Computer Science and Application 计算机科学与应用, 2022, 12(12), 2778-2787
Published Online December 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1212282

Hans汉斯

# 面向票据的OCR识别算法研究与实现

徐　倩，郭必然，贾泓波

北京信息科技大学，北京

收稿日期：2022年11月14日；录用日期：2022年12月14日；发布日期：2022年12月22日

## 摘　要

随着对票据使用的不断增多，票据的存储、管理以及票据信息的查找，逐渐变得繁琐，给人们带来困扰。通过对票据中信息的识别，发现其中圆形印章中的字符并不能准确识别，针对环形字符以及印章中文字的准确识别进行研究，实现了面向票据的OCR识别算法。使用Canny算子边缘检测、Hough变换、极坐标变换、以及确定极坐标变换起点的算法等，实现了能够按照印章中文字排列的逻辑进行变换，并成功识别出印章中所含的文字内容。实验结果表明，对印章中文字内容识别的正确率达到83.84%。

## 关键词

OCR技术，霍夫变换，印章识别，环形字符识别，极坐标转换

# Research and Implementation of Bill Oriented OCR Recognition Algorithm

**Qian Xu, Biran Guo, Hongbo Jia**

Beijing Information Science and Technology University, Beijing

Received: Nov. 14th, 2022; accepted: Dec. 14th, 2022; published: Dec. 22nd, 2022

## Abstract

**With the increasing use of bills, the storage and management of bills and the search of bill information have become cumbersome and perplexing. Through the recognition of the information in the bill, it is found that the characters in the round seal cannot be accurately recognized. Aiming at the accurate recognition of the circular characters and the characters in the seal, an OCR recognition algorithm for the bill is realized. Using Canny operator edge detection, Hough transformation, polar coordinate transformation, and the algorithm to determine the starting point of polar coordinate transformation, the transformation can be carried out according to the logic of the text ar-**

文章引用: 徐倩, 郭必然, 贾泓波. 面向票据的OCR识别算法研究与实现[J]. 计算机科学与应用, 2022, 12(12): 2778-2787.
DOI: 10.12677/csa.2022.1212282

**rangement in the seal, and the text content contained in the seal can be successfully recognized. The experimental results show that the correct rate of text recognition in seal is 83.84%.**

## Keywords

**OCR Technology, Hough Transformation, Seal Identification, Circular Character Recognition, Polar Coordinate Conversion**



## 1. 引言

随着经济不断发展，以银行代表的金融业票据量及票据种类越来越多，单纯靠人工来识别票据和印章已经不足以满足人们的需求。因此，为了提高印章的识别效率，实现面向票据的印章内容自动识别必不可少。票据的印章中通常包含了公司的单位名称、税务登记号以及其他重要信息。利用特定的提取方法和识别方法可以对印章内容进行处理，而印章的内容相对于文字识别来说具有不规则形状，例如呈圆环状或者拱形排列，而且文字排列方向不统一，这使得印章的识别工作更加困难。

OCR [1] (Optical Character Recognition 光学字符识别)技术，是指利用电子设备(例如扫描仪或数码相机)检查纸上打印的字符，通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字的过程。也就是通过扫描文本资料，然后对图像文件进行分析处理，来获取文字以及版面信息的过程。信息电子化在计算机技术发展的浪潮中，已经逐渐成为一个必然趋势，而文字又作为信息中最重要的一种载体，其电子化的程度决定了信息化的程度。

光学字符识别(OCR)这一概念，最早由德国科学家 Tausheck 在 1929 年提出。汉字识别最早是由 IBM (International Business Machines Corporation)公司的工程师 Casey 与 Nagy 实现的，他们在 1966 年发表了首篇汉字识别相关的文章，采用的是模式匹配的方式，可以识别 1000 个印刷体汉字。之后，OCR 技术得到大量研究，经过近 60 年的发展，并且随着相关技术以及算力的提升，现在已经广泛应用在各个领域。

在现阶段，这项技术已经比较成熟，应用风险也低。OCR 一般可分为手写体识别和印刷体识别，识别内容则包括汉字、英文字母、阿拉伯数字、常用标点符号等。一套 OCR 处理流程基本可分为版面分析、预处理、行列切割、字符识别、后处理识别矫正共计 5 个步骤[2]。

