

文章编号: 1001-2486 (2002) 01-0059-04

从图像中提取文字*

樊昀, 王润生

(国防科技大学 ATR 国家重点实验室, 湖南 长沙 410073)

摘要: 提出一种从图像中检测目标的思路, 并利用它提出了从图像中提取文字的新方法。该方法可以用于图像数据库内容检索, 视频检索, 以及将纸张出版物中的文本转化为电子出版物。新方法分为如下两个基本步骤, 首先检测出可能存在文字的候选区; 其次, 利用 Hough 变换方法分析候选区内扫描线灰度分布对候选区进行证实, 并对候选区的形状进行修剪, 使其尽可能紧凑地包含文字。该方法适用于检测出不同大小、字体、排列方向、灰度(颜色)的文字。由于该方法所有操作都是基于像元进行的, 所以便于硬件并行实现。所做的实验证实了方法的有效性。

关键词: 特征提取; 扩展滤波; 高斯滤波; hough 变换

中图分类号: TP391.4 **文献标识码:** A

Extracting Texts from Images

FAN Yun, WANG Runsheng

(ATR National Lab, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: A new way for detecting the targets through images is proposed, and then using it, a new method for extracting texts is proposed. This method can be used not only to index the image database and the video, but also to invert the documents in papers to electrical text format. It can be divided into two steps. In the first step, the candidate regions, which can include texts, are extracted, in the second step the candidate regions are verified by analyzing the distribution of gray value of the scan lines by Hough transform method. The shape of the extracted regions is also pruned to contain texts compactly. This new method can be used to extract texts with different sizes, orientations, fonts and gray values or colors. Because all operations of the new method are based on pixels, the new method can be realized by hardware easily. The experiments we have done justify its validity.

Key words: feature extraction; expanded filtering; gauss filtering; Hough transform

在视频节目尤其是新闻广播节目、纪录片等节目中, 会存在一些标题文本或包含于对象上的非标题文本, 这些文本为人们理解视频提供了重要而凝练的注释, 是一种可供索引的重要特征。在一段视频序列中, 文字往往对视频内容具有极强的概括力, 例如在某天的 CCTV 新闻联播中包含一段 $\times \times \times$ 总统访问以色列与巴拉克总理举行会谈的新闻节目, 历时近 3 分钟。这时若能从 4000 余帧视频画面中抽取其中含有“ $\times \times \times$ 总统与巴拉克总理举行会谈”嵌入标题的帧便能对该视频进行准确概括。

从图像中提取文本需要定位包含文本的图像区域。由于文本在字体、大小、对齐方式和排列上变化多端, 以及低对比度复杂背景的影响, 使得从图像中提取文本变得非常困难, 而且许多应用场合要求算法具有一定处理速度。

人们提出了许多提取文字的方法, 大体可以分为两类: (1) 第一类方法通过纹理分析, 如 Gabor 滤波^[1], 空间方差分析等, 提取文字区域, 该类方法具有一定的通用性, 但这类方法对于文本的字体和风格比较敏感, 存在着定位不准和算法复杂度高的缺点; (2) 第二类方法采用了连通元素分析技术^[2], 这类方法具有很高的处理速度和定位精度, 但是只适用于二值图像, 不适用于彩色和灰度图像。

我们提出的方法融合了两种方法的优点, 算法既有较高的定位精度、处理速度, 而且也适用于多值图像。用多幅复杂图像及多段视频进行试验, 结果表明, 该方法可以检测出不同大小、方向、排列

* 收稿日期: 2001-09-03

作者简介: 樊昀 (1976—), 男, 博士生。

方式、字体的文本,对弱对比度情况下的文本也比较有效。

1 算法描述

本节首先介绍候选区粗检算法;然后介绍候选区证实和候选区形状修剪算法。

1.1 粗检候选区

我们引入一种利用特征点集检测目标的方法。图像 I 中待检测目标为 T_i , $i = 1, 2, \dots, n$, 背景记为 B , 满足 $I = B + \sum_{i=1, \dots, n} T_i$ 。如果存在一个点集 $\Lambda \subseteq T_i$, 满足下述两条性质, 则称 Λ 是 T_i 尺度为 δ 的支撑特征点集, ST_i 为 T_i 尺度为 δ 的外包。

