# 第三章 自然场景条件下的文本区域定位

## 3.1 引言

在场景图像中，文本定位的目的是在场景图像中找到文本的位置，文本定位作为文本识别的先决步骤，在识别中占据重要的地位。类似我们人类，当我们关注场景时，并不是直接识别出一个个字符的，而是眼睛注意到一个整块的文本区域，然后对这一个区域进行字符的识别。文本定位的难点在于场景的复杂性和文本的多样性。场景的复杂性表现在场景中存在各种干扰，如窗子，草地，这些干扰通常和文本有着相同的纹理。另外场景中的光照变化也会导致文本区域光照不均，这会导致文本区域无法成为一个统一的区域。文本的多样性表现在文本的颜色、大小和形状不同。本章从场景文本特性出发，提出一种结合最大极值稳定区域和颜色聚类的候选区域提取算法，并引进视觉显著性来评估候选区域的重要性以达到滤除非文本的区域。

分析自然场景文本的特点可得，在一个文本区域内部，一般一个字符内部的颜色变化不大。采用聚类的方式能够使一个字符区域呈现同一种颜色，接着利用连通域分析就可以得到候选的字符区域。而对于一个字符内部颜色变化较大，或者一个文本区域内字符颜色变化大的情况，从文本的灰度变化范围考虑。在文本内部的灰度变化都比较小，而文本和背景的灰度对比度一般都很大，是属于图像中的极值稳定区域。采用最大极值稳定区域算法能够提取出区域内部灰度变化不明显但和背景对比强烈的连通部分。另外从设置自然场景文本的目的考虑，自然场景文本区域大部分是为了引起人们的视觉注意，所以在颜色，纹理等方面和邻域相比更加突出。而视觉显著性能够评估区域在视觉中的独特性和稀缺性，其可以利用来滤除非文本区域。

本章首先对原始的彩色图像进行聚类，再对聚类后图像进行二值化得到候选字符连通域，然后将彩色图像转换成灰度图像，在灰度图像上提取最大极值稳定区域，将提取的区域二值化后得到候选的字符连通域。对于这两部分候选区域进行非显著性区域滤除以及先验信息的限制，最后将候选字符连成文本行。本章方法涉及文本连通域分析，不需要大量训练数据和漫长的训练过程。而合理的显著性区域提取方法的利用能够有效的判定文本和非文本区域，所以不需要大量严格的先验信息。最后我们在公开发表的ICDAR 2003文本定位竞赛数据集上进行测试，验证了本章方法的有效性。

## 3.2 文本定位系统设计

D:\浏览器下载目录\大论文\论文材料&格式\Chapter3_figure1.emf

本文提出的整个文本定位系统如图 所示。为了充分利用场景文本中图像的信息，系统利用两种方法进行候选连通域的提取。首先采用quick shift【】进行颜色聚类，对聚类后的图像二值化后进行连通域的分析，得到候选的连通域。为了克服文本颜色变化比较大的区域无法提取到连通域，将原始图像进行灰度化，然后在灰度图像上提取最大极值稳定区域，将提取到的最大极值稳定区域赋值为1，其余区域赋值为0得到图像的二值化模板，然后对二值化模板进行连通域的分析，得到新的候选连通域。对于得到的每个连通域计算其显著性映射值，将其与整个图像的显著性均值进行比较，超过一定阈值的连通域则保留。最后依据文本的一些先验信息得到包围文本区域的包围盒。下面详细阐述候选连通域提取算法和显著性滤除。

## 3.3 基于颜色聚类的候选区域提取

一般来说，颜色在一个文本区域内部比背景的差异更小。对于一幅24位RGB彩色图，需要处理的颜色范围达到。显然，这么大范围，高精度的颜色对于文本定位这一任务来说是没有必要的。所以我们先用颜色聚类的方式来缩小颜色空间的范围，增大各个颜色间的距离，以便于后续连通域提取。

### 3.3.1 颜色聚类方法

聚类是对数据的一种聚集，类似的数据被聚成一类，聚类是一种无监督的分类方式，不需要预先的训练过程，这是这一方法的一大优点。目前比较常用的聚类方式有K-means，GMM(Gaussian Mixture Models)，Mean shift等。

