自然场景中文字的实时定位与识别

王昌旭

2016 年 3 月 1 日

1 摘要

本文展示了一种对自然场景中的文字进行实时定位和是别的方法。实时性通过将文字检测任务变为一系列极值区域的选择来实现。基于极值区域的检测器对模糊、光照、颜色和纹理变换有着极强的鲁棒性，并能处理低对比度图像。

在识别的第一阶段，我们使用对每个极值区域使用一些可在 $O(1)$ 的时间复杂度内被计算出的全新特征来计算该极值区域为一个字符的概率。只有那些具有局部极大概率的极值区域区域被选中并进入第二阶段——使用更加复杂的特征来计算更加准确的概率。然后我们使用一种十分有效且带有反馈的搜索算法将这些极值区域区域聚合成单词并进行字符分割。最后，我们使用 OCR 的方法对这些字符进行识别。

我们在两个公开数据集上对算法进行了评估。在 ICDAR 2011 数据集上，与目前所有以公开的算法相比，我们的方法取得了目前为止最佳的字符定位效果。在更具有挑战性的街景数据集上，我们的方法取得了目前为止最高的召回率。

2 研究现状

已有一系列用于解决自然场景中字符定位问题的方法被提出。如 Epstein 等人提出将图像转换为灰度图后使用 Canny 算子来检测字符边缘。对每个像素，其平行的边缘被用于计算笔画粗细，最后具有相似笔画粗细的像素悲剧和为一个个字符。该方法由于其依赖于边缘检测的效果，因此对噪声和图像模糊十分敏感。此外还有其他人提出了一种同样使用边缘检测但使

用不同连通分量的算法。这些方法都可以在 ICDAR Robust Reading 竞赛结果中找到。

只有很少数能同时解决字符定位和识别的算法被提出。比如 Wang 等人的方法使用滑动窗口来寻找独立的字符，然后使用词汇表将字符聚合成单词。这种方法可以可以处理带有噪声的数据，但是其效果却受制于词汇表的大小。车辆的车牌识别系统已经成为在视频监控领域中一个特殊的热门领域超过 10 年左右。随着先进的用于交通管理应用的视频车辆检测系统的的到来，车牌识别系统被发现可以适合用在相当多的领域内，并非只是控制访问点或收费停车场。现在它可以被集成到视频车辆检测系统，该系统通常安装在需要的地方用于十字路口控制，交通监控等，以确定该车辆是否违反交通法规或找到被盗车辆。一些用于识别车牌的技术到目前为止有如 BAM（双向联想回忆）神经网络字符识别，模式匹配等技术。应用于系统的技术是基于模式匹配，该系统快速，准确足以在相应的请求时间内完成，更重要的是在于阿尔伯塔车牌识别在字母和数字方位确认上的优先发展。由于车牌号码的字体和方位因国家/州/省份的不同而不同，该算法需要作相应的修改保持其结构完整，如果我们想请求系统识别这些地方的车牌。

还有一类基于极大稳定极值区域（MSER）对字符进行检测的方法，最终并基于 MSER 完成字符分割以进行字符识别。MSER 是极值区域的一种特殊情况，其在一段连续的阈值变化中保持面积的不变。这类方法有着一般具有十分好的效果，但是却无法在模糊或低对比度的图像上正常工作。根据 ICDAR 2011 Robust Reading 竞赛组织者提供的描述，最终竞赛的赢家正是基于 MSER 的检测算法，但是这种方法还未被公布并且它不包括字符识别。

本文提出的方法与基于 MSER 的方法的不同之处在于，它将测试所有的极值区域（并非 MSER 的一个子集）同时在保持和 MSER 相同的计算复杂度的基础上减少了内存占用，并可以达到实时级的速度。本方法借鉴 Zimermann 等人的思想，放弃 MSER 中对极大稳定的要求，并选择一种基于分类的极值区域 (CSER)。在我们的方法中，实时地选择合适极值区域的工作通过一系列级联分类器实现，并在此过程中使用了一些全新的特征，这些特征专门设计用来进行字符检测。此外，分类器将被训练输出区域为字符的概率，因此可以被用于提取一个字符的多个分割。

