

**毕 业 论 文**

**题 目** 图片验证码识别算法的设计与实现

**学生姓名：** 张扬

**学 号：** 1405140130

**所在学院：** 计算机科学与技术学院

**专 业：** 计算机科学与技术 (软件班)

**指导老师：** 白光伟 教授

2018 年 6 月

图片验证码识别算法的设计和实现

# 摘 要

验证码作为一种防止机器人使用重复手段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。现在已有各种各样的验证码形式，如基本的图片输入型验证码，新型的滑条验证码，选择验证码等等。其中对于图片验证码，各种验证码生成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。使得图片验证码能够有更强的保护能力。

随着近年来机器学习和图像识别领域的不断发展，各种新型算法不断诞生，已有成熟的算法也不断被改进和完善，针对传统的简单验证码，目前机器已经有能力破解。传统的图片验证码生成规则亟待改进。

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。针对性地给出一些特征算法并尝试分析其性能。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。

本文将详细介绍前期图片预处理与字符分割时所设计的特征性算法，包括算法设计，参数调整，性能测试。简略介绍进行字符分类及预测的支持向量机算法及参数调整。

**关键词：**图片验证码 图像处理 字符分割 性能分析 支持向量机

**Design and Implementation of Image Captcha Recognition Algorithm**

# Abstract

The captcha, as a means to prevent the robot from using repeated means to violently attempt to obtain network resources, is simple and diverse in its generation method, and is mainly used to distinguish people from machines. There are various types of captcha, such as basic picture input type captcha, new type of slide captcha, selection captcha, and so on. For the image captcha, the captcha generated by various captcha generation programs have different protection strengths. Research on captcha machine identification can provide a basis for future research on stronger captcha generation rules. Makes the image captcha more capable of protection.

With the continuous development of machine learning and image recognition in recent years, various new algorithms have been born and mature algorithms have been continuously improved and perfected. For traditional simple verification codes, current machines have the ability to crack. Traditional image captcha generation rules need to be improved.

This paper mainly focuses on a style of image captcha, and gives a series of image processing methods, including binarization, repair, and denoising. Give some feature algorithms targeted and try to analyze their performance. And specifically designed a simple and efficient and high accuracy of the character cutting method and finally use the support vector machine(SVM) for machine training and prediction.

This article will introduce the characteristic algorithm designed in the pre-image preprocessing and character segmentation in detail, including algorithm design, parameter adjustment, and performance testing. The SVM algorithm and parameter adjustment for character classification and prediction are briefly introduced.

**Keywords:**image captcha;image processing; character segmentation; performance analysis;SVM

# 目录

[摘 要 I](#_Toc514081818)

[Abstract II](#_Toc514081819)

[目录 III](#_Toc514081820)

[第一章 绪论 1](#_Toc514081821)

[1.1 课题背景及意义 1](#_Toc514081822)

[1.2 国内外研究现状 1](#_Toc514081823)

[1.3 研究思路与方法 3](#_Toc514081824)

[1.4 论文结构 3](#_Toc514081825)

[第二章 相关工作 5](#_Toc514081826)

[2.1 图片验证码机器识别处理流程 5](#_Toc514081827)

[2.2 图像处理技术 5](#_Toc514081828)

[2.2.1 二值化 5](#_Toc514081829)

[2.2.2 修复与去噪 5](#_Toc514081830)

[2.2.3 字符分割算法 6](#_Toc514081831)

[2.2.4 样本归一化（图像缩放） 6](#_Toc514081832)

[2.3 模式识别技术 7](#_Toc514081833)

[第三章 原始样本的特性分析 8](#_Toc514081834)

# 绪论

## 1.1 课题背景及意义

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式的访问，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，难以让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。

研究验证码机器识别可以为后续对验证码生成规则的改进提供数据支持，避免使用那些已经可以近乎完美解决的方法来增加识别难度，为开发新的干扰要素提供支持，使得图片验证码能够继续提供有效人机区分的服务。

本课题旨在以验证码识别为媒介学习图片处理和图片分割以及简单了解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。

