第3卷 第10期
1998年10月

中国图象图形学报
Journal of Image and Graphics

Vol. 3, No. 10
Oct. 1998

版面分割中游程平滑后的图文特征分类

张利　朱颖　吴国威

(清华大学电子工程系,北京　100084)

摘　要　游程平滑是版面分割前期常用的一种方法,它将具有较小距离的象素连通在一起形成连通元素。对这些连通元素加以分类是有效划分文本、图象和图形的前提。本文提出了利用几何特征的分类规则以及进一步细分的线结构特征检测方法。

关键词　版面分割,游程平滑,特征分类

1　引　言

游程平滑(Run－length－smearing)是版面分割前期常用的一种方法,它的目的是对版面图象中同一扫描行上黑象素点之间的距离进行检测,以便将具有较小距离的象素连通在一起,为进一步分割做准备。对于图文混合的文本图象,经游程平滑后再对连通域检测,从而从空间位置上将版面图象的内容分成一系列连通元素,连通元素之间以空白区域相分隔。在将连通元素进一步合并成区域之前,首先应对连通元素的图文类型加以区分,过程如图1所示。本文将讨论图文分类及其特征的选择。

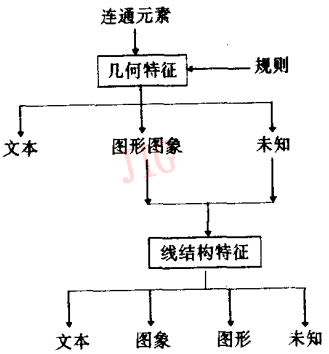


图1　连通元素的分类

2　利用几何特征的分类规则

在二值文本图象中,基于二值信息的几何特征是区分文本、图形和图象的重要特征之一。常用的几何特征包括连通元素的高度、宽度、宽高比、周长、象素点密度等。本文使用这些几何特征,设计了一组规则,对常见的连通元素加以区分。几何特征定义如下:

(1) 高度: Height = CC_bottom-CC_top。(CC是Connected Components的缩写,即连通元素)

(2) 宽度:Width=CC_right-CC_left。

(3) 宽高比:AR = Width/Height。

(4)象素点密度:Density = 黑象素点个数/面积,面积 = 连通域内部的总象素点个数。

使用的规则如下:(TEXT代表文本连通元素,NONTEXT代表非文本连通元素)

规则1　IF (HTmin1<Height<HTmax1) AND (DTmin1<Density<DTmax1) AND (ATmin1<AR<ATmax1),THEN CC→TEXT。

规则2　IF (HTmin2<Height<HTmax2) AND (DTmin2 < Density < DTmax2) AND (AR < ATmax2),THEN CC→TEXT。

规则3　IF (HTmin3<Height<HTmax3) AND (DTmin3 < Density < DTmax3) AND (AR < ATmax3),THEN CC→TEXT。

规则4　IF (HTmin4<Height<HTmax4) AND

收稿日期:1997-12-10;收到修改稿日期:1998-02-11

(DTmin4 < Density < DTmax4) AND (AR > ATmax4),THEN CC→TEXT。

规则5 IF (HTmin5<Height<HTmax5) AND (DTmin5 < Density < DTmax5) AND (AR < ATmax5),THEN CC→TEXT。

规则6 IF (HTmin6<Height<HTmax6) AND (DTmin6 < Density < DTmax6), THEN CC→TEXT。

规则7 IF (HTmin7<Height<HTmax7) AND (DTmin7<Density<DTmax7) AND (ATmin7<AR<ATmax7),THEN CC→TEXT。

规则8 IF (Height<HTmax8) AND (Width<WTmax8),THEN CC→TEXT。

规则9 IF (Width<HTmax8) AND (Height<WTmax8),THEN CC→TEXT。

规则10 IF ((Height < HTmax10a) AND (ATmin10<AR<ATmax10a)) OR ((HTmin10<Height<HTmax10b) AND (AR<ATmax10b), THEN CC→TEXT。

