• 多媒体技术•

自然场景图像中的中文文本检测算法研究

缪裕青[1,2] 刘水清[1] 张万桢[3] 欧威健[1] 蔡国永[1,2]

(1.桂林电子科技大学 计算机与信息安全学院，广西 桂林 541004；

2.桂林电子科技大学 广西可信软件重点实验室，广西 桂林 541004；

3. 桂林航天工业学院 实践教学部,广西 桂林 541004)

摘 要：为降低当前场景文本检测算法在图像背景复杂时检测中文文本的误检率，提出了一种基于自然场景图像的中文文本检测算法 TDSI。使用一系列启发式规则分别改进 MSER 算法和 SWT 算法，根据笔画宽度值的标准差过滤非文本区域。根据汉字的结构特征，即候选区域的质心、重合区域等特征将文本区域聚集成汉字。实验结果表明，针对背景较复杂的自然场景图像中的中文文本，TDSI 算法在自建图像数据库中获得了较好的准确率、召回率和 F 值。

关键字：场景文本检测；中文文本检测；最大稳定极值区域；笔画宽度变换；启发式规则；汉字结构

中图法分类号：TP391.4 文献标志码：A 文章编号：9925300

Text Detection Algorithm in Natural Scene Images

MIAO Yuqing[1,2] , LIU Shuiqing[1], ZHANG Wanzhen[1], OU Weijian [1],CAI Guoyong[1,2]

(1. School of Computer Science and information Security, Guilin University of Electronic Technology, Guilin 541004, China;

2. Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China;

3. Department of Experiential Practice, Guilin University of Aerospace Technology, Guilin 541004, China)

Abstract: To reduce the false-positive rate of text detection in complex background, a Chinese text detection algorithm in natural scene images (TDSI) is presented. Firstly, a series of heuristic rules are used to improve MSER algorithm and SWT algorithm. The standard deviation of stroke width value is used to filter non-text areas. Then Chinese characters structure is used to gather candidate character areas into single Chinese character. The location of candidate characters' center and overlap areas are used to estimate whether two candidate character areas are the same character or not. Experiments show that using self-built image database, TDSI algorithm has higher accuracy rate, recall rate and F value, for the natural scene images in more complicated Chinese environment.

Keywords: scene text detection; Chinese text detection; maximally stable extremal region; stroke width transform; heuristic rule; Chinese characters structure

0 引言

自然场景图像中的文本识别主要包括三个步骤：图像二值化、文本检测和文本识别。本文主要研究图像二值化和文本检测。其中，图像二值化常用的算法是 MSER 算法[1][2][3]（Maximally Stable Extremal Region，最大稳定极值区域）。文本检测过程常用的算法是 SWT 算法[4][5][6]（Stroke Width Transform，笔画宽度变换）。2011 年，Chen 等[7]使用 MSER 算法做预处理以改进 SWT 算法的性能。该算法较准确的提取极值区域，但对背景复杂的图像中的文字检测准确率不高。2015 年，Busta 等[8]提出一种易于使用的笔画探测器。该算法检测速度较快，检测效果较好，但当图像对比度低、图像背景复杂时，文本检测的准确率不高。当前，国内外很多学者聚焦于英文场景文本检测的研究[9][10]，对中文环境下的场景文本检测研究较少，对中文的检测效果不佳。

综上所述，在当前场景文本检测算法中，虽然能较准确的检测场景图像中的文本，但当场景图像背景较复杂时，误检率较高。此外，许多研究都是

收稿日期：20xx-xx-xx； 修订日期：20xx-xx-xx.

基金项目：广西自然科学基金（2014GXNSFAA118395）；国家自然科学基金（61363029）；桂林电子科技大学研究生教育创新计划（2016YJCX72）

作者简介：缪裕青(1966-)，女，浙江台人，博士，副教授，CCF 会员（37787M），研究方向为数据挖掘、云计算、并行与分布式计算；刘水清(1991-)，女，山西忻州人，硕士研究生，研究方向为图像数据挖掘。张万桢（1981-），女，湖北武汉人，硕士，讲师，研究方向为数据挖掘；欧威健(1990-)，男，广东省河源人，硕士研究生，研究方向为数据挖掘；蔡国永（1971-），男，广西河池人，博士，教授，CCF 会员(12524S)，研究方向为社交媒体数据处理、机器学习、可信软件。**E-mail:lsqchina647@gmail.com**

