

文章编号:1001-9081(2010)05-1259-03

基于连通域分析和支持向量机的传真图像关键词定位

蔡 锋, 刘立柱

(信息工程大学 信息工程学院, 郑州 450002)

(caifeng3222@163.com)

摘 要: 电话号码区域定位是传真图像电话号码识别中的关键技术之一。首先采用连通域分析对传真图像实现较为精确的版面分析, 形成比较完整的单词连通域, 提取单词连通域的水平穿越次数和空间分布特征, 形成51维的特征向量。采用基于正态决策树的多分类支持向量机(SVM), 来完成对传真图像电话号码区域关键词的定位。实验结果表明, 算法能够快速有效地完成关键词的定位, 具有较强的实用价值。

关键词: 连通域分析; 水平穿越次数; 空间分布特征; 支持向量机; 关键词定位

中图分类号: TP391 **文献标志码:** A

Key words location of the fax images based on connected component analysis and SVM

CAI Feng, LIU Li-zhu

(College of Information Engineering, Information Engineering University, Zhengzhou Henan 450002, China)

Abstract: Locating the telephone number region is a very important technology in telephone number recognition of the fax images. After realizing a relative precise page analysis on the fax images by adopting the Connected Component Analysis (CCA) to form comparatively whole word regions, the features of horizontal traversing times and spatial distribution were abstracted to form feature vector of fifty-one dimensions. The multi-class Support Vector Machine (SVM) based on normal decision tree was introduced to achieve the key words location. The experimental results show that the method can realize the location quickly and effectively, and it is valuable in applications.

Key words: Connected Component Analysis (CCA); horizontal traversing times; spatial distribution feature; Support Vector Machine (SVM); key words location

在信息化社会和知识经济时代, 图像通信具有广阔的市场和应用规模^[1]。面对需要管理的大量图像信息, 基于用户特征的自动筛选与检索研究逐渐成为热点。二值传真图像中的电话号码就是用户的一个特征, 基于用户电话号码自动识别的图像信息筛选与检索是本文研究的内容。

电话号码自动识别技术, 一般由电话号码区域定位、字符分割以及数字识别三个部分组成。图1是基于用户电话号码的图像筛选与检索系统的结构框架。其中的电话号码区域定位是电话号码识别中的关键技术之一, 其准确与否将直接影响到电话号码识别的准确率。实际应用中, 经译码恢复得到的传真报文, 图像质量较差, 很大程度上影响了识别效果。合理有效的电话号码区域定位算法, 可以提高识别的准确率。本文要探讨的就是基于电话号码前关键词的区域定位算法。

1 连通域特征提取

传真图像包含了大量的信息, 其内容可能包括徽标、印章、签名、表格或者图片等, 但大部分还是以文字居多。通过对传真图像进行较为精确的版面分析, 剔除面积大且黑像素密度小的连通域, 留下小的字符连通域。通过对这些连通域的特征提取, 来完成关键词的定位。

1.1 线性整形归一化

不同字符大小以及单词长度造成了连通域面积上的差异, 为满足特征提取需要相同大小字符点阵的要求, 考虑到基于线性整形归一化算法简单、易于实现, 且运算量小的特点,

对样本进行线性整形归一化, 以提高系统对不同字体、字号的稳定性, 并克服图像噪声带来的诸多影响。针对所选样本集, 通过实验发现: 归一化为 64×16 大小的字符点阵, 字符不发生像素笔划的丢失, 具有良好的适应性。

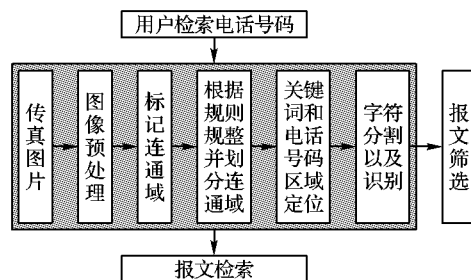


图1 基于用户电话号码的传真报文筛选与检索系统框架

下面介绍整形归一化的数学原理。假设原始图像 $f(i, j)$ 的坐标用 (i, j) 表示, 其大小为 $I \times J$, 归一化后的图像坐标用 (m, n) 表示, 其大小为 $M \times N$, 则原图像中点 (i, j) 在新图像中坐标 (m, n) 的计算公式如式(1)所示:

$$\begin{cases} m = \sum_{k=1}^i H(k) \cdot \frac{M}{\sum_{k=1}^I H(k)} \\ n = \sum_{k=1}^j V(k) \cdot \frac{N}{\sum_{k=1}^J V(k)} \end{cases} \quad (1)$$

