文章编号:1001-9081(2010)02-0449-04

基于敏感点颜色聚类 and 行聚类筛选的文本提取

刘琼¹, 周慧灿¹, 王耀南²

(1. 湖南文理学院 计算机科学与技术学院, 湖南 常德 415000; 2. 湖南大学 电气与信息工程学院, 长沙 410082)

(liuyuhan1108@126.com)

摘要:针对现有的文本提取算法不能适应复杂背景变化和文字本身的形状变化问题,提出一种基于敏感点颜色两级聚类和文本行聚类筛选的方法。新方法利用人眼视觉对颜色大幅度变化更敏感的特点,以敏感点的主要颜色作为聚类分析的依据,克服了现有阈值方法和聚类方法受背景颜色变化影响较大的问题。在此基础上,以文本行的空间排列特征为依据进行文本行筛选,以克服一般方法容易受文字形状和尺寸变化影响的缺点。实验表明,新方法对于背景的复杂变化和文字的形状尺寸变化都具有很好的适应性。

关键词:文本提取; K均值聚类; 边缘密度; 文本行聚类

中图分类号: TP391.41 **文献标志码:** A

Text extraction based on clustering colors at sensible points and clustering text-lines for text-selection

LIU Qiong¹, ZHOU Hui-can¹, WANG Yao-nan²

(1. College of Computer Science and Technology, Hunan University of Art and Science, Changde Hunan 415000, China;

2. College of Electrical and Information Engineering, Hunan University, Changsha Hunan 410082, China)

Abstract: Since the existing text extraction methods can not adapt to the variation of complex background and shape, a new method was brought forward. It was founded on two-level color clustering of sensible points and text-line clustering. Because human vision perception is more sensitive to great change of colors, the new method only selected the main colors at sensible points to cluster. The strategy could solve the problems of the existing methods based on threshold and clustering which were greatly influenced by the variation in colors of complex background. And then, the text-lines were selected according to the fact that texts always align with each other in a same text-line. That course can eliminate the influence of variation in shape and size of characters. Experimental results indicate that, the new method has good adaptability to complex change of background, and texts with different size and shape.

Key words: text extraction; K-means clustering; edge density; text-line clustering

0 引言

提取图像和视频帧中的文字正日益得到广泛的应用,如图像理解、基于内容的图像检索等。由于大多数彩色文本图像具有复杂的背景,而视频图像一般分辨率较低,这使得通常的光学字符识别(Optical Character Recognition, OCR)系统面临巨大的挑战,因为,目前成熟的OCR系统一般要求清晰的二值化图像输入。因此,在OCR识别之前必须把文本从复杂背景中提取出来。

在过去十年中,图像中的文本检测技术引起了广泛关注^[1-3],但是,如何把文字从背景中很好地分离出来仍然还有许多需要探讨的问题。全局灰度阈值方法是一种比较常见的文本分割方法,其中最具有代表性的就是Ostu阈值方法^[4],这类方法通常要求文字背景单一或者变化缓慢,当背景较复杂时效果极差。局部自适应阈值方法^[5]对背景的复杂变化有一定的适应性,但是,算法对大对比的边缘轮廓不具备分辨能力,导致大量的无关背景被保留。基于颜色聚类的方法也是目前比较常见的方法^[6],这类方法主要存在两个问题:一是合理的聚类数目确定起来比较困难;二是当文字和背景颜色都不单一时,大面积的背景对聚类效果影响较大,使得算法

缺乏适应性。也有一些关注于文字笔画特征的方法,与其他方法相比对颜色的变化更具有适应性,文献[7]提出了一种基于文字笔画的双边模型来描述文字的局部特征,使其与背景元素分开,取得了较好的效果,该方法的缺点是需要笔画宽度先验知识并且计算比较复杂,准确度依赖于文字笔画的粗细和规范程度,对文字的尺寸和形状较敏感。

总的来看,无论是基于阈值的方法还是基于颜色聚类的方法,都存在对变化的复杂背景颜色难以适应的问题;而基于文字本身特征的方法则准确性不高,不能适应文字的形状和尺寸的变化。针对这些问题,本文提出一种基于敏感点颜色两级聚类和文本行聚类分析的方法,对图像中的视觉敏感点分别进行局部和全局颜色分析,消除较大面积、次要的背景颜色对阈值选择和聚类分析的影响,再利用文字的空间排列特性实现文本行的筛选,以解决算法对文字尺寸和形状的适应性问题。

算法主要分成四个步骤:1)敏感点颜色聚类分析;2)颜色子图分离;3)子图文本行筛选;4)子图合并。

1 敏感点颜色聚类分析

在复杂的背景下,文字一般具备如下一些特性:1)文字

收稿日期:2009-08-09。

作者简介:刘琼(1964-),女,湖南常德人,教授,主要研究方向:数字信号处理、模式识别;周慧灿(1974-),男,湖南常德人,副教授,硕士,主要研究方向:多媒体技术、智能信息处理与融合;王耀南(1957-),男,江西吉水人,教授,博士生导师,博士,主要研究方向:智能控制、智能信息处理与融合。

