自然场景下基于连通域检测的文字识别算法研究

刘新瀚,钱 侃,王宇飞,朱向霄,孙知信
(南京邮电大学 物联网学院,江苏 南京 210003)

摘 要:自然场景中由于文字背景的复杂性等原因,给文字识别工作带来了极大的困难。文中提出一种边缘检测与连通域分析相结合的算法以识别自然场景中的文字,提高文字识别的准确率与召回率。首先采用 ColorRoberts 算子直接对彩色图像进行边缘检测,从而避免彩色图像转换为灰度图像过程中的信息丢失现象;然后对检测出的图像边缘进行去除长直线、去除孤立的噪声点、形态学运算的后续处理操作;最后经过连通域标记、分析,提取出文本区域。通过仿真实验,结果表明了该算法的合理性和有效性。

关键词:边缘检测;ColorRoberts 算子;连通域标记;连通域分析
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2015)05-0041-05
doi:10.3969/j.issn.1673-629X.2015.05.011

Research on Character Recognition Algorithm Based on Connected Domain Detection in Natural Scene

LIU Xin-han, QIAN Kan, WANG Yu-fei, ZHU Xiang-xiao, SUN Zhi-xin
(College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Due to the complexity of text background in natural scene, it brings great difficulties to the character recognition. In this paper, propose an algorithm combined edge detection and connected component analysis to identify the text in natural scene, which improves the accuracy and recall rate of text recognition. Firstly, the ColorRoberts operator is used to detect the edge of color image directly in this algorithm, thus avoiding the information loss in the process of conversion from a color image to a gray image. Secondly, the subsequent processing operation, such as removing the long straight lines, removing isolated noise points and morphological processing, is actualized on the image edge which has been detected. Finally, through the connected domain labeling and analyzing, the text regions are extracted. Simulation results are given, which shows that the algorithm is reasonable and effective.

Key words: edge detection; ColorRoberts operator; connected domain labeling; connected domain analysis

0 引 言

自然场景蕴含了大量的文本信息,比如,车牌号、交通标志、广告牌、登机牌。这些含有丰富语义信息的文字,对人们理解自然场景提供了便利。

自然场景中文本识别技术的核心是文本的定位与识别,其中文本识别的光学字符识别技术(OCR),经过学者的几十年不懈努力已经比较成熟[1],其针对文档图像识别有着良好的性能。而在自然场景下直接使用 OCR 进行文字识别准确率很低,这是因为在自然场景下图像更为复杂,表现在以下几个方面:

(1)文档图像通常为白底黑字,而自然场景中的文本其背景复杂,可能包含建筑物边缘等与文字结构类似的背景。

(2)文档图像中文字结构清晰、辨识度高,而自然场景中可能由于光照不均匀、拍摄角度不当、包含大量噪声等问题,使图像中文字产生变形、模糊不清等现象。

(3)文档图像中文字布局按一定规律从上到下、从左到右依次排列,而自然场景中文字排列顺序任意,在图像中位置也不确定,这在一定程度上给识别带来

收稿日期:2014-04-07 **修回日期**:2014-07-10 **网络出版时间**:2015-04-22
基金项目:国家自然科学基金资助项目(60973140,61170276,61373135);江苏省产学研项目(BY2013011);江苏省科技型企业创新基金项目(BC2013027);江苏省高校自然科学研究重大项目(12KJA520003)
作者简介:刘新瀚(1992-),男,硕士,研究方向为图像处理和模式识别;孙知信,博士,教授,研究方向为计算机网络与安全、计算机仿真、软件工程。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150422.1113.030.html

了极大困难。但是自然场景中文字也有其独特特征，比如文字边缘比较明显，文本区域通常有着相似的色调和亮度，相邻字符距离间隔、紧密性相当，这些都给文字识别工作提供了便利。

对自然场景中文本信息的提取有着广泛的应用前景，主要应用在网络安全系统、视觉感知辅助、基于内容的图像和视频资料检索、书籍及历史文献的电子化等领域。近年来国内外许多学者对此做了大量研究，虽然在文本识别的准确率和召回率方面有了明显的改善，但是相关技术仍停留在实验室阶段，还没有达到商业化应用的要求。

