件加上适当限制，或者寻找选取阈值的新方法。

参 考 文 献

[1] Watanabe, S. (渡边贞一) and CYBEST Group: Computer Graphics and Image Processing, Vol. 3, 350—358, 1974.

[2] Rosenfeld, A.: J. A. C. M., **17** (1), 146—160, 1970.

用阈值法对细胞数字图象进行区域划分

——食管癌细胞自动分类的研究专题之三

汤之永　杨存荣

（清华大学自动化系）

应用电子计算机自动识别细胞首先需将细胞样本片上的细胞图象（图 1 见封 2）由带有计算机控制的扫描显微分光光度计转换成一组反映各位置的透光度或吸光度（即光密度）大小的细胞灰度数字图象 f（图 2 见封 2）。我们采用的扫描步距为 4μ、2μ、1μ、0.5μ，光度分辨率为 7 bit，光波长约为 500 毫微米。为要准确地抽提细胞的特征，首先需对细胞灰度数字图象 f 进行找连通区的预处理，已如前文所述。预处理后，要对数字图象进行区域划分（又称边界检测）。划分的方法很多，本文介绍我们用“阈值法”进行初步试验的结果。

一、用三种直方图求“阈值”的方法

这里的“阈值”是指细胞与背景，胞核与胞浆的光密度分界值。

1. 频数直方图法　假设有一细胞灰度数字图象 f，按其灰度值分布情况大体可分为三个区域，即背景区 f_B、胞浆区 f_C、胞核区 f_N。各区域内的灰度值基本符合正态分布，而各区域的灰度均值（期望值）又是可分的。

现用 H_j 表示灰度等级，$j=0, 1, 2\cdots255$。

用 E_j 表示灰度值 Z_i 等于 H_j 的象点 e_i 的集合

$$E_j = \{e_i:\ Z_i = H_j\}。$$

定义：灰度值 $Z_i = H_j$ 的全部象点总个数（即集合 E_j 中全部元素 e_i 的个数）为灰度值 H_j 所出现的频数，记为 F_j。

F_j 是 H_j 的函数，作灰度频数直方图如图 3 (a)。

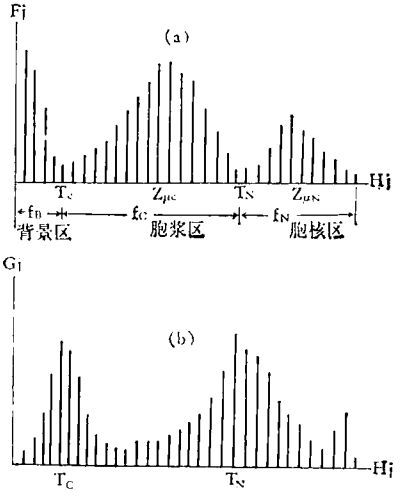

图 3　(a) 理想的灰度频数
(b) 理想的灰度差分和直方图

图中 $Z_{\mu C}$ 和 $Z_{\mu N}$ 分别相当于胞浆和胞核的灰度均值，对应直方图中两个谷的灰度值 Z_C 和 Z_N 分别选作为胞浆和胞核的“阈值”。

2. 差分和（梯度和）直方图法　假设有一细胞灰度数字图象 f，按其灰度值可分为三个区域，处于各区域内的相邻象点间的灰度差值是比较小的。而处于任一区域边界两边的相邻象点间的灰度差值则是比较大的，如图 4 所示。

现用 Z_i 表示第 i 个象点的灰度值

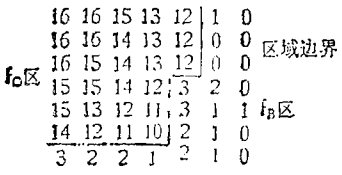

图 4

Z_i^k 表示第 i 个象点的某一邻近象点的灰度值，$k = 1, 2 \cdots 8$，如图 5 所示。

$$\begin{matrix} Z_i^1 & Z_i^2 & Z_i^3 \\ Z_i^8 & Z_i & Z_i^4 \\ Z_i^7 & Z_i^6 & Z_i^5 \end{matrix}$$