与背景一般较大的颜色或亮度差异;2)同组文字在颜色上具有一致性;3)同组文字在空间排列上成行或成列。

研究表明,人眼视觉具有对大对比度颜色频繁变化更加敏感的特点。因此,图像中的这些位置更容易引起人们的注意,而那些大面积的色块和平滑过渡的颜色区域通常会被人们所忽略,显然,文本的特性1)正符合前者的要求。从图像处理的角度来看,大对比度颜色频繁变化的区域必然是边缘密度很高的区域。另外,从观察习惯上来看,对于图像的某个局部小区域,人眼一般更关注数量上占主要的颜色而忽略一些次要的颜色,这一习惯对消除不必要的视觉干扰是有好处的。

基于上述理由,在一幅文本图像中,只要考虑在视觉敏感点处的小面积区域中占主要成分的颜色即能保证文本颜色的捕获。

1.1 敏感点选择

前面已经提到,人眼视觉的敏感点一般是边缘密度较高的区域,因此,边缘密度分析是确定敏感点的基础。边缘检测的方法很多,常见的有 Sobel 边缘检测、Canny 边缘检测和 log 边缘检测等^{[8]289-295}。实验中采用了 Sobel 边缘检测方法,对彩色图像的 R、G、B 三个子图分别检测其灰度边缘,然后用逻辑或三个子图合并,为便于描述,设合并后的二值图像为 $Edge_{RGB}$ 。

对于 $Edge_{RGB}$ 中的任意一点,计算以该点为中心, $w \times w$ 大小窗口内的平均边缘密度,实验中 $w = 30$ 。计算窗口尺寸 w 的选择比较宽松,其大小并不影响边缘密度局部极大值的位置。计算公式如下:

$$EdgeDens_{(i,j)} = \frac{1}{w \times w} \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} Edge_{RGB}(m,n) \quad (1)$$

此时, $EdgeDens$ 为灰度图像,其中的最大值意味着最高的边缘密度,最适合作敏感点。但是,在连续多个敏感点选定时存在这样的问题:如果依次选择最大值、次大值、第三大值、……,直到最小阈值分别作为敏感点,点的分布会比较紧密,并且数量较多,不利于提高计算效率。根据边缘密度的计算方法,容易得出如下结论:边缘密度最高的点周围的边缘密度应当非常接近最高点,很可能就包含了次高点、第三高点,等等。从区域分析的角度来看,如果考虑了最高点,显然该点周围的点就不用再考虑。因此,在敏感点选择时,采取间隔选择的方法,比如:最高点、第十高点、第二十高点、……,以此类推直到低于最小阈值。最小阈值来源于当纹理密度过低时,不可能有文本,不用分析。实验中发现当敏感点的数量达到 100 个时,就已经能适应绝大多数情形。图 1 给出了一幅实验图像的敏感点分布情况,可以看到,大部分的敏感点都落在文字的笔画附近。

图1 敏感点分布示例

1.2 敏感点局部和全局聚类

敏感点颜色分析的根本目的是要找出文本的颜色,根据前述的文本特性,大多数的敏感点应当落在文字的笔画附近,并且,文字的颜色具有一定的单一性,即同一组文字一般具有相同的颜色。因此,如果在敏感点附近区域内考虑两种最主要的颜色,则包含文本颜色的可能性极大。具体实现如下:

设原图像为 I ,对于已经选定的任意敏感点 (i,j) ,以 $I(i,j)$

为中心,在 I 中选取 10×10 大小的矩形检测窗口,对窗口内所有像素的颜色利用 K-means 聚类方法聚为两类,以两个聚类中心点作为 $I(i,j)$ 点附近的两种主要颜色,这一步骤称为敏感点局部聚类。应当指出,尽管 K-means 方法并不能避免聚类时的局部最优问题,但个别点聚类不准确并不影响后续的处理。

在全部敏感点的主要颜色中,同时包含了局部背景颜色和文本的颜色,由于文本颜色的一致性特点,在这些主要颜色中,文本颜色一定能占据较多的数量,但并不一定是数量最多的,这取决于背景的复杂程度,当复杂程度越高,背景颜色种类越多,而每种颜色数量越少,此时,文字的颜色更容易被检出;反之,则背景颜色单一,而数量较多,除了文字颜色之外的其他颜色数量较少,保留两类主要颜色一般也能保证文本颜色的检出。考虑到文本可能有多组的情况,为了保证算法的适应性,利用 K-means 聚类方法将全部敏感点主要颜色聚为 4 类,也可以更多,实验表明保留 4 类即能适应多种复杂情形,这一步骤称为敏感点全局聚类。

