图片验证码识别算法的设计与实现

张扬

摘要：

图片验证码作为一种防止机器人使用重复手动暴力尝试获取网络资源的一种手段，其生成方式简单且种类多样，主要用于区别人和机器。但各种验证码生产程序生成的图片验证码保护强度不一。对验证码机器识别的研究可以为日后研究更强的验证码生成规则提供基础。本文主要针对一种样式的图片验证码，给出一系列图像处理的方法，包括二值化，修复，去噪。并针对性地设计一种简单高效且有较高正确率的字符切割方法最后使用支持向量机进行机器训练和预测。

1. 引言

图片验证码技术是现代网络安全技术的一部分，用于区分访问对象是人还是机器人，以防止使程序进行大流量的有目的的访问。一般的验证码可以阻止一些简单的机械试的访问，在如今机器识别已经成熟的时代，对于简单排列的数字和字母的图片识别已经不是什么难题了。国内外已经有很多学者对验证码的识别进行研究。尝试机器识别验证码的目的不是破解而是发明更有效的，难以机器识别的验证码来保持发明验证码的初衷。验证码的识别包括图片预处理阶段，包括图片灰度化，图片二值化，图片去噪，图片增强等；字符分割阶段，也是识别验证码最难最关键的阶段，字符分割的好与坏会直接影响到下一步分类学习识别阶段；机器学习与预测阶段，这一阶段常用的分类及预测的有KNN和SVM算法。现阶段的研究表明，字符分割是验证码处理的重难点，目前还没有通用的完美的算法去解决复杂粘连字符的分割问题，有些甚至人眼也需要一定时间去区分，这是阻碍机器识别验证码的主要障碍，也成为如今验证码依然可以成为人机区分手段的基础保障。国外学者指出，当一组样本的验证码能够被机器有效识别（完全识别正确）超过10%则该验证码（系统）已经不再安全。

本课题旨在以验证码识别为媒介学习图片处理图片分割以及机器学习的一些算法，尝试设计一套简单的图片验证码识别算法。

第一节对图片样本的特性作出一些分析，第二节主要对图像预处理各步骤算法作介绍和一些性能分析，第三节主要介绍和分析字符分割算法，第四节对实际预测效果和性能作出一些测试和评价。

1. 原始样本的特性分析

针对一种特定规则生成的图片验证码设计图像处理机器识别算法是改进其生成算法以防止后续扔被机器暴力破解的前提，因此针对特征图片进行图像处理时可以避免大量嵌套一系列通用图像处理的算法而针对性地设计特征处理算法，因此需要在算法设计前期针对原始的图片样本做一些图像的特征分析。本案例选取500张样式样本作为机器学习的学习样本。

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
| 图1. 一些原始验证码图片样本 | |

通过肉眼观察上述验证码图片的样本，我们可以发现一个样本图片主要由背景，干扰，字符三部分组成，而且这类生成规则下的验证码字符元素的粘连不大甚至几乎没有，这对于图片后期的字符切割是极为有利的，同样也就预示着这类验证码极其容易受到机器的暴力破解。

* 1. 图像结构及描述的定义

|  |
| --- |
|  |
| 图2. 图像定义 |

本文中处理的图像坐标原点设在图像左上角，图像宽度方向为X轴，高度方向为Y轴对于其中的给定一点像素为方便表述及后续处理，定义像素位置向量

定义像素位置复数

。

对于输出为实数的（离散）函数 ，其二元函数表达形式 二维向量表达形式，复变表达形式 在本文中表达相同意义，并取其表述简单的形式对算法进行描述。

图像区域

其中width表示图像宽度，height表示图像高度。

* 1. 样本集中主要图像结构元素的分布

我们将样本集中所有的样本图像叠加后求平均灰度值，便可以观察到整个样本集中干扰字符等的分布状况。

对样本集做如下处理

处理后我们可以得到如下结果：

|  |
| --- |
|  |
| 图3. 叠加后的灰度平均值分布情况 |
| 图中，蓝色矩形框表示主干扰存在区域，红色矩形框表示字符的主要存在区域。  我们可以发现，有较多分布的较大干扰反而对后期字符处理影响不大，由于与字符间隔较大，它们可以很容易得被去除。因此我们主要需要处理的噪声为穿过字符存在区域较细的长干扰线。 |

* 1. 图像背景灰度值分布的前期分析

通过对验证码图片的样例进行观察可以得知图片背景灰度在图片高度方向几乎没有变化，在沿着图片宽度方向逐渐变亮。我们选取几行不经过字符主要存在区域的行作为背景采样行，用500张样本进行灰度分布的采样。

做如下操作

|  |
| --- |
|  |
| 图4. 主要背景区域采样行灰度值分布情况 |

对上述采样图像及原始样本图像进行分析可以得知，在宽度20和45位置的两个波谷是由黑色的主干扰线引起的平均灰度值下降，第49行的平均灰度出现另外四个明显波谷是由于有些字符如p,g，向下延伸较大，在第49行仍有黑色的字符像素分布引起的。四个波谷几乎恰好对应四个字符的分布位置。

略去黑色干扰线和个别字符引起的灰度值分布波动，背景的灰度值从图像左侧到右侧明显呈上升趋势，从采用图像上粗略估计背景灰度的最小值应该是不低于185 。

1. 图像预处理算法设计及性能分析
   1. 灰度图像二值化

本案例中样本灰度图像的二值化主要需要讨论两个方面，使用常规的二值化阈值确定方法还是根据样本特性设计一种更简单的二值化阈值确定算法；另一方面对样本进行二值化时，每个样本确定一个二值化阈值还是所有样本使用一个全局的二值化阈值，两者效果是否有较大的差距。