文中针对文本识别过程中的文本定位算法进行研究，在边缘检测的基础上采用基于连通域检测的算法，以适用于移动终端。

1 相关研究

大多数文本定位算法都是基于文本特征进行文本定位的，根据同一区域字符颜色的一致性提出基于连通域的文本定位方法，根据字符特有的纹理特征提出基于纹理的文本定位方法，根据字符含有大量边缘的特征，提出基于边缘的文本定位方法。

1.1 基于连通域的文本定位方法

基于连通域的文本定位方法认为文本区域具有一致的颜色，利用字符颜色与背景有一定的对比度这一特征分割图像，然后对分割后的图像进行连通域分析，得到许多个连通域，并作为候选字符连通域[2]。再利用文本的几何特征、前景点与背景点数目比等特征排除非文本连通域，从而得到文本区域[3]。

基于连通域的文本定位方法适用于对背景单一、光照均匀、字符颜色一致的文本进行定位，文本定位速度较快[4]。但是在复杂的背景中，一些类似于文本的背景图案可能会被错误地定位在文本区域，定位过程中连通域最优阈值的大小也很难确定。

1.2 基于纹理的文本定位方法

基于纹理的文本定位方法认为图像中的文本区域具有特殊的纹理，纹理产生的原因是由于字符有一定的排列方向、字符颜色和背景颜色呈周期性变化[5]。该方法的基本步骤为：首先将图像分割成若干个不重叠的子区域，再利用诸如 Gabor 滤波、小波变换等方法获得子区域的纹理特征，最后用分类器根据子区域的纹理特征对其进行分类，得到最终的文本区域。

基于纹理的文本定位方法适用于对自然场景中较小的字符进行定位，在背景简单的场景下能有效定位出文本区域[6]。但是该方法在对图像进行处理的过程中用到复杂的分类器、滤波器，致使定位耗时过长，不适宜在资源有限的移动终端上应用。

1.3 基于边缘的文本定位方法

基于边缘的文本定位方法认为自然场景中的文字与背景之间具有一定的差别，该方法利用字符具有丰富的边缘信息进行检测，可以有效地检测到字符的边缘。常用的边缘检测算子有 Robert 算子、Sobel 算子、Laplace 算子等[7]。Robert 算子具有边缘定位准确的特点，缺点是对噪声敏感；Sobel 算子在实际应用中最为常用，其采用了加权的模板系数，一定程度上能够抑制噪声；Laplace 算子是一个标量二阶微分算子，具有各向同性的性质，但该算子对噪声非常敏感[8]。自然场景中的图像为彩色图像，需要先将彩色图像转换为灰度图像后，才能利用这些算子进行边缘检测，这就不可避免地造成了彩色图像的信息丢失，因此张云鹤等[9]提出了 ColorRoberts 算子，可以直接针对彩色图像进行检测。

基于边缘的文本定位方法具有计算量小、文本检测速度快的优点。但是对于背景复杂的自然场景而言，其背景中也含有丰富的边缘信息，这些背景的边缘影响文本的准确定位。

基于以上分析，边缘检测算法文本定位速度快，适合应用于资源有限的移动终端上，但是检测出的背景边缘会影响文本定位的准确性。图像经过边缘检测后得到背景单一的二值化边缘图像，用基于连通域的文本定位方法可以排除背景边缘，得到最终的文本区域。

2 文本定位算法

2.1 算法总体设计

文中采用边缘检测与连通域分析相结合的方法对自然场景中的文字进行检测。文本定位算法流程图如图 1 所示。

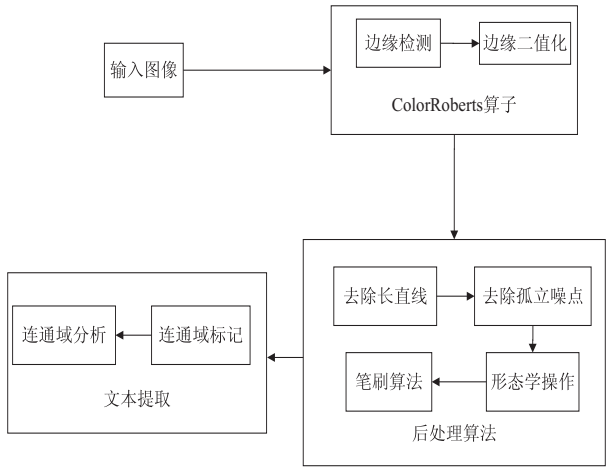


图 1 文本定位算法流程图

下面阐述文中算法：通过前面的分析可知，自然场景中文本区域有着较高的边缘密度，字符边缘与背景有着较为明显的颜色差别。边缘检测算法其时间复杂

度相对较低,运算量小,适合在硬件资源有限的移动终端上实现。传统的边缘检测算法大多是针对灰度图像进行检测,在把彩色图像转换为灰度图像的过程中会造成颜色信息的丢失,这不利于文本的准确定位。而文中采用的 ColorRoberts 算子可以直接针对彩色图像进行边缘检测,有效避免了颜色信息丢失的现象。