## 1.2 国内外研究现状

目前国内外的研究表明，对于一种生成规则的英文字母和数字只要有足够多的学习样本，通过机器学习进行分类和预测已经不是什么难题，这也就表明简单的字符或者字符数字组合的验证码已经没有什么保护效果了。

|  |  |
| --- | --- |
|  |  |
| 图1. 基本不具有保护效果的简单验证码 | |

上述字符清晰，几乎没有干扰的字符型图片验证码经过适当处理基本可以被机器完美识别。但是在中国，汉字型验证码的出现使得即使不存在干扰字符清晰可辩没有粘连也使得机器学习有些力不从心，因为过多的样本基数和不同的汉字字形对于人类来说可以简单识别，至多就是增加了些打字时间；但是对于机器学习来说，样本基数过多太过耗费对比时间，字形样式复杂不能使用简单的特征向量进行学习等等因素使得“简单”的汉字验证码依旧十分有效，这也算是在机器学习层面防止验证码破解的一种手段。

对于输入型的验证码更多的是通过增加图像处理难度来防止机器破解，如添加干扰，字符扭曲，字符粘连等等。

|  |  |
| --- | --- |
|  |  |
| 图2. 添加干扰要素的图片验证码 | |

|  |  |
| --- | --- |
|  |  |
| 图3. 形成字符粘连的图片验证码 | |

对于验证码中的图像干扰，至今没有通用的解决方案，需要根据特点来“按需设计”，有些图像验证码为了防止破解干扰要素添加得即使是人类也需要一定的辨认时间，这就使得机器处理不是那么容易了。

对于粘连的字符，机器是很难根据特征去分割的。对于粘连字符的分割目前不断有算法改进，但对于谷歌式的极度粘连的验证码至今尚未有良好分割效果的分割方案。但是过度的粘连依旧会对人类识别也造成困难。

国内外的研究表明，对于粘连字符的分割远比去除干扰要素困难的多，由于干扰要素特定的生成规则，总是可以找到一些“有所作为”的解决算法；但是对于粘连的字符，大多数的分割算法依旧是通过寻找字符间空隙来分割，虽然提出了很多可以寻找最佳切分轨迹的字符分割算法如滴水分割，惯性滴水分割等等，但是这对于那些完全粘连甚至通过字符交错来粘连的验证码处理几乎是不可能的。

对于其他形式的输入型验证码的研究如计算求解型，动画型等超出本文研究范围的再次就不多做过多赘述。

## 1.3 研究思路与方法

首先从网络上获取多组验证码样本，每组500张，分别用于训练，强化和识别测试。

将获取到的样本通过Matlab进行样本的数据分析，分析样本的图像特征并针对性地设计处理算法。

对处理完后的单字符样本先通过支持向量机得到一个初步模型，然后使用第二组样本数据对无法识别的字符部分补充进模型进行强化训练。并对改进后的模型不断调整参数使得用于识别测试的第三组样本可以达到最好的识别效果。

