Computer Science and Application 计算机科学与应用, 2020, 10(10), 1804-1819
Published Online October 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1010191

Hans汉斯

基于连通域的联锁表扫描件图表自动识别算法

方尉宇，徐尚志，李志鹏

同济大学电信学院电子与通信工程，上海
Email: lizhipeng@tongji.edu.cn

收稿日期：2020年10月6日；录用日期：2020年10月21日；发布日期：2020年10月28日

摘　要

针对目前主流图表自动识别算法不适用于大型联锁表扫描件图像的问题，本文设计了基于连通域的联锁表扫描件图表自动识别算法。该算法能通过图像处理和神经网络对扫描件中的表格及文字进行识别后，完整复现在电子表格中，并将疑似识别错误的字符及其所在单元格突出显示，方便人工复核。算法主要分为预处理、定位和识别三个部分。其中预处理部分，提出了DN-OSTU与DNG-OTSU二值化算法，通过不同内核均值滤波和图像相除、线性归一化等方法对光线不均的扫描件进行二值化，并提出基于累计概率霍夫变换PPHT的倾斜矫正算法，能快速且准确检测出倾斜角度；定位部分，采用基于连通域的定位算法定位表格方框及文字区域，并提出基于图表特征的RS方框查补算法，确保表格的完整性和各单元格定位的准确性；识别部分，使用原图提取字符制作训练集，训练卷积循环神经网络CRNN，达到较高准确率。实验中，对多家设计院提供的联锁表进行测试，实验结果表明：单元格识别准确率达到92.8%，字符识别准确率为98.74%，单图从识别到电子表输出速率均在5秒以内。本文设计的联锁表扫描件图表自动识别算法具有准确率高、鲁棒性好、识别速度快等特点，可为纸质版联锁表扫描件复现电子版从而二次开发提供有效的技术途径。

关键词

连通域，联锁表，卷积循环神经网络，图表自动识别算法

Scanning Image Recognition Algorithm of Interlocking Table Based on Connected Domain

Weiyu Fang, Shangzhi Xu, Zhipeng Li

Electronic and Communication Engineering, Tongji University, Shanghai
Email: lizhipeng@tongji.edu.cn

Received: Oct. 6th, 2020; accepted: Oct. 21st, 2020; published: Oct. 28th, 2020

文章引用: 方尉宇, 徐尚志, 李志鹏. 基于连通域的联锁表扫描件图表自动识别算法[J]. 计算机科学与应用, 2020, 10(10): 1804-1819. DOI: 10.12677/csa.2020.1010191

Abstract

Aiming at the problem that the current mainstream automatic chart recognition algorithm is not suitable for scanning images of large interlocking tables, this paper designs an automatic recognition algorithm for scanning charts of interlocking tables based on connected domains. The algorithm can recognize the tables and texts in the scanned images through image processing and neural network, and then reproduce them in the electronic form completely, and highlight the characters and their cells that are suspected of being misrecognized to facilitate manual review. The algorithm is mainly divided into three parts: preprocessing, positioning and recognition. In the preprocessing part, DN-OSTU and DNG-OTSU binarization algorithms are proposed, and the scans with uneven light are binarized by means of different kernel mean filtering, image division, and linear normalization. And a tilt correction algorithm based on Progressive Probabilistic Hough Transform is proposed, which can quickly and accurately detect the tilt angle. In the positioning part, a positioning algorithm based on connected domains is used to locate table boxes and text areas, and an RS box checking and filling algorithm based on chart features is proposed to ensure the integrity of the table and the accuracy of each cell positioning. In the recognition part, we use the original image to extract characters to make a training set, then train the convolutional recurrent neural network CRNN to achieve high accuracy. In the experiment, the interlocking tables provided by a number of design institutes were tested. The experimental results showed that the accuracy of cell recognition reached 92.8%, the accuracy of character recognition was 98.74%, and the output rate of single image from recognition to electronic watch was within 5 seconds. The automatic recognition algorithm for scanned parts of interlocking table designed in this paper has the characteristics of high accuracy, good robustness, and fast recognition speed. This can reproduce the electronic version of the scanned copy of the paper version of the interlocking table, thereby providing an effective technical approach for secondary development.

Keywords

Connected Domains, Interlocking Tables, The Convolutional Recurrent Neural Network, Automatic Chart Recognition Algorithm

1. 引言

即使在信息技术高度发达的今天，纸质文档作为传统信息保存和交流的媒介，在人们的日常生活中仍然发挥着重要的作用[1]。在铁路建设中，最常见的纸质文档便是联锁表，其记录的道岔、进路、信号机等的信息是轨道联锁的重要内容。作为信号施工图的主要部分，各设计院提供的纸质版联锁表常常需要录入到计算机中进行二次开发。为减少录入所消耗的人力物力，需要开发专用 OCR 识别算法，自动识别纸质版联锁表扫描件，并完整复现在电子表中。

目前针对特定格式的文档自动识别算法研究较多，如邮政编码自动识别、快递单据识别、车牌识别等，大型图表的识别也有一些研究，但实际应用系统较少，理论也不完善[1]。随着人工智能领域的快速不断发展，基于深度学习的图像字符识别方法算法层出不穷，文字识别准确率趋近完美。本文将针对联锁表扫描件的实际情况，结合表格识别、深度学习等方法，设计开发一套准确性高、鲁棒性强的联锁表扫描件识别算法。

2. 联锁表

2.1. 联锁表介绍

联锁表是根据车站信号平面布置图所展示的线路、道岔、信号机、轨道电路区段等情况，按规定的原则和格式编制的，能够表明车站各个信号设备间相互制约关系。其扫描件图像如图 1 所示，其尺寸通常大于 9000 * 6000。扫描过程中设置水平与垂直分辨率为 600 dpi，可扫描成三通道彩色图或单通道灰度图。较大的尺寸和较高的分辨率虽然提高了图表清晰度，但也导致了更多噪声的干扰。