图 5

表示相邻象点间的灰度“差分”的定义有许多种。举例如下：

(1) 用 $d_i^k = (Z_i - Z_i^k)\,1[Z_i - Z_i^k]$ 表示第 i 象点的“一阶差分”。

式中 $1[Z_i - Z_i^k]$ 为单位阶跃函数，取值如下

$$1[Z_i - Z_i^k] = \begin{cases} 1 & Z_i > Z_i^k \\ 0 & Z_i \leqslant Z_i^k \end{cases}$$

(2) 可用 $S_i = \sum_{k=1,6} d_i^k$ 表示第 i 象点的“X，Y 邻域差分”。

(3) 可用 $S_i = \sum_{k=2,4,6,8} d_i^k$ 表示第 i 象点的“四点邻域差分”。

(4) 可用 $S_i = \sum_{k=1}^{8} d_i^k$ 表示第 i 象点的“八点邻域差分”，等等。

若用 P_j 表示灰度值 $Z_i = H_j$ 的象点差分和 S_i 的集合 $P_j = \{S_i : Z_i = H_j\}$。

定义：灰度值 $Z_i = H_j$ 的全部象点的差分和 S_i 的总和（即集合 P_j 中全部元素 S_i 的总和）为灰度值 H_j 的差分和，记为 G_j。

$$G_j = \sum_{S_i \in P_j} S_i$$

G_j 是 H_j 的函数，可作一灰度差分和（梯度和）直方图如图 3(b)。

图中两个峰所对应的灰度值可分别选作为胞浆和胞核的“阈值” Z_C 和 Z_N。

3. 平均差分和直方图法 从实际的临床细胞图象所作得的频数直方图和差分和直方图如图 6(a)、(b) 可看出：由于前述定义的 $G_j = \sum_{S_i \in P_j} S_i$，所以 G_j 值的大小除了与 $Z_i = H_j$ 的象点的灰度差分 S_i 的值的大小有关外，还与 $Z_i = H_j$ 的象点个数（即属于 P_j 的 S_i 的个数）即频数 F_j 有关。为了反映某一灰度值 H_j 的象点与其邻点间的灰度差的一种平均性能，现提

