

基于颜色和笔画特征的文本分割算法

黄百钢 李俊山 胡双演

(西安二炮工程学院指挥自动化系 西安 710025)

摘要 自动提取图像中的文本对图像视频检索具有重要意义。提出了一种基于颜色和笔画特征,应用无监督聚类方法进行复杂背景下的文本分割算法。首先在对文本行图像增强的基础上,应用颜色约减和直方图确定文本颜色。然后提取颜色和笔画特征,应用k-均值聚类算法分割出文本和背景像素。最后应用后处理优化分割结果。实验表明,该算法具有较好的分割效果。

关键词 文本分割,笔画特征,无监督聚类,文本颜色

中图分类号 TP391.4 **文献标识码** A

Text Segmentation Using Color and Stroke Features

HUANG Bai-gang LI Jun-shan HU Shuang-yan

(Xi'an Research Inst. of High-tech., Xi'an 710025, China)

Abstract A novel text segmentation approach based on unsupervised clustering using color and stroke features was presented. Firstly, the possible text and background color is estimated on the enhanced text line image by color reduction and histogram calculation. Then color and stroke features were extracted and text and background pixels are determined by K-mean clustering algorithm. At last, segmentation result is optimized by post-processing. The performance of our approach is demonstrated by experimental results for a set of images with Chinese and English text.

Keywords Text segmentation, Stroke feature, Unsupervised classification, Text color

近年来我国数字高清电视、互联网迅速发展,数字图像、视频也随之迅猛增长,对其高效管理和检索成为日益重要的问题。视频图像中的文本往往与视频图像内容紧密相关。自动识别视频帧、图像中的文本对基于内容的信息检索具有重要意义。

1 自动文本分割技术

一般将视频图像自动文本识别按顺序分为4个步骤^[1],即:文本检测、文本定位、文本分割和文本光学字符识别(OCR)识别。文本分割的任务是从文本定位得到的图像中去掉背景像素,留下文本像素,形成白底黑字的二进制图像,该图像可以直接送入OCR软件,从而自动识别出文本。文本分割的效果直接影响OCR软件识别率。字体大小、文字颜色、文本背景、语言种类等不同,使自动文本分割变得困难。

在文本分割方面研究人员已经提出了许多方法。对于背景相对比较简单的图像,局部阈值的方法是有效的^[2,6],然而当背景复杂时该类方法会将部分背景像素错分为前景像素。当前景具有相近的颜色时,颜色聚类的方法比较有效^[3,4]。然而当前景颜色差别较大时,效果不理想。文献^[5,9]应用混合高斯模型(GMM)或马尔科夫随机场模型(MRF)取得了比较好的效果,然而前景颜色模型的建立,使该方法复杂化,分割效果也与学习样本有关。

本文中,利用HSI颜色空间的色度和亮度分量作为颜色特征,笔画滤波器的响应作为笔画特征,通过无监督聚类的方法进行文本分割。避免了文献^[2]方法的阈值确定问题,也避免了文献^[5]方法的学习样本选取问题,与文献^[3]的方法比较而言,在特征选取上具有较大差别,比较适合于中文文本的分割。试验表明本方法对不同字体、颜色、大小的文本都可较好分割。本文算法分以下几步:(1)图像增强预处理;(2)文本颜色估计;(3)特征提取;(4)无监督聚类;(5)基于连通成分的后处理。

2 算法描述

本文算法输入为一幅包含一行文本的图像,该图像由文本定位算法而得到,试验中使用文献^[1]的定位算法。输出为白底黑字的二进制图像。

2.1 图像增强预处理

试验表明太小或太大的字符高度对OCR识别都不利,字符高度在100像素左右时OCR可以取得比较好的识别效果^[4],因此本算法中先用双3次插值将输入图像保持长宽比例不变再缩放到高度为100像素左右。这样的预处理也对本算法的后续步骤有益。

2.2 文本颜色估计

采用与文献^[4]中类似的方法来估算文本颜色。一幅文

到稿日期:2008-08-25 返修日期:2009-02-06 本文受国家自然科学基金(60772151)资助。

黄百钢(1971-),男,博士研究生,副教授,主要研究方向为图像视频检索等,E-mail:huangbaigang029@sina.com;李俊山(1957-),男,教授,博士生导师,主要研究方向为图像处理等;胡双演(1979-),男,博士,讲师,主要研究方向为图像处理等。