针对场景图像中的英文进行检测，少有针对中文的检测。针对这些问题，本文提出一种基于自然场景图像的中文文本检测算法 TDSI（Text Detection Algorithm in Natural Scene Images）。TDSI 算法将 MSER 和 SWT 两种算法的优势相结合，既使用 MSER 算法去掉大量干扰信息，又使用 SWT 算法根据候选区域的笔画宽度值区分文本区域和非文本区域。通过本文提出的改进 MSER 算法和改进 SWT 算法过滤掉大量非文本区域。最后根据汉字结构将文本区域聚集成单个汉字，再将其聚合成文本行。

1 TDSI 算法

1.1 算法流程

针对图像背景复杂时对中文文本的检测效果差的问题提出改进算法 TDSI。该算法首先使用启发式规则改进 MSER 算法和 SWT 算法。然后使用改进的 MSER 算法对目标图像进行预处理，得到二值图像，即文本候选区域；然后使用改进的 SWT 算法将非文本区域过滤掉；最后根据汉字的结构特征，将候选区域聚集成汉字，再将之聚集成文本行。TDSI 算法流程如下：

（1）通过 MSER 算法得到最稳定极值区域即候选文本区域。使用启发式规则过滤掉部分明显的非文本区域；

（2）通过 SWT 算法得到笔画宽度图像。运用相应的启发式规则将非文本区域过滤掉，得到文本区域；

（3）根据汉字的结构特征聚集成中文单字；

（4）把汉字聚集成文本行，使用矩形框进行渲染。

算法流程图如图 1 所示。

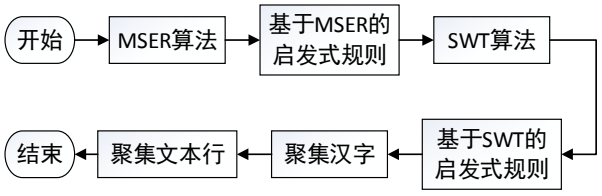

图 1 算法流程图

1.2 基于 MSER 算法的启发式规则

通过 MSER 算法得到的最大稳定极值区域是一些不规则图形，不方便提取特征。一个候选区域的特征包括位置、长宽和质心等，通过对最大稳定极值区域进行椭圆拟合，可以较易地得到这些特征。最大稳定极值区域既包括文本区域，也包括非文本区域，对椭圆拟合后的最大稳定极值区域使用启发式规则可以将部分明显的非文本区域过滤掉。TDSI 算法使用的基于 MSER 的启发式规则包括：

（1）候选区域面积

候选区域中面积非常小的一般不是文本区域，需对其进行过滤。当将候选区域的面积阈值定为 20 时，结果最优，如式（1）：

$$ResultMSER1=\{MSER_i|AreaMSER_i\}>20 \quad (1)$$

（2）椭圆拟合后的长宽比

汉字笔画有的短粗、有的细长，如果拟合后的椭圆特别细，近似一条直线，说明该区域一定不是文本区域，需将长宽比大于一定阈值的区域过滤掉。当将阈值定为 5 时，结果最优，如式（2）：

$$ResultMSER2=\left\{ ResultMSER1_i \mid \frac{LongAxis_i}{ShortAxis_i} \le 5 \right\} \quad (2)$$

其中，长宽比是指拟合后的椭圆的长轴与短轴之比，$LongAxis_i$ 是拟合椭圆长轴的长度，$ShortAxis_i$ 是拟合椭圆短轴的长度，i 表示最大稳定极值区域的个数。

（3）拟合椭圆与最大稳定极值区域的面积比

拟合椭圆是对最大稳定极值区域的拟合，其面积与最大稳定极值区域存在一定差异。如果最大稳定极值区域是非文本区域比如树叶，最大稳定极值区域的面积与拟合椭圆面积差异不大。相反，如果最大稳定极值区域是文本区域，其面积与拟合椭圆面积差异较大。根据该规则，将拟合椭圆的面积与最大稳定极值区域的面积之比太小的区域过滤掉。当阈值取 1.35 时，结果最优，如式（3）：

$$\mathrm{Re}\,sultMSER3=\{\mathrm{Re}\,sultMSER2_i \mid \frac{AreaEllipse_i}{AreaMSER_i} \ge 1.35\} \quad (3)$$