1.3 颜色子图分离

通过敏感点全局聚类,4 种主要的颜色被检出,但此时并不能确定哪些颜色是文本颜色。实际上,情况一般比较复杂,比如某一种颜色可能在一个区域是文本而在另一个区域却不是,因此必须进一步进行文本筛选。由于已经得到 4 种主要关心的颜色,其他颜色则可以不作考虑,另外,为了后续算法的方便,从原图中将 4 个颜色子图分离出来,方法如下:

$$Sub_I_n(i,j) = \begin{cases} 1, & \|I(i,j) - color_n\| \leq thresh \\ 0, & \|I(i,j) - color_n\| > thresh \end{cases} \quad (2)$$

其中: Sub_I_n 表示第 n 个颜色对应的子图, $n = 1, 2, 3, 4$; $thresh$ 表示颜色容差,实验中取 60。子图分离效果如图 2 所示。

(a) 原图像

(b) $n=1, color_1=(247,189,168)$

(c) $n=2, color_2=(213,25,27)$

(d) $n=3, color_3=(189,55,75)$

(e) $n=4, color_4=(171,5,151,164,25)$

图2 子图分离示意图

2 文本行筛选

2.1 初步筛选

由于文本提取是文本区域粗略定位以后的工作,因此,在4个颜色子图中,有文字的子图中一般文字会分布在靠近子图中间的位置,而子图边缘上取值为1的点分布则较少,如图2(c)所示。相反,如果一个子图有大量的背景,则边缘上取值为1的点分布则较多。基于上述原因,当边缘上有较多1值点时(一般不低于总边长的1/4),该子图是背景颜色,去除(例如,图2中,子图(e)将被舍去)。

在剩下的子图中,进一步去除其他干扰像素,步骤如下:1)典型相关分析(Canonical Correlation Analysis, CCA)连接成分分析,并标记所有连通区域^{[8]270-273};2)去掉与边缘相连的区域;3)去掉过小和过大区域。

经过上述步骤,大面积背景将被去除,文字则被保留,如图3所示。对于粗略定位以后的文本区域来说,其文字部分应当占用图像相当的比例。因此,去干扰后的子图如果1值点过少(低于全图面积的5%),则不应当是文本,去掉该子图,显然图3(a)应该舍去。

图3 剩余子图进一步去除干扰像素

2.2 文本行聚类筛选

多数的文本判别方法更加关注文字本身的形状、纹理等特征,而较少注意文字在空间位置上的分布特点,由于缺乏合适的文本行特征表述方法,通常该特点只能用于辅助性判别^[9],其作用也比较有限。一般认为,对同一行文字,其高度、宽度和重心高度应当是一致的,但事实上,在没有准确分割的情况下,一般文字的宽度并不能保证一致,对于西方文字更不一致;而由于明显的行间隔和文字结构上的原因,同一行文字高度一致的情况更多一些;此外,无论是完整的文字,还是部分的笔画,同一行文字的重心纵坐标在总体上是一致的。

基于上述分析,选择连通区域的高度、连通区域的重心纵坐标作为特征对所有连通区域进行K-means聚类分析以寻找文本行。但是,实验发现,采用上述特征进行文本行聚类时效果并不理想,如图4(a),“中”字的下半部被错误地聚类,导致底部文本行类内方差显著增加,换言之,同一行内的连通区域重心波动幅度过大,类内方差不足以锁定文本行。分析发现,重心合理的波动幅度与文字的尺寸是相关的,显然,当文字尺寸较大时,重心的波动幅度可以更大一些;当文字较小时,允

许的重心波动幅度也较小。因此,在文本行聚类时同时引入相对重心纵坐标(即重心纵坐标/连通区域高度)参加文本行聚类,取得了很好的效果,如图4(b)。大量的聚类实验表明:当类内平均方差低于4时(选用cityblock距离),该类簇即为文本行。

图4 引入相对重心纵坐标前后的聚类效果

需要说明的是,为了有效减少聚类元素的数量,提高聚类准确性,在文本行聚类之前对已标注的连通区域作了进一步处理:1)寻找连通区域的最小外接矩形区域;2)合并搭接矩形区域;3)重新进行CCA;4)再次寻找连通区域的最小外接矩形区域。上述步骤有利于合并文字内的离散笔画。

在文本行聚类时,聚类数目选择是必须要考虑的问题,本文利用行像素变异直方图估计可能的文本行数,并留出一簇的余量以提高算法的适应性,但是余量面对非常规整的文本行时,可能出现过分类的情形,算法考虑了这种情况,当两簇的类间距离不大于4时,合并两簇。图6给出了行像素变异直方图示例,在单行总像素数的10%处提取左穿和右穿标记线的点对数,即为估计的文本行数。

