文档图像的二值化算法综述

李　倩
(中国传媒大学 广播电视数字化工程中心,北京 100024)

摘要:文档图像的二值化是光学字符识别(OCR)的基础,本文在实验的基础上通过对现有的二值化算法进行研究分析,综合比较了期望灰度法、Otsu 方法、迭代最优方法、Niblack 方法、平均梯度法和四叉树分解方法,分析了几种算法的优缺点,并对其发展趋势进行了简要的论述。

关键词:光学字符识别;二值化
中图分类号:TP391.43　**文献标识码**:A　**文章编号**:1673-4793(2008)04-0066-05

Summary of Binarization Algorithms for Document Images

LI Qian
(Engineering Center of Digital Audio & Video, Communication University of China, Beijing 100024)

Abstract: The binarization is the key problem in OCR. In this paper, the existing methods are analyzed, including expected gray value method, otsu method, recursive optimal method and division based method. A comprehensive comparision based on the experiments is carried out to show their advantages and drawbacks, the developing trends is also discussed briefly.

Key words: OCR; binarization

1　引言

作为信息的最重要载体,电子文档处理的研究引起人们极大的兴趣。在任何文档处理系统中,预处理极为重要,其效果好坏会严重地影响其它模块的工作。特别是灰度图像二值化效果的好坏,对识别效果以及其后的一切处理都有相当大的影响。原因之一是,任何物理传感输入都是灰度图像,文档处理系统的大多数模块却仅仅处理二值图像,图像二值化是必不可少的。

根据运算范围的不同,文档图像的二值化方法分为全局方法和局部方法,全局方法根据文档图像的直方图和灰度空间分布确定一个阈值,以此实现灰度文档图像到二值化图像的转化,典型的全局算法包括平均灰度法,Otsu 方法,迭代最优算法等,局部阈值通过考查每个像素点的邻域来确定阈值,比全局阈值具有更广泛的应用,常用的局部阈值方法有 Niblack 方法,Bernsen 方法,平均梯度法等。另外还有很多其他方法例如基于熵和基于模糊集合的方法等。

收稿日期:2008-03-03
作者简介:李倩(1984-),女(汉族),河南西平人,中国传媒大学硕士研究生。Email:Liqian@gmail.com。

2　二值化算法

2.1　全局阈值化方法介绍

期望灰度值法[1]:设图像的尺寸 $M \times N$,其灰度取值为 $L_1, L_2, L_3 \cdots\cdots L_N$,使用随机变量 X 表示每个像素点的灰度值。图像的灰度分布情况可以用概率分布描述,设各灰度级出现的概率分别为:

$p_1 = p(L_1), p_2 = p(L_2), \ldots, p_N = p(L_N)$ 且有

$$\sum_{n=1}^{N} p_n = 1 \tag{1}$$

灰度期望值是灰度图像的一个重要统计量,以它为阈值可以使黑像素和白像素的灰度值均等,

$$\mu_{thresshold} = \sum_{n=1}^{N} L_n p_n \tag{2}$$

该算法对于简单图像效果好,而且计算复杂度最低,但是对于亮度不均匀文档图像效果比较差。

Otsu 方法[2]:又称为最大类间方差的方法,于 1979 年提出,是一种自适应的阈值确定方法。它根据图像的灰度特性,将图像分成背景和目标两部分,背景和目标的方差越大,说明两部分的差别越大,因此类间最大方差的分割意味着错分概率最小。

对于灰度图像 $I(x,y)$,前景和背景的分割阈值记为 T,属于前景的像素点数占整个图像的比例为 ω_0,其平均灰度为 μ_0;背景像素点数占整个图像的比例为 ω_1,其平均灰度为 μ_1,图像的总平均灰度记为 μ,类间方差为 g。

假设图像 $M \times N$ 的背景较暗,图像中像素灰度小于阈值 T 的像素个数记为 N_0,像素灰度大于阈值 T 的像素个数记为 N_1,则有:

$$\omega_0 = \frac{N_0}{M \times N} \tag{3}$$

$$\omega_1 = \frac{N_1}{M \times N} \tag{4}$$

$$N_0 + N_1 = M \times N \tag{5}$$

$$\omega_0 + \omega_1 = 1 \tag{6}$$

$$\mu = \omega_0 \mu_0 + \omega_1 \mu_1 \tag{7}$$

$$g = \omega_0 (\mu_0 - \mu)^2 + \omega_1 (\mu_1 - \mu) \tag{8}$$

对于灰度图像而言,分别以每个灰度值为阈值计算对应的类间方差,使类间方差最大化的灰度值即为阈值。

该算法对于简单的文档图像具有较好的效果,而且运算速度较快,目前的应用非常广泛。

最优阈值方法[3]:又称为逼近迭代算法,这种方法的原理是将直方图用两个或多个正态分布的概率密度函数来近似的方法,阈值取为对应两个或多个正态分布的最大值之间的最小概率处的灰度值,其结果是具有最小误差的分割。这里的误差包括两部分:将目标误认为背景而被剔除,将背景、噪声归为目标。