其中，$AreaEllipse_i$ 是拟合椭圆的面积，$AreaMSER_i$ 是最大稳定极值区域的面积。

（4）图像边界像素交集

场景图像中的文本区域一般不会出现在图像的边界位置，因此将含有图像边界像素的最大稳定极值区域过滤掉，如式（4）：

$$ResultMSER4=\{ResultMSER3_i|ResultMSER3_i \cap edge=\varnothing\} \quad (4)$$

其中，$edge$ 是图像的边界像素。

1.3 基于 SWT 算法的启发式规则

使用 SWT 算法得到的笔画宽度图像，包括文本区域和非文本区域。通过基于 SWT 算法的启发式规则将部分非文本区域过滤掉，便于将文本区域

聚集成汉字。使用的启发式规则包括：

（1）同一幅图像中汉字的笔画宽度值基本保持不变，即一个候选区域的笔画宽度值与图像的平均笔画宽度值差距较小。而标准差就是用于衡量一组数据中某个数据与其平均值的差异程度的指标。也即当某个区域笔画宽度值的标准差较小时，该区域为文本区域；而标准差较大时，则该区域为非文本区域。把笔画宽度值的标准差大于 5.2 的区域认为是非文本区域，将其过滤掉，如式（5）：

$$ResultSWT1=\left\{ResultMSER4_i \mid \sqrt{\frac{1}{N}\sum_{j=1}^{N}(SWT_j-\mu)^2}<5.2\right\} \quad (5)$$

其中，N 表示一幅图像中候选区域的个数，SWT_j 是一幅图像中第 j 个候选区域的笔画宽度值，μ 是一幅图像的笔画宽度值的算术平均值。

（2）在同一幅图像中，一般相邻文本字号一致，其笔画宽度值相差不大。如果候选区域邻域像素的笔画宽度值与当前像素的笔画宽度值相差较大，说明该区域是非文本区域，需将之过滤掉。当邻域像素的笔画宽度值与当前像素的笔画宽度值之比小于 3 时，效果最佳，如式（6）：

$$ResultSWT2=\{ResultSWT1_i \mid \frac{NeiSW_i}{CurSW_i}<3\} \quad (6)$$

其中，$NeiSW_i$ 是邻域像素的笔画宽度值，$CurSW_i$ 是当前像素的笔画宽度值。

（3）将笔画宽度值限定在（20,300）之间，过滤掉笔画宽度值过大或过小的区域。如果笔画宽度值过小，一般是小的点或极细的线条，而不是字符区域，应该被过滤掉；而在拍摄的自然场景图像中，大多文字笔画宽度不会很大，需过滤掉笔画宽度值过大的区域，如式（7）：

$$ResultSWT3=\{ResultSWT2_i \mid 20<SW_i<300\} \quad (7)$$

其中，SW_i 是笔画宽度值。

1.4 针对中文场景的改进

在英文中大部分字母都是由一个完整的部分构成，只有“i”由两部分构成。但由于“i”上方的点很小，即使丢失也不影响最终结果。相对而言，汉字复杂多变，包括上下结构、左右结构、全包围结构、半包围结构和品字形结构等，结构与结构之间互不相连。如果不对其进行处理，当图像中的文本行走向是水平方向，并且有汉字是上下结构时，就无法将文本聚合成文本行；反之亦然。因此要先将候选区域聚集成汉字，再将汉字聚合成文本行。

由于单个汉字各结构间的距离一定小于相邻汉字间的距离，根据该规则可以将候选区域组合成汉字。首先计算两两候选区域间的距离，从距离最小的两个开始，判断这两个候选区域是否满足以下规则：