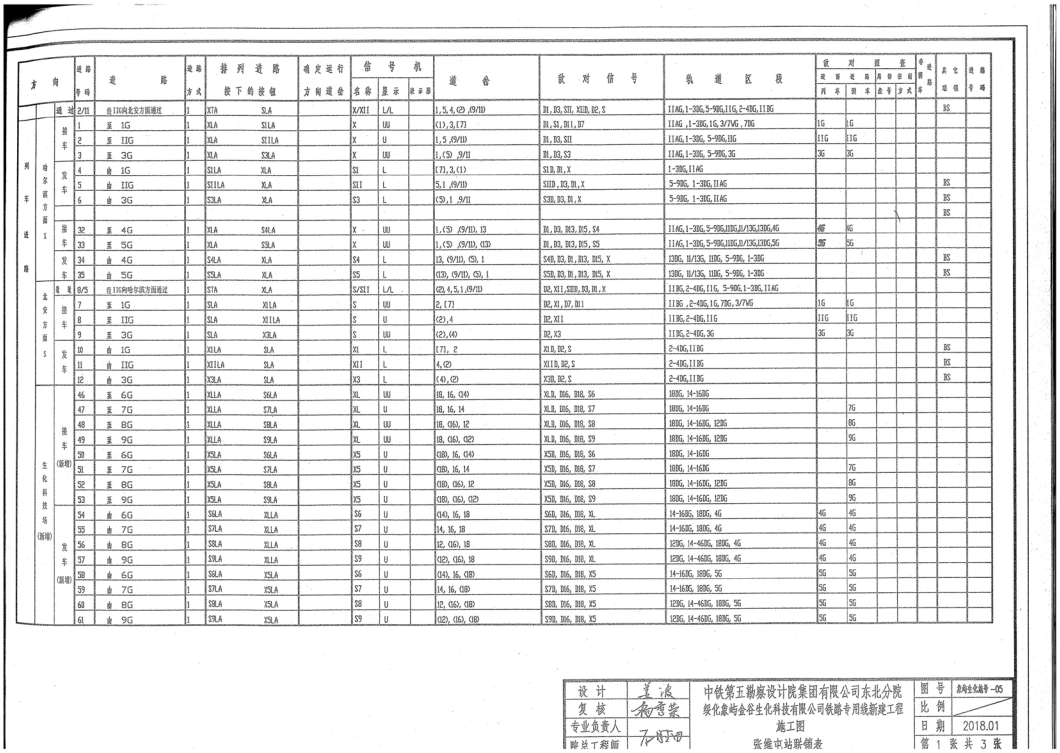


Figure 1. Scanned image of track interlocking table
图 1. 联锁表扫描件示意图

2.2. 联锁表特征

各设计院提供的纸质版联锁表经过不同扫描仪扫描出的图像具有“结构类似，图像各异”的特征，即不同扫描件的图像结构特征基本类似，但图像背景、字符与框线位置却有很大差异，加大了自动识别的难度。

2.2.1. 结构类似

扫描件图像结构特征包括：1) 表格结构特征，需要识别并复原的表格占据图像绝大部分面积，提取表格框架时，需要排除右下角图注部分表格。2) 排版特征，主表中文字排版以横排为主，只有前几列存在竖排文字；用神经网络识别时，需要进行横竖排判断并转换。3) 文本特征，除表格标题都是中文外，表中内容多以英文、数字或部分标点符号(逗号、短横线)为主；同一扫描件字体型号基本统一，有少量后加字符的字体略有不同；不同设计院提供的联锁表字体差异明显，使后续制作训练集、训练神经网络增加了难度。

2.2.2. 图像各异

图像各异既体现在单表内各单元格之间，也体现在不同联锁表的图像背景、字符与框线位置、字体

型号等方面。如图 2 所示，分别截取了六张联锁表扫描件中某一单元格图像进行对比。左边三幅均为灰度图像：左 1 图像中，边框线段与字符均由密集、不连续黑色像素点构成，图像右半部分受扫描光线影响，噪声密集；左 2 图像中，边框线段与字符轮廓完整，但噪声点较多，分布比较均匀；左 3 图像中，线段与字符轮廓中均有若干白色断点，且出现人为勾画笔迹，通过观察发现，不同联锁表中均有类似勾画笔记的情况，因此后续字符定位算法需要排除勾画干扰，准确定位前面的字符区域。右边三幅均为彩色图像，较之灰度图像，噪声较少，图像质量更高，但也出现新的结构特征：右 1 图像中，字符区域整体靠近下框线，且逗号与框线相交，在字符定位中，需要考虑字符与框线粘连情况；右 2 图像中，部分字符被中横线划掉，旁边标注上新的、字迹更粗大的字符，此类情况只较少存在部分联锁表中，因此本文主要讨论正常字符的识别，对带横线字符识别略有涉及。此外，当考虑带横线字符识别时，需要设计此类格式的电子表还原算法；右 3 图像中，表格整行被长横线划掉，可以预想到，该横线会对表格框架的识别产生巨大影响，因此在提取框架、定位单元格的时候，需要充分考虑此种干扰。

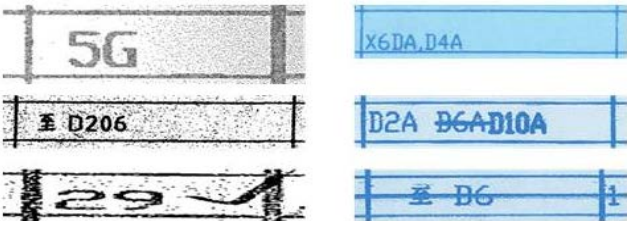


Figure 2. Comparison of scanned images of different interlocking tables
图 2. 不同联锁表单元格对比图

3. 基于形态学的联锁表扫描件自动识别算法设计

3.1. 识别算法流程

本文在基于形态学基础上，结合循环卷积神经网络完成联锁表扫描件自动识别算法的设计。自动识别算法流程由四部分组成：图像预处理模块、定位模块，字符识别模块以及最后的复查标注输出模块。其中图像预处理模块是对纸质版联锁表经扫描仪扫描成 TIF 或者 JPG 格式的图像进行处理，具有缩小图像尺寸、倾斜检测和校正、去除干扰噪声、二值化处理等功能；定位模块首先对二值化后的图像进行表格方框定位，并利用联锁表固有特征的先验知识进行查框补框后，再进一步定位、提取单元格内字符区域，从而减少需要识别的图像区域；然后，字符识别模块利用训练好的深度学习网络模型对各单元格文字区域依次识别。最后，复查标注输出模块将识别的文字按原图序排列到电子表格中，并根据识别概率将疑似错误的字符进行改色显示，以提示需要对该部分进行人工复核。算法流程见图 3 所示。

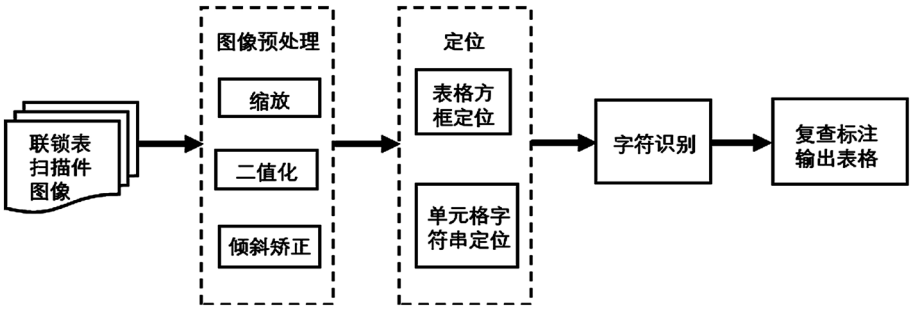


Figure 3. Recognition algorithm flowchart
图 3. 识别算法流程图

3.2. 图像预处理——DN-OTSU 二值化算法与基于 PPHT 的倾斜矫正算法

通过扫描仪将纸质联锁表转换为电子版图像(TIF 或 JPG 格式)后，在计算机中存储的是一幅不理想的原始多值图像。外界影响如光线明暗不均、页面折叠阴影、扫描仪客观因素噪声等问题，导致图像清晰度下降，必将对后续的文字区域分割产生负面影响，进而降低神经网络的识别准确率。且在对图像进行去噪和二值化等预处理过程中需要谨慎对待，因为可能会对字符形态有一定的影响[2]。因此，整个文字识别流程中最为核心的环节便是对图像进行预处理，高质量的预处理结果与高识别准确率成正相关。

本文设计了一种适应性强、对文字干扰少的预处理算法，为后续字符识别提供完整、可靠、低噪声的字符信息，提高整体识别率。首先要对原始图像进行灰度化和缩放处理，以大大减少有效图像数据量，加快后续算法处理速度；然后比较了几种传统二值化算法在扫描件上的实际应用效果，并结合均值模糊与高斯模糊等，设计了 DNG-OTSU (Divide and Normalize OTSU)与 DN-OTSU (Divide and Normalize with Gaussian Blur OTSU)二值化算法，分别用于表格方框定位和文字区域识别，提高了方框定位准确性的同时，有效避免了预处理过程中字符形态改变带来的误识别率。最后鉴于扫描时可能存在的图像倾斜问题，结合图表竖线特征，提出了基于 PPHT 的倾斜矫正算法，对图像进行校正处理，以避免倾斜直接影响后续表格单元格定位；