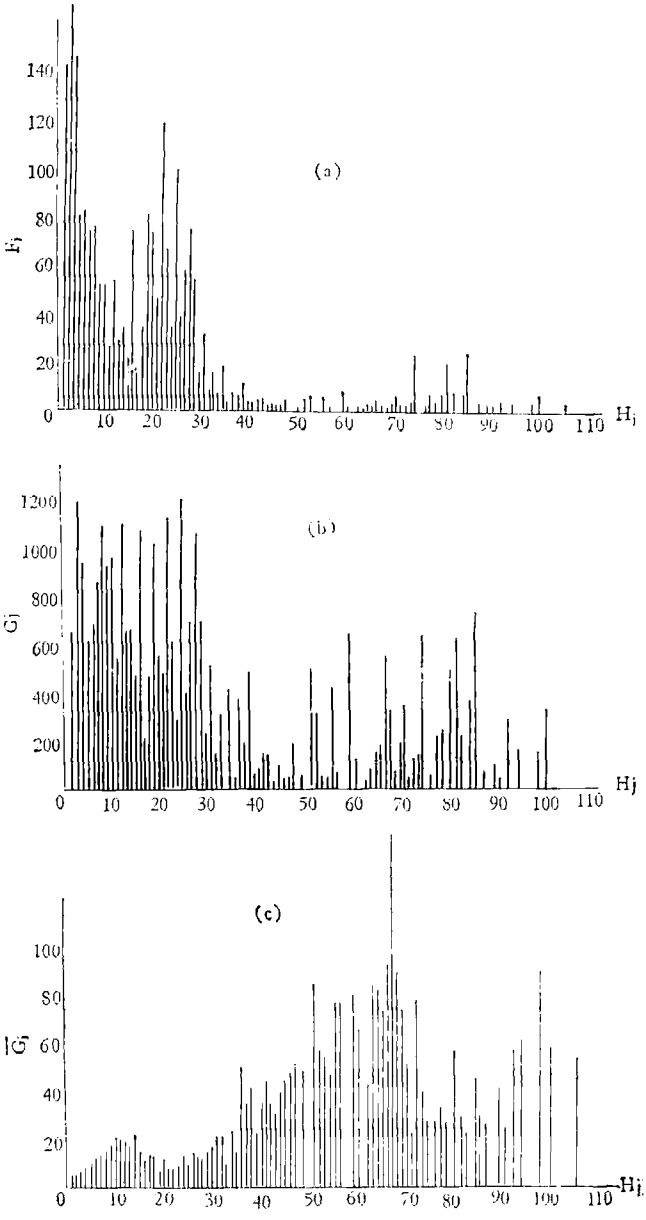

图 6 (a) 频数数直方图 (b) 差分和直方图
(c) 平均差分和直方图

出一种平均差分和 $\bar{G}_i$ 的概念。

$$\bar{G}_i = G_i / F_i。$$

$\bar{G}_i$ 也是 H_i 的函数，并可作一平均差分和直方图，如图 6(c)。

选平均差分和直方图的峰所对应的灰度值 Z_C 和 Z_N 作为胞浆和胞核的"阈值"。

二、用"阈值法"进行细胞图象区域划分的初步试验：

1. 求 Z_C 和 Z_N 阈值时，美国 TICAS 系统采用频数直方图法[1]，日本 CYBEST 系统则采用差分和直方图法[2]。鉴于美国和日本的经验，我们先采用了以下几种方法进行试验比较，以便寻找适合我国情况的方法。

求胞浆阈值 Z_C 时，是在灰度为 0—9 的范围内试验比较了以下五种方法：

方法(一)：在频数直方图上选第一个谷所对应的灰度值作为 Z_C。

方法(二)：在差分和直方图上选第一个峰所对应的灰度值作为 Z_C。

方法(三)：在差分和直方图上选最大差分和值 $(G_i)_{max}$ 所对应的灰度值作为 Z_C。

方法（四)：在平均差分和直方图上选最大平均差分和值 $(\bar{G}_i)_{max}$ 所对应的灰度值作为 Z_C。

方法(五)*：选一固定阈值 $Z_C = 4$。

2. 求胞核阈值 Z_N 时，是在灰度大于 12 的范围内试验比较了以下六种方法：

方法(一)：在差分和直方图上选最大差分和值 $(G_i)_{max}$ 所对应的灰度值作为 Z_N。

方法(二)：在平均差分和直方图上选最大平均差分和 $(\bar{G}_i)_{max}$ 所对应的灰度值作为 Z_N。

方法(三)**：由差分和直方图上挑选出一组较大的差分和值 $\{G_i\}$，并一一求其平均差分和 $\{\bar{G}_i\}$，再在这一组平均差分和值 $\{\bar{G}_i\}$ 中选最大者所对应的灰度值作为 Z_N。

方法(四)**：由平均差分和直方图上挑选出一组较大的平均差分和值 $\{\bar{G}_i\}$，并一一求其所对应灰度值的总差分和值 $\{G_i\}$，再在这一组差分和值 $\{G_i\}$ 中选最大者所对应的灰度值作为 Z_N。

方法（五)***：对差分和直方图进行平滑处理，然后挑选出若干个局部的峰，并分别对这些局部的峰求其平均差分和，再从这些平均差分和中选最大者所对应的灰度值作为 Z_N。

方法（六)***：对频数直方图进行平滑处理，然后选最低频数（即谷值）所对应的灰度值作为 Z_N。

三、各种方法的试验结果：

按上述各种方法选出的阈值 T_C 和 T_N，对细胞数字图象进行划分所得的细胞和胞核面积

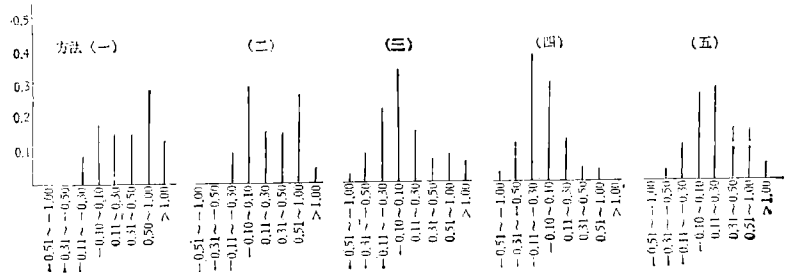

图 7　细胞面积误差 ε_{SC} 分布图（4μ）

S_C 和 S_N，与实际面积（真值)**** S_{C0} 和 S_{N0} 比较，其相对误差用

$$\varepsilon_{S_C} = \frac{S_C - S_{C0}}{S_{C0}} \% \quad 和 \quad \varepsilon_{SN} = \frac{S_N - S_{N0}}{S_{N0}} \%$$