1）K-means

K-means是这其中简单又最常用的算法，k表示的是聚类中心的个数，每个类是由每个类的聚类中心来表示。k-means包括四个步骤：

1、随机选择初始的聚类中心

2、把每个目标分配给最近的中心（最近度量指的是目标到中心的欧式距离）。

3、根据上一步聚好的类，重新计算聚类中心（所有点到上一步中心的平均值）。

4、重复2直到，聚类中心不在发生变化

K-means 的优点是简单快速并且能够很容易的实现，其缺点是K值需要人工设定，另外对于离群点太过敏感。

2）GMM

GMM用于聚类是基于这样的假设即数据空间的数据是由多个高斯模型生成的。GMM解决的问题是根据数据空间的数据来估计多个高斯模型的参数，然后根据这些参数确定高斯模型，一旦高斯模型确定了，那么数据的聚类中心也就确定了。

假设数据空间中的数据符合高斯混合模型即



其中 是聚类中心的个数，代表的是高斯模型之间的权值并且符合 。其中符合下式



其中 和 代表的是第个聚类中心的均值和方差。那么需要根据数据来估计GMM的参数为 。估计GMM的参数的算法为EM算法，EM算法分为两步，在E步假设参数是已知的，根据最大似然来估计每个高斯模型的权值。在M步对E步得到的结果重新估计参数。然后不断迭代的使用EM步骤，直到收敛。

GMM用于聚类的优点是最后得到的结果是数据属于某个类的概率，其缺点同样是先确定聚类中心的个数。

3）Mean shift

Mean shift的目标是根据给定的数据寻找概率密度的局部最大值。Mean-shift主要包括四个步骤：

1、随机选择若干个感兴趣区域

2、计算感兴趣区域数据的中心

3、移动区域到新的中心

4、不断计算直到收敛

Mean shift的数学定义如下



其中为核函数加权下的Mean shift代表的是均值向量移动的方向，是感兴趣区域的中心， 表示的是核函数， 表示的是带宽。在算法中，首先确定Mean shift的起点为感兴趣区域的中心，然后计算Mean shift 向量的终点式【】中的第一项。然后将感兴趣区域的中心移动到向量的终点，重新计算Mean shift，当式【】满足时时，结束移动，这时中心就收敛到数据空间中局部最大值。



Mean shift的思想是将数据点分配给隐含概率密度函数的某个模型。它的优点是聚类的类别数不需要预先知道并且聚类的结构可以是任意的，它的缺点是计算复杂度太高。

4）Quick shift

Quick shift是由Mean shift[9]改进而来的。Quick shift改进这一缺点，它不需要使用梯度来寻找概率密度的模式，而仅仅是将每个点移动到使概率密度增加的最近的点来获得。公式如下



其中代表的是特征空间中点的下一个位置， 代表的是两点之间的距离， 是核函数，一般选择高斯核函数，是特征空间中点的个数。通过不断移动，所有点连成了一颗树，再通过一定的阈值将树分割成一个森林，这样森林里的每棵树就是一个聚类。特征空间是一个五维空间，包含转换到Lab空间的三个颜色分量和两个空间位置信息。

Quick shift继承了Mean shift的优点，不需要指定聚类中心，但同时改进了其速度慢的缺点。在实验中，我们从效果，时间等各个方面比较了上述四种聚类方法，实验结果如下

在实验中，首先采用Quick shift算法对图像进行聚类。每个像素都有一个相对应的类别标签，一般认为图像中整个字符区域都有相似的颜色。从图【】中可以看出，经过聚类后，颜色的类别数大大减少了，从而增大了字符区域和背景的对比度。得到聚类图像后，在图像上进行二值化两次，两次处理是为了获得亮文本和暗文本。获到聚类图像的二值化结果之后，对它们进行连通域分析，就得到文本区域的候选区。

基于颜色聚类的候选区域提取结果如【】所示，从图中可以看出经聚类后如图(a)，将图(a)的结果经过灰度化后，再将灰度范围[0,255],用颜色蓝到红之间映射可以得到结果图(b)(在彩色情况下显示)，从图(b)中可以看出，暗文本区域占据是蓝色范围，其对应的背景占据的是红色区域。亮文本区域占据的是红色区域，而对应背景占据的为黄色区域。于是通过设定阈值为灰度范围中值，暗文本图(c)是通过将高于中值的区域赋值为1，低于中值的为0。亮文本图(d)则反之。