(1) 记 $N(x, y)$ 为 (x, y) 的邻域, 若 $(x, y) \in \Lambda$, 则可依据 $N(x, y)$ 的性质确定点 $(x, y) \in \Lambda$, $N(x, y)$ 的性质与目标 T_i 直接相关。

(2) Λ 在 T_i 中依 δ 稠密 $\delta > 0$, 如果 $\delta(x_0, y_0) = \{p = (x, y) \mid \|p - p_0\| < \delta, p_0 = (x_0, y_0)\}$, 有 $T_i \subseteq ST_i = \bigcup_{(x, y) \in \Lambda} \delta(x, y)$ 。

利用支撑特征点集检测目标 T_i 的方法可以分为如下 3 步:

- (1) 检测得到支撑特征点集 Λ 。
- (2) 选取合适的 δ , 由 Λ 得目标 T_i 的外包 ST_i 。
- (3) 依据 T_i 的其他特性证实 ST_i 。

设图像 I 中有多个待检测的文字块 T_i , $i = 1, 2, \dots, n$ 。对 T_i 作如下假设:

(1) T_i 的边缘是有一定的幅度的阶跃型边缘, 这些边构成的集合为 Λ_i 。

(2) Λ_i 是 T_i 依某一尺度 δ 的支撑特征点集。

(3) Λ_i 在 T_i 中有一定的密度, 若 $\forall (x_0, y_0) \in \Lambda_i$, 则 $\|\{(x, y) \mid (x, y) \in \delta(x_0, y_0) \cap (x, y) \in \Lambda_i\}\| > N_\delta$, 其中 N_δ 是 δ 的函数, $\|\cdot\|$ 表示集合元素个数。分: (1) 检测阶跃边集合 Λ_i ; (2) 构造外包 ST_i 两步检测候选区。为避免误检, 采用文[3]的展开滤波器法(Expand filter), 依据图 1 所示的四个方向的 5×5 模板 Φ , $i = 1 \dots 8$, 设计滤波器 Θ 。



(a) 垂直向模板, (b) 水平向模板, (c) 135 度向模板, (d) 45 度向模板

(a) vertical template (b) horizontal template (c) 135° directional template (d) 45° directional template

图 1 扩展滤波器的四方向模板

Fig. 1 The four templates of expand filter

下面简要描述展开滤波器法的基本原理: 若将一滤波器 θ 与 $M \times M$ 图像 s 卷积得响应 c , 寻找的模板特征位于图像中的 (x_l, y_l) , 有 $[s] \theta = c$, $[s]$ 是图像 s 的分块循环卷积矩阵。定义 DSNR 为:

$$\text{DSNR} = 10 \log \frac{c^2(L)}{\left(\sum_{j=1}^M c^2(j)\right) - c^2(L)} \quad (1)$$

最大化 DSNR, 得:

$$\theta = k^{-1} [[s]^T [s]]^{-1} s_l \quad (2)$$

其中 $k = \theta^T s_l / \theta^T [[s]^T [s]] \theta$ 为比例常数, 将模板 ϕ 平移到位置 l 得到的信号为 $s_l = \phi_l + \lambda$ 。若噪声为加性, 则有 $[s] = [\phi] + [\lambda]$, 得:

$$E\{\theta\} = k^{-1} [R_{\phi} + R_{\lambda}]^{-1} \phi_l \quad (3)$$

R_{ϕ} 和 R_{λ} 分别是模板和加性噪声的自相关矩阵。 $\Theta(\mu, \nu)$ 为 θ 的傅氏变换, 则有:

$$\Theta(\mu, \nu) = \frac{\overline{\Psi(\mu, \nu)}}{\|\Psi(\mu, \nu)\|^2 + S_{\lambda}(\mu, \nu) / S_{cc}(\mu, \nu)} \quad (4)$$