在进行边缘检测以后,图像中不可避免地产生了噪声,这些噪声严重影响了文本定位的准确性。文中对图像中的噪声进行去除长直线、去除孤立的噪声点、形态学运算的后续处理操作,以减少图像中的噪声对文本定位产生的影响。

经过后续处理操作的图像,其背景单一,符合连通域检测的要求。最后利用连通域标记算法(二次扫描算法)对去除噪声后的二值图像进行标记,再用先验知识进行连通域分析来剔除这些非文本区域,得到最终的文本区域。

2.2　ColorRoberts 算子检测算法

ColorRoberts 算子的基本思想为:ColorRoberts 算子融合了 Roberts 和 LOG(Laplacian-Gauss)两种算子,并结合像素点之间的彩色值欧氏距离综合考虑进行边缘检测[9]。具体检测流程图如图 2 所示。

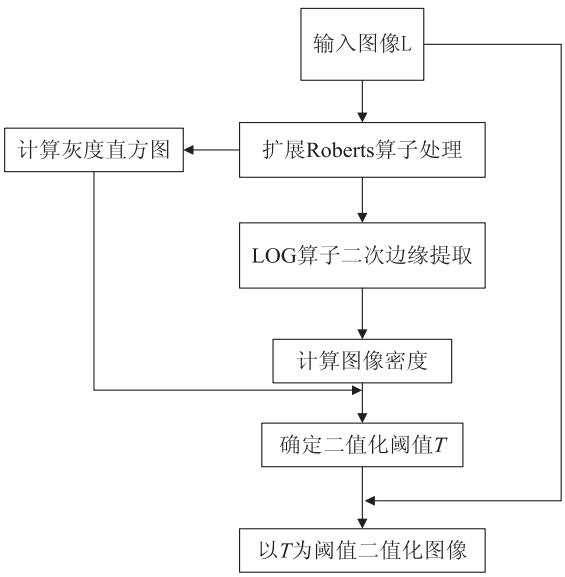


图 2　ColorRoberts 算子检测流程图

2.3　后处理算法

2.3.1　去除长直线

自然场景中的图像,经过 ColorRoberts 算子边缘检测后,二值化图像中会出现水平或竖直的长直线。这些长直线大多是背景图像的边缘,给下一步的文本定位工作带来影响,所以必须找到一种有效的方法去消除这些长直线。长直线的去除步骤[10]如下:

(1)由于自然场景中的文本一般位于图像的中心方位,为了减少计算量,将二值化图像的最上列、最下列、最左列、最右列像素点设为背景点(灰度值为 0)。

(2)在垂直方向长度为 3 个像素范围内的区域,检测水平方向上连续的前景点个数。如果个数大于图像水平宽度的五分之一,则判定其为水平长直线,则将该连续前景点的灰度值置为 0。

(3)在水平方向长度为 3 个像素范围内的区域,检测竖直方向上连续的前景点个数。如果个数大于图像竖直宽度的五分之一,则判定其为竖直长直线,则将该连续前景点的灰度值置为 0。

2.3.2　去除孤立的噪声点

文中定义白色为前景点,黑色为背景点,所谓孤立的噪声点就是孤立的前景点。如果某前景点 M,在它的 8 邻域内其他前景点个数小于 $N*255$(N 为阈值,值为 5),则该前景点为孤立的噪声点。在图像二值化过程中会不可避免地引入噪声点,噪声点的有效去除将增加文本定位的准确性。

孤立的噪声点去除方法如下:按照从左往右、从上到下的原则,依次遍历二值图像中的前景点,检测其 8 邻域内其他前景点个数。若 8 邻域内存在其他前景点,且前景点个数大于 $N*255$,则跳过该点,继续检测下一前景点。若 8 邻域内其他前景点个数小于 $N*255$,说明该点为孤立噪声点,将其灰度值由 255 变为 0,之后继续检测下一前景点。