图5 行像素变异直方图文本行数估计

最后,将所有余下的子图中检出的文本行在图2对应子图中重新取样,然后合并为一幅图像,完成文本提取,图6为最终提取结果。

图6 提取结果

3 文本提取实验

实验选取了来源于Internet和视频捕获的近400幅包含文字的图像,在文本提取之前,先利用文献^[10]的方向密度算法进行文本区域粗定位,得到不同分辨率的文本图像块共计382幅。然后,利用本文方法实施文本提取,部分具有代表性的实验结果如图7~9。

从图7可以看到,当图像背景非常复杂时,Ostu 阈值方法基本上无法区分背景和前景文字,抗干扰能力极差;而本文算法则取得了比较理想的效果,但是,一些有用的点如:“新”字的点和网址中的点由于过小,伴随干扰被去掉了,在后续研究中还需进一步解决。

图7 与 Ostu 阈值方法比较

在图8中,当文字有多种颜色时,普通的聚类方法将文字外沿的白色作为背景去掉,与此同时,上方的白色文字也被去除,丢失了重要的文字信息;而本文方法则保持了完整的文字信息,表现出良好的适应能力。

图8 与普通聚类方法比较

普通聚类方法和阈值方法在提取文字时主要依据文字和背景在颜色上的差异,当图像质量较低、颜色差异不大时(如图9(a)),则性能很差;而文献[5]的方法,是不依赖文字颜色的方法,该方法利用笔画特征来区分背景和文字,对颜色变化有很强的适应能力,图9(b)给出了提取效果,但是,由于该方法需要笔画宽度先验知识,并不实用,图中可以看到,文字中较宽的笔画也被错误地去掉了,表明算法对笔画宽度过于敏感。图9(c)为本文算法的提取结果,尽管仍保留了少量背景干扰,但文字信息保存完整,算法不需要任何先验知识,实用性高得多。

本质上,本文方法是综合了文字颜色信息和排列形状信息的方法,两类信息相互补充,保证了算法无论是在复杂的颜色变化情形下还是在小对比度、低质量图像环境中都能够取得较好的提取效果,表现出极强的适应能力。

图9 与文献[5]方法比较

4 结语

基于敏感点颜色两级聚类和文本行聚类分析的方法,有效地克服了传统的阈值方法不能适应复杂的背景变化的困难;解决了一般聚类方法容易受大面积背景颜色影响的问题;利用文本行特征取代文字本身特征进行文本筛选的方法,有效地把形状信息结合到文本提取过程中,使文本提取的准确性和适应性明显提高。实验表明,基于敏感点颜色两级聚类和文本行聚类分析的方法,对非常复杂的背景、对比度不明显的文字和质量较差的视频图像都具有极高的适应性。此外,新方法给出的文本行特征的有效表示方法,在其他相关场合也有一定的参考意义。

参考文献:

- [1] WU V, MANMATHA R, RISEMAN E M. Finding text in images [C]// Proceedings of ACM International Conference on Digital Libraries. New York: ACM, 1997: 1-10.
- [2] CHEN D T, BOURLARD H, THIRAN J-P. Text identification in complex background using SVM[C]// Proceedings of Conference on CVPR. Washington, DC: IEEE Computer Society, 2001, 2: 621-626.
- [3] YE XIQIANG, HUANG QINGMING, GAO WEN, *et al.* Fast and robust text detection in images and video frames[J]. Image and Vision Computing, 2005, 23(6): 565-576.
- [4] OTSU N. A threshold selection method from gray-scale histograms [J]. IEEE Transactions on System, Man, and Cybernetics, 1979, 9(1): 62-66.
- [5] DAWOUD A, KAMEL M S. Iterative multimodel subimage binarization for handwritten character segmentation[J]. IEEE Transactions on Image Processing, 2004, 13(9): 1223-1230.
- [6] CHEN D T, ODOBEZ J M, BOURLARD H. Text detection and recognition in images and videos frames[J]. Pattern Recognition, 2004, 37(3): 595-608.
- [7] YE XIQIANG, CHERIET M, SUEN C Y. Stroke-model-based character extraction from gray-level document images[J]. IEEE Transactions on Image Processing, 2001, 10(8): 1152-1161.
- [8] GONZALEZ RC, WOODS RE EDDINS SL. 数字图像处理: MATLAB 版[M]. 吴刚, 等译. 北京: 电子工业出版社, 2004: 270-273, 289-295.
- [9] GAO JIANG, YANG JIE. An adaptive algorithm for text detection from natural scenes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE Computer Society, 2003, 2: 84-89.
- [10] 刘琼, 周慧灿, 王耀南. 基于亮度分级和方向密度的无监督文本定位[J]. 计算机应用, 2008, 28(6): 1523-1526.