$S_{\lambda}(\mu, \nu)$ 、 $S_{cc}(\mu, \nu)$ 是噪声 λ 和系数 $c(x, y)$ 的功率谱密度。假设 $S_{\lambda}(\mu, \nu) = \sigma_{\lambda}^2$ 、 $S_{cc}(\mu, \nu) = \sigma_c^2$ 。将设计得到的 θ 与图像 I 卷积得 TI 。 TI 是由 I 检测出的阶跃边缘幅度图。将 TI 与尺度为 δ 的 Gauss 核 G 卷积得 GTI ，其中 G 定义为：

$$G(x, y) = \frac{1}{2\pi\delta^2} \exp\left(-\frac{x^2 + y^2}{2\delta^2}\right) \quad (5)$$

若 $TI(x, y) > Th$ ，则有 $(x, y) \in \Lambda_i$ ，其中 Th 为边缘幅值门限。选取 δ ，使得 Λ_i 依尺度 δ 在 T_i 中稠密，则有下列结论成立：若 $(x, y) \in ST_i$ ，则有 $GTI(x, y) > \frac{N\delta Th}{2\pi\delta^2} \exp(-1/2)$ 。证明略。

综上所述，得如下的候选区粗检算法：

步骤 1 $TI = I \circledast$ ，其中 \circledast 是相关算子。

步骤 2 $GTI = GTI * G$ ，其中 $*$ 是卷积算子。

步骤 3 对 GTI 取门限输出 $O(x, y)$ ，其中 $O(x, y)$ 满足下式，

$$O(x, y) = \begin{cases} I(x, y) & GTI(x, y) > \frac{N\delta Th}{2\pi\delta^2} \exp(-1/2) \\ 0 & \text{其他} \end{cases} \quad (6)$$

分析 $O(x, y)$ 中的非 0 区的连通性，得到候选区，并输出该候选区的最小外接矩形。

1.2 候选区证实

利用文字区域的其他特征证实粗检出的候选区。注意到在图像中，彼此邻近的文字在大小、字体、排列以及前景和背景通常大致相同，如果沿文字排列方向，做扫描线穿过候选区，则在扫描线上的图像灰度(颜色)排列上通常呈现出多个连续的宽度大致相同，峰(谷)值也大致相同的峰(谷)；灰度呈有规则的起伏排列。而对于背景来说，则通常没有这种性质。统计扫描线灰度分布的峰(谷)宽 w ， $w_1 < w < w_2$ ，峰(谷)值 l ， $l_1 < l < l_2$ 的联合分布，得到 $F_{w, l}(w, l)$ 。若该扫描线穿过文字，则在 $F_{w, l}$ 中应存在比较明显的峰。若设 P_w, P_l 为 $F_{w, l}$ 在 w 向和 l 向上的均值， V_w, V_l 为 $F_{w, l}$ 在 w 向和 l 向上的标准差， V 为总标准差， $T_i = 10$ ，扫描线方向选为候选区区域的主轴方向，分别定义如下：

$$P_w = \frac{\sum_{w=w_1}^{w_2} w \sum_{l=l_1}^{l_2} F_{w, l}(w, l)}{\sum_{w=w_1}^{w_2} \sum_{l=l_1}^{l_2} F_{w, l}(w, l)}, \quad P_l = \frac{\sum_{l=l_1}^{l_2} l \sum_{w=w_1}^{w_2} F_{w, l}(w, l)}{\sum_{w=w_1}^{w_2} \sum_{l=l_1}^{l_2} F_{w, l}(w, l)} \quad (7)$$

$$V_w = \left[\frac{\sum_{w=w_1}^{w_2} (w - P_w)^2 \sum_{l=l_1}^{l_2} F_{w, l}(w, l)}{\sum_{w=w_1}^{w_2} \sum_{l=l_1}^{l_2} F_{w, l}(w, l)} \right]^{1/2}, \quad V_l = \left[\frac{\sum_{l=l_1}^{l_2} (l - P_l)^2 \sum_{w=w_1}^{w_2} F_{w, l}(w, l)}{\sum_{w=w_1}^{w_2} \sum_{l=l_1}^{l_2} F_{w, l}(w, l)} \right]^{1/2} \quad (8)$$

$$V = (V_w + V_l) / 2, \quad (9)$$

综上所述，可得如下的候选区证实算法：

步骤 1 对候选区内所有扫描线的灰度分布进行统计，得分布 $F_{w, l, t} = \sum_{l=l_1}^{l_2} F_{w, l, t}$ ，其中 t 为扫描线标号， $F_{w, l, t}$ 为第 t 根扫描线灰度模式分布。

