Computer Science and Application 计算机科学与应用, 2017, 7(11), 1059-1066
Published Online November 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.711120

Hans汉斯

Research on Automatically Identifying Algorithm and Application about Handwritten CAPTCHA of Kinds of Diverse Transformation

Chuncai Wang[1], Yuanyuan Sun[2]

[1]Department of Science and Technology, Changchun University of Science and Technology, Changchun Jilin
[2]Engineering Research Department, Changchun WHY-E Science and Technology Co., Ltd., Changchun Jilin
Email: wangchuncai@why-e.com.cn, sunyuanyuan@why-e.com.cn

Received: Nov. 3rd, 2017; accepted: Nov. 16th, 2017; published: Nov. 23rd, 2017

Abstract

Research on technique of automatically identifying CAPTCHA can promote people to identify the readability of verifying the code further and strengthen the difficulty that the machine identifies and raises a network safety thus. Currently aiming at the CAPTCHA identification methods are basically used optical character recognition (OCR) method to identify the standard characters written by the machine. The paper puts forward color CAPTCHA identifies to mainly include color verification code binary by threshold, the connect district segmentation of the handwritten character list, the near district of the same character list links, use convolution neural network to train a character and to identify handwritten character. The paper realization obviously surpasses an identifying of OCR result. The result shows CAPTCHA of the website that is basically passed by the website test; the website can automatically identify CAPTCHA of the website.

Keywords

The Color Completely Automated Public Test to Tell Computers and Humans Apart (CAPTCHA) Binary by Threshold, The Connect District Segmentation, The Distance Near District Link, Use Convolution Neural Network (CNN) to Train a Character, Identify a Character

一种多样变换的手写验证码自动识别算法的研究及应用

王春才[1]，孙媛媛[2]

[1]长春理工大学，计算机科学技术学院，吉林 长春

文章引用：王春才，孙媛媛. 一种多样变换的手写验证码自动识别算法的研究及应用[J]. 计算机科学与应用，2017, 7(11): 1059-1066. DOI: 10.12677/csa.2017.711120

[2]长春市万易科技有限公司工程研究中心，吉林 长春
Email: wangchuncai@why-e.com.cn, sunyuanyuan@why-e.com.cn

收稿日期：2017年11月3日；录用日期：2017年11月16日；发布日期：2017年11月23日

摘　要

研究验证码自动识别技术可以进一步提升人识别验证码的可读性，增强机器识别的难度，从而提高网络安全性。针对目前提出的验证码识别方法基本都是采用光学字符识别(OCR)方法对机器写的标准字符进行识别，本文提出了一种多样变换的手写验证码自动识别算法，对彩色验证码进行识别主要包括彩色验证码的二值化、手写字符的区域分割、同一字符的区域连接、使用卷积神经网络对手写字符进行训练、手写字符识别。本文的实现结果明显优于OCR的识别结果。结果表明通过该网站的测试，基本上能自动识别该网站的验证码。

关键词

彩色验证码二值化，区域分割，相近区域连接，单个字符卷积神经网络训练，单个字符的识别

Copyright © 2017 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着互联网的极度开放，网络爬虫技术可以肆意获得任意公开的数据，网络的安全问题也日益突出，其中设置人类可读容易机器破解难的验证码在一定程度上可以阻止肆意的网络爬虫技术。验证码CAPTCHA 起源于美国卡内基梅隆大学的一个科研项目，原意为全自动区分计算机和人类的图灵测试是一种区分用户是人类还是计算机的公共自动程序。当用户要对网络站点进行用户注册、用户登录或非登录式留言回复时，均需先正确填写网站提供的验证码，并通过网站服务器验证后，才能顺利完成各种操作。验证码技术在增强网络安全方面起到了重要作用。

