# 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，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，得到候选的连通域。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。

### 3.3.1 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。

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。

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。

数学表达。。。。。

Mean-shift

quick shift是由mean shift[9]改进而来的。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。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信息。

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。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。这样图像灰度化后经过类似MSER的处理即对灰度图进行两次二值化处理，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图3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。

## 3.4 基于最大极值稳定区域的候选区域提取

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。MSER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。MSER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。

MSER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。最大极值稳定区域算法能够提取跟背景亮度对比强烈的文本，但如果文本跟背景亮度相差不大或者图像存在模糊时，其效果会下降很多。如图2所示图(a)背景和前景对比鲜明，MSER提取的效果很好，文本区域明显。图(c)背景复杂，提取出的MSER区域将文本区域和背景混肴在一起

## 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。

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。

图 4基于显著性的区域滤除

Fig. 4 Region filter based on saliency

这里的显著性映射是采用论文[10]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。



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

为了验证本文方法的有效性，我们采用的是公开发表的ICDAR 2003文本定位竞赛数据集[7]。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



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[11]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[2]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。

### 3.6.2 文本定位结果及分析

由结果可以看出，使用Otsu算法分割效果很好，目标能完全分割出来，分析原因是因为弱小目标在图像局部中所占比例较大。

## 3.7 本章小结

红外弱小目标检测中，需要首先对图像进行预处理，使用预处理后的图像与原图像做差分运算就能使目标信息得到极大地增强，然后使用合适的分割算法就能将弱小目标从背景中提取出来。本章在预处理阶段介绍了中值滤波、最大中值滤波、均值滤波、最大均值滤波、自适应中值滤波、高通滤波与Top-Hat滤波，在图像分割阶段介绍了改进的局部Otsu二值图像分割和基于四叉树的图像分裂合并分割算法。并对部分预处理算法和分割算法进行了仿真与实验，实验结果显示均值滤波最能提高目标局部信噪比，Top-Hat滤波最能提高目标局部对比度，而且Top-Hat算法处理时间最短，很适合用来对图像进行背景抑制。而分割算法中基于局部Otsu分割算法和区域分裂合并算法则都能达到很好的分割效果，为以后的研究奠定了基础。