3.2.1. 灰度化处理和缩放

若扫描件是彩色图像，包含大量无用信息，为降低计算量，提高识别效率[3]，有必要对其进行灰度化。采用 OPENCV 计算机视觉库提供的读图函数，自动将图像转为单通道灰度图像。而此时图像尺寸依旧过大，需要在不改变字符形态、减少噪声的基础上，将其缩小到四分之一。根据联锁表扫描件图像特征，在试验过多种缩放算法后，决定采用高斯金字塔算法。使用高斯金字塔下采样一次，将原始图像作为 G0，即高斯金字塔的第 0 层，采用(5 * 5)高斯核卷积，然后对其卷积后的图像下采样去除偶数行和偶数列，得到高斯金字塔上一层图像 G1，其长宽仅为 G0 的二分之一，面积仅为 G0 的四分之一。

3.2.2. DN-OTSU 与 DNG-OTSU 二值化算法

预处理中，最为关键的步奏便是对联锁表灰度图像进行二值化处理[4]，目的是将 256 色灰度图像按照某个阈值，转换成仅含有 0 (黑色)和 255 (白色)的二值图像，即前景黑色字符和背景全白[5]。好的二值化算法能有效保留前景字符图像轮廓，弱化背景带来的干扰，为后续识别环节打好基础。因此需要根据联锁表扫描件的图像特征选择或设计合适的二值化算法。

目前，较为常用的二值化方法有固定阈值、自适应阈值和 OTSU 大津法等。固定阈值顾名思义就是直接设定某一阈值，对图像进行二值化计算，通常适用于图像质量较好，光线条件均匀或光线变化不多的情况。实际使用中，图像容易受到光线变化影响，固定阈值不能很好的将目标与背景分离，这时候需要自适应阈值方法。自适应阈值会根据一小片区域的灰度值来动态计算该区域的阈值，使最后的输出更合理。而 OTSU 更适用于图像直方图出现双峰时，其算法将遍历 0 到 255 阈值，计算前景(大于阈值部分)与背景(小于阈值部分)的类间方差值，最大方差值对应的阈值即为最佳阈值。

图 4 便是采用不同二值化方法对联锁表图像处理的结果，为更好对比效果差异，小图均展示联锁表图像的右上角区域。(a)为原图，扫描件右上角区域由于纸张褶皱或者光线原因，灰度值较低，背景偏暗，字符与背景对比度较低，需要重点考虑此区域二值化效果。在图(b)、(c)、(d)中，原图(a)中较亮区域前景图像轮廓均得到较好保留。对比发现此区域二值化效果最佳的为图(d)，字符形态最为清晰，且表格边线更为均匀。但容易看出，三种方法均未能有效解决原图中阴影部分问题。

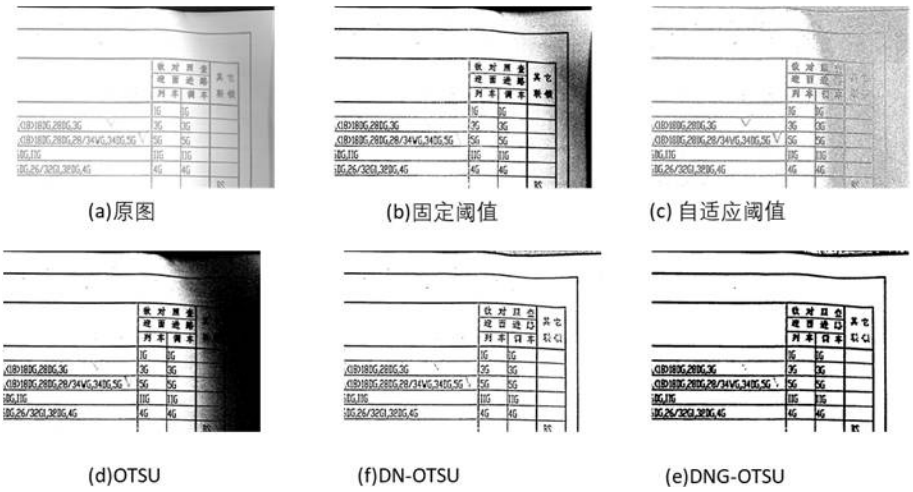


Figure 4. Comparison of binarization results
图 4. 二值化效果对比

为解决联锁表扫描件阴影区域二值化问题，本文提出 DN-OTSU 算法，即在 OTSU 大津法二值化前加上图像相除和归一化等步奏。其算法首先分别对原图进行卷积核为(3,3)、(23,23)的均值模糊得到矩阵 Blur1、Blur2，再将两个矩阵的值除以 255，然后将 Blur1 与 Blur2 算术相除，注意矩阵相除中，需要设置分母为 0 或者相除出现无限大时，该值为 0。相除得到的矩阵再利用归一化函数，将各点的数值线性缩放到 0 到 255 的整数。最后采用 OTSU 对得到的图像矩阵进行二值化。如图(f)，可以看出，该算法既能有效消除明亮不均匀带来的影响，又较好保留了前景字符和边框线的形态，符合预期目标。本文将在后续识别过程中采用 DN-OTSU 二值化后的扫描件图像。

在图表识别中，需要根据表格的横竖线确定表格框架，从而定位每个单元格，再进行字符识别等。因此对图像二值化后若能得到完整、清晰、均匀的横竖线，有助于后续提取表格框架。本文在 DN-OTSU 算法基础上加入高斯模糊过程，即 DNG-OTSU 算法。该算法在 DN-OTSU 二值化扫描件得到图(f)以后，再使用卷积核为(15,15)的高斯模糊，最后再一次利用 OTSU 算法进行二值化，得到图(e)。对比图(f)，表格框线更加清晰，线段宽度均匀且比原线更宽。但使用高斯模糊后的缺点也很明显，字符形态更加“圆润饱满”。字符间距较小时，相邻字符易粘连；文本行靠下时，易与单元格框线粘连，进而影响后续的字符定位和识别效果。因此本文将在后续倾斜矫正和提取表格框线过程中采用 DNG-OTSU 二值化后的扫描件图像。

3.2.3. 基于累计概率霍夫变换 PPHT 的倾斜校正算法

联锁表扫描件获取的过程中，由于可能移动采集设备或者扫描仪校准等因素，扫描件可能出现一定程度的倾斜。当倾斜度较大时，会严重影响表格方框定位和字符区域定位等步骤的实现效果，因此需要对图像进行倾斜矫正。本文根据联锁表图像特征，提出了检测角度准确、旋转图像快速的基于累计概率霍夫变换(Progressive Probabilistic Hough Transform, PPHT)的倾斜矫正算法，分成两个步奏：倾斜角度检测和图像旋转校正。

a. 倾斜角度检测

角度作为图形的基本属性，经常被人们用来描述图像的某些特征或者位置关系。因此，在图像处理中，人们对角度检测做了大量工作。主要的检测方法有：通过检测图像直线计算角度的霍夫变换法[6] [7]、通过侧面投影直方图的变差计算倾斜角度的水平投影法[8] [9]、通过计算傅立叶空间中密度最大方向的傅立叶变换法[10]、基于连通域计算向量方向的 k-近邻聚类法[11]，以及通过字符特征点进行聚类和直线拟合计算倾斜角度的图像直线拟合法[12]。

考虑到算法复杂度和联锁表图像的尺寸特征，计算时间代价较高的水平投影法、傅立叶变换法和 k-近邻聚类法均不在考虑之中。由于联锁表属于规则的图表文档，可通过简单的图像处理，得到表格框线的横线图或者竖线图，因此本文使用霍夫变换法检测直线角度计算表格倾斜角度。

