文章编号:1001 - 9081(2007)04 - 1020 - 03

图像识别预处理在扫描病案自动分类中的应用

龙雅琴,古乐野,柳 岸
(中国科学院 成都计算机应用研究所,四川 成都 610041)
(longyaqin118@126.com)

摘 要:利用扫描后的病案图像的特征进行预处理,加快识别和归档的效率。首先用大津法对图像进行二值化,然后用 Radon 进行倾斜检测,最后用数学形态学开运算减少干扰,用投影法框定待识别标题位置。

关键词:二值化;Radon 变换;数学形态学开运算;投影
中图分类号: TP391.41 **文献标识码**:A

Application of image preprocessing in auto-classification of scanned hospital documents

LONG Ya-qin, GU Le-ye, LIU An
(*Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan* 610041, *China*)

Abstract: In order to make recognition and classification more effective, the scanned hospital documents were preprocessed based on their characteristics. At first, the image was binarized by using Ostu method. Then the skewness of the image was dectectd by Radon transform. Finally, mathematical morphology open operator was used to reduce the disturbances and the projecting method was used to segment the header of the image.

Key words: binarization; radon transform; mathematical morphology open operator; projection

0 引言

随着科技的发展,不少系统要求实现档案管理电子化,但对于以前的历史档案多为纸质文件,需要对其进行扫描后保存为数字图片,再分类管理。比如医院就存在大量的历史纸质病案文件,通过扫描将纸质病案文件变为数字图片文件后,由于历史病案数量大,一份病案中包含的各种单据数目和种类繁多,如果进行人工归档,比较费时费力。可以利用光学字符识别(Optical Character Recognition,OCR)技术来完成扫描病案的自动归档。为了提高效率和准确率,只需要对病案资料中的部分内容,即病案资料的页标题进行识别,而不用整页识别。为了正确的提取病案资料中的待识别信息,需要对病案图像进行预分析处理。

通过对病案资料的仔细分析发现,所有的图像均有较粗的页标题线或表格线,且要识别的部分均在图像上端的 1/4 之内,这为转正图像和快速准确的提取与归档相关的部分内容提供了十分有利的条件。

本文设计的扫描病案自动分类系统的工作流程如下:

扫描图片→图像二值化→倾斜校正→版面分析,框定识别部分→OCR→自动归档

1 扫描图像二值化

由于扫描得到的多是灰度图片,所以需要先将图片进行二值化。本文采用大津法[1]对图像进行二值化。

1.1 大津法原理

记 $f(i,j)$ 为 $M \times N$ 图像 (i,j) 点处的灰度值,灰度级为 m,不妨假设 $f(i,j)$ 取值 $[0,m-1]$。记 $P(k)$ 为灰度值为 k 的频率,则有:

$$P(k) = \frac{1}{MN}\sum_{f(i,j)=k} 1$$

假设用灰度值 t 为阈值分割出的目标与背景分别为 $\{f(i,j) \leqslant t\}$ 和 $\{f(i,j) > t\}$,于是:

目标部分比例:

$$\omega_0(t) = \sum_{q \leqslant i \leqslant t} P(i)$$

目标部分点数:

$$N_0(t) = MN\sum_{0 \leqslant i \leqslant t} P(i)$$

背景部分比例:

$$\omega_1(t) = \sum_{t < i \leqslant m-1} P(i)$$

背景部分点数:

$$N_1(t) = MN\sum_{t < i \leqslant m-1} P(i)$$

目标均值:

$$\mu_0(t) = \sum_{0 \leqslant i \leqslant t} iP(i)/\omega_0(t)$$

背景均值:

$$\mu_1(t) = \sum_{t < i \leqslant t \backslash m-1} iP(i)/\omega_1(t)$$

总均值:

$$\mu = \omega_0(t)\mu_0(t) + \omega_1(t)\mu_1(t)$$

大津方法指出求图像最佳阈值 g 的公式为:

$$g = \mathrm{Arg}\max_{0 \leqslant t \leqslant m-1}[\omega_0(t)(\mu_0(t) - \mu)^2 + \omega_1(t)(\mu_1(t) - \mu)^2]$$

该式右边括号内实际上就是类间方差值,阈值 g 分割出

收稿日期:2006 - 09 - 27;**修订日期**:2006 - 11 - 23

作者简介:龙雅琴(1981 -),女,重庆人,硕士研究生,主要研究方向:图像处理; 古乐野(1960 -),男,重庆人,研究员,博士生导师,主要研究方向:嵌入式系统; 柳岸(1982 -),男,山东人,硕士研究生,主要研究方向:图像处理.