3 简介

真实场景中的字符定位和识别在许多计算机视觉技术的应用中都是十分关键的一个环节，如基于文本的图像搜索、街景应用中对商店的标记识别以及虚拟现实系统，因此一直是计算机世界研究的热点。在过去的几年中，举行了许多相关竞赛，但即使是目前最好的方法在也只在 ICDAR 2011 竞赛中取得了 62% 的定位准确率，并且该数据集并不能代表真实场景（数据集中的单词全部是水平的，并在图像用占据主要位置，没有投影变换或比较显著的噪声）。为了进行车牌识别，需要以下几个基本的步骤：1) 牌照定位，定位图片中的牌照位置；2) 牌照字符分割，把牌照中的字符分割出来；3) 牌照字符识别，把分割好的字符进行识别，最终组成牌照号码。车牌识别过程中，牌照颜色的识别依据算法不同，可能在上述不同步骤实现，通常与车牌识别互相配合、互相验证。1) 牌照定位自然环境下，汽车图像背景复杂、光照不均匀，如何在自然背景中准确地确定牌照区域是整个识别过程的关键。首先对采集到的视频图像进行大范围相关搜索，找到符合汽车牌照特征的若干区域作为候选区，然后对这些侯选区域做进一步分析、评判，最后选定一个最佳的区域作为牌照区域，并将其从图像中分离出来。2) 牌照字符分割完成牌照区域的定位后，再将牌照区域分割成单个字符，然后进行识别。字符分割一般采用垂直投影法。由于字符在垂直方向上的投影必然在字符间或字符内的间隙处取得局部最小值的附近，并且这个位置应满足牌照的字符书写格式、字符、尺寸限制和一些其他条件。利用垂直投影法对复杂环境下的汽车图像中的字符分割有较好的效果。3) 牌照字符识别方法主要有基于模板匹配算法和基于人工神经网络算法。基于模板匹配算法首先将分割后的字符二值化并将其尺寸大小缩放为字符数据库中模板的大小，然后与所有的模板进行匹配，选择最佳匹配作为结果。基于人工神经网络的算法有两种：一种是先对字符进行特征提取，然后用所获得特征来训练神经网络分配器；另一种方法是直接把图像输入网络，由网络自动实现特征提取直至识别出结果。实际应用中，车牌识别系统的识别率还与牌照质量和拍摄质量密切相关。牌照质量会受到各种因素的影响，如生锈、污损、油漆剥落、字体褪色、牌照被遮挡、牌照倾斜、高亮反光、多牌照、假牌照等等；实际拍摄过程也会受到环境亮度、拍摄方式、车辆速度等等因素的影响。这些影响因素不同程度上降低了车牌识别的识别率，也正是车牌识别系统的困难和挑战所在。为了提高识别率，除了不断地完善识别算法还应该想办法克服各种光照条件，使采集到的图像最利于识别。

在图像中定位文字区域很可能是一个十分耗费计算资源的任务，因为一幅图像中一共有 2^N 个可能为文字区域的子集（N 为像素数量）。因此主流文字定位算法在解决此问题是可分为两种思路。

一类方法通过滑动窗口将搜索区域限制在图像的一个子区域内。这类方法将对图像的搜索复杂度降至了 cN 级别，其中 c 是一个常数，代表算法所需处理不同放缩比、长宽比、旋转等变换的种类。

另一类方法通过连通分量分析将像素聚合成字符区域，这类方法假定属于同一个字符的像素有着相似的属性。连通分量分析方法又根据使用属性的不同分为许多种（如颜色、比划粗细等）。基于连通分量的方法的优点在于其算法复杂度不依赖于文字区域的属性（如尺寸、旋转、字体）等，并且提供了对字符的分割操作以方便进行 OCR。但是这类方法也有其缺点，具体来讲就是对会连通分量产生改变的干扰十分敏感，如干扰、阻隔等。