综上所述，本文实现了面向票据的 OCR 识别算法，解决了对票据上圆形印章和拱形文字的识别，给人们对于票据信息的获取和保存带来便利。

## 2. 相关工作

### 2.1. 研究背景

通过对 OCR 技术简单的测试，不难发现 OCR 对数字和英文字符的识别效果是普遍高于对中文的识别效果的，之所以会出现上述情况，与中文字体的复杂形状有着直接的关系。传统的 OCR 技术，能够使用模式匹配、支持向量机或者浅层神经网络等方法，针对文字噪声少、设计高性能的特征向量，可以得到很高的准确度，但是当传统的 OCR 技术用于有着大量噪声、复杂的中文文字或者非线性排列的文字识别时，识别效果较差[3]。传统 OCR 标注如图 1 所示：

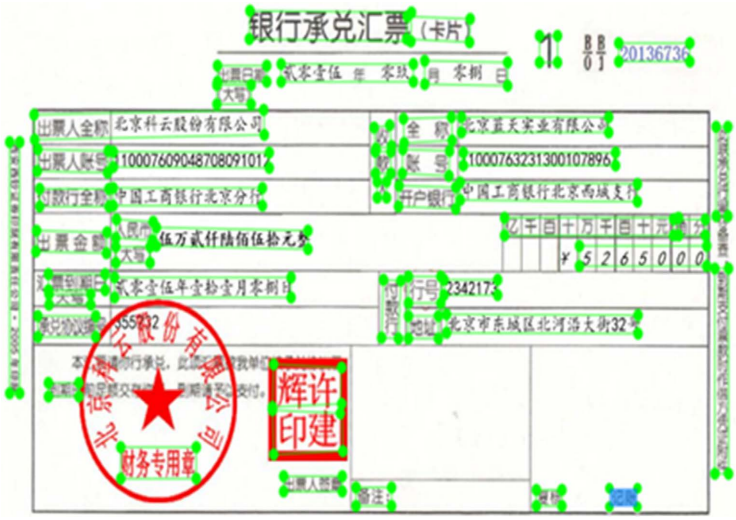


**Figure 1.** Traditional OCR annotation
**图 1.** 传统 OCR 标注

传统的 OCR 文本检测依赖于一些浅层次的图像处理方法或者图像分割方法以及一些复杂繁琐的后处理技术进行文字定位，因为这些技术的使用，传统 OCR 所处理的对象往往局限于成像清晰、背景干净、字体简单而同时又排列规整的文档图像[4]。

### 2.2. 国内外研究现状

现阶段，印章的提取与识别技术主要停留在通过提取票据印章与预留印章对比来判定印章真伪的阶段。比如汪伊函提出的基于小波分析的书画印章图像识别方法。利用小波分析方法对书画印章图像进行降噪处理，并在小波神经网络分类器中实现对书画印章图像的识别[5]。戴峻峰、杨天、熊文心等人根据印章元通常成圆环状排列的特点，克服印章文字方向不统一的问题，利用 CTPN (Detecting Text in Natural Image with Connectionist Text Proposal Network) + CRNN (Convolutional Recurrent Neural Network)网络进行文字的检测与识别[6]。陈娅娅、刘全香、王凯丽、易尧华等人提出基于深度残差网络(ResNet)和迁移学习的古印章文本识别方法[7]。张倩、郝红光、韩星周等人利用卷积神经网络 VGGnet 对印章印文分类识别作为一种辅助方法应用于印章印文自动识别中[8]。

以上方法主要通过检测印章颜色与背景色的差异进行印章提取后再通过各种算法与预留印章进行对比，从而对印章进行识别。本文主要研究方向为对票据中的印章内容按文字排列的逻辑进行识别。

## 3. 实验概述

本文的研究主要基于下述几个方法，来实现最终对印章中拱形文字的识别。实验流程图如图 2 所示。

### 3.1. 基于颜色通道分离法获取图像中印章

通过对图片进行分析发现，字的颜色是黑色，印章的颜色是红色，纸的颜色是灰色(接近白色)。前景和背景在颜色上存在差异，于是本文通过颜色特征将前景和背景分离。主要技术路线如下：