目前的验证码识别在国内外已成为热门领域，其理论体系日渐完善。Per-Olal 对验证码提取特征并用神经网络训练方法对一种简单验证码进行识别。2005 年 Edward 等通过将字符旋转至水平并对单个字符应用小波滤波结果作为特征，以接近 100%的识别率破解了验证码。同年 Kumar 和 Patrice 的研究表明，在验证码识别中，验证码图形字符分割比识别更难，一旦正确将字符逐个分开，运用机器学习算法就可以轻易解决识别问题。然而至今没有一种通用有效方法解决字符分割难题。2008 年 Jeff Yan 等以高达 90%以上的识别率破解了微软 2007 年使用的验证码[1]，对于有单个字符的干扰线的字符去除干扰线识别提出了解决的方法，对于这种长的和字符具有相同的像素宽度的、字符颜色一致的干扰线目前没有很好的解决方法。2005 年程治国，刘允才[2]提出一种通用的去除文字图像中干扰线的算法利用图形学的理论，采用改进的最短路径算法和方向偏移算法检测干扰线。非粘连的正常排列有一定空间字符的分割可以通过二值化后的图像像素的竖直投影直方图来分割字符，但是对于一些扭曲变形斜着排列的字符就不适用，但可以用基于上下连通区域法分割字符，CFS (Color flooding segment)对于彩色填充分割算法适用于具有区域连通性的验证码。粘连字符分割可以采用滴水算法进行字符的分割[3]，有效避免了字符的过分割。

汪洋等[4]采用最近邻 KNN 算法破解了少数银行网站的验证码。殷光、陶亮[5]提出了一种 SVM 验证码识别算法通过采用支持向量机比模板匹配算法识别准确率高。陈超、毛坚桓等[6]采用了卷积神经网络对铁路货运网站的验证码进行识别，使用深度学习方法对字符集进行训练优化网络参数。一些大的开发平台提出使用深度学习直接对网站验证码进行学习，对于这种手写各种形状标记验证码训练样本的数据量巨大。

2. 彩色验证码二值化

本文主要针对一些手写扭曲、抖动变换的验证码进行研究，包括某些字符涂抹的颜色点分布不均和同一字符大小不一、扭曲、抖动不同等特征。验证码图像如图 1 所示其分辨率统一为 160*70 像素，验证码特点彩色带干扰线、人眼不好识别，具有任意手写、抖动的特征，每个单个字符的形状特点不同，见图 1。

对彩色验证码识别前，需将验证码码图像预处理，包括彩色图像灰度化与灰度图像二值化，如果直接对彩色图像自适应阈值二值化，可能会过滤掉某一个亮度和背景相差不大的字符信息，本文采用了红、绿、蓝三通道分别二值化，并取每个通道的最大值作为二值化的图像，使用此种方法对彩色图像进行二值化没有过滤掉有用的字符，见图 2。

3. 彩色验证码字符分割

彩色验证码二值化处理后，采用 CFS 区域连通对字符进行分割，以提取图像中每个字符作为卷积神经网络层输入。根据图像的特点本文采用区域连通性算法对验证码进行分割。本文提到的区域连通算法以种子填充算法为基础，通过算法预先设定种子点，从该种子点出发，经过搜索标记验证码的所有像素点，将与预定义的性质相似如像素一致，找到种子点周围的 8 邻域像素一致的点加入到该种子队列中，来生长这些区域。区域连通算法也可通过沿着水平 x 轴依次扫描所有的像素点，找到字符像素，标记为 1，找到标记为 1 的像素的上、下、左、右、四个对角线八邻域标记为 1，依次标记水平字符像素点。纵向依次扫描水平像素点，找到字符标记像素点依次进行标记，发现周围八邻域按照标记的像素点值进行标记，再对已标记的像素点重新进行扫描进行合并同一区域的字符。对分割后字符进行归一化，通过查找单个字符的上下左右边界来存储单个字符的位置。