（1）如果两个候选区域有重合部分，说明这两个区域是同一个汉字的两部分；

（2）如果两个候选区域的质心坐标近似重合，说明这两个区域是同一个汉字的两部分；

（3）如果两个候选区域间的距离小于等于两个候选区域的长宽平均值，可能是同一汉字的两部分；

（4）如果两个候选区域的像素值相差不超过 30，可能是一个汉字的两部分；

（5）如果两个候选区域的笔画宽度值相差不超过 100，可能是一个汉字的两部分；

若满足，则将两个候选区域组合成一个汉字。然后根据距离从小到大依次进行组合，直到没有符合条件的候选区域为止，这样就将候选区域组合成一个个汉字。

文本行中的汉字一般都在同一条直线上，这些汉字质心的纵坐标（或横坐标）大小相差不大，每个汉字的最高点的纵坐标（最左侧点的横坐标）、最低点纵坐标（最右侧点的横坐标）都大致相同。根据这些特性，将汉字聚合成文本行。

2 实验与结果分析

2.1 数据库构建和标注

目前大多数公开的自然场景图像数据库都是基于英文环境，少数是中英环境混合，但没有完全基于中文环境的自然场景图像数据库。为测试 TDSI 算法的性能，构建了一个完全基于中文环境的自然场景图像数据库。其中图像内容主要涉及路标、交通警示语、标语、横幅等。这些图像背景复杂，具有不同的颜色、字体、字号、光照、对比度等，比较适合做算法测试。

根据 ICDAR（International Conference on Document Analysis and Recognition，文档分析与识别国际会议）2013 比赛[11]的要求，为每张图像添加标注。每张图像的标注内容和格式为“图像编号 矩形最左上角点的坐标的 X 值 矩形最左上角点的坐标的Y值 矩形最右下角点的坐标的X值 矩形最右下角点的坐标的 Y 值”。

2.2 实验结果及分析

2.2.1 文本区域对比

(a) 原图

(b) Chen 算法实验结果图

(c) Busta 算法实验结果图

(d) TDSI 算法实验结果图

图 2 TDSI 算法与 Chen 算法、Busta 算法实验结果对比图：a：原图；b：Chen 算法实验结果图；c：Busta 算法实验结果图；d：TDSI 算法实验结果图

实验使用 Busta[8]算法、Chen[7]算法和 TDSI 算法做对比。实验过程中，TDSI 算法忽略所有字符数目少于 3 和包含非法字符的文本区域。实验结果如图 2 所示。图 2(a)为数据库中任意抽取的两张原图，图 2(b)为由 Chen 算法得到的实验结果图，图 2(c)为由 Busta 算法得到的实验结果图，图 2(d)为由 TDSI 算法得到的实验结果图。其中蓝色矩形框框出的部分即为算法检测到的文本区域。在 2(b)中把第一张图中的商品错误识别成文本区域，第二张图只检测出部分文本区域。在 2(c)中第一张图检测出部分文本区域，把一部分商品错误识别成文本区域，把第二张图背景中的人错误识别成文本区域。在图 2(d)中文本区域定位基本正确，说明针对背景复杂的自然场景图像中的中文，TDSI 算法比 Chen 算法、Busta 算法有明显优势。

2.2.2 检测结果对比

准确率、召回率和 F 值取自建图像数据库中所有图像检测结果的平均值。实验结果采用 ICDAR 文本定位竞赛的评价标准[10]。检测结果对比如表 1、图 3 所示：

表 1 检测结果对比

方法	准确率	召回率	F 值
TDSI 算法	0.713	0.896	0.794
Busta 算法	0.529	0.873	0.659
Chen 算法	0.580	0.667	0.620

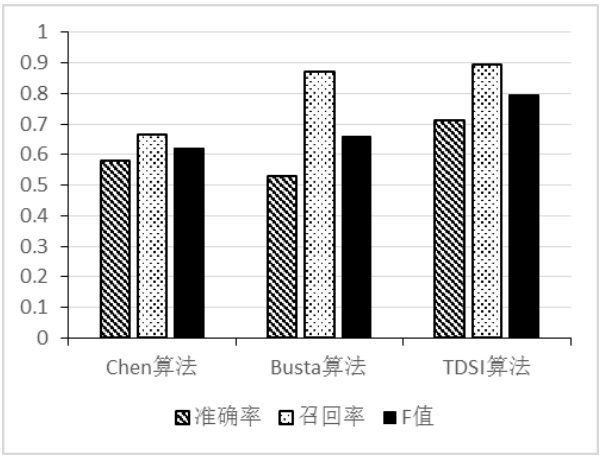

图 3 算法检测结果对比

由表 1 和图 3 可知，TDSI 算法的准确率、召回率和 F 值均最高。Busta 算法的准确率最低，召回率和 F 值较高。Chen 算法的召回率和 F 值最低，准确率较高。

