别分类结果的影响。傅京孙[4]曾指出，在识别工作中"特征数 N 要尽可能的小，特征若找得对，N 可能很小，如果不对，N 可能很大还不能识别。"本文正是从考虑识别对象的物理特点入手，分析了由于正常细胞与癌细胞形态结构的主要差异会产生相应的频谱分布差异，因而可以利用此差异形成的最佳特征进行分类。实践表明，选用了好的有效特征，还可以降低光学系统的要求，例如毋需将物体放入液门，设备可大大简化，信噪比的要求也可适当降低不至明显影响分类效果，这些对实际的识别工作都是很有利的。

此外，我们还研究了增加特征数目对分类效果的影响。

二、数据获取及特征选择

从医院检验室的涂片中，随机取了122个癌细胞，又取了150个正常细胞。我们用了其中32个癌细胞，40个正常细胞作为训练样本。其余90个癌细胞和110个正常细胞作为待识别样本。

所用的光-电混合功率谱分析系统(示意)见图1。

将经过显微摄影得到的黑白细胞底片放在输入平面上，就可利用相干光学系统获得细胞图象的功率谱。由楔环探测器的输出可得到32个环及32个楔的功率谱数据。为了消

收稿日期：1984年8月31日；收到修改稿日期：1984年12月18日

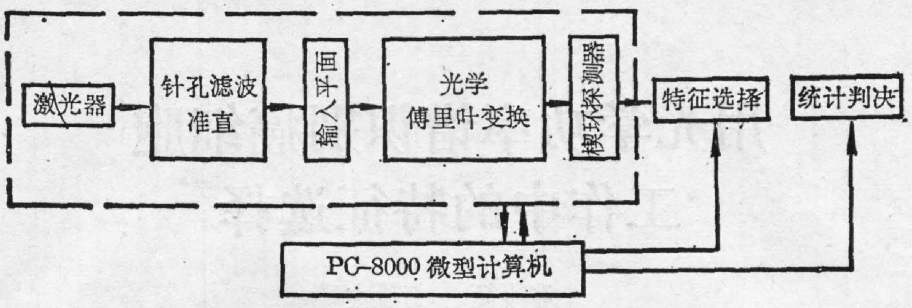

图 1 光-电混合功率谱分析系统

Fig. 1 Optical-digital hybrid system for power spectrum analysis

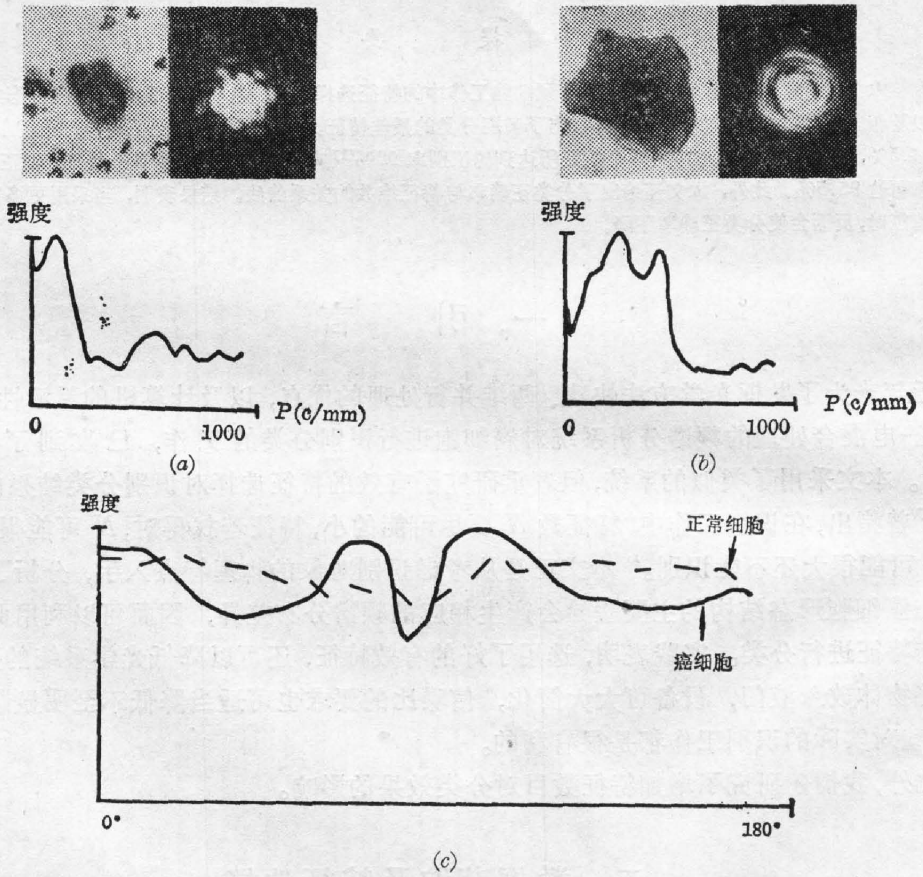

图 2 细胞、功率谱及功率谱扫描曲线

Fig. 2 Cells, power spectra and scanning curves of spectra

(*a*) Cancer cell, its spectrum and radial scanning curve of the spectrum; (*b*) Normal cell, its spectrum and radial scanning curve of the spectrum; (*c*) Angular scanning curves