将所涉及的算法及其他部分整合成java应用程序，并且组织适当的用户图形界面。

## 1.4 论文结构

本文主要研究针对一种样式的图片验证码，设计适当的图像处理算法，字符分割算法，归一化处理算法，并且设定适当的参数使用支持向量机对输入图片进行机器字符识别。

本文的组织结构以及各章节的主要内容如下：

第1章为绪论，主要论述课题的研究背景与意义，同时分析了国内外的研究现状，并介绍了论文的主要工作与组织结构。

第2章为相关工作，主要介绍与本文密切相关的一些研究工作，包括图像处理字符分割相关理论的了解以及用于分类预测的机器学习模型的了解。

第3章为原始样本的特性分析，主要介绍从肉眼直观上能获取到的样本信息的分析以及描述本文对图像结构的一些定义。

第4章为图像处理算法设计与性能分析，主要介绍针对性算法（二值化，边缘修复，去噪）的设计以及这些算法的实际效果测试。

第5章为字符分割及单字符样本的处理，主要介绍字符分割算法的设计及效果测试，以及分割完的单字符样本大小归一化算法。

第6章为字符分类与预测算法，简单介绍使用的支持向量机中参数的设定以及预测效果的测试。

第7章为程序架构介绍，主要简单介绍程序各部分的组成。

第8章总结本文研究成果，并且展望下一阶段的研究方向。

# 相关工作

## 2.1 图片验证码机器识别处理流程

图片验证码的识别主要依赖于图像处理技术和模式识别技术。图像处理技术主要包括灰度化，二值化，去噪与修复，字符切割，样本归一化等。

模式识别技术主要包括字符样本特征的提取，机器训练与识别等。

下图为一个验证码识别系统的基本处理流程：

|  |
| --- |
|  |
| 图4. 验证码识别处理流程图 |

## 2.2 图像处理技术

本小节主要介绍本文用到的图像处理各阶段（二值化，修复与去噪，字符分割，归一化）的一些通用算法。

### 2.2.1 二值化

二值化有全局阈值二值化和局部阈值二值化两种基本算法。全局阈值即整幅图像使用唯一的阈值将图像进行二值化处理；局部阈值即将图像进行分块，对图像的每一区块使用不同的阈值进行二值化，这种二值化方法适用于图像背景照度不均匀或者图像构成极为复杂的图片。

全局阈值二值化主要有OSTU，灰度拉伸法,直方图双峰法等；局部阈值二值化主要有Bernsen算法等。

对于这些常见的二值化算法的具体描述，本文暂不做过多赘述。

### 2.2.2 修复与去噪

图像修复与去噪是图像处理中的重要环节，主要是针对图像中的一些特征给出一系列变换算法达到一定目的的过程。

在模板卷积法处理中如有正方形窗口中值滤波去椒盐噪声；高斯平滑增强图像边缘等。

通过频域处理的如巴特沃斯低通滤波来防止图像出现“振铃现象”等。

通过形态学处理如开运算来去除细小连通成分，闭运算填充细小空缺等。

这些通用的基本图像处理算法给本文后续去噪算法的设计提供了思路，并且可以作为本文去噪算法效果的对比对象。对于这些算法的具体描述，这里不做过多赘述。

### 2.2.3 字符分割算法

字符分割是图像分割的一个子集，它可以使用图像分割的基本方法，也可以尝试针对字符图像的特征单独设计。

对于没有粘连的字符进行分割，我们可以尝试使用图像分割中的区域生长，找出封闭的字符形状区域将这些区域分割开。

与此同时，区域生长算法需要了解的“信息”过多，也就这较多的迭代次数，对于没有粘连的字符，比如车牌字符的分割，清晰银行卡号的分割完全可以使用简单而且效率很高的垂直分割法。

对于没有粘连的验证码，也同样可以使用垂直分割，如果验证码有粘连，那么需要考虑更为复杂的分割算法如分水岭分割算法，滴水分割算法等。

本文下一章将会简单介绍待处理的验证码样本，届时可以看出本文的样本支持使用垂直分割的一些要求但不完备，因此需要基于垂直分割重新设计或者补充算法来达到一个较好的分割效果。

### 2.2.4 样本归一化（图像缩放）

经过分割处理后的单字符样本由于图像规格不一样，不能直接提取特征向量输入，因此需要对图像进行归一化（缩放）。

通常图像的缩放针对的是比例缩放，如果是不锁定纵横比的缩放，通常情况是将图像沿着纵向和横线分别缩放处理。

其中图像放大主要采用插值算法，具体有线性插值，多项式插值，高斯插值等。根据对结果图像平滑度和清晰度要求的不同可以选取不同的插值方法（函数）。

图像缩小主要采用下采样或者平均法，下采样即在原图像上等比例采样部分像素点，使得这些像素点正好满足新图像的大小要求；平均法即先进行下采样，并对下采样点周围将被忽略的像素点进行求平均，以平均值来填充新图像。

## 2.3 模式识别技术

模式识别，就是通过计算机用数学技术方法来研究模式的自动处理和判读。最终进行字符识别通常是有监督的机器学习，当然我们前期可以通过无监督的聚类将大量的数据样本进行预分类。

通常情况下对于需要监督的模式识别，结构越复杂需要的训练样本越多，同时识别预测的正确率也就越高。

一般情况下用于字符识别的模型有KNN,SVM,BP,CNN(卷积神经网络)等。前两者的模型相对于后两者简单的多，当然后两者当训练数据足够多时有着极高的识别精度。

KNN与SVM在字符识别领域已经有着相当高的正确率，它们经常被用于字符分类和识别，由于这两者的模型较为简单，训练所需的样本容量不需要像卷积神经网络那么庞大，适合于测试。

CNN模型中有非常成熟且在2000年前就早已商用的LeNet-5数字与英文识别模型。该模型早在上世纪末就用于银行支票手写体的识别可见CNN模型的识别精度之高。

# 第三章 原始样本的特性分析