收稿日期: 2009-10-30; 修回日期: 2010-01-04。

作者简介: 蔡锋(1984-), 男, 江苏泰州人, 硕士研究生, 主要研究方向: 网络传真数据处理; 刘立柱(1949-), 男, 河北邢台人, 教授, 博士生导师, 主要研究方向: 传真通信, 网络数据处理。

其中函数 $H(i)$ 和 $V(j)$ 分别表示特征点在 X 轴和 Y 轴上的投影特征函数。从式(1)可以看出,归一化后图像的特征在垂直方向和水平方向具有相同的密度。

线性归一化的投影特征函数为:

$$\begin{cases} H(i) = 1, & i = 1, 2, \dots, I \\ V(j) = 1, & j = 1, 2, \dots, J \end{cases} \quad (2)$$

代入式(1),即可得到线性归一化的计算公式为:

$$\begin{cases} m = i \times \frac{M}{I}, & i = 1, 2, \dots, I \\ n = j \times \frac{N}{J}, & j = 1, 2, \dots, J \end{cases} \quad (3)$$

图2为将样本用线性整形归一化方法归一化为 64×16 大小的字符点阵的示意图。



图2 单词连通域的线性整形归一化处理

1.2 连通域特征

选取空间分布特征和水平穿越次数特征作为连通域特征。

空间分布特征^[2]:对于一幅归一化为 64×16 像素大小的图像,选取定义的 $x_1 \sim x_{44}$ 共44个特征,如图3所示。这些特征分别定义为:将 64×16 的矩形方阵均分为32个区域,即每个区域是 8×4 的像素大小,其中, $x_1 \sim x_{32}$ 是连通域中落入这32个区域的黑像素个数除以每个区域的总像素数所得到的每个区域中的黑像素密度向量; $x_{33} \sim x_{36}$ 为水平块4个区域的密度向量; $x_{37} \sim x_{44}$ 为垂直块8个区域的密度向量。

空间分布特征是结构特征和统计特征相结合的产物,增强了特征的抗干扰性。

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32

图3 空间分布特征示意图

水平穿越次数特征:图像中水平方向像素点由黑至白或者由白至黑的次数。对一幅归一化为 64×16 像素大小的图像,分别从图像的 $n/8$ 高度处(n 为整数,且 $n = 1, 2, \dots, 7$),进行水平扫描,求得 $x_{45} \sim x_{51}$ 共7个特征。该特征主要用来区分长短词条连通域,以完成词条的聚类,降低分类超平面形成误差。

2 基于连通域和 SVM 的传真图像关键词定位

通过大量的传真图像统计发现,用户电话号码前面通常会跟有“传真”或者“电话”的英文或者俄文单词的大小写全名或者缩写(只针对英文、俄文和中日韩以及一些东南亚的传真报文),诸如“Fax”、“Tel”、“PHONE”等,因此电话号码区

域定位便归结到对这些词条的定位。

2.1 基于连通域的传真报文版面分析

版面分析是文档图像信息处理的关键,它的任务是利用计算机自动对文本图像进行分析和处理,将文本图像版面划分为文本、表格、图像等不同类别的属性区域。传真图像作为文本图像,精确的版面分析有助于文种识别、表格提取、徽标检索等应用研究的需要。其研究方法主要分为两种:自顶向下和自底向上。自顶向下利用的是文本图像的全局信息,而自底向上则是重视局部信息,经过像素→字符连通域→连通域合并划分的顺序,是一个由局部到整体的过程。本文采用文献[3]提出的方法,即采取以自底向上为主并结合一些统计信息的方法。

首先,对于一幅传真图像,假定已经过了预处理,包括二值化、干扰噪声滤波以及倾斜校正,算法的下一步是根据连通域的宽度和面积等特征来进行聚类分析。因此在进行连通域分析之前,先对以下特征进行统计:字符的最大宽度($MaxWidthOfCharacter$),字符的最小宽度($MinWidthOfCharacter$),文字行之间的最大间隔($MaxMarginOfLines$),文字行之间的最小间隔($MinMarginOfLines$),同一行之中单词之间的最大间隔($MaxMarginOfWords$),同一行之中单词之间的最小间隔($MinMarginOfWords$),单词的最大面积($MaxArea$),单词的最小面积($MinArea$),单词的最大宽度($MaxWidth$),单词的最小宽度($MinWidth$)。