据 Busta[8]介绍，Busta 算法提取的文本区域比 MSER 算法多，而 Chen 算法使用 MSER，因此 Busta 算法的召回率比 Chen 算法高。又因为 Chen 算法使用笔画宽度值过滤大部分非文本区域，而 Busta 算法没有使用任何方法过滤非文本区域，所以 Busta 算法的准确率没有 Chen 算法高，即 Busta 算法的误检率较高。

TDSI 算法使用改进的 MSER 算法提取的文本区域比 Busta 算法多，因此 TDSI 算法的召回率比 Busta 算法高。另外，TDSI 算法使用改进的 SWT 算法过滤非文本区域，而 Busta 算法没有过滤非文本区域，因此 TDSI 算法的准确率比 Busta 算法高。虽然 Chen 算法同时使用 MSER 和 SWT 算法，但 Chen 算法只对 MSER 算法进行改进。而 TDSI 算法分别对 MSER 算法和 SWT 算法都做了改进，且根

据汉字的结构特征进行了改进，因此 TDSI 算法的准确率比 Chen 算法高。总体上，TDSI 算法的检测结果比 Chen 算法和 Busta 算法都好。

3 结束语

针对图像背景复杂时大多数场景文本检测算法的误检率较高，且很少有算法专门针对中文文本进行检测的问题，本文提出了基于自然场景图像的中文文本检测算法 TDSI。使用一系列启发式规则分别对 MSER 算法和 SWT 算法进行改进，将改进的 MSER 算法和改进的 SWT 算法相结合，过滤非文本区域。然后根据汉字的结构特征将文本区域聚集成汉字，再将之聚集成文本行。实验结果表明，对于背景复杂的场景图像，TDSI 算法对中文的处理效果较好，能较准确地检测出文本区域，对中文文本检测的准确率、召回率和 F 值均较高。

参考文献:

[1] Xiao C, Ji L, Gao C, et al. Fast and Accurate Text Detection in Natural Scene Images[M]// Intelligence Science and Big Data Engineering. Image and Video Data Engineering. Springer International Publishing, 2015.

[2] Liu J, Su H, Yi Y, et al. Robust text detection via multi-degree of sharpening and blurring[J]. Signal Processing, 2015, 124(C):259-265.

[3] Liu J, Su H, Yi Y, et al. Robust text detection via multi-degree of sharpening and blurring[J]. Signal Processing, 2015, 124(C):259-265.

[4] Yao C. Detecting texts of arbitrary orientations in natural images[J]. 2012, 157(10):1083-1090.

[5] Zhang Y, Lai J, Yuen P C. Text string detection for loosely constructed characters with arbitrary orientations[J]. Neurocomputing, 2015, 168(C):970-978.

[6] SLIU Ya-ya，YU Feng-qin，CHEN Ying. Scene Text Localization Based on Stroke Width Transform[J]. Journal of Chinese Computer Systems,2016, 37(2):350-353(in Chinese). [刘亚亚, 于凤芹, 陈莹. 基于笔画宽度变换的场景文本定位[J]. 小型微型计算机系统, 2016, 37(2):350-353.]

[7] Chen H, Tsai S S, Schroth G, et al. Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions[C]// IEEE International Conference on Image Processing. IEEE, 2011:2609-2612.

[8] Buta M, Neumann L, Matas J. FASText: Efficient Unconstrained Scene Text Detector[C]// IEEE International Conference on Computer Vision. IEEE, 2015:1206-1214.

[9] Zhong G, Cheriet M. Tensor representation learning based image patch analysis for text identification and recognition[J]. Pattern Recognition, 2015, 48(4): 1211-1224.

[10] Tian S, Bhattacharya U, Lu S, et al. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients[J]. Pattern Recognition, 2016, 51(C):125-134.

[11] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 robust reading competition[J]. 2013, 2(2-3):1484-1493.

联系方式：
邮编：541004
地址：广西省桂林市七星区金鸡路 1 号桂林电子科技大学
联系人：刘水清
联系电话：13393507399