在本文中，我们提出了一种端到端的实时文字定位和识别算法，该方法在标准数据集上取得了目前最佳的效果。实时性通过将文字检测任务变为一系列极值区域的选择来实现。基于极值区域的检测器对模糊、光照、颜色和纹理变换有着极强的鲁棒性，并能处理低对比度图像。这种算法的复杂度是 $O(2pN)$，其中 p 表示使用的通道（投影）数。

在识别的第一阶段，我们使用对每个极值区域使用一些可在 $O(1)$ 的时间复杂度内被计算出的全新特征来计算该极值区域为一个字符的概率。只有那些具有局部极大概率的极值区域区域被选中并进入第二阶段——使用更加复杂的特征来计算更加准确的概率。然后我们使用一种十分有效且带有反馈的搜索算法将这些极值区域区域聚合成单词并进行字符分割。

此外，我们提出一种全新的梯度幅度投影来检测图像边缘并计算极值区域。进一步的测试表明使用梯度投影后，极值区域检测器能检测出 94.8% 的字符。系统架构包含三个相异部分：室外部分，室内部分和通信链路。室外部分是安装摄像头在拍摄图像的不同需要的路口。室内部分是中央控制站，从所有这些安装摄像头中，接收，存储和分析所拍摄图像。通信链路就是高速电缆或光纤连接到所有这些相机中央控制站。几乎所有的算法的开发程度迄今按以下类似的步骤进行。一般的 7 个处理步骤已被确定为所有号牌识别算法共有。它们是：触发：这可能是硬件或软件触发。硬件触发是旧的方式，即感应圈用于触发和这个表述了图像通过检测车牌的存在何时应该被捕获。硬件触发现在在操作上在许多地方被软件触发取代。在软件触发，图像分为区，通过图像对于分析的车辆的检测的执行。图像采集：硬件或软

件触发启动图像捕捉设备来捕捉和存储图像来进一步的分析。车辆的存在：这一步是只需要如果在确认一定时间间隔后触发完成不需要知道车辆存在于捕获的图像中。这一步背景图像与捕获的图片作比较，并检测是否有任何重大改变。如果没有，拍摄的图像被忽略，否则进入到下一个步骤。寻找车牌：此步骤是在捕获的图像中定位车牌。一些技术的可用于这一步，例如颜色检测，特征分析，边缘检测等。在捕获的图像中的任何倾斜是纠正在这一步。一旦车牌已被定位，图像即准备进行字符识别。字符分割：分割可以通过检测浓到淡或者淡到浓的过渡层。车牌中的每个灰色字符产生了一个灰色带。因此，通过检测类似灰度带每个字符可以被分割出来。识别过程：这是光学字符识别的一步。一些技术可以被用于到这一步包括模式匹配，特征匹配和神经网络分类。发布过程：这是应用程序的特有的一步。根据应用此步骤可保存已被检测出来的车牌用于交通数据收集，尝试匹配号牌与被盗车辆数据库或在停车场中为认可停车的车辆打开汽车门等等。车牌识别系统有两种触发方式，一种是外设触发，另一种是视频触发。外设触发工作方式是指采用线圈、红外或其他检测器检测车辆通过信号，车牌识别系统接受到车辆触发信号后，采集车辆图像，自动识别车牌，以及进行后续处理。该方法的优点是触发率高，性能稳定；缺点是需要切割地面铺设线圈，施工量大。视频触发方式是指车牌识别系统采用动态运动目标序列图像分析处理技术，实时检测车道上车辆移动状况，发现车辆通过时捕捉车辆图像，识别车牌照，并进行后续处理。视频触发方式不需借助线圈、红外或其他硬件车辆检测器。该方法的优点是施工方便，不需要切割地面铺设线圈，也不需要安装车检器等零部件，但其缺点也十分显著，由于算法的极限，该方案的触发率与识别率较之外设触发都要低很多。1) 间接法: 指通过识别安装在汽车上的 IC 卡或条形码中所存储的车牌的信息来识别车牌及相关信息。IC 卡技术识别准确度高，运行可靠，可以全天候作业，但它整套装置价格昂贵，硬件设备十分复杂，不适用于异地作业; 条形码技术具有识别速度快、准确度高、可靠性强以及成本较低等优点，但是对于扫描器要求很高。此外，二者都需要制定出全国统一的标准，并且无法核对车、条形码是否相符，也是技术上存在的缺点，这给在短时间内推广造成困难。2) 直接法: 基于图像的车牌识别技术属于直接法，是一种无源型汽车牌照智能识别方法，能够在无任何专用发送车牌信号的车载发射设备情况下，对运动状态车辆或静止状态车辆的车牌号码进行非接触性信息采集并实时智能识别。与间接法识别系统相比，首先，这种系统节省了设备安置及大量资金，从而提高了经济效