如果合并标记的字符区域为 4 个，直接进行卷积神经网络的输入端进行样本训练测试。

如果合并标记的字符区域大于 4 个，单个小的字符噪声应进行删掉，删掉后的字符区域为 4 个再进行卷积神经网络的输入端进行样本训练测试。

比较相邻两个区域的最小距离，如果相邻两个区域的最小两个点之间的距离小于 s1，两个区域可以合并，如果相邻两个区域的最小两个点之间的距离大于 s1 并且小于 s2，其中 s1 < s2，可能是字符 i 或者 j，如果相邻两个区域的最小两个点之间的距离大于 s2，两个区域就可能是两个字符。s1 和 s2 通过选取 10,000 张图片，40,000 相邻两个字符的最近的两个点之间距离比较，挑选出由 i 和 j 的字符的点和无点连续的线最近的两个点之间距离比较，发现完整相邻两个字符之间的最短距离会最大为 s2，i 和 j 之间的点和无点连续的线最近距离会相对小为 s1，相邻字符之间的最短距离小于 s1，把分开的字符合并，见图 3。

4. 彩色验证码卷积神经网络(CNN)模型

CNN 是一个多层的神经网络[7]，每层由多个平面组成，每个平面有多个独立神经元，每个特征图通过交叠滑动方法，在离散上计算局部和，其操作等价于连续上的卷积，即卷积网络 Yann 等提出了基于 CNN 的文字识别系统 LetNet-5 Patrice 等提出了简化的 CNN，见图 4。