最优阈值方法的处理步骤:

(1)计算图像最小灰度值 Z_{min} 和最大灰度值 Z_{max},令阈值初值为

$$T^0 = (Z_{min} + Z_{max})/2 \tag{9}$$

(2)根据阈值将图像分为背景和目标两部分,求出两部分的平均灰度值 Z_0 和 Z_1,计算 Z_0 和 Z_1 的平均值

$$Z_0 = \frac{\sum_{I(i,j) \le T^k} I(i,j)}{\#I(i,j) \le T^k}$$

$$Z_1 = \frac{\sum_{I(i,j) > T^k} I(i,j)}{\#I(i,j) > T^k} \tag{10}$$

(3)计算新阈值

$$T^{k+1} = (Z_0 + Z_1)/2 \tag{11}$$

如果 $T^k = T^{k+1}$ 或者达到设定的最大迭代字数就结束,否则转步骤(2)。

该算法能较好区分图像的前景和背景,但是会导致图像一些细微信息的丢失。

2.2　局部阈值方法介绍:

Niblack 方法[3]:基于局部均值和局部标准差,它的基本公式如下:

$$T(x,y) = m(x,y) + k^* s(x,y) \tag{12}$$

对于图像 $I(x,y)$,在 (x,y) 处的阈值 $T(x,y)$ 由局部均值 $m(x,y)$ 和局部标准差 $s(x,y)$ 决定,k 表示调整系数,通常设为 -0.2。在 Niblack 方法中,窗口大小的选择非常重要,既要小到能保持足够的局部细节又要大到能抑制噪声。Niblack 方法能很好地保持图像细节,对于清晰的文档图像能够提供很好的二值化结果,但是在一些退化的文档图像中会保留一些不必要的细节。在最初的 Niblack 方法中,k 值是固定的,但是对于不同的图像,通常需要根据图像的灰度分布情况自动调整 k 的值才能取得较好的结果,因此后来提出很多改进的算法,其中一个是

Zhang 和 Tan[4] 提出了一个改进的 Niblack 方法，基本公式是

$$T(x,y)=m(x,y)[1+k(1-\frac{s(x,y)}{R})] \quad (12)$$

k 和 R 都是经验常量，改进的 Niblack 方法使用 k 和 R 来减少对噪声的敏感度。

本文实现了另外一种改进的 Niblack 方法，它不仅考虑图像的局部统计特性也考虑图像的全局统计特性，具体的步骤如下：

(1)计算全局均值 m_g，全局标准差 δ_g。

(2)计算局部均值 m_l 和局部标准差 δ_l。

(3)计算 k 值

$$k=-0.3^{*}\frac{(m_g(i,j)*\sigma_g(i,j)-m_l(i,j)^{*}\sigma_l(i,j))}{max(m_g(i,j)^{*}\sigma_g(i,j),m_l(i,j)^{*}\sigma_l(i,j))} \quad (13)$$

从实验结果可以看出本方法，当窗口选择比较小的时候能够很好地保存细节，但是二值化的结果会有很多的阴影，当窗口选择比较大时，二值化的结果会比较好。

平均梯度值法[5]：Niblack 方法的一个变种，它基于局部均值和均不平均梯度。

灰度图像 $I(x,y)$ 的梯度定义为：

$$\nabla I(x,y)=\left[\frac{\delta I(x,y)}{\delta x},\frac{\delta I(x,y)}{\delta y}\right] \quad (14)$$

平均梯度算法，具体步骤如下：

(1)使用梯度算子与灰度图像卷积

$$\begin{pmatrix}-1&1&1\\-1&-2&1\\-1&1&1\end{pmatrix}$$

获得梯度图像 $G(x,y)$。

(2)以 3×3 为窗口计算平均灰度，分别对灰度图像的每个像素点进行二值化。

通过实验结果表明这种方法能很好反映文档图像中字的轮廓。

2.3 分解方法介绍：

经过实验，上述的各种二值化方法只能用于一类文档图像，对于一些退化比较严重图像，上述的方法都不能得到理想的二值化结果。

分解方法[6]：大概思想是利用四叉树将灰度图像进行递归分解，直到可以根据子图像的特征进行分类，分别对不同的子图像块进行不同的处理。

大概流程如下：

(1)输入灰度图像 $I(x,y)$ 为子图像

(2)计算灰度图像的对比度 contrast

(3)如果 contrast 小于设定阈值或者图像大小已经小于 64×64，则使用四叉树的方法进行分解，然后重复(1)，否则进行(4)