本文所说的连通域指的是八连通域,从一个黑像素出发,判断周围八个方向是否存在相邻的黑像素点。查找连通域的方法描述如下:1)连通域标记,确定各连通域的矩形边界,并保存为边界参数链表,本文采取文献[4]提出的方法,综合了线标记和区域增长标记的优势,做到了一次性扫描标记,并不受标记的区域性状影响,具有很好的鲁棒性;2)连通域搜索,连通域的搜索顺序为逐扫描线自左至右、自上而下;3)连通域排序,使同一字符的连通域在链表中处于相邻的位置。

英文中有一些字母,如“i”、“j”由上下两部分组成,因此形成了两个连通域,其链表位置是不相邻的,应重新对链表进行排序。徽标和图像则由于组成成分比较多,不像字符具有规则性,一般会由许多个重叠的连通域组成。因此在连通域标记完毕之后,要进行连通域的合并,首先根据行间最小间隔($MinMarginOfLines$)进行上下相邻两个连通域的合并,这样每一行内的连通域只剩下左右相邻一个状态。

由于算法的最终目的是对关键词条“Fax”等的正确区域定位,经大量统计发现,报文中影响关键词条准确定位的情况主要是由于一些标点符号的影响,主要有:“fax:”,“Tel/Fax”,“fax.”,同时由于传真图像质量问题而存在的一些噪声点,这对关键词条连通域的正确形成有很大的影响。而实际传真图像中有一些文本经常包含在表格和边框内,为此统计三个连通域特征:连通区域面积、连通区域密度以及连通区域宽高比。连通区域密度指的是连通区域黑像素数与整个连通域的总像素数之比,表格、边框以及反斜杠呈现明显的低密度特征,而下划线等呈现较大的宽高比,一些噪声点形成的连通域则存在面积小的特点。设定面积阈值 A 、密度阈值 D 和宽高比阈值 R 。凡面积小于 A 、密度小于 D 以及宽高比大于 R 的连通域就进行置白处理,这样处理可以消除标点符号和一些小

噪声点的影响。

然后根据 MinMarginOfWords 合并左右相邻连通域以得到完整的连通域。接着根据连通域的面积 MaxArea 和宽度 MaxWidth 进行聚类,将传真报文版面分为文字、徽标、表格、图片等,同时完成根据面积对划分为文字的连通域进行的标记。

2.2 基于决策树 SVM 分类器的电话号码前关键词定位算法

支持向量机(Support Vector Machines, SVM)^[5]是根据统计学习理论提出的一种机器学习方法。它基于结构风险最小化原则,不过分依赖样本的数量和质量,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势。其基本思想是通过由内积函数定义的非线性变换将输入空间映射到高维特征空间,然后在高维空间中寻找使训练数据分类间隔最大的广义最优超平面来实现特征的分类。

基于决策树 SVM 分类器^[6]的电话号码前关键词定位算法,以传真图像中用户电话号码前的关键词条为训练样本,要求应尽可能多地囊括这种关键词条的样本,主要针对的是俄文和英文的词条,选取 TEL, Tel, ТЕЛ, Тел, FAX, Fax, ФАКС, Факс, PHONE, ТЕЛЕФОНА 共 10 个关键词条样本,提取其水平穿越次数和空间分布特征,通过训练样本集得到最优分类函数,并通过训练好的支持向量机对关键词连通域进行定位,具体过程如下。

1) 选取训练样本,生成训练矩阵。从传真图片库中,手工裁剪出关于 10 个关键词条的样本,其中每个关键词条各 5 个样本,共 50 个样本,统计其宽高比,求得一个阈值(2.14)。然后手工裁剪出 50 个非关键词条样本,宽高比限定在(2, 14),对其进行线性整形归一化,并提取两类特征,以构成 50 × 51 的训练矩阵 E 。

2) 样本特征归一化。计算样本特征最大值 X_{\max} 和最小值 X_{\min} ,将特征值 X 与 X_{\min} 之差除以 X_{\max} 与 X_{\min} 之差,即可将特征值归一化到区间 $[0, 1]$ 。

3) 构造和训练多分类支持向量机。采用基于“正态决策树”的多分类支持向量机^[7],选取径向基(Radial Basis Function, RBF)核函数。将训练矩阵作为支持向量机的输入参数,先经一级分类器将样本分为从 TEL 到 ФАКС 的八个特征词条和类似宽高比的短样本,以及 PHONE、ТЕЛЕФОНА 两个特征词条和类似宽高比的长样本。然后分别训练短词条样本和长词条样本之间的分类器,得到该训练集下的最优分类函数 $F(x)$,以区分出非关键词条和关键词条。如图 4 所示。