益; 其次，由于采用了先进的计算机应用技术，所以可提高识别速度，较好地解决实时性问题; 再次，它是根据图像进行识别，所以通过人的参与可以解决系统中的识别错误，而其他方法是难以与人交互的。直接法一般有图像处理技术，传统模式识别技术及人工神经网络技术。1) 图像处理技术: 运用图像处理技术解决汽车牌照识别的研究最早始于 80 年代，但国内外均只是就车牌识别中的某一个具体问题进行讨论，并且通常仅采用简单的图像处理技术来解决，并没有形成完整的系统体系，识别过程是使用工业电视摄像机拍下汽车的工前方图像，然后交给计算机进行简单的处理，并且最终仍需要人工干预，例如车辆牌照中省份汉字的识别问题，1985 年有人利用常见的图像处理技术方法提出汉字识别的分类是在抽取汉字特征的基础上进行的，根据汉字的投影直方图选取浮动闭值，抽取汉字在竖直方向的峰值，利用树形查表法进行汉字的粗分类; 然后根据汉字在水平方向的投影直方图，选取适当闭值，进行量化处理后，形成一个变长链码，再用动态规划法，求出与标准模式链码的最小距离，实现细分米完成汉字省名的自动识别。2) 传统模式识别技术。传统模式识别技术指结构特征法，统计特征法等。90 年代，由于计算机视觉技术的发展，开始出现汽车牌照识别的系统化研究。1990 年 AS.Johnson 等运用计算机视觉技术和图像处理技术实现了车辆牌照的自动识别系统。该系统分为图像分割、特征提取和模板构造、字符识别等三个部分。利用不同闽值对应的直方图不同，经过大量统计实验确定出车牌位置的图像直方图的闽值范围，从而根据特定闽值对应的直方图分割出车牌，再利用预先设置的标准字符模板进行模式匹配识别出字符。3) 人工神经网络技术。近几年来，计算机及相关技术发达的一些国家开始探讨用人工神经网络技术解决车牌自动识别问题，例如 1994 年 M.M.M.FANHY 等就成功地运用了 BAM 神经网络方法对车牌上的字符进行自动识别，BAM 神经网络是由相同神经元构成的双向联想式单层网络，每一个字符模板对应着唯一个 BAM 矩阵，通过与车牌上的字符比较，识别出正确的车牌号码。这种采用 BAM 神经网络方法的缺点是无映解决识别系统存储容量和处理速度相矛盾的问题。

4 本文提出的算法

4.1 极值区域

我们首先定义一幅图像 $\mathbf{I}$ 为一个映射 $\mathbf{I}: \mathcal{D} \subset \mathbb{N}^2 \to \mathcal{V}$，其中 $\mathcal{V}$ 一般为 $\{0, \cdots, 255\}^3$（即一个色彩空间）。然后，我们定义图像的一个通道为 $\mathbf{C}: \mathcal{D} \to \mathcal{S}$，其中 $\mathcal{S}$ 为一个全序集并且存在一个映射 $f_c: \mathcal{V} \to \mathbf{S}$ 将像素值映射到该全序集。我们定义 A 为邻接关系 $A \subset \mathcal{D} \times \mathcal{D}$，常见的邻接关系有 4-邻接和 8-邻接，在本章的实现中我们使用 4-邻接关系。