除激光器输出波动以及各底片不同背景所产生的噪声，64 个数据均对总能量进行了归一化[3]，并在以此作为原始数据的基础上进行特征选择。图 2 为两类细胞的图谱照片及功率谱分布曲线。

在模式识别中，因为任何统计判决都必然以良好的特征选择为基础，所以，特征选择得当与否往往是识别成败的关键。由于识别对象的千差万别，对不同识别对象必须考虑用不

同的特征抽取方法，很难有统一的方法和理论，而且有些重要特征并不是原始测量数据的线性函数而是高阶的非线性函数，"因此在实际工作中，有效的特征主要都是通过设计者的直觉而找到的。"[5]

在宫颈细胞的识别中，无论用数字方法或光学方法，对于癌细胞常考虑以下病理特征：(1)核增大；(2) 核染色增深；(3) 核浆比倒置；(4) 核畸形；(5) 核内染色质出现粗颗粒，染色质不均匀；(6) 整个细胞呈畸形[6]。而正常细胞的核较小且呈圆形或卵圆形。以上 6 个癌细胞病理特征中有 5 个涉及核的问题，可见核在区别正常细胞和癌细胞中起着主要作用。就癌细胞而言，由于核增大，使得入射光场在核区透过底片的能量较大，衍射中心亮区的宽度变窄，又由于核染色增深，核浆比倒置，故使其能量主要集中在零频附近，因而频谱分布接近于大孔衍射。此外，核的畸形导致谱的方向性较强。就正常细胞而言，由于核较小，所以透过底片的总能量较小，衍射中心亮区的宽度较宽，频谱分布接近于小孔衍射。此外，由于核呈圆形或卵圆形，所以其谱的方向分布比较均匀，如图 2 所示。图中径向输出曲线，由于楔环探测器外环的面积比内环大，所以造成前几环的输出值较小，致使径向输出曲线往往在开始时是下凹的。通过以上统计的综合分析，我们选择了 8 个特征，列于表 1。

表 1　用于识别分类的 8 个特征

Table 1　Eight features used for classification

T_1	$\|F(\rho)\|^2$ 在某两个小频率区间内的值之差
T_2	$\rho^2\|F(\rho)\|^2$ 在 10～100 周/mm 范围内的方差
T_3	$\|F(\rho)\|^2$ 在 10～100 周/mm 范围内的方差
T_4	$\rho^2\|F(\rho)\|^2$ 在 10～100 周/mm 范围内的斜率
T_5	$\|F(\theta)\|^2$ 的方差
T_6	$\|F(\theta)\|^2$ 的最大和最小值之差
T_7	功率谱总能量
T_8	$\rho^2\|F(\rho)\|^2$ 在 10～40 周/mm 范围内的斜率

表 2　两类细胞 8 个特征的平均值、标准差以及单个特征的分类阈值和分类正确率

Table 2　Classification parameters using eight features

特　征	正常细胞特征平均值	正常细胞特征标准差	癌细胞特征平均值	癌细胞特征标，准　差	分类阈值	分类正确率 (%)
T_1	-5.76774×10^{-3}	0.0232463	0.0554079	0.0221798	0.0251	94.5
T_2	6.58567×10^{-3}	3.83045×10^{-3}	1.62542×10^{-3}	7.12316×10^{-4}	2.72×10^{-3}	88
T_3	2.00942×10^{-3}	1.65268×10^{-3}	9.74157×10^{-3}	6.91771×10^{-3}	3.40×10^{-3}	83.5
T_4	0.066455	0.0217047	0.0317898	8.66329×10^{-3}	0.04590	85
T_5	8.51153×10^{-4}	1.26181×10^{-3}	2.91732×10^{-3}	6.47517×10^{-3}	9.2×10^{-4}	62.5
T_6	0.0154458	7.79779×10^{-3}	0.0235289	0.0194568	0.02040	63.5
T_7	7926.43	4854.61	14843.2	18671.2	12540	59
T_8	0.0192025	4.29126×10^{-3}	0.0183318	5.54918×10^{-3}	0.01835	55.5