1) 读取原始图像 A；

2) 提取图像的红色通道；

3) 计算 B 的统计直方图 C，确定最佳阈值，通过这个阈值，就可以将前景像素提取出来，计算灰度值见公式(1)：

$$Ir(x,y)=\begin{cases}0, \text{ if } Ir(x,y)\le \text{Threshold}\\ 255, \text{ otherwise}\end{cases} \tag{1}$$

其中，$Ir(x,y)$是坐标$(x,y)$处的像素点的红色分量的灰度值，Threshold 为阈值。

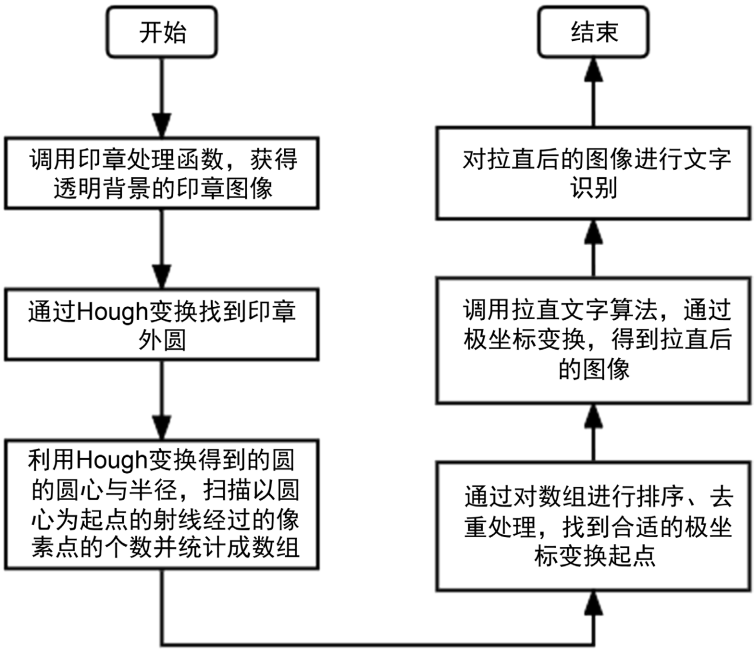


**Figure 2.** Experimental flow chart
**图 2.** 实验流程图

4) 根据阈值，对 B 进行二值化，得到最终图片 D。

通过颜色通道分离后印章如图 3 所示：


**Figure 3.** Separate seal through color channel
**图 3.** 通过颜色通道分离印章

## 3.2. 基于 Canny 算子的边缘检测算法

图像的边缘是指图像局部区域亮度变化显著的部分，边缘检测主要是图像的灰度变化的度量、检测和定位。

Canny 边缘检测是一种使用多级边缘检测算法检测边缘的方法。主要步骤如下：

1) 使用高斯滤波器，以平滑图像，滤除噪声。高斯滤波使用的高斯核是具有$x$和$y$两个维度的高斯函

数，且两个维度上标准差一般取相同，形式见公式(2)：

$$G(x,y)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{2}$$

2) 计算图像中每个像素点的梯度强度和方向。使用 Sobel 算子计算像素梯度：Sobel 算子是两个 3 × 3 矩阵，$S_x$ 用于计算图像 $x$ 方向像素梯度矩阵 $G_x$，$S_y$ 用于计算图像 $y$ 方向像素梯度矩阵 $G_y$，$G_x$ 和 $G_y$ 计算公式见公式(3) (4)：

$$G_x=S_x*I=\begin{bmatrix}-1 & 0 & +1\\ -2 & 0 & +2\\ -1 & 0 & +1\end{bmatrix}*I \tag{3}$$

$$G_y=S_y*I=\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ +1 & +2 & +1\end{bmatrix}*I \tag{4}$$