(4)根据子图像中笔划的平均梯度获取子图像中的笔划方向，再通过共生矩阵提取子图像的纹理特征判断子图像的类型，分为三种：背景，模糊子图像块和清楚子图像块，不同类型的子图像使用不同的二值化方法。

子图像分类[7]：在使用分解方法对图像进行二值化的过程中，首先要对每个子图像块进行分类，正确的分类有助于得到很好的效果，错误的分类可能会得到不正确的结果。其分类的步骤如下：

(1)计算子图像块方向，分解方法使用八个梯度算子与子图像卷积，计算子图像的平均梯度，取平均梯度最大的算子对应方向为子图像的方向，八个梯度算子如下：

$$G0(0^{\circ})=\begin{pmatrix}-1&1&1\\-1&-2&1\\-1&1&1\end{pmatrix}$$

$$G1(45^{\circ})=\begin{pmatrix}1&1&1\\-1&-2&1\\-1&1&1\end{pmatrix}$$

$$G2(90^{\circ})=\begin{pmatrix}1&1&1\\1&-2&1\\-1&-1&-1\end{pmatrix}$$

$$G3(135^{\circ})=\begin{pmatrix}1&1&1\\1&-2&-1\\1&-1&-1\end{pmatrix}$$

$$G4(180^{\circ})=\begin{pmatrix}1&1&-1\\1&-2&-1\\1&1&-1\end{pmatrix}$$

$$G5(225^{\circ})=\begin{pmatrix}1&-1&-1\\1&-2&-1\\1&1&1\end{pmatrix}$$

$$G6(270^{\circ})=\begin{pmatrix}-1&-1&-1\\1&-2&1\\1&1&1\end{pmatrix}$$

$$G7(315^{\circ})=\begin{pmatrix}-1&-1&1\\-1&-2&1\\1&1&1\end{pmatrix}$$

它们对应的方向如下:

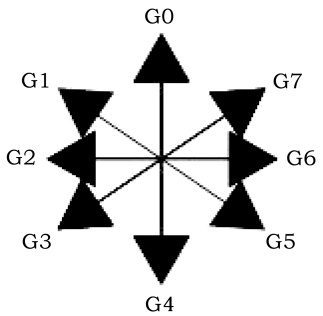

图 1　梯度算子对应的方向

(2)根据子图像块的方向计算共生矩阵 *GLCM*,根据共生矩阵计算子图像块的特征,以此作为分类的依据。

当子图像的方向为 *G*0 时,共生矩阵定义为:

$$[(i,j-d)L(i,j)L(i,j+d)]$$

当子图像的方向为 *G*1 时,共生矩阵定义为

$$\begin{bmatrix} L & L & L & L & (i-d,j+d) \\ L & L & L & L & L \\ L & L & (i,j) & L & L \\ L & L & L & L & L \\ (i+d,j-d) & L & L & L & L \end{bmatrix}$$

当子图像的方向为 *G*2 时,共生矩阵定义为

$$\begin{bmatrix} (i-d,j) \\ L \\ (i,j) \\ L \\ (i+d,j) \end{bmatrix}$$

当子图像的方向为 *G*3 时,共生矩阵定义为

$$\begin{bmatrix} (i-d,j-d) & L & L & L & L \\ L & L & L & L & L \\ L & L & (i,j) & L & L \\ L & L & L & L & L \\ L & L & L & L & (i+d,j+d) \end{bmatrix}$$

通过如上定义的偏移,计算子图像的共生矩阵 GLCM。

(3)计算子图像特征 ES(Edge Strength)进行分类,

$$ES^2 = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 \times GLCM(i,j) \qquad (14)$$

分类的标准见表 1:

表 1　子图像特征分类标准

特征 ES	子图像类型	二值化方法
ES < 0.5	背景	直接设为白色
0.5 ≤ ES < 5	模糊子图像	图像增强后,期望灰度值法
5 ≤ ES	清楚子图像	Otsu 方法

实验证明,该法在处理背景和字符清楚的简单文档图像中,计算速度慢,效果与其他方法的差别不明显,在处理严重退化的文档图像时,分解方法表现出了优越的性能。

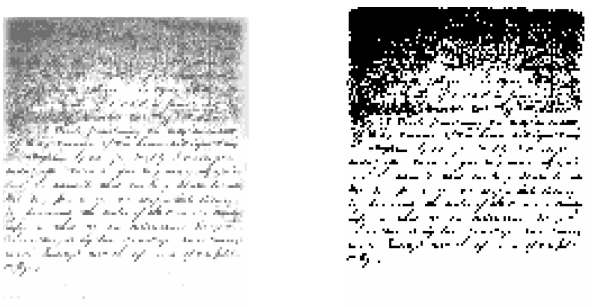

图 2　原始图像(左)期望灰度值方法二值化(右)