的目标和背景两部分构成了整幅图像,而目标取值 $\mu_0(t)$,概率为 $\omega_1(t)$、背景取值 $\mu_1(t)$,概率为 $\omega_0(t)$,总均值为 μ,根据方差的定义即得该式。因为方差是灰度分布均匀性的一种度量,方差值越大,说明构成图像的两部分差别越大,当部分目标错分为背景或部分背景错分为目标都会导致两部分差别变小。因此使类间方差最大的分割意味着错分概率最小,这便是大津方法的真正含义[2]。

1.2　实际应用

由于大津法计算量大,所以采取了以下等价的公式:

$$g = \mathrm{Arg}\max_{0\leqslant t\leqslant m-1}\omega_0(t)\omega_1(t)[\mu_0(t)-\mu_1(t)]^2$$

而且为了提高效率,对图像进行 1/4 均匀采样。

2　倾斜校正

由于印刷或扫描过程的偏差,可能导致图片倾斜,这直接影响了后面的识别和浏览效果,所以需要检测倾斜角度,并转正图像。文献[3]提出了一种基于投影的倾斜检测方法,但主要针对文本图像,病案图像可能仅有少量文本,还有很多图片及书写符号,直接采用这种方法来检验倾斜角度效果不好,精度也不高。但是文献[3]给我们提供了一条思路。通过对于病案资料的分析发现,所有的病案资料都有较粗的标题分割线或者框格线,所以可以用直线检测的方法来得到图片倾斜的角度,然后转正。Radon 变换是一种基于投影思想的变换,可以用于直线检测,且精度较高[4,5]。

2.1　Radon 变换原理

Radon 变换的原理是将原始图像通过线积分的形式变换到另外一对参数域内,有许多表达方式[5],下式便是其中的一种:

$$\tilde{g}(r,\theta) = \int_{+\infty}^{-\infty}\int_{+\infty}^{-\infty} g(x,y)\delta(\rho - x\cos\theta - y\sin\theta)\,\mathrm{d}x\mathrm{d}y$$

式中,δ 函数被称为冲激函数,定义为:

$$\delta(x) = \begin{cases}\infty\ (\text{离散时为 1}), & \text{当 } x = 0\\ 0, & \text{当 } x \neq 0\end{cases}$$

δ 函数内的表达式为直线参数方程,表明该变换是沿着该直线进行积分的。用 Radon 变换可以实现 Hough 变换的功能,可以用它来检测图像中的直线的方向。计算不同的 $\tilde{g}(r,\theta)$,找出 Radon 变换数值最大的值,它所对应的 θ 值代表了图像中最长直线的方向。

2.2　歪斜病案的快速检测

为了提高检测效率,对图像进行了一些处理。首先,充分应用了病案图像的先验知识,所要找的标题线或者表格线均在图像上端的 1/4 之内,这里以医嘱记录为例,如图 1。

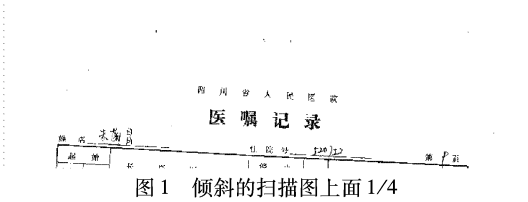


图 1　倾斜的扫描图上面 1/4

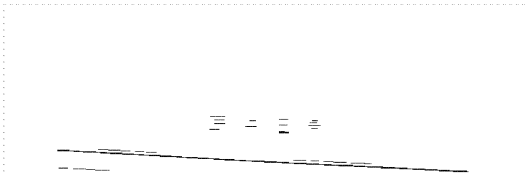


图 2　将图 1 作去干扰处理结果

所以只用针对图片的这部分做 Radon 变换。其次,对于上面截取的部分,为了减少周围字体对待检测线的干扰,按以下方式消除表格中的文字和图像[6]:先搜索每一行中的连续黑色像素串,如果长度小于某一阈值,则将这一黑像素串变为白色像素,这样几乎留下的就是标题线和表格线及较长的横线,如图 2。

再对剩下的部分作 Radon 变换,找到 θ 值,进行图像转正。图 3 为图 2 作 Radon 变换后的结果。

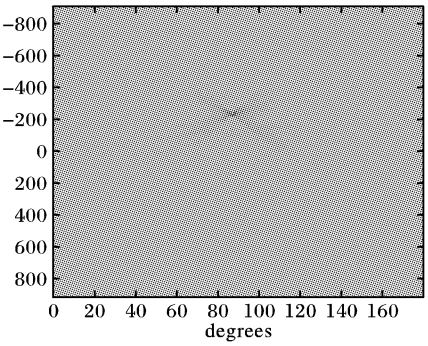


图 3　将图 2 作 Radon 变换的结果

图中清楚可见在 *degree* = 87 的地方有一个点,说明此处变换数值最大,即图 2 中最长的倾斜的表格线与垂直的 y 轴的夹角为 87°,与 x 轴夹角为 3°。知道倾斜角度后就可以将图像进行转正。