其中 $I$ 为灰度图像矩阵，$*$ 表示互相关运算，由公式(5)可得到梯度强度矩阵 $G_{xy}$。

$$G_{xy}=\sqrt{G_x^2+G_y^2} \tag{5}$$

3) 应用非极大值(Non-Maximum Suppression)抑制，以消除边缘检测带来的杂散响应。

4) 应用双阈值(Double-Threshold)检测来确定真实的和潜在的边缘。

5) 通过抑制孤立的弱边缘最终完成边缘检测。

通过边缘检测后输出如图 4 所示：

**Figure 4.** Canny edge detection
**图 4.** Canny 边缘检测

### 3.3. 基于 Hough 变换检测印章中圆形位置

本文中主要通过 Hough 变换找到印章对应的圆，主要步骤如下：

1) 首先选取圆上任意一点 A，获取 Canny 算子检测边缘时就已经获得的梯度方向，并设定步长。

2) 沿该梯度方向以一定步长选取像素格作为圆心 B，并计算 AB 之间的距离，作为半径。记录圆心坐标，半径。

3) 对圆上每一点重复上述操作。

4) 统计所有圆心坐标和半径，获得得票数最多的圆心坐标和半径，作为输出。

5) 绘制一个遮罩来盖住印章中心无意义图案，便于后期文字拉直的处理，处理后如图 5 所示：


**Figure 5.** Hough transformation and draw mask
**图 5.** Hough 变换并绘制遮罩

### 3.4. 基于极坐标变换完成拱形文字拉直

对圆和椭圆的旋转文字进行极坐标到直角坐标转换。拉直分为两部分，第一部分为寻找极坐标变换的起点角度；第二部分应用极坐标变换拉直图像。具体步骤如下：

1) 已知圆的半径为 $r$，则圆的周长为 $2\pi r$ (单位：像素)。取需要计算像素点的半径条数 $n=2\pi r$，统计该条半径上的二值化后实体像素点，得到 $V=[e_1,e_2,\cdots,e_n]$，如图 6、图 7 所示：


**Figure 6.** Radial lines as voids
**图 6.** 视作空隙的径向线

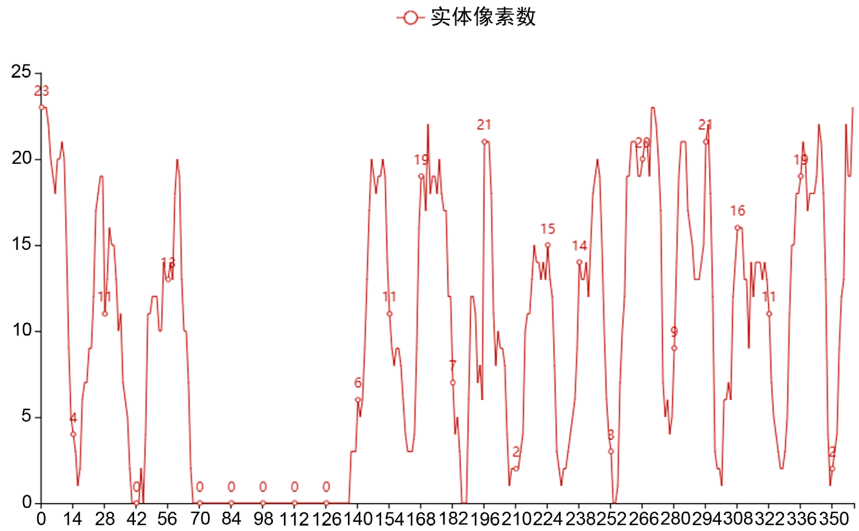


**Figure 7.** Line chart of statistical radial pixels
**图 7.** 统计各径向像素点的折线图

2) 将 $v$ 中元素去除重复值的元素，从小到大排序后，得到长度为 $k$ 的数组 $U=[d_1,d_2,\cdots,d_n]$。

3) 取 $U$ 中第 $i$ 个数 $d_i$ 的值，其中 $i$ 的值由公式(6)确定：

$$i=\begin{cases}\left\lceil\frac{k}{2}\right\rceil, k<10\\ 5, k\geq 10\end{cases} \tag{6}$$

与 $V$ 中与每个元素的值进行对比。如果元素的值小于 $d_i$ 则标记为 1，代表该元素对应半径的像素点少，视作文字间隙。否则标记为 0。由公式(7)：

$$p_j=f\left(e_j\right)=\begin{cases}1, e_j<d_i\\ 0, e_j\geq d_i\end{cases} \tag{7}$$