其中，$|F(\rho_i)|^2=\int_0^{\pi}d\theta\int_{\rho_i}^{\rho_i+\Delta\rho}\rho\cdot|F(\rho,\theta)|^2d\rho$，$|F(\theta_j)|^2=\int_{\theta_j}^{\theta_j+\Delta\theta}d\theta\int_0^{\rho_0}\rho\cdot|F(\rho,\theta)|^2d\rho$ 分别是第 i 环和第 j 楔的输出，ρ，θ 分别为径向(空间频率)坐标和角坐标，$F(\rho,\theta)$ 为谱面上 (ρ,θ) 处的光振幅。T_1 是本文新选择的特征，它是在低频部分某两个小频率区间上 $|F(\rho)|^2$ 的差值。这个特征综合地反映了癌细胞与正常细胞功率谱的差异，是一个关键性特征。其余特征是文献[1]、[2]中所采用过的，但本文在实验中对 T_5、T_6 是未加光阑而获取的。这两个特征反映了由于癌细胞核畸形所造成的两类细胞角向分布的差异。这 8 个特征的平均值及标准差列于表 2。

三、分类方法及结果

首先，我们对每一个特征进行分类，其阈值由训练样本决定，所得结果见表 2 中最后一列。可以看出，T_1 的分类正确率最高，这说明该特征很好地描述了两类细胞本质上的差异。然后对多个特征进行分类。为了减小误差，对所选择的特征先进行正规化处理[6]，方法如下：

$$T_{\text{正规}}^{(N)}=\frac{T^{(N)}-T_{\min}^{(N)}}{T_{\max}^{(N)}-T_{\min}^{(N)}}\quad(1\leqslant N\leqslant 8),\tag{1}$$

式中 $T_{\max}^{(N)}$，$T_{\min}^{(N)}$ 分别是训练样本中第 N 个特征的最大值和最小值。正规化后的特征所构成的每个特征矢量，$T^{t}=(T_1,\cdots,T_N)^t$，就代表 1 个细胞图象，其中 t 表示转置。为了说明不同的变换会给出类似结果，下面我们用两种线性判决函数对细胞进行分类。

1. 最小二乘最小距离判决法

这种判决方法的基本思想是找一个线性变换，使原来比较分散的各类图象样本经过变换后能够相对地集中，这样就可以在变换后的空间中用最小距离进行判决分类。

这种方法的判决函数为[6]

$$g_1(T)=a_{11}T_1+a_{12}T_2+\cdots+a_{1N}T_N+a_{1N+1},\tag{2}$$

$$g_2(T)=a_{21}T_1+a_{22}T_2+\cdots+a_{2N}T_N+a_{2N+1},\tag{3}$$

其中 $\boldsymbol{T}=(T_1,\cdots,T_N,1)^t$ 是扩充了的特征矢量，系数 $a_{ij}(1\leqslant i\leqslant 2,\ 1\leqslant j\leqslant N+1)$ 是变换矩阵 $\boldsymbol{A}=[a_{ij}]_{2,N+1}$ 的元素，变换矩阵 $\boldsymbol{A}$ 为

$$\boldsymbol{A}=\boldsymbol{S}_{VT}\cdot\boldsymbol{S}_{TT}^{-1},\tag{4}$$

式中 $\boldsymbol{S}_{VT}$ 是矢量 $\boldsymbol{V}$ 和 $\boldsymbol{T}$ 的互相关矩阵，$\boldsymbol{S}_{TT}$ 是矢量 $\boldsymbol{T}$ 的自相关矩阵，$\boldsymbol{V}^{(1)}=(1,0)^t$，$\boldsymbol{V}^{(2)}=(0,1)^t$ 是预先给定的两个 2 维矢量。由 $g_1(T)\gtrless g_2(T)$ 就可以判决 T 属于第一类或第二类。由此计算所得的最佳识别率和特征维数的关系见表 3 中的**方法 1** 和图 3 的实曲线。

表 3 两种方法正确识别率和特征维数的关系

Table 3 Correct-recognition rate versus dimension of feature vectors for the two methods

	特征维数	1	2	3	4	5	6	7	8
方法 1	最佳分类正确率(%)	94.5	95	95.5	95.5	94.5	92	92	90.5
方法 2	最佳分类正确率(%)	94.5	95	95	94.5	93.5	92.5	92	91

2. 霍特林迹判据法

这种判决方法的基本原则是寻找一个 $M\times N$ 维的变换矩阵($M<N$)，使原始 N 维特征矢量 $\boldsymbol{T}^1=(T_1, \cdots, T_N)^t$ 经变换后映射到 M 维特征空间中，同时满足两类间的离散最大，同类间的离散最小的条件。

变换矩阵 $\boldsymbol{A}$ 由矩阵 $\boldsymbol{S}_2^{-1}\cdot\boldsymbol{S}_1$ 的前 M 个本征矢量构成[3]：

$$\boldsymbol{A}=(\phi_1, \phi_2, \cdots, \phi_M)^t, \qquad (5)$$

其中矩阵 $\boldsymbol{S}_2^{-1}\boldsymbol{S}_1$ 的本征矢量 $\boldsymbol{\phi}_i(1\leqslant i\leqslant N)$ 按其本征值递降的顺序排列，即 $\lambda_1>\lambda_2>\cdots>\lambda_N$。