2.3.3　形态学操作

在图像二值化、去除长直线、去除孤立噪声点过程中,都可能造成字符笔画的断裂。这给文本定位工作带来了困难,有可能使一个字符被定位在两个不同的文本区域,必须要对图像做进一步处理才能提高定位的准确性,所以文中引入了形态学操作。经过形态学操作,图像中断裂的字符笔画可以粘连,并使图像得到简化。形态学的基本操作[11]包括膨胀、腐蚀、开运算、闭运算等。

膨胀是使物体边界向外扩展,可以通过膨胀操作连接两个较近的区域。也可以去掉物体内部细小的空洞。其定义为:

$$R = M \oplus N = \{(p,q) \mid N_{pq} \cap M \neq \varnothing\}$$

其含义为,二值图像 R 是 N 腐蚀 M 后产生的,它是满足以下条件的点 (p,q) 的集合:如果 N 的原点平移到点 (p,q),那么它与 M 的交集非空。

腐蚀可以使边界向内收敛,消除二值化图像中的边界点,也可以用来消除小的噪声点。其定义为:

$$H = M \otimes N = \{(p,q) \mid N_{pq} \subseteq M\}$$

其含义为,二值图像 H 是 N 腐蚀 M 后产生的,它是满足以下条件的点 (p,q) 的集合:如果 N 的原点平移到点 (p,q),那么 N 将完全包含于 M 中。

先腐蚀后膨胀的过程称为开运算,开运算可以消除细小的噪声区域,使大物体边缘变得平滑。先膨胀

后腐蚀的过程称为闭运算,闭运算可以填补物体的细小空洞,使邻近区域相连。

文中形态学操作为,在水平方向和竖直方向进行 3 个像素的开运算,用以消除细小的物噪声区域;在水平方向和竖直方向进行 3 个像素的闭运算,用以连接断裂的字符笔画。

2.3.4 笔刷算法

二值图像在经过形态学操作处理后,仍存在字符笔画断裂的现象。文中引入一种笔刷算法[12],可以使字符邻近的笔画相连接。这里采用横向笔刷刷图,公式如下:

$$\mathrm{BL}(s,t)=\begin{cases}1, \sum_{m=-M/2}^{m=M/2} L(s+m,t) > 0\\ 0, \sum_{m=-M/2}^{m=M/2} L(s+m,t) = 0\end{cases}$$

式中, $\mathrm{BL}(s,t)$ 是经笔刷算法处理后的图像; $L(s,t)$ 为待处理图像; M 为横向笔刷的宽度。

2.4 文本提取

2.4.1 连通域标记

文中采用的连通域标记算法为二次扫描算法[13]。定义几个变量: $\mathrm{BI}(m,n)$ 表示二值化图像像素点 (m,n) 的灰度值,1 表示前景点,0 表示背景点。$P(m,n)$ 为二维数组,用来记录像素点 (m,n) 的连通域标号,在第一次扫描后记录的是像素点的临时连通域标号,在第二次扫描后记录的是像素点的最终连通域标号。$C(I)$ 为一维数组,其存储的是属于同一个连通域的子连通域的共同连通域标号,其中 I 为子连通域的临时连通域标号。

下面对算法进行介绍:

步骤一:用 4-邻域规则对二值图像 $\mathrm{BI}(m,n)$ 进行第一次扫描,并用二维数组 $P(m,n)$ 进行像素点的标记。同一个连通域内被标记的像素点属于等价标号的像素点,必须将这些不同数值的等价标号统一为同一数值的等价标号,此时的等价标号为共同连通域标号。其方法为:用一维数组 $C(I)$ 存储子连通域的共同连通域标号,一维数组的值就是共同连通域标号,其下标为子连通域的临时连通域标号。对于等价标号的处理,重复扫描一维数组 $C(I)$,将同一个连通域中,子连通域的共同连通域标号改为同一数值。

步骤二:此步骤将实现连通域的合并。扫描数组 $P(m,n)$,将像素点临时连通域标号替换为共同连通域标号,使连通域得以合并。合并后,二维数组 $P(m,n)$ 中的连通域标号就是最终得到的连通域标号。