本文采用 CNN 进行验证码训练和识别。

Figure 1. CAPTCHA image
图 1. 彩色验证码图像

Figure 2. Result of CAPTCHA binary by threshold
图 2. 彩色验证码二值化结果

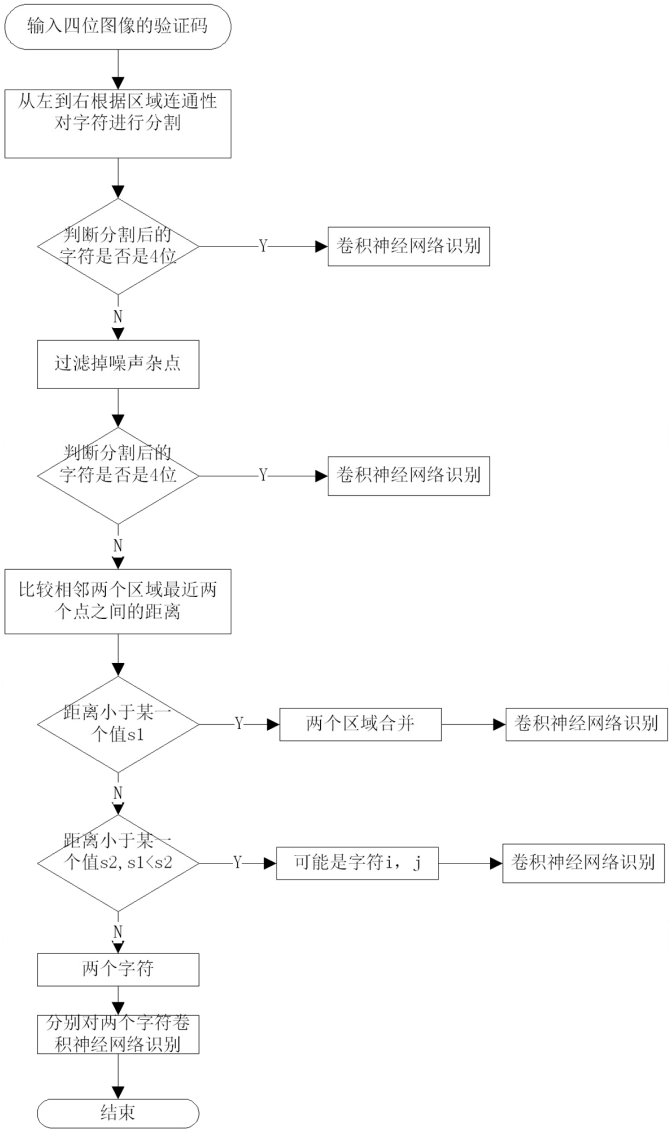

Figure 3. Connect district segmentation of CAPTCHA
图 3. 彩色验证码连通区域分割

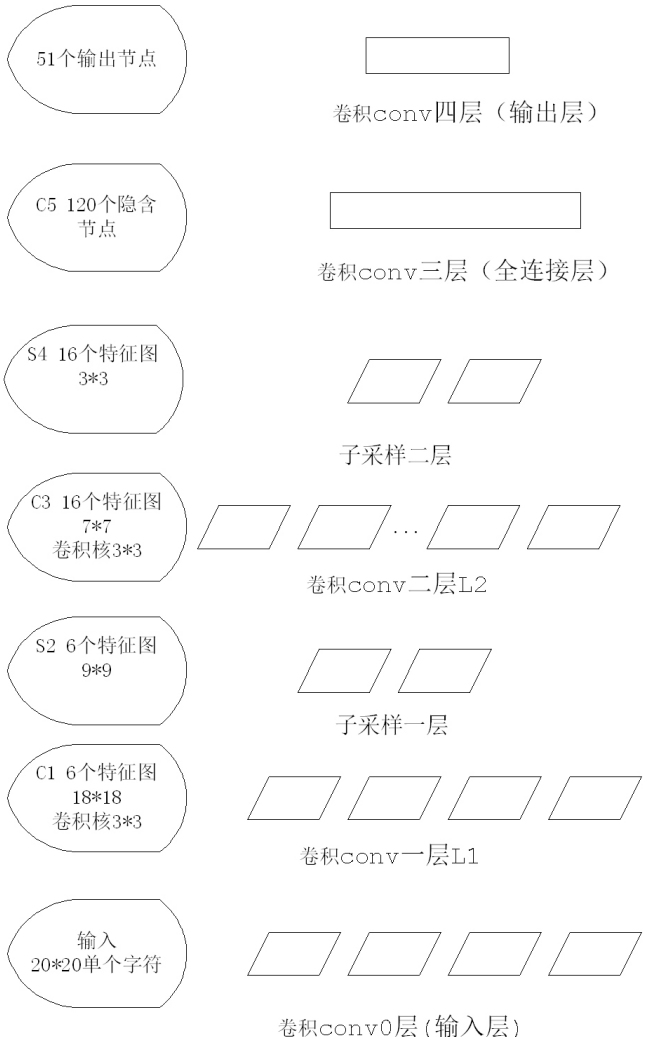

Figure 4. Topology structure of CNN
图 4. 卷积神经网络的拓扑结构

验证码 CNN 拓扑结构如图 4 所示图中自下至上分别为卷积 0 层、子采样层、卷积四层。各层神经元与连接数如表 1 所示。

C1 层是一个卷积层，由 6 个特征图 Feature Map 构成。特征图中每个神经元与输入为 3 × 3 的邻域相连。特征图的大小为 18 × 18，这样能防止输入的连接掉到边界之外(20 − 3 + 1 = 18)。C1 有 60 个可训练参数(每个滤波器 3 × 3 = 9 个 unit 参数和一个 bias 参数，一共 6 个滤波器，共(3 × 3 + 1) × 6 = 60 个参数)，共 60 × (18 × 18) = 19,440 个连接。

S2 层是一个下采样层，有 6 个 9 × 9 的特征图。特征图中的每个单元与 C1 中相对应特征图的 2 × 2 邻域相连接。S2 层每个单元的 4 个输入相加，乘以一个可训练参数，再加上一个可训练偏置。每个单元的 2 × 2 感受野并不重叠，因此 S2 中每个特征图的大小是 C1 中特征图大小的 1/4 (行和列各 1/2)。S2 层有 12 (6 × (1 + 1) = 12)个可训练参数和 2430 (9 × 9 × (2 × 2 + 1) × 6 = 2430)个连接。

C3 层也是一个卷积层，它同样通过 3 × 3 的卷积核去卷积层 S2，然后得到的特征图就只有 7 × 7 个神经元，但是它有 16 种不同的卷积核，所以就存在 16 个特征图。C3 中每个特征图由 S2 中所有 6 个或者几个特征图组合而成。C3 的前 6 个特征图以 S2 中 3 个相邻的特征图子集为输入。接下来 6 个特征图