规则11 IF (HTmin11<Height<HTmax11) AND (ATmin11 < AR < ATmax11), THEN CC→TEXT。

规则12 IF (Height > HGmax12) AND (AR<AGmax12),THEN CC→NONTEXT。

规则13 IF (Height>HGmax13) AND (AR>AGmax13),THEN CC→NONTEXT。

规则14 IF (HHmin14<Height<HHmax14) AND(DHmin14<Density<DHmax14) AND (AR>AHmax14),THEN CC→TEXT。

规则15 IF (Height > HGmIN15), THEN CC→NONTEXT。

规则16 IF (Density < DGmin16) OR (Density>DGmax16),THEN CC→NONTEXT。

对上述规则的设计依据如下准则:针对连通元素的常见类型设计过滤原则,如文本字符串、水平或垂直的分割线、文本行中的标点等,逐层过滤。每一种类型的连通元素在其几何特征上与其它类型相异,因而选择能区分开该类型连通元素和其它类型连通元素的几何特征,来设计该种类型的过滤原则。在规则使用的顺序上,对出现频率高的连通元素过滤的原则在前,不常出现的连通元素,其过滤规则排在后。将大部分连通元素经过前几条规则过滤掉,只对少数未知连通元素使用后续规则,以减少处理时间。

在上述规则中,规则1用于过滤文本行中高度、密度和宽高比比较集中的字符串,这部分字符串在连通元素中占大部分。规则2和规则3用于过滤高度较小、长度较短、密度较高的字符串。规则4用于过滤高度较小、长度较长、密度较高的字符串。规则5用于过滤高度较大、密度较低的字符串。规则6用于过滤字符排列稀疏、密度很低的字符串。规则7用于过滤字连通元素中很长的字符串。规则8和规则9用于过滤文本行中的孤立点形成的连通元素,如 i、j 中的点及部分标点符号。规则10和规则11用于过滤文本行中尺寸较小的单个字符形成的连通域,如单独的数学符号、标点符号等。规则12和规则13分别用于过滤图象中较长的垂直和水平直线形成的连通元素,这种直线在排版时通常用来分隔相邻的文本块。规则14用于过滤标题中的大字体。规则15用于过滤一定高度以上的图形图象所形成的连通元素。规则16用于过滤点密度甚低和甚高的图形图象形成的连通元素。

上述规则中参数的选择是对多种杂志、期刊的印刷体文本统计的结果,对印刷体英文文本图象具有一定的稳定性,不随图象分辨率的变化而变化。经过规则判别后,大部分元素被分为文本类(TEXT)和非文本类(NON-TEXT),即图形图象元素,但是仍有小部分元素不满足任何一条规则,被拒识为未知类型(Unkonwn),下面将讨论的特征作第二次分类。

3 线结构特征分类

文献[2]提出了"连通值"(connectivity value)的特征,用于区分图形和图象区域。图形和图象的压缩方法不同,因而作这种区分是有必要的。

我们认为二值图象中文本、图形与图象区域的最主要的差别是内部形状结构的差别。构成文本区域的主要内容是具有一定形状的字符,构成图形区域的主要内容是各种形状的线条,图象区域中象素分布具有一定连续性,线条信息相对较少。针对上述差别,利用形态学中的击中与击不中变换(Hit-or-Miss-Transform)来获取区域中的线条信息。击中与击不中变换定义为:

$$A \otimes B = (A \Theta E) \cap A^{C} \Theta F)$$

A 是离散二维欧氏空间的子集,$B=(E,F)$是一个结构元素对,A^C 是 A 的补集。

击中击不中变换在一次运算中同时捕获到图象内外标记,当且仅当 E 平移到某一点可填入 A 的内

部,F 平移到该点可填入 A 的外部时,该点才在击中击不中变换的输出中。

本文设计了如下 8 个结构元素对,分别用于检测水平、垂直、45°、135°四个方向的线结构信息。其中 1 代表结构元素对中第一个结构元素基元,0 代表结构元素对中第二个结构元素基元,X 不属于结构元素对。结构元素原点在中心。之所以只设置 4 个方向的线结构元素对,是因为在印刷版面中存在在着大量的水平或垂直线用于分割栏目,而且在图形中水平和垂直的线条也占绝大多数。