在第一次扫描过程中,若某一像素点 $\mathrm{BI}(m,n)=1$,且在它的 4 - 邻域中,上邻域 $\mathrm{BI}(m,n-1)=1$,左邻域 $\mathrm{BI}(m-1,n)=1$,也即这三个像素点都属于同一个连通域的目标像素点。当 $C(P(m,n-1)) \neq C(P(m-1,n))$ 时,则上邻域和左邻域的共同连通域标号的值不一致,遍历一维数组 C,按如下方式处理等价标号冲突:

```
temp1 = C(P(m,n - 1));
for i = 1:1:newlable
if( C(i) = =temp1)
C(i) = C(P(m - 1,n));
end
end
P(i,j) = P(m - 1,n);
```

2.4.2 连通域分析

在进行连通域标记后,矩形框中的区域就是备选文本区域。而这些区域中一部分是非文本区域,需要根据一些先验知识来剔除这些非文本区域,以排除干扰。为此,制定如下约束条件,不满足以下条件的则视为非文本区域:

(1)矩形框中,前景点的像素个数与矩形框中总像素个数之比大于 40%;

(2)矩形框中像素个数小于图片总像素个数的 80%;

(3)矩形框的宽度与高度之比在 1/15 到 15 之间;

(4)矩形框的宽度(高度)大于图像的宽度(高度)的 1/24 且小于 1/5。

在非文本区域被剔除后,图像中的文本区域矩形框可能存在彼此之间覆盖或者相交的现象,需要将这些矩形框进行连通域合并。对于矩形框覆盖现象,保留大的矩形框去掉小的矩形框。对于矩形框相交现象[14],若两个矩形框 D_1 和 D_2 满足如下公式,则进行矩形框合并操作:

$$\left[\frac{S(D_1 \cap D_2)}{\min(S(D_1),S(D_2))} > 0.2\right] \cup \left[\frac{S(D_1 \cap D_2)}{\min(S(D_1),S(D_2))} > 0.1 \cap \frac{\max(S(D_1),S(D_2))}{\min(S(D_1),(D_2))} > 10\right]$$