我们定义图像 I（或通道 C）的一个 Region 为 $\mathcal{D}$ 的一个连续子集（所为连续，是指 $\forall p_i, p_j \in \mathcal{R}\, \exists p_i, q_1, q_2, \cdots, q_n, p_j : p_i A q_1, q_1 A q_2, \cdots, q_n A p_j$。我们定义 Region 边界 $\partial\mathcal{R}$ 为那些与 Region $\mathcal{R}$ 邻接却不属于 $\mathcal{R}$ 的像素的集合，即 $\partial\mathcal{R} = \{p \in \mathcal{D}$
$\mathcal{R} : \exists q \in \mathcal{R} : pAq\}$。现在，我们定义 *极值区域* (极值区域) 为那些边界像素值比内部像素高许多的 Region，写成数学语言即 $\forall p \in \mathcal{R}, q \in \partial\mathcal{R} : \mathbf{C}(q) > \theta > \mathbf{C}(p)$，其中 θ 为极值区域的阈值。

一个阈值为 θ 的极值区域 r 可以由多个或个阈值为 $\theta - 1$ 的极值区域和值为 θ 的像素和并集：$r = (\bigcup u \in R_{\theta-1}) \cup (\bigcup p\ in\mathcal{D} : \mathbf{C}(p) = \theta)$ 构成，其中 $R_{\theta-1}$ 表示阈值为 θ_1 的极值区域。该性质指出极值区域间有一种包含关系，一个极值区域可以包含一个或多个后继极值区域（或没有后继，如果它只包含具有相同值的像素）和唯一的前驱极值区域。

在本文中，我们考虑 RGB 和 HSI 色彩空间，并且额外使用一个 *亮度导数*通道，其中每个像素的导数通过该像素及其邻域像素的最大亮度差来表示：

$$\mathbf{C}_\nabla(p) = \max_{q \in \mathcal{D}: pAq} \|\mathbf{C}_\mathbf{I}(p) - \mathbf{C}_\mathbf{I}(q)\|$$

实验验证表明 85.6% 的字符可通过在一个通道上的极值区域检测，94.8% 的字符区域可以通过所有通道检测。一个字符被认为被成功地检测到，如果极值区域的边界矩形和真实字符边界矩形有 90% 的重合。在我们提出的方法中，我们结合使用亮度（I）、亮度导数（∇）、色度（H）和饱和度（S）通道进行实验，并在运行时间和定位准确率之间取得了最佳的平衡。

4.2 可增量计算描述子

能够对极值区域进行快速分类的关键在于能快速对每个区域计算其描述子作为分类器的特征。正如 Zimmerman 和 Matas 在他们论文中所提出的，我们可以使用一类特殊的描述子，这类描述字可以根据极值区域见的包含关系逐步递增地计算得出。

我们使用 $R_{\theta-1}$ 表示阈值为 $\theta-1$ 的极值区域。一个极值区域 $r \in R_\theta$ 科表示为一系列阈值为 $\theta-1$ 的极值区域的并集并加上一些值为 θ 的像素。我们进一步假设对于每个阈值为 $\theta-1$ 的极值区域 $u \in R_{\theta-1}$ 其描述子 $\phi(u)$ 已知。为了计算描述子 $\phi(r), r \in R_\theta$，我们必须结合那些组成 r 的极值区域 $u \in R_{\theta-1}$ 的描述子和值为 θ 的像素，即 $\phi(r) = (\oplus\phi(u)) \oplus (\oplus\psi(p))$，其中 $\oplus$ 表示对描述子进行结合的算子，$\psi(p)$ 被称为初始化函数，用于计算给定像素 p 的描述子。我们将那些存在 $\psi(p)$ 和 $\oplus$ 的描述子称为**可增量计算的**。