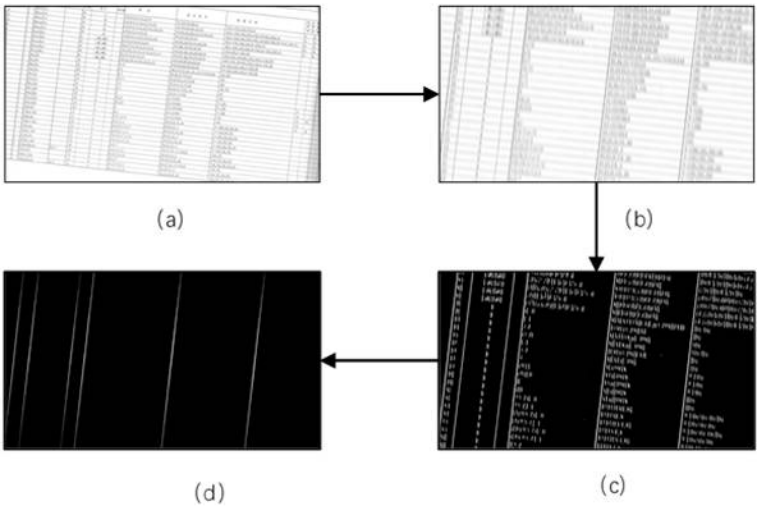


Figure 5. Detect tilt angle by vertical line
图 5. 竖线检测倾斜角

以图 5 为例，为消除图(a)中横线和文本行的干扰，对该图像进行两次卷积核为(3,33)均值模糊得到图(b)，再利 DNG-OTSU 算法进行二值化并取反(白底黑线变成黑底白线)，得到图(c)中只含稍微倾斜的竖向线段。在计算机视觉库 OPENCV 中，支持三种不同的霍夫变换：标准霍夫变换(SHT)，多尺度霍夫变换(MSHT)，累计概率霍夫变换(PPHT)。其中，PPHT 由 SHT 改进而来，其算法执行效率高，计算时间代价最小。利用 PPHT 检测图(c)中直线线段，然后过滤掉较短线段，只保留高度差超过图像一半高度的长线段得到图(d)。接着计算各线段长高之比的反正切函数，累和后求均值作为图像倾斜角度。需要注意的是，当角度大于 90 度时，(180 度——倾斜角度)才是直线相对竖直方向的偏离角度。

b. 图像旋转校正

图像的旋转校正可以看成图像中像素点空间位置改变，需要空间变换和灰度级差值两个步奏的算法。像素通过变换映射到新的坐标位置，当新的位置为非整数坐标时，就需要灰度级差值将映射的新坐标匹配到输出像素之间。本文采用“最近邻插值”插值方法，令输出像素的灰度值等于映射最近的位置像素。

图像的旋转，对应着各像素点通过某一函数变换到新的位置，可通过像素矩阵的仿射变换表示。以平移变换为例，假设某像素位置坐标为 (x,y)，通过线性函数变换到新的坐标位置 (x',y')：

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{1}$$

再引入齐次坐标，在原有二维基础上，增加一个维度，进一步可以将等式化简为：

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = M \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

$$M = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

变换矩阵M表示的就是两幅图片的仿射变换关系。因此只要知道了变换矩阵，就能将图片进行平移、旋转、缩放等仿射变换。

以图 6 为例，若将图像在原点向右旋转Θ角度，则旋转矩阵(没有平移)为：

$$M = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{4}$$

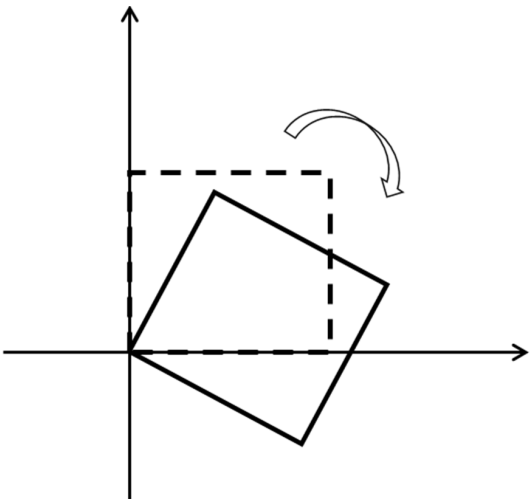


Figure 6. Rotation schematic diagram
图 6. 旋转示意图

若图像在点(C_x, C_y)进行旋转，则旋转矩阵(有平移)为

$$M = \begin{bmatrix} \alpha & \beta & (1-\alpha)C_x - \beta C_y \\ -\beta & \alpha & \beta C_x + (1-\alpha)C_y \end{bmatrix} \tag{5}$$

其中$\alpha = scale \cdot \cos\theta$，$\beta = scale \cdot \sin\theta$，scale 为图像缩放系数。

当按原图尺寸进行图像旋转时，得到的新图会缺失掉四个边角的部分像素信息。针对这一问题，新图就需要更大的尺寸，以此确保图像旋转后能够保持完整信息。如图 7 所示，倾斜角为 Θ，则旋转后新图像的尺寸为$(W*\cos(\theta)+H*\sin(\theta)$，$H*\cos(\theta)+W*\sin(\theta))$，旋转产生的空白用白色填充。

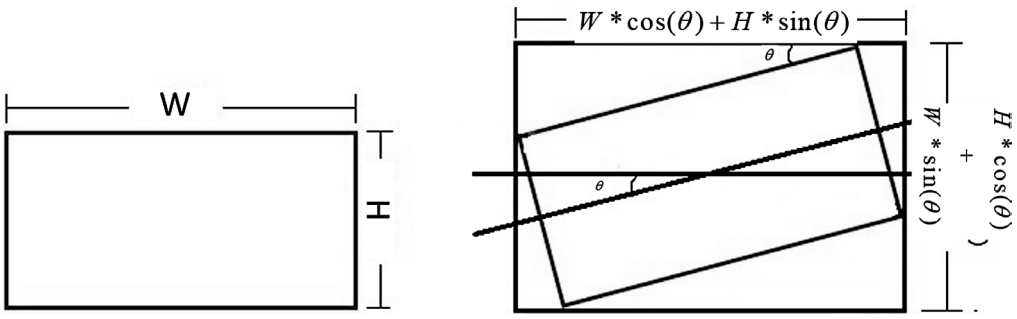
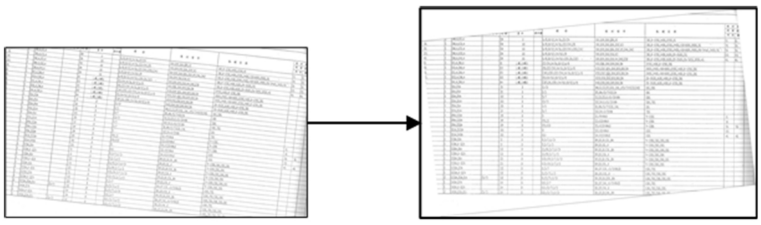


Figure 7. Keep the original content and enlarge the size
图 7. 标准试验系统结果曲线

通过计算旋转矩阵 M 和新图像尺寸后，旋转实际效果如图 8 所示，矫正后的图像字符形态与全部信息都得到了保留，符合预期目标。



倾斜矫正前　　倾斜矫正后

Figure 8. Schematic diagram of tilt correction
图 8. 倾斜矫正示意图

3.3. 基于连通域的表格方框定位算法

表格方框准确定位是单元格字符准确提取的必要前提。联锁表通常打印在 A3 或更大的纸张上，尺寸较大，使用过程中容易褶皱，在人工扫描时也就容易平铺不均匀，从而导致图像中表格线易弯曲、甚至错位的情况发生。因此常见的表格检测方法如霍夫变换方法[13]、投影法等并不适用于此类大型图表扫描件表格方框定位。