表示，各种方法的上述相对误差分布情况示于图 7—10。若误差的期望值越接近零，方差越

* 固定阈值 $Z_C = 4$ 是根据 78 年作的 100 多个数字图象和细胞照片相比较后得出的统计结果。

** 方法(三)和(四)是差分和与平均差分和直方图两种方法的组合。

*** 方法(五)和(六)是对直方图先作一些平滑预处理后再进行阈值选择的方法。

**** 实际面积 S_{C0} 和 S_{N0} 是用已知面积的小方格计算放大的细胞照片的方法获得的。

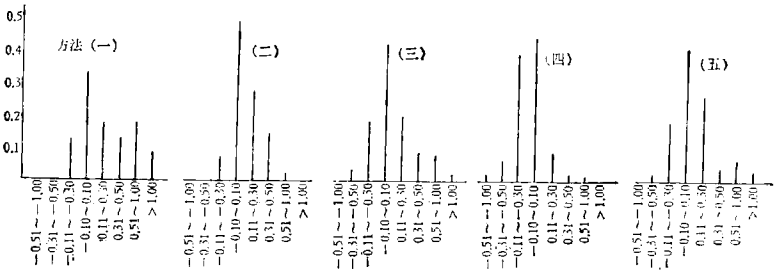

图 8　细胞面积误差 ε_{SC} 分布图（1μ）

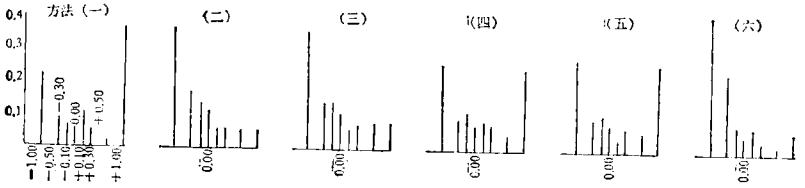

图 9　胞核面积误差 ε_{SN} 分布图（4μ）

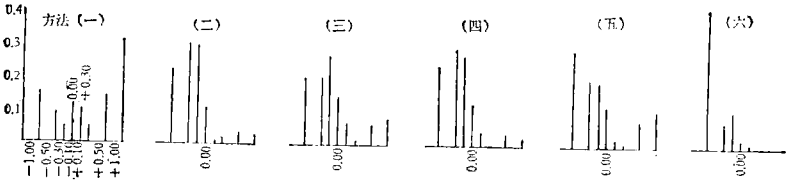

图 10　胞核面积误差 ε_{SN} 分布图（1μ）

小，则说明这种方法所选的阈值越准确。

部分比较结果列于表 1 和表 2，由表 1 可知在求 T_C 时，采用平均差分和直方图法（方法四）较好。由表 2 可知在求 T_N 时，采用平均差分和直方图法（方法(二)）略好些，组合的方法(四)和(三)也与方法(二)接近。

表 1　按细胞阈值 T_C 划分的细胞面积相对误差 $|\varepsilon_{SC}|\leqslant 50\%$ 者占百分数的比较表

年份	步距	个数	方法(一)	(二)	(三)	(四)	(五)
78	4μ	66	58%	68%	94%	98.5%	85%
	1μ	40	75%	97.5%	100%	100%	100%
79	4μ	78			80%	92%	77%
	2μ	32			94%	97%	78%
	1μ	77			87%	97%	84%
	0.5μ	13			100%	100%	100%
共计	4μ	144			86%	95%	81%
	1μ	117			92%	98%	90%

表 2　按胞核阈值 T_N 划分的胞核面积相对误差 $|\varepsilon_{SN}|\leqslant 50\%$ 者占百分数的比较表

年份	步距	个数	方法(一)	(二)	(三)	(四)	(五)	(六)
78	4μ	73	45%	48%	40%	47%	50%	40%
	1μ	45	73%	82%	84%	84%	58%	24%
79	4μ	77	34%	59%	61%	40%	32%	43%
	2μ	32	38%	69%	66%	81%	28%	25%
	1μ	77	21%	66%	56%	64%	52%	21%
	0.5μ	13	15%	92%	77%	92%	54%	0%
共计	4μ	150	39%	53%	51%	43%	41%	42%
	1μ	122	40%	72%	66%	71%	54%	22%