得到 $O=[p_1,p_2,\cdots,p_n]$。

4) 找到其中最大的连续为1的标记段的中点元素，该元素代表的弧度转化为角度即为变换的角度起点。

① 为保证比较的起点不截断连续，寻找起点偏移量 offset，如图 8 所示：

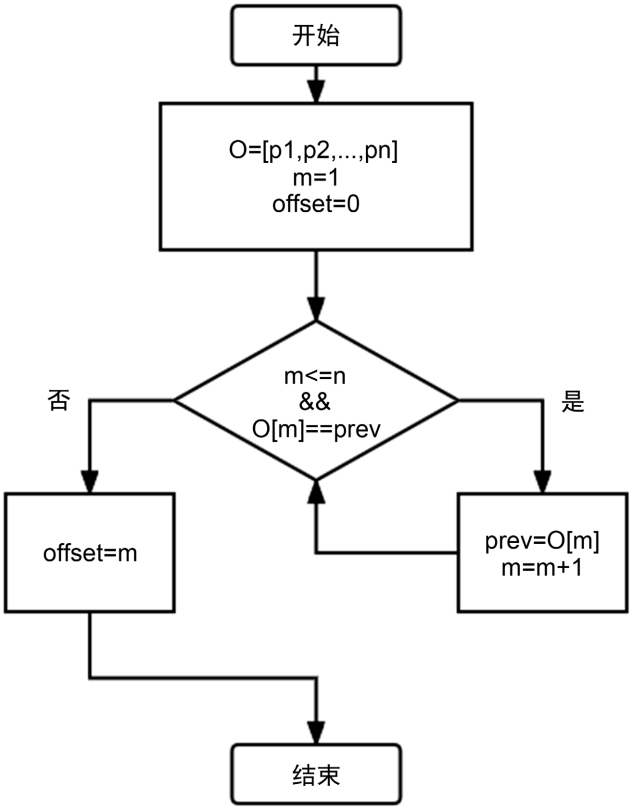


**Figure 8.** Flow chart of finding offset starting point
**图 8.** 寻找偏移起点流程图

② 从 $p_{\text{offset}}$ 开始，寻找数组最大连续为 1 段的起点 index 和长度 length。变换起点 start 由公式(8)确定：

$$\text{start}=\left(\text{index}+\text{length}\right)\%n \tag{8}$$

5) 确定起始变换点后对圆形印章进行极坐标变换，获得拉直后的图像，拉直后的图像如图 9 所示：


**Figure 9.** Image after straightening
**图 9.** 拉直之后的图像

### 3.5. 对拉直之后的内容进行文字识别

文字识别采用了基于 DB (Differentiable Binarization)，EAST (Efficient and Accuracy Scene Text)，SAST (Static Application Security Testing)算法的检测模型和基于 Rosetta，CRNN (Convolutional Recurrent Neural Network)，RARE (Robust text recognizer with Automatic Rectification)，SRN (sequence recognition network)，STAR-Net (Spatial Transformer + CRNN)算法的识别模型。采用国内开源 OCR 框架的 PaddleOCR 官方提供的检测模型 ch_ppocr_server_v2.0_det 和识别模型 ch_ppocr_server_v2.0_rec，得到文字内容如图 10 所示：