虽然颜色聚类可以取的较好的结果，但是对于文本区域颜色相差较大的情况，会遗漏候选连通域。这时需要利用文本区域的另一个特性来提取更多的连通域。

## 3.4 基于最大极值稳定区域的候选区域提取

对于颜色变化较大的区域，本章从文本的灰度变化情况考虑，提出用最大极值稳定区域提取算法来获得更多的连通域。最大极值稳定区域是由Matas[6]等人提出的一种仿射特征区域提取算法。最大极值稳定区域特征是区域内部的灰度几乎不变但是和背景的对比十分强烈，它是通过尝试多重阈值来进行处理的，那么提取到的区域能够在多重阈值下保持形状不变。

最大极值稳定区域先将图像转换成灰度图像，然后在一定的阈值下将图像转换成一系列的二值图像，随着亮度阈值的增加或者减少，区域不断的出现、生长和合并。两个不同阈值间的区域变化不超过一定阈值就能够被认为是稳定的。最大极值稳定区域的数学定义：定义图像为区域到灰度的映射 ,其中满足全序结构。定义像素间的邻接关系。则图像中的区域 可定义为图像上满足连接关系的连通子集，即对于任意点，有下式成立



其中。定义的边界 为



对于 和 ，有成立，则称 为极大值区域，反之为极小值区域。对于一组相互嵌套的极值区域 。如果其面积变化率



在 处取得局部最小值，则称 为最大极值稳定区域。

最大极值稳定区域能够同时提取图像中最大极值稳定区域和最小极值稳定区域，最小极值稳定区域是在灰度图像反转后提取的。在得到极值稳定区域后，我们将稳定区域赋值为1，将其余区域赋值为0。这样就得到MSER的二值化模板。对二值化模板进行连通域分析，就得到了候选的连通域了。如图【】所示图(a)虽然文本区域颜色不一致，但是背景和前景灰度对比鲜明，MSER提取的效果很好。

## 3.5 基于视觉显著性的滤除

上面的两个方法能够把大部分的文本检测出来，但是同时也引进较多的非文本区域，另外我们也需要把单独的字符连成文本词，这样有利于后续的处理。

对于较多非文本区域的问题，从显著性区域考虑，一方面文本定位应用一般是为场景字符识别做基础，于是当我们拿着智能手机或者摄像机来获取这些图像时，一般会基本对准字符，以便使字符落在镜头里。另一方面从人们设置场景文本目的出发。场景中的文本集中于海报，广告牌，店名，提示，警告等，所有这些都是为了引起人们的注意，所以文本大部分是场景中的显著区域。从这两方面出发，我们可以利用目前较好的显著区域检测方法来滤除非文本的连通区域。

视觉显著性和相关的认知神经学在过去几十年里已经获得广泛的研究，人类的视觉注意机制能够将有限的认知资源聚集于场景中重要的刺激而抑制那些不重要的。在计算机视觉领域，有关显著性的研究在于提出模拟人类视觉注意机制的模型。现有的显著性检测的方法有基于层次的方法、基于频谱的、基于熵的和基于图像对比度的方法。基于层次的方法采用的是多尺度的图像处理，然后不同层次的结果聚集起来成为最后的显著性映射。基于频谱的方法是把图像分解成Fourier 或者Gabor频谱，显著性映射通过选择那些重要的频谱系数来构成的。基于图像对比度的方法思想是整幅图像或者指定的子窗口的均值像素亮度值被利用来计算图像中每个像素的对比度，对比度用来模拟为图像的显著性。

这里的显著性映射是采用论文【】的方法即将图像抽样成一些随机的感兴趣区域。采用这种方法的理由主要是这一方法仅需要调节一个参数，运算时间中等，并且能够在原图上进行操作。其流程如图5所示，对于一副图像如RGB（）,第一步用高斯滤波器进行滤波并将RGB空间转换成Lab空间；第二步随机生成n个窗口，对于每个窗口，计算出面积与灰度和的比，即



然后依据下式计算窗口中每个像素的显著性映射



第二步是在三个通道上分别进行的，所以最后的显著性映射要采用欧式距离把颜色空间的显著值融合在一起。



为了将显著性用于滤除非文本区域，我们首先要计算出整幅图像的显著性均值。在得到候选的连通域后，计算连通域所包围的原图部分的显著性均值。如果这部分均值大于整幅图像的，那么就保留相应的连通域，否则丢弃。整个过程如图4所示。