四、讨　论

1. 对各种不同扫描步距的灰度数字图所作的直方图作一比较，可看出：步距越小，取样点越多，直方图上零点出现得少，连续性好，阈值也易选得较为合适。反之，直方图上零点甚多，呈断续状，峰谷不明确，阈值也不易选合适。从

试验结果比较可知 0.5μ 时误差最小，4μ 时误差最大。

造成直方图稀疏的原因直接与灰度分辨率和空间分辨率的选择有关。若灰度级分得越细，而扫描步距取得越粗，则取样点越少，因而按灰度分布的直方图一定越稀疏。若把两种分辨率都选得比较高，其结果一定越准确。但这会使信息量大大增加，计算机的存贮容量和计算时间随之都大大增加。因此准确度要求不同，应合理选择不同的两种分辨率。例如日本根据粗筛机的要求选择的是 64 个灰度级，4μ 步距的粗扫描。而美国为了精确研究某些特征采用了 0.25μ 步距的精扫描。

2. 进行多区域划分时，若能先验给出各区域灰度的分布范围，从而可分别在各灰度段内再用适当的方法进行阈值选择，效果好得多。例如日本的 CYBEST 系统就曾把胞浆的阈值选择在光密度为 15—29 之间。而把胞核阈值选择限在 30—55 之间[2]。

我们根据约 120 个细胞的 300 多个数字图的统计结果，先验地把细胞阈值 Z_C 的选择限在 0—9 的范围内，效果较好。而核阈值 Z_N 的分布范围比较广，1978 年的统计主要在 20—40，而 1979 年的细胞样品又在 10—100 以上，无法确定一合适的范围，为此我们只限定 Z_N 的选择下限为 12，未限上限。看来核阈值选择总效果不如细胞阈值选择的效果好。

为要先验地给出各区域的灰度分布范围，必须要求：①细胞样品的染色与制备标准化。②灰度数值的获取规格化。目前这两方面做得还不尽符合要求，一批与另一批样品间的差别较大，难于给出一个统一的核阈值 Z_N 的选择范围。

3. 通过具体分析细胞图象和直方图，可以发现胞浆和胞核区域内部的灰度分布是不均匀的，往往分有层次，或有颗粒，反映在直方图上则是各区段内出现多峰多谷的现象。这直接影响阈值的选择。究竟如何判断哪些是有意义的峰和谷？还待进一步研究。

（四）由于上述一些原因，求核阈值尤其困难，譬如求细胞阈值 Z_C 比较好的平均差分和法用于选胞核阈值 Z_N 时，往往选到某些频数少，而差分和又大的核内颗粒点，从而将这些核内颗粒误划作胞核区，造成较大的误差。若能先验地给出被识细胞的核面积下限，则选 Z_N 时就可能有效地避免上述原因引起的误差。一般说个别颗粒或核仁的面积小于最小的核面积。

从表 3 看出限制核面积下限与不限核面积下限效果是不同的。

表 3

$\varepsilon_{SN}\leqslant\pm50\%$ 者所占百分数 步距 / 方法	4μ	2μ	1μ	0.5μ
方法(二)（未限制核面积下限）	6.5%	3.1%	36.4%	30.7%
方法(二)（限制核面积下限）	58.5%	69%	66%	92%

日本采用的是差分和直方图法。我们用这种方法进行试验时，发现决定 G_i 大小的两个因素 S_i 和 F_i 中，频数 F_i 的影响往往较大。从图 6 (a) 与 (b) 对照可看出，胞核阈值 Z_N 选在 50—70 较合适，但较大的差分和 G_i 值却出现在对应于频数较多的 H_i = 15—30 段内，从而按方法(一)所选的 Z_N = 24，由于 Z_N 选得偏小，使核面积划得偏大了。为了既考虑到区域边界处的相邻象点间的平均差分和较大，又考虑一定的象点出现概率，我们采用两种直方图组合的方法方法(三)和(四)，其效果较好。