3　二值化实验

通过多次实验,期望灰度值法,Otsu 法和最优方法的效果差别不大,对简单文档图像处理效果很好,但对退化图像的处理效果不佳;Niblack 方法能够保持图像细节,尤其在窗口比较小的时候,这种现象更加明显;平均梯度法能够很好显示字的轮廓,如果字迹比较粗,可以比较清楚看到字的边缘;分解方法在处理简单文档图像时优势不明显,但是在处理退化文档图像时效果很好。以下各图为各个算法二

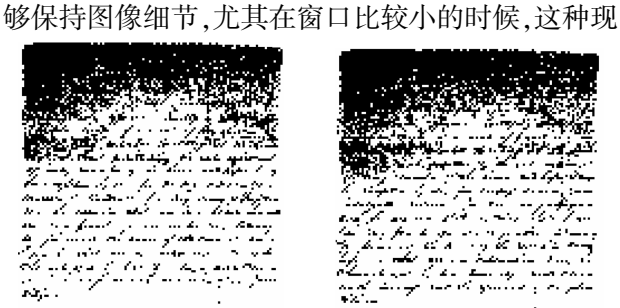

图 3　Otsu 方法二值化(左)和 Optimal 方法二值化(右)

值化效果截图：

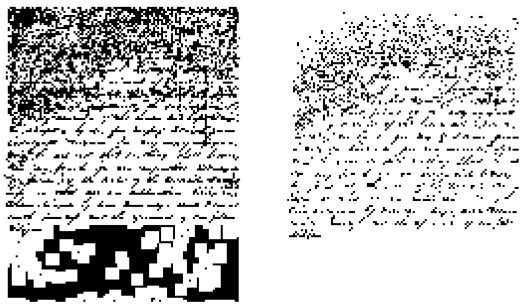

图 4 Niblack 方法二值化(左)和平均梯度方法二值化(右)

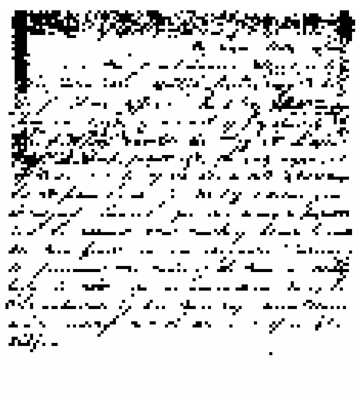

图 5 分解方法二值化

4 结论

全局阈值的方法简单,对于目标和背景明显分离,直方图分布呈双峰的图像效果良好,当文档图像包含有噪声或者光照不均匀时,全局阈值方法受到限制。局部阈值通过考查每个像素点的邻域来确定阈值,由于图像的局部性质受到光照不均匀等因素的影响较小,因此局部阈值比全局阈值具有更广泛的应用。

灰度期望值算法、Otsu 算法、最优阈值算法都适合处理简单图像,而且计算复杂度最低,但是对于退化严重的文档图像处理效果比较差。

Niblack 法、平均梯度法能很好地保持图像细节,但是在一些退化的文档图像中会保留一些不必要的细节。而且受到窗口选择比大小的限制。对于一些退化比较严重图像,同样不能得到理想的二值化结果。

分解法具有更强的适应性,在处理简单文档图像时没有明显优势,但在处理严重退化的文档图像时,分解方法表现出了优越的性能,缺点为计算速度稍慢。

文档图像的二值化是光学字符识别(OCR)的基础,虽然现在已经有很多的二值化算法,但是都只能针对某些具体的应用环境,本文实现了各个典型算法,在实现的过程中发现现有方法仍然存在很多问题,局限性强,步骤复杂,分类不精确等,所以二值化算法仍然具有很大的改进余地,这也是下一步要进行深入的工作。

参考文献

[1] 高永英,张利,吴国威. 一种基于灰度期望值的图像二值化算法[J]. 中国图像图形学报,1999,6(4):524 - 527.

[2] 齐丽娜,张博. 最大类间方差法在图像处理中的应用[J]. 无线电工程,2006(7).

[3] N B Rais. Adaptive Thresholding Technique for Document Image Analysis.

[4] W Niblack. An introduction to digital image processing[M]. Prentice Hall,1986.

[5] G Leedham. Comparison of some thresholding algorithms for text background segmentation.

[6] Y Chen, G Leedham. Decompose algorithm for thresholding degraded historical document images.

[7] S Wang. Texture Feature Extraction Using Gray Level Gradient Based Co - occurrenceMatrices.

(责任编辑:宋金宝)