水平方向的 2 个结构元素对:

$$\begin{bmatrix}1&1&1\\0&0&0\\X&X&X\end{bmatrix}\text{及}\begin{bmatrix}X&X&X\\0&0&0\\1&1&1\end{bmatrix}$$

垂直方向的结构元素对:

$$\begin{bmatrix}X&0&1\\X&0&1\\X&0&1\end{bmatrix}\text{及}\begin{bmatrix}1&0&X\\1&0&X\\1&0&X\end{bmatrix}$$

45°方向的结构元素对:

$$\begin{bmatrix}X&X&0&X&X\\X&1&X&0&X\\X&X&1&X&0\\X&X&X&1&X\\X&X&X&X&X\end{bmatrix}\text{及}\begin{bmatrix}X&X&X&X&X\\X&1&X&X&X\\0&X&1&X&X\\X&0&X&1&X\\X&X&0&X&X\end{bmatrix}$$

135°方向的结构元素对:

$$\begin{bmatrix}X&X&0&X&X\\X&0&X&1&X\\0&X&1&X&X\\X&1&X&X&X\\X&X&X&X&X\end{bmatrix}\text{及}\begin{bmatrix}X&X&X&X&X\\X&X&X&1&X\\X&X&1&X&0\\X&1&X&0&X\\X&X&0&X&X\end{bmatrix}$$

利用上述 8 个结构元素对,以及对文本图象中未知区域进行击中与击不中变换,取输出图象的并集 $g(x,y)=\bigcup f(x,y)\otimes B_n$。

其中,B_n 为上述 8 个结构元素对,则图象 $g(x,y)$ 满足下面 2 个特征:

(1) 图象 $g(x,y)$ 属于图象 $f(x,y)$。

(2) 只有当图象 $f(x,y)$ 中的某一点 (x,y) 处存在水平、垂直、45°、135° 4 个方向的线段信息时,图象 $g(x,y)$ 的象素值才为 1。

线结构特征值定义为:

LSV = 图象 $g(x,y)$ 的面积/图象 $f(x,y)$ 的面积。

其中图象面积定义为图象中取值为 1 的象素点数目。LSV 的取值范围为[0,1]。通过对文本图象中的文本、图象、图形区域统计,结果表明 LSV 是区分文本、图形、图象的一个有效特征。与前面介绍的 CHV 系数相比,不仅能区分出 3 种类型的连通元素,而且 3 种类型的 LSV 值各自相对集中,差别明显。

表 1　三种连通元素类型的 LSV 值分布

区域类型	文本	图形	图象
LSV 取值范围	0.25～0.4	>0.8	<0.1

4　实验结果及结论

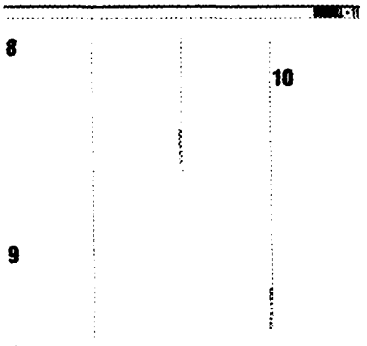
经过几何特征的规则分类和线结构特征的二次分类,文本连通元素和图象连通元素分别进入各自的后续处理过程。我们用此分类方法在 20 余种英文杂志的版面分割中进行了应用,结果是成功的。图 2 给出了一幅图象的分类结果,该图象出自美国《The Sporting News》杂志。可以看出,分类只是把版面分成了零散的文本区、图形区和图象区,为了达到版面分割的目的,还需利用其它知识进一步合并。