此外我们采用字符的先验信息如大小、长宽的比例来滤除，但没有像一般连通域分析方法那样引入众多参数来进行滤除。有了前面的基础，我们可以仅仅用大小和长宽比这两个先验信息来滤除。而字符连成文本词，我们也没有采用训练的方式获取字符间的连接关系来得到文本词，而是采用形态学的膨胀腐蚀来实现。显著性滤除和文本行构造如图6所示，图(a)表示显著性映射的灰度图，同样我们将灰度范围用颜色蓝到红进行映射得到图(b)，从图(b)中可以看出文本区域属于红色区域，代表其显著值高，而背景部分属于蓝色，代表其显著值低。将图3得到亮暗文本经过显著性滤除后可以得到图(c)，对图(c)进行水平方向膨胀，然后根据文本行的长宽比进行滤除可以得到图(d)。

图 5显著性提取流程图

Fig. 5 The flow of saliency extraction

## 3.6 实验结果与分析

### 3.6.1 文本定位的数据集和评价标准

为了验证本章方法的有效性，我们采用的是公开发表的ICDAR 2003文本定位竞赛数据集。该数据集包含两个部分，一部分用来训练，一部分用于测试，由于没有采用学习的方式，我们利用测试部分的数据进行实验。测试部分数据包含251张来自不同场景，不同光照和不同像素大小的图片。评价标准采用的是正确率和召回率，可以用下式表示



其中，正确率等于正确估计的目标个数除以总共估计的目标总数。召回率为正确估计的目标个数除以图像中原有的目标总数。因为定位系统标记的方框无法和人类标记的一模一样。为了使评价标准更加合理，ICDAR2003文本定位竞赛重新定义了正确率和召回率。首先定义两个矩形的重合度为两倍的交叉面积除以各自面积的和即



其中，是矩形的面积。对于一系列的矩形，某个矩形与其的最佳匹配如下



新的正确率和召回率可定义如下





另外综合了正确率和召回率的 测量可定义如下



### 3.6.2 文本定位结果及分析

表格 1列出不同算法在ICDAR2003数据集上的性能，上一栏是文献的方法，文献是采用连通域分析的方法，文献中定义了六个不同类型的先验信息如文本区域的大小，区域的交叉比例，区域轮廓梯度等来滤除非文本区域。有了显著性的评估，本文仅采用两个更加合理的先验信息来滤除非文本，分别为文本区域的大小和高宽比，本文设定如下

 (14)

 (15)

其中 代表的是文本连通域的大小，  代表图像的宽度， 代表图像的高度。

文献是采用训练的方法，这一方法需要在多个尺度上对图像进行处理，而且必须预先有训练的样本。本文方法能在原图上进行处理，无需训练的数据和流程。从表格中可以看出下一栏是一些参加竞赛的算法。前两种方法也都是基于训练的，采用的分类器为支持向量机，不同的是提取的特征，分别为边缘特征和纹理特征。后面两种是基于连通域分析和先验信息滤除。从表中可以看出本文算法正确率达到了68%，召回率60%优于其他的算法。部分定位效果如图7所示，我们分别选取了几种不同的情况，分别为大字体，侧面角度，草地干扰，一般情况，以及窗子干扰。

## 3.7 本章小结

自然场景条件下的文本区域定位是文本识别的基础，在这一章的主要目的是尽可能多的提取到文本区域。在前一章对场景文本图像进行深入分析的基础之上，我们利用自然场景文本颜色、灰度与背景差异的特性来提取候选的文本区域。首先对于场景文本的原始彩色图像，从颜色角度分析，本章提出了基于聚类的候选文本区域提取算法。在对目前常用的颜色聚类方法进行调研，分析，比较的基础上，选用Quick shift进行颜色聚类，在聚类后的图像上进行连通域分析得到一部分候选区域。从另一个角度，场景文本灰度和背景的对比上分析，本章采用了最大极值稳定区域提取算法来提取新的候选文本区域。这两种方法虽然能够很好的提取大部分区域但是同时也引进了非文本区域。利用场景文本的设置目的，本章采取显著性映射来评估区域的重要性以达到滤除非文本的目的。最后在公开的数据集ICDAR2003上验证本文方法的有效性，并和其他方法进行比较和分析。