以图 6 的细胞为例，列表 4 比较说明之。

表 4

方法	(一)	(二)	(三)	(四)	(五)	(六)
Z_N	24	67	59	59	66	104
ε_{SN}%	+230%	−25%	−14%	−14%	−22%	−100%

美国采用的是频数直方图法。我们起初看到频数直方图中零点较多（当扫描步距大时尤为突出），不易判断有效的谷，为此预先对直方图进行多次平滑处理，然后再去找最低谷，即方法(六)。现根据图 6 的细胞为例，发现最低谷往

往落到 H_i 较大的属于核内某些深灰度级上，这是因为核内个别颗粒的存在引起灰度跳变大，介于其中的许多灰度出现的频数为零，虽经平滑处理，但频数的最低谷也常落到此灰度段中。这是造成方法(六)出现较大的负的 $\varepsilon_{SN}\%$ 的主要原因。若能将方法(六)找频数谷的方法限在胞浆和胞核两大频数包峰之间，或能先验地给出核阈值选择的范围以及核面积的下限，估计会大大改善这种方法的效果。

由于以上各种因素，细胞图象其各区域的灰度分布基本上是非等灰度分布，使得用"阈值法"划分各部分区域受到限制，尤其核阈值 Z_N 的选择带来较大的误差，我们第一阶段的试验结果不够理想。但阈值法的优点是比较简单，因此今后我们拟继续研究，以提高其准确性，为细胞的诊断和分类等后续环节作好准备。

参 考 文 献

[1] Wied, G. L. et al.: *Acta Cytol.*, **12**, (2) 180—204, 1968.

[2] Tanaka, N. (田中昇) et al.: *Acta. Cytol.*, **21**,(1) 85—89, 1977.

食管上皮细胞自动分类过程中的特征选择

——食管癌细胞自动分类的研究专题之四

陈传涓　楼 恒

(中国科学院生物物理研究所)

为满足计算机分析细胞的特殊要求，采用了前文介绍的细胞涂片制备方法。通过数据输入，细胞数字图象的预处理和区域划分等步骤，已将待分析的单个细胞的胞浆和胞核分别提取出来。在此基础上可以根据临床细胞学家的经验和广泛的数学模型二者尽可能多地抽取胞浆与胞核的形状与结构特征。

究竟哪些特征对鉴别诊断或判决分类有意义？如何从大量已抽取的特征中进行选择？这一特征选择问题在模式识别的概率统计方法中是一个重要组成部分，例如美国 TICAS (Taxonomic Intra-Cellular Analytic System) 系统程序包内的子程序 DSELECT 子程序即专司特征选择的功能[1,2]，日本 CYBEST (Cyto-Biological Electronic Screener by Toshiba) 系统基础研究的大量内容也涉及特征选择[3]。

本文采用了依赖于 F 统计量临界值选取的多自变量及多因变量双重筛选逐步回归方法。该方法在选择对分类判决最有效的特征的同时给出由这些特征线性组合而成的判决函数。在数据获得方式尚未完善和特征抽提不够广泛的情况下，对所用训练样品得到了较为满意的分类效果。

一、材料与方法

细胞原始数据均用 OPTON SMP 光学扫描显微镜记录，空间分辨率分别为 4.0 微米和 1.0 微米，光度分辨率为 7 比特，波长 500 毫微米。被记录的数据以纸带形式经快速穿孔输出并转入国产 013 计算机原始数据库。经过特征抽提子程序对原始数据库中各细胞数据进行信息压缩后，再存入特征数据库。本项工作所用特征数据直接引自特征数据库。