连通域是指图像中具有相同像素且位置相邻的前景像素点的图像区域。在表格图像二值化后，边框完整的单元格轮廓清晰，不考虑字符粘连的情况下，只需检测矩形轮廓便能得到单元格位置信息。因此本文考虑基于连通域的方法，设计定位表格方框算法。算法首先利用数学形态学的方法[14] [15]，去掉图像中的字符干扰，得到完整的框线图，再对其进行轮廓的拓扑计算，得到连通域后，计算能够包围各连通域的最小矩形，并将矩形绘制在新图中，构建起表格框架。

如图 9 所示，由于表格一般是由横竖线组成，因此检测单元格连通域前，需要排除文本行干扰，获得图像横向、竖向的直线图。利用数学形态学的方法，分别定义两种结构元素对前述 DNG-OTSU 二值化图像(图 9a)进行闭运算，得到表格的横线图和竖线图。两种结构元素为：1) 横向结构元素，用以消除字符和竖线，只留横线，因此设定其尺寸宽度为超过字符最大宽度的某值；2) 竖向结构元素，用以消除字符与横线，只留竖线，因此设定其尺寸高度为超过字符最大高度的某值。接着定义新的横向、竖向结构元素：新横向结构元素尺寸宽度变为原横向尺寸宽度的 1/4，高度在原横向尺寸高度上适当增加；新竖向

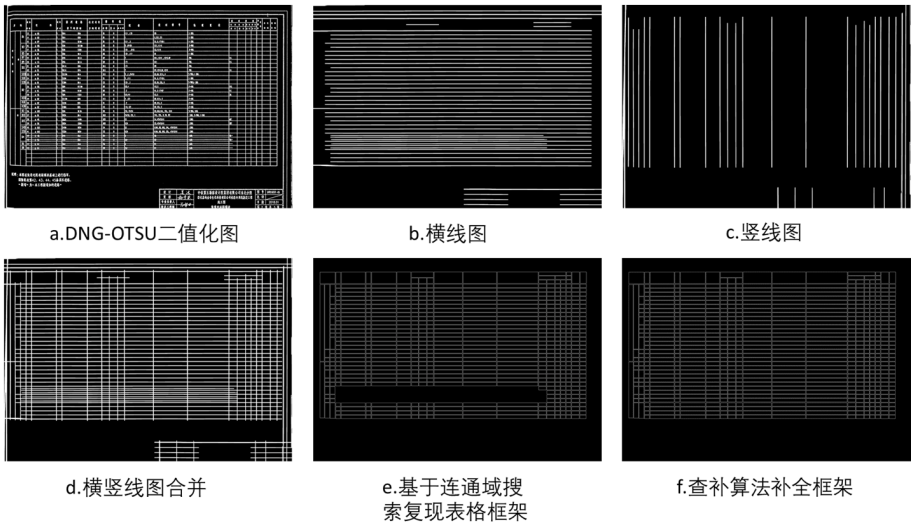


Figure 9. Table positioning process based on connected domain
图 9. 基于连通域的表格方框定位流程

结构元素尺寸高度变为原竖向尺寸高度的 1/2，宽度在原竖向尺寸宽度上适当增加；然后，用以上两种结构元素分别对得到的横线图和竖线图做开运算，得到横竖线的延长图 9b 和图 9c，延长部分不超过最小单元格宽度或高度，两图相加得到表格框架图 9d。

利用上述方法最大限度地还原了表格的框架，但图像中的表格线宽窄不均，仍有因褶皱导致的线段断裂，从而存在非闭合单元格或者缺失单元格的现象。因此需要在框架基础上，查漏补缺，重绘并定位每个单元格。在表格框架图上，可以通过 OPENCV 库函数 findContours()寻找连通域，得到所有边线完整、闭合单元格的轮廓信息，再利用库函数 boundingRect()得到包覆此轮廓的最小正矩形，返回的位置信息即单元格的起始横纵坐标以及矩形长宽。在 Python 语言中，可通过复合列表存储全部矩形的位置大小信息，表示为$\left[\left[X_0,Y_0,W_0,H_0\right],\left[X_1,Y_1,W_1,H_1\right],\cdots,\left[X_n,Y_n,W_n,H_n\right]\right]$，最后一项$\left[X_n,Y_n,W_n,H_n\right]$表示第 N + 1 个矩形所在的 X、Y 的坐标，X 轴上宽 W，Y 轴上高 H；将列表按照 X 的值由小到大排序，根据图像中表格各列首行与尾行之间 X 坐标差值，设定合理的单列判断差值 ΔX，判断任意两单元格是否属于同一列，则只需计算 X 坐标之间差值是否小于 ΔX。据此，可以将列表中的位置信息按照列的顺序转入字典形式存储。字典的键为每列的 X 坐标(可以是同一列任一单元格 X 值)，每个键存放的值为该列下的全部单元格信息。图 9e 便是基于连通域搜索复现的表格框架，但由于联锁表存在整行被中横线划掉的情况，导致划掉行单元格连通域过小，而被算法忽略掉，因此需要分析字典中的位置信息，利用表格固有特征补全框架。

3.4. 基于图表特征的 RS 方框查补算法

前述获得的字典将图像表格信息全部转为数据信息，分析数据便能轻易得到图表的各项特征，如各列宽度，整表高度、行数等，这些特征又拿来作为补全、拆分单元格的依据。据此设计了基于图标特征的 RS (replenish or split)方框查补算法，对字典中的数据进一步分析和补充，达到补全表格方框的目标。

观察联锁表分布可知，可将图表分为表头和表内。表头相对固定，在 Y 轴上起始位置统一。表内部分，按照行数可以分为非均匀列和均匀列；非均匀列在表内左部分，列数通常小于 5 列，各列行数不均，但各单元格水平上框线均与右侧水平框线相接。均匀列在表中占据绝大部分，各列行数、行高一致。根据这些特性，分别对表头、非均匀列和均匀列进行遍历、补框(判断上下两单元格间距是否过大，通过计算间距与均匀行高的倍数关系，补充一个或多个均匀行单元格，且单元格之间留有标准缝隙)和拆框(判断单个单元格是否过高，需要拆分成两个单元格，且单元格之间留有标准缝隙)等操作，进一步完善表格，极大的保障了表格框架的完整性和各单元格定位的准确性。图 9f 便是利用 RS 方框查补算法，在前者基础上补全框架，得到图 9a 的完整表格，并且每个单元格位置信息都存储在字典中。

3.5. 基于连通域的字符区域定位算法

在单元格中，文本行通常只占据靠左区域，为了更好识别文本行，需要进一步对字符区域进行定位并提取。本文利用 CRNN 深度卷积神经网络实现字符识别功能，该网络具有泛化能力强、无需字符分割、无需提取字符特征等特点。因此当单元格框线与字符重叠时，不再需要设计复杂的消隐表格线算法，只需在训练样本中加入带有下划线的样本，模拟框线与字符重叠情况。

由前述预处理过程可知，在完成表格方框定位后，图像将从 DNG-OTSU 二值化图像变更为字符更为清晰的 DN-OTSU 二值化图像。如图 10，取字典中某值，按照位置信息定位并取出单元格图 10a，字符区域靠左，与底线相隔较近，且有部分粘连。如直接按照表格方框定位算法，查找连通域后计算包覆各连通域的最小正矩形，容易受到上下框线干扰。因此，首先将单元格上部分某范围(根据上框线范围合理