式中,矩形框 D 的面积表示为 $S(D)$,两个矩形框相交区域表示为 $D_1 \cap D_2$。

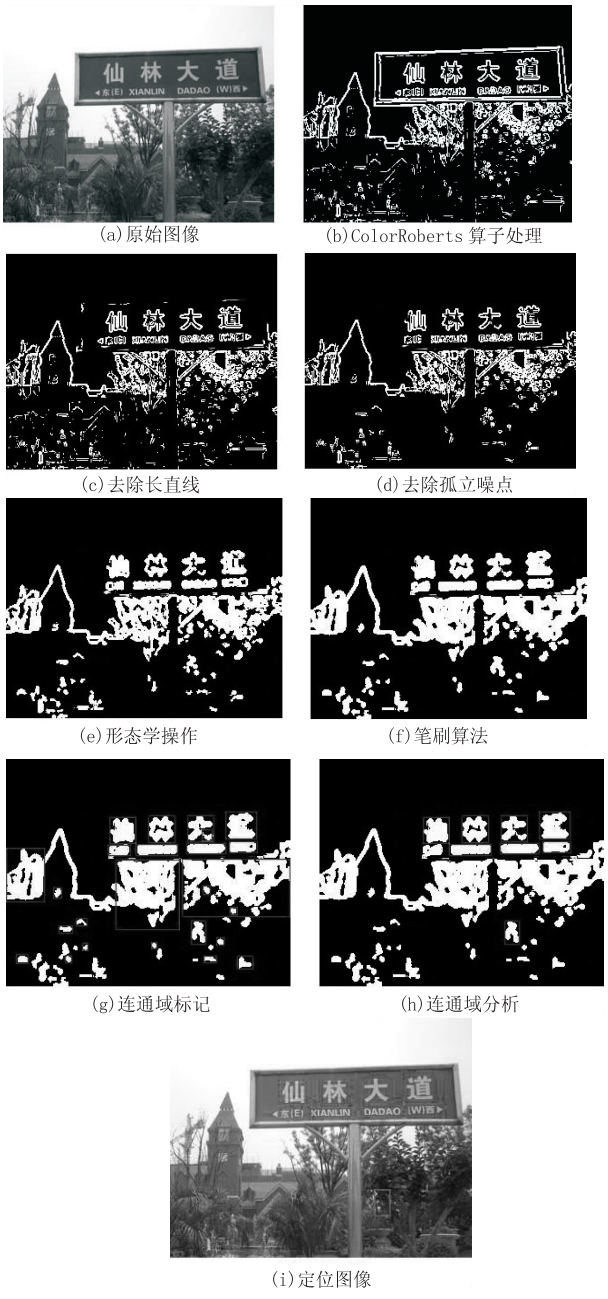
3 仿真实验

3.1 仿真环境

为了验证文中算法的性能,利用 Matlab 进行了仿真实验。

3.2 实验结果与分析

原始图像经过上述算法处理后,每一步的处理结果如图 3 所示。



(a)原始图像　(b)ColorRoberts 算子处理

(c)去除长直线　(d)去除孤立噪点

(e)形态学操作　(f)笔刷算法

(g)连通域标记　(h)连通域分析

(i)定位图像

图 3　仿真实验结果

以上实验图像表明,文中算法对自然场景中的文本有较好的定位效果。尤其是 ColorRoberts 算子融合了 Roberts 和 LOG 两种算子,能针对彩色图像直接进行边缘检测,很好保留了图像的彩色信息,综合效果更好;通过后处理算法能较好地抑制图像中噪声的产生;最后通过连通域的标记与分析能够准确地定位出大部分文字所在区域。

文中算法也有需要改进的地方,当文字出现倾斜、光照不均匀、文本区域有阴影等情况时,识别的准确性会大大降低,这也是下一步研究工作的重点。

4　结束语

文中针对自然场景下基于连通域检测的文字识别算法进行研究,在现有算法的基础上,将边缘检测与连通域分析相结合,能快速有效地定位出文本区域在图像中的位置,为后续文本的正确识别打下基础。

参考文献:

[1] Thillou C, Ferreira S, Gosselin B. An embedded application for degraded text recognition[J]. EURASIP Journal on Advances in Signal Processing, 2005(13): 2127-2135.

[2] Pei S C, Chuang Y T, Chuang W H. Effective palette indexing for image compression using self-organization of Kohonen feature map[J]. IEEE Transactions on Image Processing, 2006, 15(9): 2493-2498.

[3] 林孜阳,穆　雪,吴凯锋,等. 基于连通域的快速文字图像分割算法[J]. 计算机光盘软件与应用,2014(22):89-90.

[4] Jain A K, Yu B. Automatic text location in images and video frames[J]. Pattern Recognition, 1998, 31(12): 2055-2076.

[5] Clark P, Mirmehdi M. Combining statistical measures to find image text regions[C]//Proc of 15th international conference on pattern recognition. [s. l.]: IEEE, 2000: 450-453.

[6] Jain A K, Bhattacharjee S. Text segmentation using Gabor filters for automatic document processing[J]. Machine Vision and Applications, 1992, 5(3): 169-184.

[7] 付　辉,吕　磊,苟　芳,等. 边缘检测算法分析与实现[J]. 科技传播,2014(21):210-212.

[8] 甘金来. 图像边缘检测算法的比较研究[D]. 成都:电子科技大学,2005.

[9] 张　引,潘云鹤. 复杂背景下文本提取的彩色边缘检测算子设计[J]. 软件学报,2001,12(8):1129-1135.

[10] 汪文芳. 基于移动终端的自然场景文本定位和识别[D]. 西安:西安电子科技大学,2011.

[11] 章毓晋. 图像处理与分析[M]. 北京:清华大学出版社,1999.

[12] 李昭早. 自然场景文本区域定位[D]. 西安:西安电子科技大学,2006.

[13] 罗志灶,周赢武,郑忠楷. 基于数组型并查集的连通域标记算法[J]. 杭州师范大学学报:自然科学版,2011,10(1):86-91.

[14] Cai M, Song J, Lyu M R. A new approach for video text detection[C]//Proc of international conference on image processing. [s. l.]: IEEE, 2002.

自然场景下基于连通域检测的文字识别算法研究

作者： 刘新瀚， 钱侃， 王宇飞， 朱向霄， 孙知信， LIU Xin-han， QIAN Kan， WANG Yu-fei， ZHU Xiang-xiao， SUN Zhi-xin

作者单位： 南京邮电大学 物联网学院,江苏 南京,210003

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2015(5)