步骤 2 若 $V/(t_2 - t_1) < VT$ (VT 是预先设置的门限)，且 $P_l > PT$ (PT 是预先设置的门限)，则判定该候选区内可能存在文字，转入步骤 3；否则，舍弃该候选区。

步骤 3 检测 $F_{w, l}$ 的峰，确定峰(谷)值 l_0 和峰(谷)宽 w_0 ，对候选区内扫描线 t ，若 $F_{w, l, t}(w_0, l_0) = 0$ ，则舍弃扫描线 t ；否则转步骤 4。

步骤 4 确定 (w_0, l_0) 在扫描线 t 的存在范围 $Q_1 = [q_0, q_1]$ 。

步骤5 对保留下的候选区内的扫描线作连通分析, 剔除孤立扫描线, 并排除扫描线数 $t_1 < T_l$ 的候选区。输出结果。

2 实验结果

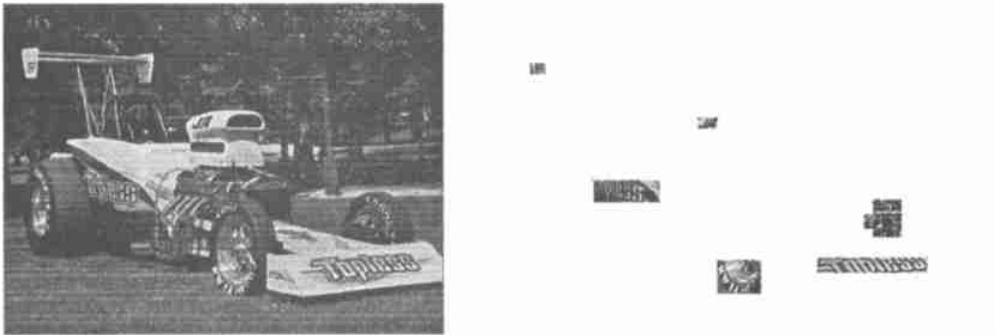
部分实验结果如图2、图3所示, 参数选为 $\sigma = 1.8$, $Th = 130$, $N_\delta = 5$, $VT = 10$, $PT = 150$ (峰) 30 (谷)。实验结果表明, 算法可以提取同一幅图像中字体不同、灰度不同、排列方向不同的文字, 对文字与背景对比度较差的情况也有一定的效果。



(a) 原始图像 (b) 提取出的文字
(a) Original image (b) extracted texts

图2 实验结果 1

Fig. 2 The experimental results 1



(a) 原始图像 (b) 提取出的文字
(a) Original image (b) extracted texts

图3 实验结果 2

Fig. 3 The experimental results 2

3 结论

本文提出一种从图像中检测目标的方法, 并将它用于提取图像中的文字。实验结果证实了该文字提取方法的有效性, 而且由于算法是基于像素的, 便于硬件并行实现。在未来的工作中, 要进一步完善该算法, 提高检测的准确率, 并拓宽它的适用范围。

参考文献:

- [1] Anil K. Jain and Sushil Bhattacharjee. Text Segmentation Using Gabor Filters for Automatic Document Processing [J]. Machine Vision and Applications, 1992, 5: 169-184.
- [2] Tan, C L Text Extraction Using Pyramid [J]. Pattern Recognition, 1998, 31 (1): 63-72.
- [3] K. Raghunath Rao, Jezekiel Ben-Arie. Multiple Template Matching Using the Expansion Filter [J]. IEEE Trans CSVT, 1994, 4 (5).
- [4] 王润生. 图像理解 [M]. 长沙: 国防科技大学出版社, 1994.
- [5] 孙即祥. 数字图像处理 [M]. 石家庄: 河北教育出版社, 1991.