设定)设置为全白，再查找连通域，如图 10b，统计可得字符区域在 Y 轴上起始位置。同理，再将底部某范围设置为全白，通过连通域确定字符区域 X 轴上起止位置，如图 10c。默认字符区域最下方为单元格底线，为方便展示定位，按照位置信息将矩形画在图中。如图 10d 所示，尽管图 10c 中因为设置了最小矩形的条件，导致没有圈出逗号，但字符区域仍完整被方框包围，且消除了上下框线的影响。

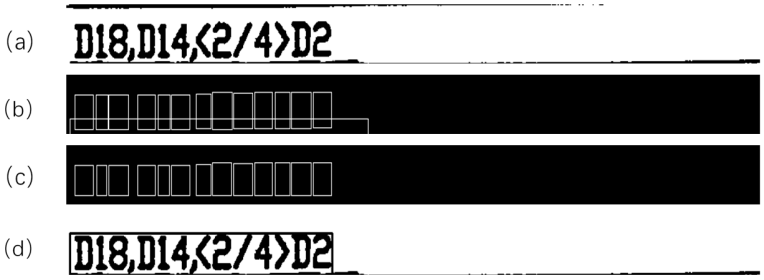


Figure 10. Cell character positioning
图 10. 单元格字符定位

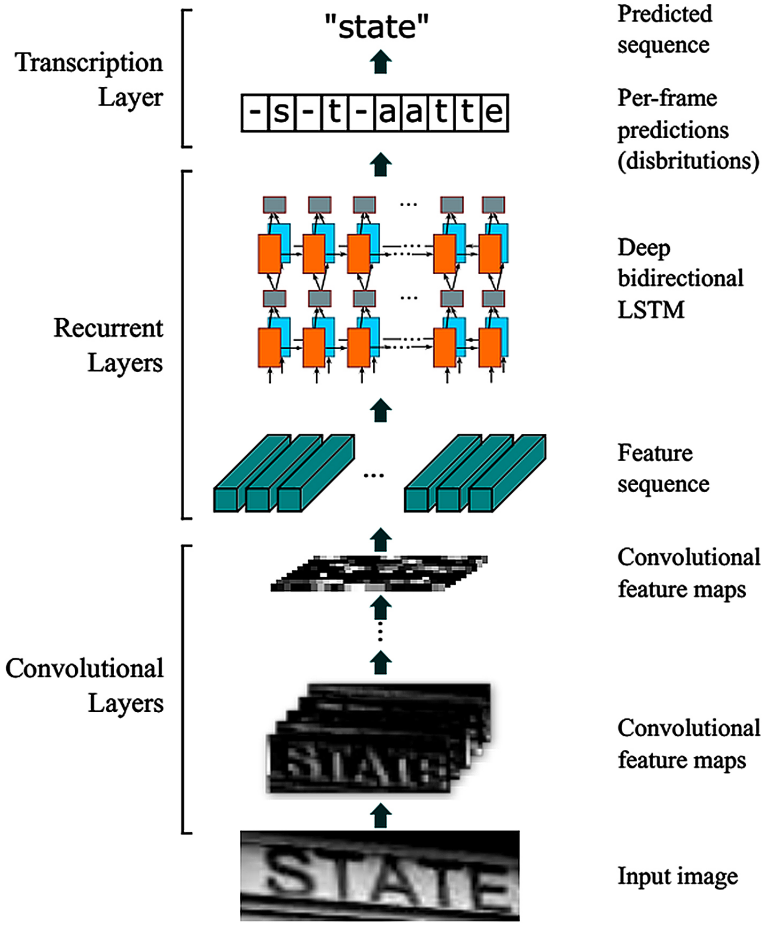


Figure 11. CRNN structure model diagram [19]
图 11. CRNN 结构模型图[19]

3.6. 识别模型-CRNN 网络

在多数图表识别系统中，影响识别准确率的原因主要有两个方面：一是字符分隔准确率。绝大部分识

别网络需要将字符区域按字分隔逐一识别，当字符之间粘连、字符与单元格边框线重叠或者汉字有偏旁部首时，字符分隔准确率大幅降低。二是人为设计字符特征提取有一定的局限性，单一的设计特征在字体变化、图像模糊[16]、背景干扰[17]、字符与单元格边框线重叠的时候泛化能力就会迅速下降，鲁棒性不够。

考虑到上述两种干扰因素，本文采用目前较为流行的场景文字识别模型 CRNN (Convolutional Recurrent Neural Network)循环卷积神经网络[18] [19]，可识别较长的文本序列[20]。CRNN 对单元格字符区域图像进行不定长字符串识别，不同于传统的字符识别算法，CRNN 不需要对整行的字符串进行分割后单字识别，而是直接整行循环卷积识别。这样就能避免了在字符分割时产生的误差，大大提高了识别准确率。其网络结构如图 11 所示。

从图可知，其网络结构从下到上分为三个部分：

1) 卷积层，通常采用 7 层 CNN (如 VGG16)，从输入图像中提取特征序列，其中输入图像要求缩放到相同高度，本文设置为 32；按照从左到右顺序，依次从特征图中生成向量，每个特征向量代表的是图像在一定宽度上的特征，本文设定这个宽度为 1，即单个像素。

2) 循环层，即一个 stack 形深层双向 LSTM 循环神经网络，作用是预测从卷积层获取的特征序列的标签(真实值)，得到所有字符的 softmax 概率分布，分布即长度为字符类别数的向量，再输入到 CTC 层。

3) 转录层，使用 CTC (Connectionist Temporal Classification)，通过去重整合等步骤把从循环层输出的标签分布翻译成最终的识别结果；CTC 连接在 RNN 网络最后一层进行序列学习和训练。对于某长度为 T 的序列，每个样本点在 RNN 网络的最后一层都会输出一个 softmax 向量，代表该样本点的预测概率，所有这些样本点的概率传递给 CTC 模型后，输出最可能的标签，然后通过去除空格和去重操作，得到最终的序列标签[21]。

4. 识别结果与讨论

4.1. 训练集制作

本文采用 Ubuntu + Anaconda 开发环境，利用 Pytorch 开发框架实现文献[20]提出的 CRNN 网络模型。搭建好 CRNN 网络模型后，为更好识别联锁表图像，需要制作专用的训练集，公开的训练集如 Caffe-ocr 中文合成数据、Synthetic Data for Text Localisation 等不适用于本识别系统训练。通常需要识别的联锁表扫描件由各地方设计院提供，字体上存在一定差异。本项目共收集到 60 份不含带横线字符的联锁表和 10 份带横线字符的联锁表。

首先考虑不带横线字符训练集。经观察存在几种稍有差异的字体，且提供联锁表的单位没有提供准确的字体型号。因此并不能通过网络下载指定字体来设计程序自动生成训练集，采取原图取字、合成样本的方式进行训练。将 60 份联锁表随机分成两组，一组 50 份用于字体采集，另外一组 10 份用作识别测试。字体采集将对各联锁表进行 DN-OTSU 二值化，再用前述连通域方法，获得包围连通域轮廓的最小矩形，从而将各独立字符从图像中提取出来。综合各图，可以得到 37 种常用字符，包括部分字母、数字、标点符号等；以及 60 种非常用字符，主要是联锁表标题汉字等。对于极少出现的字符，如某图中仅出现一次的地名等，这类字符并不放在本次训练样本中，即使是导致误识别，对整图识别率影响不大，在实际应用中也不存在影响。对 50 幅图采集结束后，选出上述 97 种字符样本。对剩下的字符图像，通过计算字符图像与各样本图像的欧氏距离完成粗略分类，然后再人工对各字符样本集进行检查。最后遍历各图，去掉字符两边多余空白，再按比例缩放到高度为 40、宽度不定的标准单个字符图像。最终得到拥有 97 类、126,794 张单个字符图像的字符库。