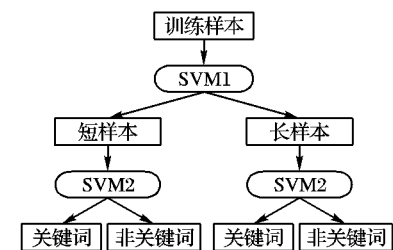


图4 基于正态决策树的多分类支持向量机

4) 关键词条区域定位。对一幅已做好版面分析的传真图像,对标记好的字符连通域进行宽高比的计算,对满足域值(2, 14)的进行线性整形归一化,然后进行两类特征的提取,构成 51 维目标特征矢量。将目标特征矢量输入支持向量机,完成关键词的分类与定位。

3 实验结果与分析

为了检验基于连通域与 SVM 的传真图像关键词定位效果,选取了包含表格、徽标、文字、印章、手写签名在内的 100 幅传真图像。经过人工统计,共含有各类关键词条共 188 个。使用多分类 SVM 进行定位,以检索精度和检索召回率来评价。检索精度定义为检索出的有效图像数与检索出的图像总数的比率,而检索召回率定义为检索出的有效图像数与图像库中总有效图像数的比率。共 188 个关键词条,其中 FAX 占主要部分,最后算得检索精度为 86.6%,召回率为 93.1%。

应该指出:精度不是很高,主要是由传真图像本身的特性所决定,报文所存在的噪声问题对连通域的标记和特征的提取有着很大的影响。噪声的影响有两点:一是由于噪声使得关键词连通域划分不准确,而使得后面在筛选连通域以及进行特征提取时出现了漏定位;二是噪声改变了关键词的空间特征,将不是关键词的空间特征变得相近,关键词连通域的特征空间则与样本集大相径庭,这时候就出现了漏定位和错定位。错定位还有一种可能,就是报文中非电话号码前存在关键词。同时,短词条的定位效果优于长词条,这与归一化尺寸有关,短词条的宽高比更加接近归一化比例,这能够更好地保持原始信息,有利于特征的有效提取。如何更好地整形归一化,如何结合传真图像版面特征更快更精确地获得字符连通域,以及如何结合更多的特征形成一个特征的优化组合,是要继续研究的工作。

4 结语

本文通过对传真图像连通域的分析,对传真图像实现了较为精确的版面分析,提出了以 51 维由单词连通域的水平穿越次数以及空间分布特征形成的特征向量,来定位电话号码前关键词的方法。定位过程中,通过自底向上的连通域分析和一些先验知识,实现版面的精确分析和对文本区域单词的连通域标记。针对训练样本不好、搜集数量较少的问题,采取能够在小样本集下能够取得较好性能的多分类 SVM,完成了对关键词的定位。实验结果表明,该方法能够对关键词取得较好的定位效果,具有较强的实用性。

参考文献:

- [1] 刘立柱. 传真图像和传真信号处理原理与技术[M]. 北京: 国防工业出版社, 2006.
- [2] 吴谨, 邱亚. 基于空间分布特征的手写体数字识别[J]. 武汉科技大学学报, 2004, 27(2): 176 - 178.
- [3] 魏宏喜, 高光来. 一种基于连通域的蒙古文文档图像版面分析方法[J]. 内蒙古大学学报: 自然科学版, 2007, 38(5): 586 - 590.
- [4] 高红波, 王卫星. 一种二值图像连通区域标记的新算法[J]. 计算机应用, 2007, 27(11): 2776 - 2777.
- [5] WANG F, VUURPIJ L, SCHOMAKER L. Support vector machines for the classification of western handwritten capitals [C]// IWFHR 2000: Proceedings of the seventh International Workshop on Frontiers in Handwriting Recognition. [S. l.]: Nijmegen Institute for Cognition and Information, 2000: 167 - 176.
- [6] 尹叶飞, 吴秀清. 基于决策树 SVM 分类器的感兴趣区域定位方法[J]. 计算机仿真, 2007, 24(1): 209 - 212.
- [7] 程娟, 平西建, 周冠玮. 基于多特征和 SVM 的文本图像版面分类方法[J]. 数据采集与处理, 2008, 23(5): 569 - 574.