显然，我们可以通过将阈值 θ 从 0 逐步累加之 255 的方法来计算出所有极值区域的描述子，即计算值为 θ 的像素的描述子 ψ 并重用那些阈值为 $\theta-1$ 的区域的描述子 ϕ。注意，这种性质指出我们只需要在内存中保留前一阈值所对用极值区域区域的描述子，因此这种方法相较于基于极大稳定极值区域的方法将极大程度减少内存占用。更进一步，如果我们假设初始化函数 ψ 和结合算子 $\oplus$ 具有常数级别的计算复杂度，则计算所有极值区域区域的算法复杂度仅有 $O(N)$。

在本文中，我们使用下述描述子:

- **面积** a: 极值区域区域的面积（即像素数量）。其初始化函数为一个常数 $\psi(p) = 1$，结合算子 $\oplus$ 为数值加法。
- **边框** $(x_{min}, y_{min}, x_{max}, y_{max})$: 即极值区域边框的右上角和左下角。对于坐标为 (x, y)d 的像素 p,其初始化函数为四元组 $(x, y, x+1, y+1)$,结合算子 $\oplus$ 为 (min, min, max, max)。区域的长和宽可通过 $x_{max} - x_{min}$ 和 $y_{max} - y_{min}$ 计算得到。
- **周长** p: 即极值区域边缘的长度。初始化函数 $\psi(p)$ 通过新加入值为 θ 的像素的位置来绝对周长的改变量，结合算子 $\oplus$ 为数值加法。$\psi(p)$ 的时间复杂度为 $O(1)$，因为一个像素最多只有四个邻居。
- **欧拉数** η: 欧拉数是二值图像的一种拓扑特征，为连通域数目和孔洞。

- **水平交叉点数** c_i：用一个长度为图像高度的向量来保存对应行像素在属于极值区域与不属于极值区域之间转变的次数。初始化函数的的值由在阈值 $\mathbf{C}(p)$ 下像素 p 的左右邻接像素的存在与否来定。结合算子 $\oplus$ 为按元素做加法。$\psi(p)$ 的计算复杂度是个常数（每个像素在水平方向至多只有两个邻居），并且按元素的加法可以也具有常数复杂度，因为假定使用的数据结构的随机访问和两端插入操作的复杂度为 $O(1)$(如双端队列)。

4.3 级联分类器

在我们提出的方法中，每个通道被分别进行迭代（原始通道和反色通道）然后检测极值区域。为了减少极高的假阳性率以及减少的极值区域，只有那些被分类器认为十分可能是字符的极值区域区域被保留。为了提高计算性能，分类阶段被分为两阶段进行。

在第一阶段，阈值从 0 逐步累加至 255，对每个极值区域 r 计算其可增量计算描述子并作为特征送入分类器，得到该极值区域为字符区域的条件概率 $p(r\|\ )$。概率 $p(r\|\ )$ 在贯穿所有阈值的极值区域级联推倒中备注总，并且只有那些具有局部极大概率的极值区域会被选中（即局部极大概率大于全局阈值 p_{min} 并且局部极大和局部极小的差大于 Δ_{min}）。

在本文中，我们使用一个基于决策树的 AdaBoost 分类器并使用特征：

- **长宽比** (w/h)
- **compactness**$(\sqrt{a}/p)$
- **孔洞数** $(1-\eta)$
- **水平交叉点特征** $(\hat{c} = median(\mathbf{c}_{\frac{1}{6}w}, \mathbf{c}_{\frac{3}{6}w}, \mathbf{c}_{\frac{5}{6}w}))$

因为只有 $\mathbf{c}$ 的一个固定子集被使用，所以具有常数时间复杂度。分类器的输出通过对数几率回归得到概率分布函数 $p(r\|\ )$。在实验中我们使用参数 $p_{min} = 0.2$ 和 $\Delta_{min} = 0.1$ 来获得更高的召回率 (95.6%)。

在第二阶段，通过第一阶段的极值区域被分为字符和非字符两类，并使用了有更多欣喜但也更耗费计算资源特征。在本文中，使用了一个具有 RBF 核的 SVM 分类器。该分类器除了上述第一阶段用到的特征外，还是额外使用了如下特征：