合成训练样本，主要通过字符组合与噪声叠加的方式来合成，背景统一为白色 255，每个样本固定为 10 个字符(含标点符号)，图像长高固定为(280, 32)。为确保样本多样化且与联锁表单元格图像近似，

从字符库中随机选取 10 个字符图像进行组合，随机字距，随机字符串上下左右边距(保证图像左空白较少，且字符串整体靠近或相接于下边缘)，随机产生下边线干扰(模拟框线干扰，产生的边线长度和宽度随机，但要随机合理)。组合完成后，随机加入高斯噪声和模糊处理。为更好的模拟扫描件二值化后，字符可能出现的断点、不连续情况，随机反转样本中 100 个像素点的值，即 0 变成 255、255 变成 0。在字符库中，常用字符的字符图像占绝大多数，如果采用随机抽取 10 个图像组成 1 个样本的方法组成样本库训练，会导致非常用字符在样本中占比太小，网络训练时容易过拟合。为解决样本不均衡问题，将字符库分开成常用字符库和非常用字符库，然后分别在常用和非常用字符库中产生 80 万和 20 万个训练样本。最后按照 9:1 的比例随机分配到训练集和验证集，并制作图片名与标签对应的 TXT 文件。

然后考虑带横线字符训练集。在 10 份的带横线字符联锁表中，共计 31 种带横线字符，仅含部分英文和数字字符。由于横线会划过单个或多个字符，且带横线字符字体与不带横线字符字体一致，因此不考虑原图取字，将从上述字体库中，取出需要的 31 种字符，再利用图像处理技术，计算字符高度，为字符加上与原图特征一致的中横线。具体训练集参数见表 1。

观察带横线字符的联锁表知，表内仅部分出现短横线划掉字符情况，故同时存在只含正常字符、只含带横线字符和两种混合(如图 2 右 2)的单元格。为更好接近原图单元格实际情况，考虑设计混合字符训练集。在上述两种字体库中，从正常字符中抽取 X 个图像，X 在 0~10 之间取值；再从带横线字符库中取 10~X 个图像；最后再从这 10 个图像中随机选择 0~2 个替换成逗号图像；按照前述合成样本格式，将 10 个字符图像合成 1 个样本。具体训练集参数见表 1，部分样本示意图见图 12。

Table 1. Three ways training set
表 1. 三种方式训练集

	联锁表份数	字体获取方式	字符种类	字符图像数量	合成训练集样本	合成验证集样本	实物样本
无横线字符	60 份	原图采集	97 种(37 种常见)	126,794 张	90 万张	10 万张	500 张
带横线字符	10 份	无横线字符+横线	31 种	50,310 张	90 万张	10 万张	100 张
混合字符		上述合并	128 种	177,104 张	180 万张	20 万张	100 张

Figure 12. Part of the sample diagram
图 12. 部分样本示意图

4.2. 训练及识别效果

训练时，合成的样本图像长高为(280, 32)，满足 CRNN 网络输入图像为固定高度的要求。采用 Adam

作为优化方法，学习速率设置为 0.0005，权重衰减设置为 0.0001。使用 CTC 损失函数，并设置批量大小为 64，共 20 个 epoch 的训练。在每个 epoch 的开始，将训练集数据随机打散。

混合训练集训练损失下降速率曲线如图 13 所示，模型的 loss 曲线下降较快，曲线出现的小尖峰代表着新一轮 epoch 的开始。该图也证明 CRNN 模型在混合合成样本集中的有效性。另外两个训练集的训练 loss 下降图与之类似，无较大差异。

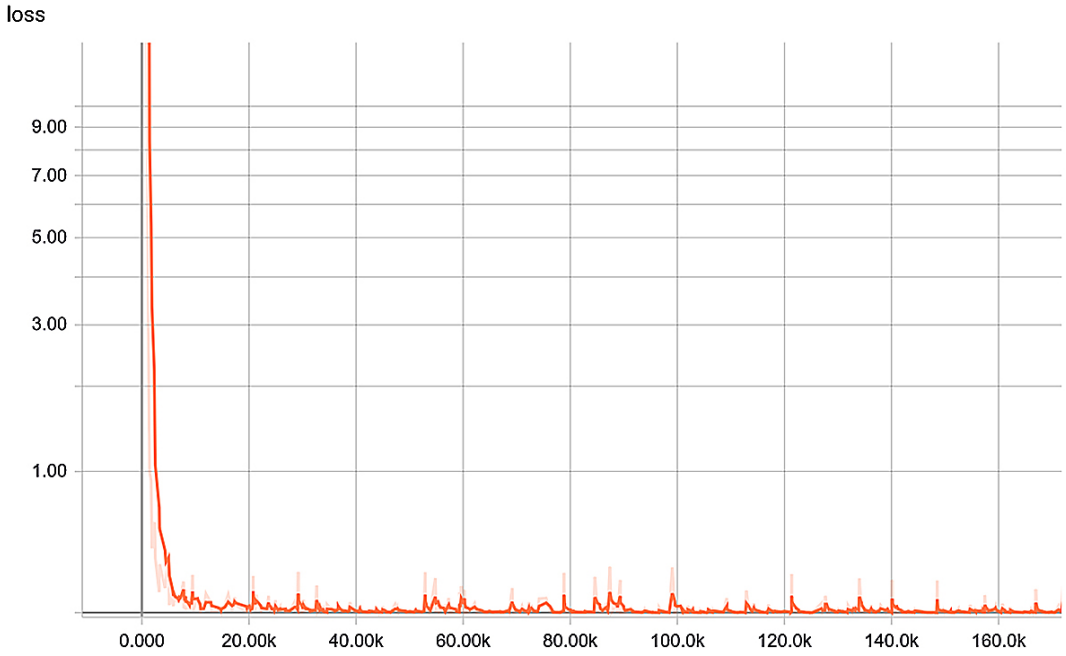


Figure 13. Training loss drop graph
图 13. 训练 loss 下降图

规定一张图像上所有数字或符号完全识别准确记录为正确，其他则记录为错误[22]。三个训练集训练的网络对应识别结果如表 2 所示。

Table 2. Recognition result
表 2. 识别结果

	合成测试集	实物测试集
无横线字符	99.2%	92.8%
带横线字符	98.6%	83.0%
混合字符	97.3%	53.0%

主要讨论无横线字符识别效果，最终在 10 万张验证集(合成)中准确率为 99.2%。将前述 10 份用作识别测试的联锁表图像进行切割，随机选取 500 张单元格图像测试该模型。共准确识别 464 张，单元格识别准确率 92.8%；其中共计字符数 6000 个，漏识别 26 个，均为不完整逗号(印刷导致逗号结构不全)；已识别 5974 个字符中，误识别 75 个字符，字符识别准确率为 98.74%。另外统计了噪声干扰导致的误识别，在 15 张图像中，出现噪声被误识别为逗号、字母等情况。

在实物测试集中，混合字符的识别情况最低，且容易出现漏识别问题。一方面是我们合成样本与实际单元格图像仍有部分差异，字符在图像中随机性不足，需要更多努力在如何合成更多样化、更实际化的样本；

另外一方面是模型泛化能力不够，提取的带横线字符与正常字符的特征不够普适性，容易出现识别误差。

4.3. 复查标注输出表格

将联锁表扫描件复现在电子表中后，由于存在一定的识别误差，需要人工进行核对。如果按照单元格依次比对，肯定会耗时耗力。所以需要在电子表中，突出显示那些易出现识别错误的单元格和字符。在识别过程中，CNN 将图片的特征提取出来后采用 RNN 对序列进行预测，最后通过一个 CTC 的翻译层得到最高条件概率的序列，即输出预测值。我们设定如果单个字符在字典中各标签的最大概率与第二大概率差值小于某值时，该字符可能出现识别错误的情况，某值由大量识别测试经验得到。据此，可以将易错字符在电子表中用红色标记，并将其所在单元格用黄色背景突出显示。