目前共使用 18 种特征，每种特征含义如下：

特征 1：NA，细胞核面积，以扫描点个数为单位。

特征 2：NA%，胞核面积在整个细胞面积内所占百分比 (NA/NA + CA)。

特征 3：NTE，核内各点光密度总和(总光

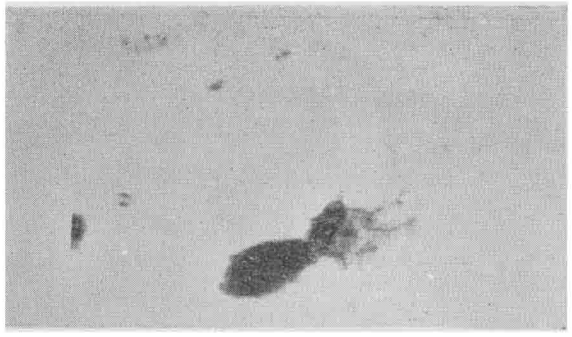

图 1　食管癌细胞的显微放大象

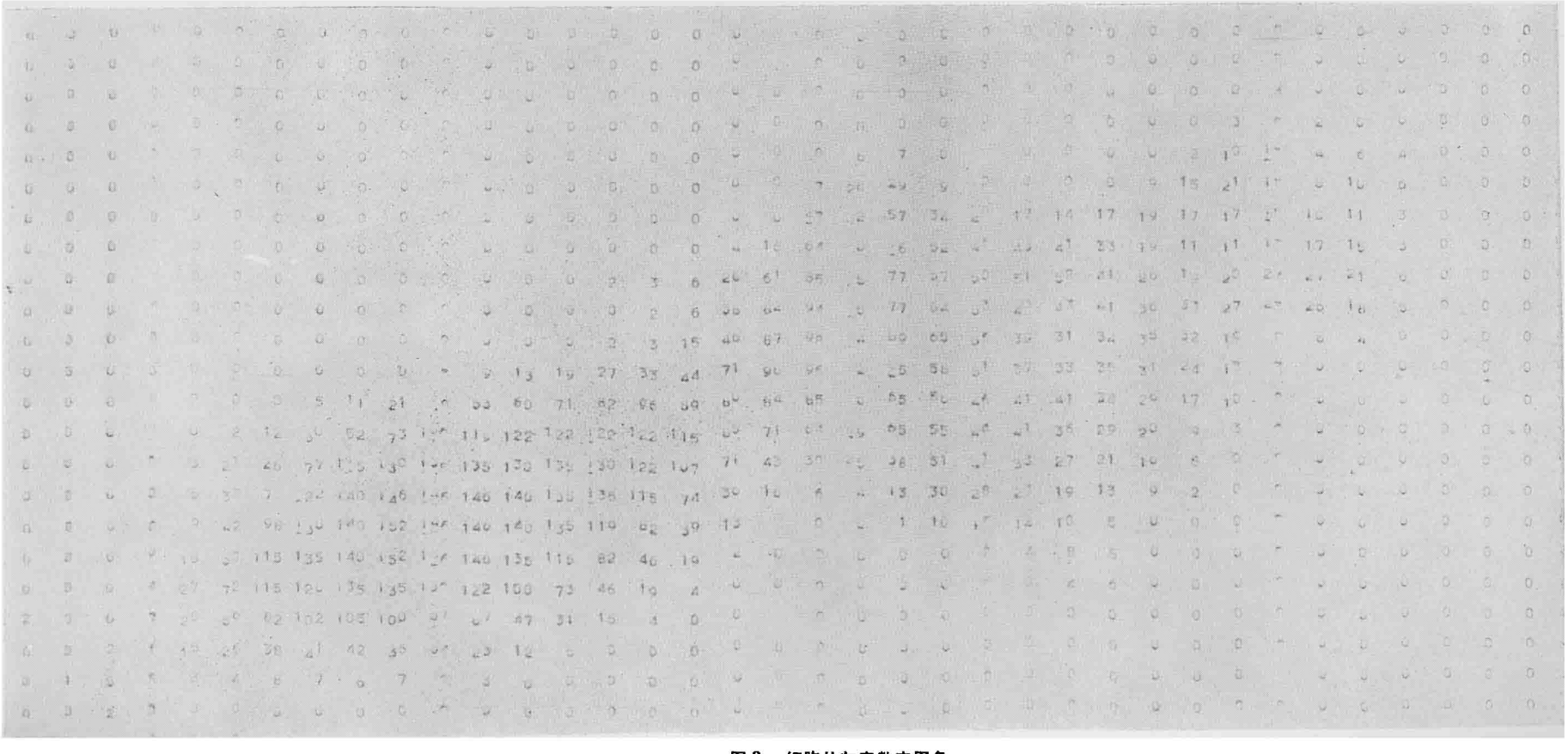

图 2　细胞的灰度数字图象