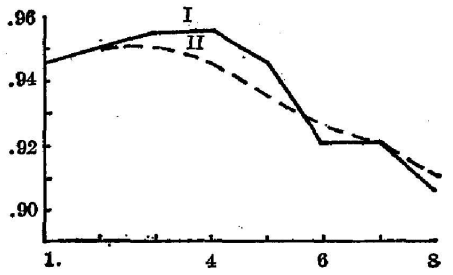

图 3　正确识别率和特征维数的关系曲线

Fig. 3　Curves of CRR versus DF for the two methods

类间散布矩阵

$$\boldsymbol{S}_1=\sum_{i=1}^{2} p_i(\mu_i-\mu_0)(\mu_i-\mu_0)^t。 \qquad (6)$$

类内散布矩阵

$$\boldsymbol{S}_2=\sum_{i=1}^{2} p_i K_i, \qquad (7)$$

$$\mu_0=p_1\cdot\mu_1+p_2\cdot\mu_2,$$

式中的 K_1, K_2, μ_1, 和 μ_2 分别为两类细胞特征矢量的协方差矩阵和平均值矢量，p_1, p_2 为两类细胞的先验概率。

从实际计算的结果看出，对应于最大本征值的本征矢量，在识别分类中起了决定性作用，因此采用了简单的阈值判决法进行识别分类。对不同特征组合进行识别时所得到的最佳分类结果见表 3 中的**方法 2** 和图 3 中的虚曲线。

四、结　束　语

1. 从表 2 和表 3 可以看出，用单个最佳特征 T_1 就能达到 94.5% 的分类正确率，而由 T_1 和其它两个特征组合后能达到 95.5%。这说明将特征选择和被识别对象的形态特征相联系，有助于选择到关键性的识别特征，从而可由较少的特征达到更佳的分类效果。

2. 从图 3 可以看出，用两种不同判决函数所获得的分类正确率与特征维数的对应关系具有相同的趋势：即当特征数目为 3～4 个时，分类正确率最高；特征数目若再增加，正确率却反而低降。当特征数达 6 个时识别率仅为 92%。这说明在模式识别中，适当的选择特征数目是十分重要的。

本工作中，王植恒同志提供了许多宝贵意见。在细胞知识的咨询和细胞底片的制备过程中，得到四川医学院刘永宗医生、陈世和医生，成都市第二妇产科医院刘丽医生，四川大学生物系研究生陈放同志的热情支持和帮助，在此一并致谢。

参　考　文　献

[1] B. Pernick and R. E. Kopp; *Applied Optics*, 1978, **17**, No. 1 (Jan), **21**.

[2] 张以谟，黄智；《光学学报》，1984, **4**, No. 4 (Apr), 368.

[3] H. Stark; *Application of Optical Fourier Transform*, (Academic Press, New York, 1982), 469; 472.
[4] 傅京孙;《自动化》, 1980, **4**, No. 1, 3.
[5] 福永圭之介;《统计图形识别导论》, (科学出版社, 1978), 316.
[6] 程民德等著;《图象识别导论》, (上海科学技术出版社, 1983), 2; 24.

Feature selection in cancer cell recognition by means of optical power spectra

ZHANG SHAOYING AND GUO LURONG
(*The Physics Department of Sichuan University, Chengdu*)

(Received 31 August 1984; revised 18 December 1984)

Abstract

In this paper we discuss the feature selection in screening cervical cells by means of optical power spectra. After analyzing morphological characteristics of normal cells and cancer cells and features of corresponding spectra, we have found the key for recognition. This feature alone can give a correct-recongition rate up to 94.5%, higher than the rate of 92% obtained before by using six features in a similar experiment[1]. We also present the relation between the correct-recognition rate and the dimension of feature space. The results indicate a fall-down of the correct-rate when the feature number is increased improperly.

更 正

本刊 **5**, No. 1, p. 8 的参考文献[5]中应加:
徐至展等;《中国科学》, 1983, (A), No. 2 (Feb), 178.

5, No. 1, p. 19 的收撑日期应为: 收稿日期

5, No. 1, p. 23 的 **Abstrac** 中应删去:
"This paper suggest a…Good results are obtained."全段

5, No. 1, 封底的 CONTENTS 中 Present of optics in China 应为:
Present state of optics in China

5, No. 2, p. 175 中 加速向高束流、高束质量发展 应为:
加速器向高束流、高束质量发展

5, No. 2, 封三表中"峰值功率"栏第 5～10 格中的数据单位"mW"应为: MW

谨向读者致谦!

作者、编者