3　版面分析,框定识别部分

版面分析非常重要,只有分析正确,找准了待识别部分,才能识别出正确的信息。文献[7,8]提出了基于神经网络和多层次可信度的版面分析方法,两种方法处理的精度虽然较高,但第一种方法需要大量训练,第二种方法处理步骤比较复杂,时间消耗都比较大,我们力求寻找一种简单高效的方法。通过再次分析病案图像的先验知识发现,待识别的部分是该页病案的标题,该标题一定在标题线或最上面的表格线上,即肯定在图像上部 1/4 以内,且字体比其附近的字体大,即笔画粗,可以用数学形态学[9]的方法来去掉附近的小字体和其他噪声的干扰,然后将处理过的这部分通过横竖投影来确定其位置。

3.1　数学形态学开运算

数学形态学中定义了膨胀和腐蚀两个基本运算:

膨胀:

$$A \oplus B = \{x;(-B+x)\cap A \neq \varnothing\}$$

腐蚀:

$$A \ominus B = \{x;B+x \subseteq A\}$$

开运算为它们的组合,即:

$$A \circ B = (A \ominus B) \oplus B$$

其中 A 为被运算图像,B 为结构元素。开运算具有消除比结构元素小的图案和噪音的作用[10],这里利用它来消除标题字周围的小字体字符和其他噪声干扰。图 4 为转正后的图,图 5 为进行了开运算后的图。

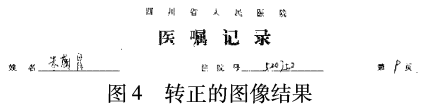


图 4　转正的图像结果

3.2　投影获取范围

首先通过水平投影来确定图片中标题的水平位置,如图

6。这时可以对原图在相应的水平范围内作垂直投影,进一步缩小识别区域,最后将确定的识别区域图像送与识别,图 7 为分割出来的图像。识别结果即为该扫描页的类型,这为分类入库提供了资料。

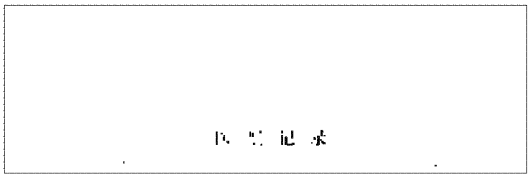


图 5 将图 2 作开运算结果

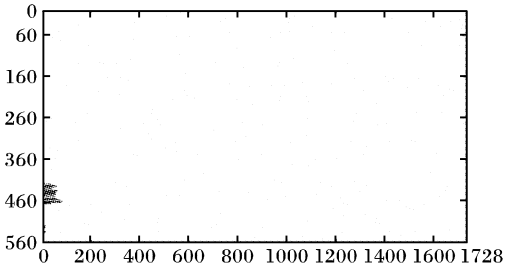


图 6 将图 5 作水平投影结果

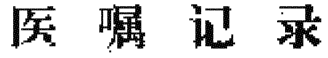


图 7 分割后的图像

4 结语

主要通过大津法对扫描的灰度图像进行动态二值化,根据图像特征选择 Radon 变换法对图像进行倾斜角检测转正,最后利用数学形态学开运算和投影方法分割出待识别部分,送与识别,提高了识别处理和归档的速度,并且大量的节约了人力。由本文设计思路开发的软件正在开发应用之中。

参考文献:

[1] OSTU N. A Threshold Selection method From Gray-Level Histogram [J]. IEEE Trans. on System Man Cybernet, SMC-8, 1978: 62 - 66.

[2] 付忠良. 图像阈值选取方法——Otsu 方法的推广[J]. 计算机应用,2000,20(5):37 - 39.

[3] 谷口庆治. 数字图像处理——应用篇[M]. 北京:科学出版社,2002.

[4] 孙文方,赵亦工. 基于有限 Radon 变换的图像纹理方向的检测[J]. 计算机应用, 2005,25(12):233 - 234.

[5] 彭钧. 基于 Radon 变换的噪声图像内规则形状目标的识别[J]. 微计算机信息, 2003,(7):84 - 85.

[6] 彭健,汪同庆,杨波,等. 一种单色表格的快速分析方法[J]. 计算机工程,2002,28(11):212 - 214.