Table 1. Neuro cell and connection
表 1. 各层神经元与连接数

卷积层	神经元数	权值数	与上层连接数
输入层	400	0	0
卷积一层	1944	60	19440
卷积二层	784	556	27244
全连接层	120	17400	17400
输出层	51	51	51

以 S2 中 4 个相邻特征图子集为输入。然后的 3 个以不相邻的 4 个特征图子集为输入。最后一个将 S2 中所有特征图为输入。这样 C3 层有 6 ×(3 × 9 + 1) + 6 ×(4 × 9 + 1) + 3 ×(4 × 9 + 1) + (9 × 6 + 1) = 556 个可训练参数和 151,600 (7 × 7 × 556 = 27,244)个连接。

S4 层是一个下采样层，由 16 个 3 × 3 大小的特征图构成。特征图中的每个单元与 C3 中相应特征图的 2 × 2 邻域相连接，跟 C1 和 S2 之间的连接一样。S4 层有 32 个可训练参数(每个特征图 1 个因子和一个偏置 16 × (1 + 1) = 32)和 720 (16 × (2 × 2 + 1)× 3 × 3 = 720)个连接。

C5 层是一个卷积层，有 120 个特征图。每个单元与 S4 层的全部 16 个单元的 3 × 3 邻域相连。由于 S4 层特征图的大小也为 3 × 3 (同滤波器一样)，故 C5 特征图的大小为 1 × 1(3 − 3 + 1 = 1)：这构成了 S4 和 C5 之间的全连接。C5 层有 17,400 (120 × (16 × 3 × 3 + 1) = 17400 由于与全部 16 个单元相连，故只加一个偏置)个可训练连接。

经分析：验证码字符 0-9 大小字符 A-Z (无 I)，小写字母为 a、b、d、e、f、g、h、i、j、l、m、n、q、r、t、u 与大写字母差别很大，共选择 51 个字符进行样本训练及测试。本文通过查找输出字符输出概率的最大值来判定输出的字符。

5. 实验验证

采集的样本量过多，本文只选择一小部分样本进行训练和测试每个字符采集的分割样本都在 100 个以上，共有样本 5100 个，其中随机选择 80%的样本作为训练样本，20%的测试样本，在样本分割正确的情况下，基本上都能识别出字符。测试环境平台使用 Intel 酷睿 i3 处理器和 Windows 7 操作系统，在训练次数为 50 次，训练集和调整值的误差已经几乎为 0。

由表 2 可见，通过本文采用的方法在没有彩色干扰线的情况下，大部分字符基本上都都能识别出来，在目前已有的训练和测试样本 5100 个小样本的情况下，字符 6 和字符 b、字符 9 和字符 g、字符 A 和字符 4 一些比较相近的手写体各色各样可能会存在误识别的情况，在实际大量验证码测试的情况下，在存在彩色干扰线的情况下四个字符分割识别准确率能达到 50%以上，为继续丰富训练测试样本库，再进一步的验证中在应用程序中建立 51 个文件夹用来存放已识别的字符，找到识别错误的字符放到正确的样本库中，建立大的样本模型库。未来需要进一步对有彩色的干扰线的字符进行分割、训练、识别，也可以使用深度学习的方法不分割字符直接对验证码进行字符标记，前期的字符标记需要采集并进行大量的标记、增加人为标记的工作量，对卷积神经网络的参数进行调整、训练，每一个输出的都是四个 51 个字符的最大概率的输出。

通过对该验证模型在训练次数为 50 次，蓝色是训练集，红色是调整值集合。第一幅图是训练次数为 50 次，训练集和测试集实际对象的真实值，第二幅图训练次数为 50 次，迭代 1 次时，训练集和调整值的误差在 50 次很大，第三幅图训练次数为 50 次，迭代 5 次时，训练集和调整值的误差在 50 次已经几乎为 0，见图 5。