本行图像中可能存在许多种颜色,文本像素的颜色也可能有差异。首先按照颜色之间的距离将输入文本行图像的颜色约简到 Ncolor 种。然后计算输入图像中间 4 行的归一化颜色直方图和输入图像四周像素的归一化颜色直方图,并求二者之差。该差值直方图的最大值对应的颜色为文本颜色,最小值对应的颜色为背景颜色。设得到的文本颜色为 c_t ,背景颜色为 c_b 。

2.3 特征提取

根据观察,文本和背景大多在亮度上差别较大,如果亮度相近则颜色色调有明显差别,这主要是为了让人能从背景中区别出文本。本文中采用 HSI 颜色空间中的亮度(I)和色调(H)分量作为颜色特征。从 RGB 颜色到 HSI 颜色空间的亮度和色调的计算公式如下^[7]:

$$I = (R + G + B) / 3 \quad (1)$$

$$H = \begin{cases} \theta(B \leq G) \\ 360 - \theta(B > G) \end{cases} \quad (2)$$

其中

$$\theta = \arccos \left\{ \frac{[(R - G) + (R - B)] / 2}{[(R - G)^2 + (R - G)(G - B)]^{1/2}} \right\}$$

为了后续处理,将 I 和 H 的值归一化到 [0, 1] 的范围。

为了利用文本中的笔画信息,用文献[8]提出的笔画滤波器对输入图像滤波,将其响应系数作为笔画特征。像素点 (x, y) 在方向 α 和尺度 d 的笔画滤波器响应为:

$$R_{\alpha, d}(x, y) = \frac{|\mu_1 - \mu_2| + |\mu_1 - \mu_3| + |\mu_2 - \mu_3|}{\max(\delta_1, \delta_2, \delta_3, 10)} \quad (3)$$

其中 $\alpha \in \{0, \pi/4, \pi/2, 3\pi/4\}$ 为方向, $d \in \{1, 3, 5\}$ 为尺度, $\mu_i, \delta_i (i=1, 2, 3)$ 为笔画滤波器所考察的 3 个局部区域的均值和方差,详见文献[8]。取各个方向和尺度的最大响应作为该点的最终响应值,即

$$R(x, y) = \max_{\alpha, d} R_{\alpha, d}(x, y) \quad (4)$$

为了后续处理对 $R(x, y)$ 也进行归一化处理。

这样由亮度(I)、色调(H)和笔画滤波器响应(R) 3 个特征组成特征向量 F , 即 $F = \{I, H, R\}$ 。

2.4 无监督聚类

在图像检索研究中, K-均值聚类算法是一种常用的聚类算法^[10], 它可将像素点分为 k 类。在应用中利用 K-均值聚类算法将像素点分为文本和背景两类, 即 $k=2$ 。聚类所用的特征为 2.3 节所述的特征向量 F 。设文本类的聚类初始值为 f_t , 背景类的聚类初始值为 f_b 。理想的文本类的像素点的颜色应该具有上文确定的文本颜色 c_t , 设其亮度为 $c_{t,I}$, 色调为 $c_{t,H}$, 笔画滤波器响应(R)应当为最大值, 即归一化后的值 1, 即理想的文本特征向量为 $f_t = [c_{t,I}, c_{t,H}, 1]$ 。理想的背景类的像素点的颜色应该具有上文确定的背景颜色 c_b , 设其亮度为 $c_{b,I}$, 色调为 $c_{b,H}$, 笔画滤波器响应(R)应当为最小值, 即归一化后的值 0, 即理想的背景特征向量为 $f_b = [c_{b,I}, c_{b,H}, 0]$ 。文本类的聚类初始值为 f_t 应当与理想的文本特征向量 $f_t = [c_{t,I}, c_{t,H}, 1]$ 距离最近, 即:

$$f_t = \min_{x, y} ((f(x, y) - f_t)^2) \quad (5)$$

背景类的聚类初始值为 f_b 应当与理想的背景特征向量 $f_b = [c_{b,I}, c_{b,H}, 0]$ 距离最近, 即:

$$f_b = \min_{x, y} ((f(x, y) - f_b)^2) \quad (6)$$

其中 (x, y) 为像素点坐标。

算法步骤如下:

Step1 按式(5)、式(6)计算 2 个类别的初始聚类中心;

Step2 将待分类的模式特征矢量集中的模式逐个按最小欧氏距离原则划分给两类中的某一类;