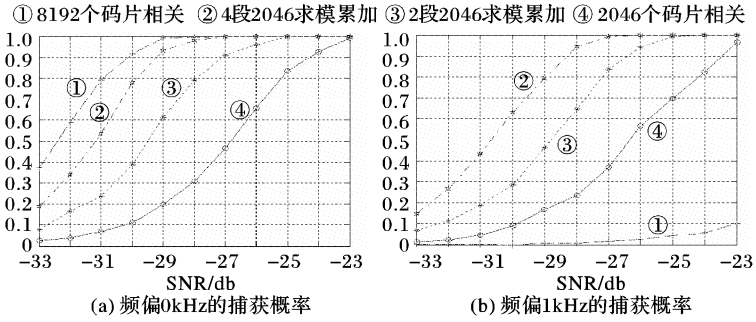
[7] 徐兆军,业宁,王厚立. 基于神经网络的版面分析[J]. 计算机应用, 2004,24(12):274 - 275.

[8] 陈明,丁晓青,吴佑寿. 多层次可信度指导下的自底向上的版面分析算法[J]. 模式是识别与人工智能,2003,16(2):198 - 203.

[9] 李凤慧. 基于数学形态学的图像噪声处理[J]. 信息技术,2006,30(6): 45 - 46,142.

[10] 杨波,汪同庆,叶俊勇,等. 文档图像的版面分析一基于数学形态学的方法[J]. 小型微型计算机系统, 2003,24(9):1673 - 1676.

(上接第 1019 页)



(a) 频偏0kHz的捕获概率 (b) 频偏1kHz的捕获概率

图 7 捕获概率仿真结果

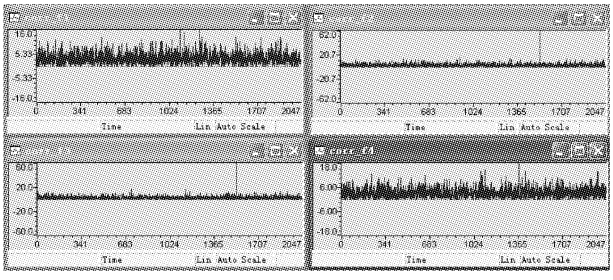


图 8 DSP 各通道捕获结果

以 8800MIPS 速度处理信息。本系统实现平台的核心器件为 TI 公司的 TMS320C6416,主频工作在 600MHz,它使用了两级高速缓存,片内 RAM 容量达到 1Mbyte,提供 64 个独立 EDMA 通道,以及 Mcbsp、EMIF 等多种集成外设,满足多种工程应用。本系统数据存储采用 16 位字长 Q.15 格式,C 语言编程,合理使用 CCS 的优化选项、C6416 特有的优化指令和内联函数,并通过合理安排数据存储位置使 cache 缺失流水化,尽可能的优化程序,提高程序效率。实现中使用多个 EDMA 通道完成数据搬移配合 CPU 运算,缩短运行时间。最终在单片 c6416 的硬件平台下完成一次长码捕获的时间为 32s。图 8 为信噪比 - 25db、频偏 - 3kHz 各通道的捕获结果,相关长度为 4 段 4092 个样点求模累加,可见,通道②和通道③可以捕获到明显的相关峰。

4 结语

随着 FFT 在扩频码捕获上的广泛应用以及数字信号处理芯片运算能力的不断提高,采用 DSP 芯片进行扩频信号的快速捕获越来越受到人们的重视,其优点是可以在保证捕获性能的条件下有效地节约硬件资源。本文描述了一种基于 FFT 的频域串行 - 伪码并行长码捕获的改进方法,并在以 c6416 DSP 芯片为核心的硬件平台上实现,该方法适用于大多普勒频偏、低信噪比下的扩频通信系统,实验结果表明具有捕获速度快、捕获性能好、实现结构简单等优点。

参考文献:

[1] 徐定杰,石吉利. 动态环境下基于 FFT 实现伪码快速捕获[J]. 中国航海,2003,(2):1 ~4.

[2] 陈辉. 伪码在大动态多普勒条件下的快速捕获[J]. 无线电技术,2005,(33):37 - 42.

[3] 周三文,黄龙,卢满宏. FFT 在高动态扩频信号捕获中的应用[J]. 飞行器测控学报,2005,24(2):61 - 64.

[4] 胡建波,杨莘元,卢满宏. 一种基于 FFT 的高动态扩频信号的快速捕获方法[J]. 遥测遥控,2004,25(6):19 - 24.

[5] 王浩,易大江,王飞雪,等. 超长 PN 码延迟 - 等待直接捕获方法[J]. 通信学报,2006,27(1):99 - 102.

[6] 朴虎哲,冯永新,潘成胜. GPS 信号中 P 码直接捕获技术研究[J]. 沈阳理工大学学报,2005,24(3):56 - 60.