Table 2. Part of CAPTCHA's image recognition result
表 2. 部分验证码图像识别结果

测试图像	本文方法	Tesseract OCR 软件	文献[5] SVM	文献[6]卷积神经网络
	4FHA	qfh4	4FH4	4FHA
	YxUK	Y*UK	YXUK	YxU2
	mxWV	M*WV	MXWV	mxWV
	gwgd	9wd	gW9b	9w9d

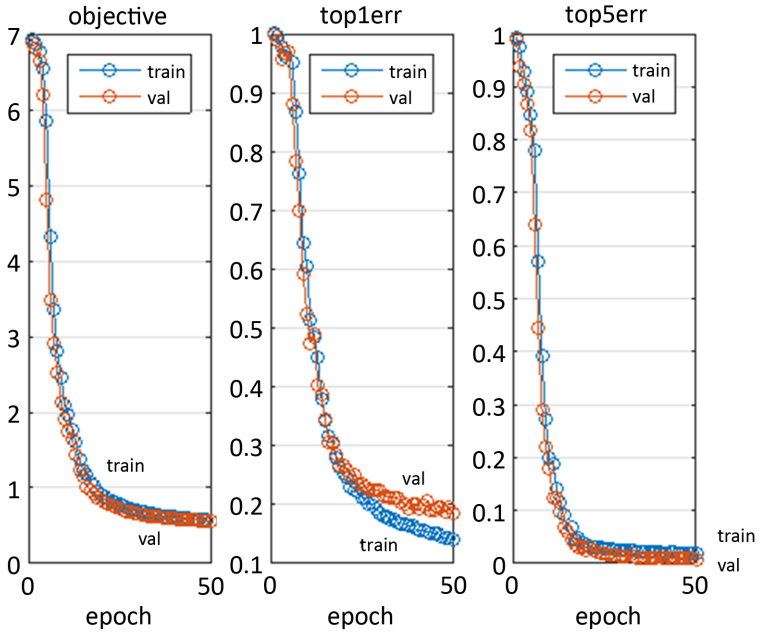

Figure 5. Training result of CNN
图 5. 卷积神经网络的训练结果

6. 实验结论

本文主要针对抖动的、扭曲的、涂色不均匀的类似手写的验证码进行研究，包括彩色验证码的二值化、彩色验证码的区域连通分割，使用深度的卷积神经网络对字符进行训练和识别，通过对小样本量建立训练测试模型，基本上能正确分割识别出验证码的字符，未来主要使用该算法采集大的样本进行训练测试，建立大的样本训练测试模型，提高识别的准确率。目前对于带彩色的干扰线和和字符颜色一致的验证码还没有很好的解决办法，未来研究的方向是对干扰线和字符颜色一致的验证码进行自动识别。

基金项目

吉林省科技发展计划项目重大科技成果转化项目 20170301005GX。

参考文献 (References)

[1] Yan, J., El Ahmad, A.S., *et al*. (2008) A Low-Cost Attack on a Microsoft CAPTCHA School of Computing Science. Newcastle University, UK.

[2] 程治国, 刘允才. 一种通用的去除文字图像中干扰线的算法[J]. 上海交通大学学报, 2005, 39(8): 1288-1291.

[3] 李兴国, 高炜, 黄江林. 基于滴水算法的验证码中粘连字符分割方法[J]. 计算机工程与应用, 2014, 50(1): 163-166.

[4] 汪洋, 许映秋, 彭艳兵. 基于 KNN 技术的校内网验证码识别[J]. 计算机与现代化, 2017(2): 93-97.

[5] 殷光. 基于 SVM 的验证码识别算法研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2010: 1-54.

[6] 陈超, 毛坚桓, 刘寅. 基于卷积神经网络的铁路货运网站验证码识别[J]. 指挥信息系统与技术, 2016, 7(4): 91-96.

[7] 田怀川. 基于神经网络的图形验证码识别及防识别的研究与应用[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2010: 1-73.

Hans汉斯

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org