Step3 计算重新分类后的各类中心, 并更新;

Step4 如果各类中心不再变化, 算法结束, 否则转 Step2。

经过聚类后, 确定了文本像素和背景像素, 将文本像素变为黑色, 背景像素变为白色, 即得到聚类结果二进制图像。

2.5 基于连通成分的后处理

在这一步中主要利用文本像素区域的尺寸、长宽比例和位置消除不可能是文本的区域。如果一个文本区域面积过小, 则被认为是噪声像素, 应该删除。实验中删除面积小于 20 像素的文本区域。对于长宽比例过小或过大的区域也应该删除, 因此删除比例小于 0.2 或大于 10 的文本区域。文本区域一般不与图像边缘区域连接。因此, 和图像边界连接的像素达到该文本区域的高或宽的 60%, 则将此区域从文本区域中删除。

3 实验分析

为了验证本文算法, 选取了 200 幅文本行图像进行测试。在这些文本中有 1150 个汉字、968 个英文字母。这些图像的背景比较复杂, 包含的文本大小、字体和颜色差别较大。将分割的结果图像送入清华紫光文通 OCR 软件 TH-OCR V9.0 进行识别, 得到识别结果。图 1 给出了 3 幅示例图像。图 1(a) 中“然”字与“书”字之间有一块背景与前景颜色接近, 而且面积较大, 文献[3]算法分割结果中“然”字与“书”留下两块黑块, 导致 OCR 软件将“然”字错误识别为“鹅”字。本文算法分割结果中去掉了这些黑块, 得到正确的识别结果。图 1(b) 中文本两部分颜色不同, 前部分为黄色, 后部分为白色, 文献[3]算法分割结果没能处理好黄色文本, 导致 OCR 软件不能识别

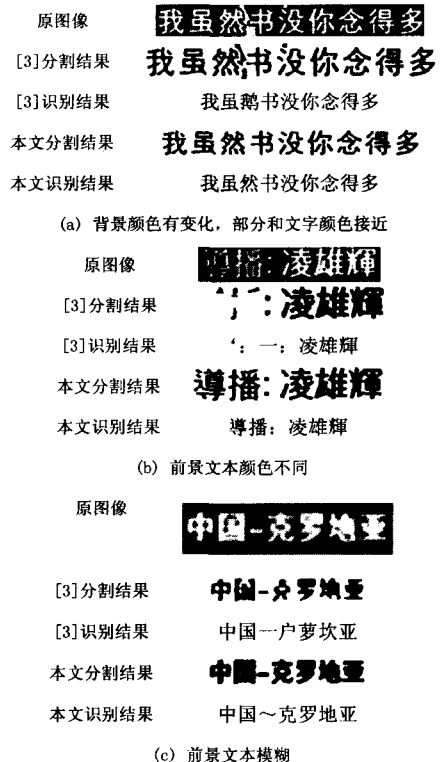


图 1 文本分割示例

出“簿播”二字。本文算法对两种颜色的文本都能较好处理,识别结果正确。图 1(c)中前景文本比较模糊,背景比较复杂。文献[3]算法分割结果比较残缺,导致中间几个字识别错误;本文算法分割结果比较完整,识别结果基本正确。从图 1 可以看出本文算法分割文本比较准确。

为了定量评价本文的算法,采用字符识别率作为评价指标。字符识别率(CRR)是 OCR 软件正确识别的字符数与文本行图像中所有字符数的百分比。设字符识别率为 CRR,OCR 软件正确识别的字符数为 N_1 ,文本行图像集中所有字符数为 N ,则

$$CRR = \frac{N_1}{N} \quad (7)$$

N_1 和 N 是从实验中由人工统计而来的。

表 1 比较了本文算法在使用不同特征时的字符识别率,由表 1 可知在对图像进行增强预处理后,同时使用颜色和笔画特征时可取得最高的识别率。

表 1 本文算法在使用不同特征时字符识别率的比较

特征向量	中文 CRR	英文 CRR	综合 CRR
颜色(未增强)	70.3%	72.4%	71.6%
颜色+笔画(未增强)	78.2%	80.1%	79.3%
颜色(增强)	75.2%	78.3%	76.5%
颜色+笔画(增强)	83.1%	85.2%	84.7%