但是,本文中有关线的检测也是有缺陷的,因为它只是对 4 个常见方向的线信息检测,而对其它方向的线却不能很好地检测,从而导致对图形的分类错误。解决的方法是利用 HOUGH 变换,但其代价是运算量比较大,花费时间较多。此外,由于有关文本行列的参数是通过对整个版面进行统计的,有可能会把印刷字体较大的题目和段落的首字符划为图形类,这也需要在后续的工作中加以纠正。

(a) 原图

(b) 文本部分

(c) 图象部分



(d) 图形部分

图 2 一个分类的例子

参考文献

1 Hinds S C, Fisher J L. D'Amato D P. A Document Skew Detection Method Using Run-length Encoding and the Hough Transform, Proc. 10th Int. Conf. Pattern Recognition.

2 Fan K -C, Liu C -H, Wang Y -K. Segmentation and classification of mixed text/graphics/images documents. Pattern Recognition Letters 15, 1994:1201～1209.

3 Dougherty E R. An Introduction to Morphological Image Processing. Rochester, New York, 1982.

4 Fan K -C, Wang, L -S. Classification of Document Blocks Using Density Feature and Connectivity Histogram. Pattern Recognition Letters 16, 1995:955～962.

张 利 1987年毕业于清华大学无线电电子学系，1992年获清华大学电子工程系硕士学位后留校任教。已出版著作2部，发表论文10余篇。现主要从事图形图象教学和科研任务，研究方向是图文版面自动分割、图象监控及图象压缩与传输等。

朱　颖　1994年毕业于清华大学电子工程系,1996年获清华电子工程系硕士学位后赴美留学,现为美国普林斯顿大学博士研究生。

吴国威　教授,1958年毕业于清华大学电子工程系,从事的科研工作曾多次获得部委级奖励,并获国家发明专利,现已发表论文数十篇。研究方向为信号与图象处理、计算机视觉、图象识别和人工智能。

Classification of Connected Documents Gotten by Run-length Smearing in Document Segmentation

Zhang Li, Zhu Ying, Wu Gouwei

(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084)

Abstract Run-length-smearing is often used in document segmentation. The most important thing after run-length smearing is to classify the connected components to text, image or graphics. One rule using geometrical feature to classify the components is presented here. A method used to detect line structure in a binary image is also given out in this paper.

Keywords Document segmentation, Run-length smearing, Classification

---

1999年度《CT理论与应用研究》征订

《CT理论与应用研究》杂志于1987年创刊,自1992年1月经国家科委批准为中央级刊物,在国内外公开发行。国内统一刊号为CN11-3017/P,国际标准刊号为:ISSN1004-4140。

由国家地震局地震科学联合基金、中国体视学学会CT理论及应用专业委员会主办。本刊刊登国内外计算机层析成象技术(简称CT)研究方面的新成果,有关各种放射源的成像技术的新方法,用于医学、地球物理、地震、工业、无损探伤、石油采矿层析成像理论与应用研究的新成果与综述性文章,逐步将CT理论与应用方面的知识系统化,促进国内外的学术交流。每篇论文必需中英文对照标题,中英文摘要作者姓名、邮编、地址。该刊国内外从事计算机层析成像技术及应用研究的科技工作者、医疗CT方面的科技人员、医务人员、有关大专院校的教师、研究生、大学生科研教学之用。本刊顾问有:陈运泰,王乃彦等院士,及中华医学的专家刘赓年,李松年教授等。有德国和香港的顾问,在全世界发行。主编:郭履灿研究员。

本刊为季刊,16开本,正文51页,封1-4。每本定价6.00元,全年4期26元(含邮资),邮局汇款可直接寄到　100081　北京市民族学院南路5号《CT理论与应用研究》编辑部收。本刊编辑部可以开寄正式收据。银行汇款请汇往:北京市复兴路63号地震科学联合基金会,邮政编码:100036,银行帐号为:中国工商银行北京翠微路分理处144420-82《CT理论与应用研究》编辑。