- **孔洞面积比** a_h/a：其中 a_h 代表 ER 区域内孔洞的面积（像素数）。
- **凸包面积比** a_c/a：其中 a_c 为 ER 区域凸包的面积。
- **外轮廓拐点数** κ：代表 ER 区域边界凹角与凸角的变化数目。一个字符一般只含有数量比较少的外轮廓拐点（$kappa < 10$），而非字符区域（如草）则含有大量的外轮廓拐点。

我们注意到，以上所有特征都是放缩不变的，但不是旋转不变的，因此我们的训练集中需要包含具有不同旋转角度的字符。

5 实验

我们使用大约 900 个正例和 1400 个负例来训练上述方法中的分类器，这些样本由人工从 ICDAR 2003 训练数据集中提取（用于训练级联分类器）和从字体库中生成（用于训练 OCR）。我们在两个数据集上使用相同参数测试了我们提出的算法。

5.1 ICDAR 2011 数据集

ICDAR 2011 Robust Reading 竞赛数据集包含了 1189 个单词和 6393 个字母，共 255 幅图像。使用 ICDAR 2011 比赛的评估标准，本方法在文本定位问题上达到了 64.7% 的召回率，73.1% 的准确率和 68.7% 的 f-测度。

本方法相交 ICDAR 2011 Robust Reading 竞赛的获胜者在召回率上有显著的提升，但是准确率（73%）却相对获胜者（83%）则更差，因此总的 f-测度（69%）结果不及竞赛获胜者（71%）。值得注意的是，ICDAR 2011 竞赛作为开放式竞赛，作者只需要提供他们方法的输出结果。

单词识别的结果无法与其他已知方法进行对比，因为端到端的文本定位和识别并非 ICDAR 2011 Robust Reading 竞赛的一部分并且在数据集上没有其他方法提供文本识别结果。

5.2 街景文本数据集

街景文本数据集（SVT）包含 647 个单词和 3796 个字母，共 249 张从谷歌街景中提取的图像。本数据集更具有挑战性，因为图像中文本具有不同的朝向、字体大小差异更大，并且图像包含噪声。真实数据的格式同

ICDAR 2011 数据集也有所不同——标注只覆盖部分单词。再已被标注的单词中，本方法达到了 32.9% 的召回率（评估方法与上节相同）。文本定位的准确率（19.1%）并不能算作一个衡量标准，因为标注不完整。值得注意的是，许多错误的检测是由于图像中的水印导致的，这也侧面证明了本方法在对抗噪声和低对比度时的鲁棒性。

本方法只能侧面的同 Wang 等人提出的方法，他们使用了不同的评估标，得到 f-测速度为 41%（召回率 29%，准确率 67%）。此外，Wang 等人的方法需要使用词汇表来对图像中的文本进行定位，而我们提出的方法在检测文本时不需要任何关于文本内容的先验知识，因此不会受词汇表的限制。

6 结论

本文提出了一种端到端的实时文字定位和识别算法。在分类的第一阶段，我们使用一系列全新的可在 $O(1)$ 复杂度内被计算出的特征来计算极值区域为字符的概率，并且只有具有局部极大概率的极值区域被选中并进入第二阶段，在第二阶段我们使用一些更加耗费计算资源的特征进行更准确地分类。实验指出，包括全新的梯度幅度投影后极值区域可以覆盖 94.8% 的字符。本算法在 800×600 的图像上平均计算时间为 0.3s（使用普通 PC）。

本算法在两个公共数据集上进行了评估。在 ICDAR 2011 数据集上，该方法在所有已公开的算法中取得了目前最好的字符定位效果（召回率 64.7%，准确率 73.1%，f-测度 68.7%），并且我们是在 ICAD 2011 Robust Reading 比赛数据集中第一个提交端到端字符识别结果的系统（召回率 37.2%，精度 37.1%，f-测度 36.5%）。

在更具有挑战性的街景文字数据集上，文字定位的召回率为 32.9%。但是，我们并没有可以进行直接比较的结果，因此也无从得知结果优劣。