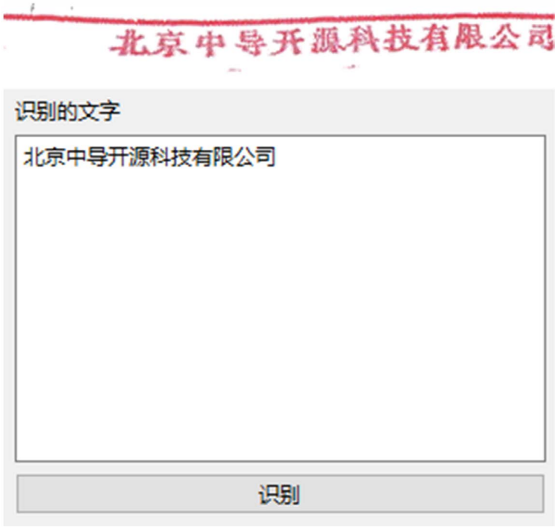


**Figure 10.** Text content after recognition
**图 10.** 识别后的文字内容

## 4. 实验结果及分析

对具有 1114 张图片的测试集进行测试，其中识别内容正确的张数有 934 张，正确率为 83.84%。

对其中识别错误的内容进行分类分析。其中分析结果如图 11 所示：

其中：

a) 图像本身分辨率小于 100 px 的图片；

b) 颜色过于淡薄的图片；

c) 印章底色干扰项多的图片；

d) 印章中文字笔画少的图片；

e) 印章中有双层圈进行干扰的图片；

f) 印章中文字排列紧密成环的图片；

g) 由于环境光导致图像辨识度过低的图片。

本实验的创新性在于可以通过找到极坐标变换起点，按照印章中文字排列的逻辑进行变换，并成功识别出印章中所含的文字内容。

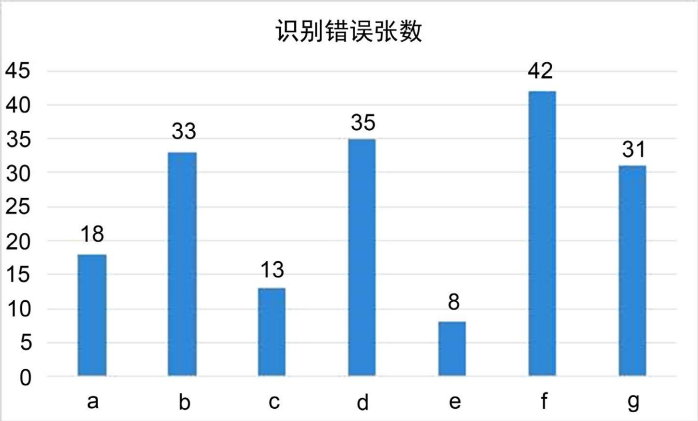


**Figure 11.** Identification result error analysis histogram
**图 11.** 识别结果错误分析柱状图

本文通过实验对比直接进行极坐标变换以及通过确定变换起点进行变换两种情况下的识别正确率，如表 1 所示：

**Table 1.** Comparison of test results
**表 1.** 检测结果对比

| 使用方法 | 识别数量 | 识别正确率 |
|---|---|---|
| 直接进行变换 | 1114 | 77.05% |
| 确定起点后变换 | 1114 | 83.84% |

在对实验结果进行分析后，得出本实验还存在着一些有待改进的地方，在后续的维护和优化中会进一步对算法进行完善，以提高算法的正确率。

## 5. 结束语

本文针对常规 OCR 识别在对印章内容进行识别过程中，出现的由于文字形状不规则无法正确识别印章内容的情况，提出了面向票据的 OCR 识别算法；针对类似拱形或圆形文字无法识别的问题，使用了可以由极坐标变换来将文字拉直的算法，便于后续识别；针对无法确定极坐标变换中变换起点的问题，提出了可以由统计各径向线上的像素点个数进行比较来定位极坐标变换起点的算法；使得能够较为精准地识别出印章内容。同时也为后续票据的存储、管理以及票据信息的查找提供一定的技术基础。

## 基金项目

由北京信息科技大学大学生创新创业训练计划项目——计算机学院(5112210832)支持。

## 参考文献

[1] 王文华. 浅谈 OCR 技术的发展和应用[J]. 福建电脑, 2012, 28(6): 56+92.

[2] 梁林森. 基于 OCR 技术的医疗收费票据自动录入系统研究[J]. 电力设备管理, 2021(4): 198-199.

[3] 杜训祥. 基于卷积神经网络的图像中文 OCR 识别纠错方法及系统的研究[J]. 江苏通信, 2021, 37(1): 109-112.

[4] 王阳, 李振东, 杨观赐. 基于深度学习的 OCR 文字识别在银行业的应用研究[J]. 计算机应用研究, 2020, 37(S2): 375-379.

[5] 汪伊函. 基于小波分析的书画印章图像识别方法[J]. 信息与电脑(理论版), 2022, 34(14): 89-91.

[6] 戴俊峰, 杨天, 熊闻心. 基于极坐标转换的中文印章文字识别[J]. 计算机工程与设计, 2021, 42(11): 3174-3180.

[7] 陈娅娅, 刘全香, 王凯丽, 易尧华. 基于 ResNet 和迁移学习的古印章文本识别[J]. 计算机工程与应用, 2022, 58(10): 125-131.

[8] 张倩, 郝红光, 韩星周. 利用 VGGnet 对印章印文分类识别的适用条件研究[J]. 通信技术, 2019, 52(7): 1639-1642.