表 2 对本文算法和文献[3]算法的结果进行比较,可见本文算法的中文文本识别率明显提高,这得益于本文算法利用了文本的笔画特征。

表 2 本文算法与文献[3]性能的比较

算法	中文 CRR	英文 CRR	综合 CRR
本文算法	83.1%	85.2%	84.7%
文献[3]算法	79.6%	84.5%	82.6%

结束语 本文提出了一种利用颜色和笔画特征的无监督聚类方法来进行复杂背景下的文本分割。首先在对文本行图像增强的基础上,应用颜色约简和直方图确定文本颜色。然后提取颜色的亮度、色度分量和笔画特征,并应用 K-均值聚类算法分割出文本和背景像素。最后应用后处理优化分割结果。利用 OCR 软件对分割结果的字符识别率作为评价指标,

试验表明本文算法比已有的一些方法具有更好的性能。下一步工作要将本算法用于基于语义的图像视频检索。

参考文献

- [1] Lyu M R, Song J Q, Cai M. A comprehensive method for multilingual video text detection, localization, and extraction [J]. IEEE Transactions on Circuit and System on Video Technology, 2005, 15(2)
- [2] Tsai C M, Lee H J. Binarization of color document images via luminance and saturation color features [J]. IEEE Trans. on Image Processing, 2002, 11(4)
- [3] Gllavata J, Ewerth R, Freisleben B. Text detection in images based on unsupervised classification of high-frequency wavelet coefficients [C] // International Conference on Pattern Recognition. 2004
- [4] Wernicke L R. A Localizing and Segmenting Text in Images and Videos [J]. IEEE Transact. on Circuits and Systems for Video Technology, 2002, 12(4): 256-258
- [5] Odobez J M, Chen D. Robust Video Text Segmentation and Recognition with Multiple Hypotheses [C] // Proc. of IEEE International Conference on Image Processing 2002. Vol. II, Rochester, New York, 2002: 433-436
- [6] Gllavata J, Ewerth R, Freisleben B. Finding Text in Images via Local Thresholding [C] // Proc. of the 3rd IEEE Int'l Symposium on Signal Processing and Information Technology. Darmstadt, Germany, 2003
- [7] 冈萨雷斯. 数字图像处理 [M]. 北京: 电子工业出版社, 2006: 233-240
- [8] Liu Qifeng, Jung Cheolkon, Kim Sangkyun, et al. Stroke Filter for Text Localization in Video Images [C] // IEEE International Conference on Image Processing. Oct. 2006: 1473-1476
- [9] 付慧, 刘峡壁, 贾云得. 图像中多语种文本提取的高斯混合建模方法 [J]. 计算机研究与发展, 2007, 44(11): 1920-1926
- [10] 陈世亮, 李战怀, 袁柳. 一种基于概念层次的图像检索方法 [J]. 计算机科学, 2008, 35(4): 139-141
- [4] Baylor S, Corbett P, Park C. Efficient method for providing fault tolerance against double device failures in multiple device systems [P]. U. S. Patent 5,862,158. 1999
- [5] Blaum M, Brady J, Bruck J, et al. EVENODD: an efficient scheme for tolerating double disk failures in RAID architectures [J]. IEEE Transactions on Computers, 1995, 44: 192-202
- [6] Blaum M, Roth R M. On lowest density MDS codes [J]. IEEE Transactions on Information Theory, 1999, 45: 46-59
- [7] Hafner J L, Deenadhayan V, Kanungo T, et al. Performance metrics for erasure codes in storage systems [R]. RJ 10321. IBM Research, San Jose, CA, 2004
- [8] MacWilliams F J, Sloane N J A. The Theory of Error-correcting Codes [M]. The Netherlands; Northolland, Amsterdam, 1977
- [9] Plank J. A tutorial on Reed-Solomon coding for fault-tolerance in RAID-like systems [J]. Software, Practice and Experience, 1997, 27: 995-1012
- [10] Reed I S, Solomon G. Polynomial codes over certain finite fields [J]. Journal of the Society for Industrial and Applied Mathematics, 1960, 8: 300-304
- [11] Xu L, Bruck J. X-code: MDS array codes with optimal encoding [J]. IEEE Transactions on Information Theory, 1999, IT-45: 272-276
- [12] Corbett P, English B, Goel A, et al. Row-diagonal parity for double disk failure [C] // Proceedings of the Third USENIX Conference on File and Storage Technologies. 2004: 1-14

(上接第 266 页)