实际效果如图 14 所示，从上到下第 1、3、5 单元格为联锁表扫描件二值化图像，依次识别后写入到电子表中得到图中第 2、4、6 单元格。人工复核时，会重点比对黄色背景的单元格，如第 2 单元格中，将第一单元格的“I”误识别为“]”；第 6 单元格中，尽管“1”标记为红色，人工核对后，并没有错误。混合单元格如图 6 中第 3 和第 5 单元格所示，带横线字符能还原在电子表中。但第 3 单元格中，字符“S”因为与前横线相交，在第 4 单元格中被识别成带横线字符。

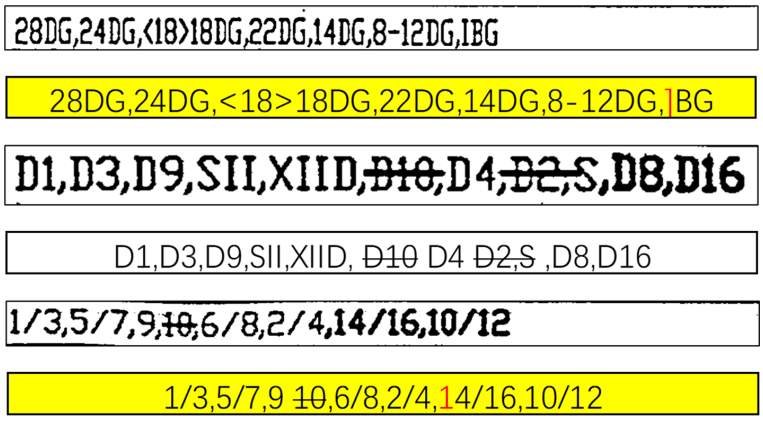


Figure 14. The effect of writing into the electronic watch after recognition
图 14. 识别写入电子表效果对比

5. 结论

本文设计的基于连通域的联锁表扫描件图表自动识别算法，能通过图像处理和神经网络对扫描件中的表格及文字进行识别，完整复现在电子表格中，并将疑似识别错误的字符及其所在单元格突出显示，方便人工复核。针对扫描件扫描过程中存在的光线不均、噪声复杂的情况，提出了 DN-OSTU 与 DNG-OTSU 二值化算法，在图像 RGB 空间进行两种内核的均值滤波和图像相除、归一化等，使得二值化后的图像噪声干扰少、字符形态保留完整；针对联锁表扫描过程出现的倾斜问题，提出了检测倾斜角度快速且准确的基于 PPHT 的倾斜矫正算法，保证后续还原表格框架以及定位的准确性；针对联锁表扫描件尺寸较大，表格线易弯曲、甚至错位的情况，提出了基于连通域的定位算法定位表格方框及文字区域，并提出基于图表特征的 RS 方框查补算法，确保表格的完整性和各单元格定位的准确性；识别中针对多设计院的联锁表无具体字体型号，采用原图提取字符制作训练集，训练卷积循环神经网络 CRNN，达到较高准确率。实验中，对多设计院提供的联锁表进行测试，单元格识别准确率达到 92.8%，字符识别准确率为 98.74%。经多次测试，单幅联锁表扫描件转为电子版表格时间控制在 5 秒以内。

综上，该算法具有准确率高、鲁棒性好、识别速度快等特点，可为纸质版联锁表扫描件复现电子版

从而二次开发提供有效的技术途径，也可进一步推广到大型表格扫描件的自动识别领域。

基金项目

这项工作得到了国家自然科学基金的资助，资助号为 61773290 和 71571107。

利益冲突

作者声明没有利益冲突。

参考文献

[1] 谢亮. 表格识别预处理技术与表格字符提取算法的研究[D]: [硕士学位论文]. 广州: 中山大学, 2005.

[2] 张达峰. 基于深度卷积神经网络的文字识别算法研究[D]: [硕士学位论文]. 贵阳: 贵州大学, 2019.

[3] 李业富, 赵玉刚, 姜文革. 基于图像处理的陶瓷产品缺陷识别研究[J]. 现代制造工程, 2014(5): 109-112.

[4] 王科俊, 冯伟兴. 中文印刷体文档识别技术[M]. 北京: 科学出版社, 2010:12-13.

[5] 刘昱. 印刷体表格识别的研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2013.

[6] Yu, B. and Jain, A.K. (1996) A Robust and Fast Skew Detection Algorithm for Generic Documents. *Pattern Recognition*, **29**, 1599-1629. https://doi.org/10.1016/0031-3203(96)00020-9

[7] Jiang, H.F., Han, C.C. and Fan, K.C. (1997) A Fast Approach to the Detection and Correction of Skew Documents. *Pattern Recognition Letters*, **18**, 675-686. https://doi.org/10.1016/S0167-8655(97)00032-9

[8] Akiyama, T. and Hagita, N. (1990) Automated Entry System for Printed Documents. *Pattern Recognition*, **23**, 1141-1154. https://doi.org/10.1016/0031-3203(90)90112-X

[9] Ishitani, Y. (1993) Document Skew Detection Based on Local Region Complexity. *Proceedings of* 2*nd International Conference on Document Analysis and Recognition* (*ICDAR*' 93), Tsukuba Science City, 20-22 October 1993, 49-52. https://doi.org/10.1109/ICDAR.1993.395784

[10] Postl, W. (1986) Detection of Linear Oblique Structures and Skew Scan in Digitized Documents. *Proceedings* of 8*th International Conference on Pattern Recognition*, Paris, October 1986, 687-689.

[11] Jiang, X.Y., Bunke, H. and Widmer-Kljajo, D. (1999) Skew Detection of Document Images by Focused Nearest-Neighbor Clustering. *Proceedings of the* 5*th International Conference on Document Analysis and Recognition*, 629-632.

[12] Cao, Y., Wang, S.H. and Li, H. (2003) Skew Detection and Correction in Document Images Based on Straight-Line Fitting. *Pattern Recognition Letters*, **24**, 1871-1879. https://doi.org/10.1016/S0167-8655(03)00010-2

[13] Ballard, D.H. (1981) Generalizing the Hough Transform to Detect Arbitrary Shapes. *Pattern Recognition*, **13**, 111-122.

[14] Ye, X., Cheriet, M., Suen, C.Y., et al. (1999) Extraction of Bankcheck Items by Mathematical Morphology. *International Journal on Document Analysis and Recognition*, **2**, 53-66. https://doi.org/10.1007/s100320050037

[15] Ye, X., Cheriet, M. and Suen, C.Y. (2001) A Generic Method of Cleaning and Enhancing Handwritten Data from Business Forms. *International Journal on Document Analysis and Recognition*, **4**, 84-96. https://doi.org/10.1007/s100320100056

[16] 张敏, 于剑. 基于划分的模糊聚类算法[J]. 软件学报, 2004, 15(6): 858-868.

[17] 龙建武, 申铉京, 陈海鹏. 自适应最小误差阈值分割算法[J]. 自动化学报, 2012, 38(7): 1134-1144.

[18] Szegedy, C., Liu, W., Jia, Y.Q., et al. (2015) Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 July 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[19] Shi, B., Bai, X. and Yao, C. (2016) An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 2298-2304. https://doi.org/10.1109/TPAMI.2016.2646371

[20] 冯海. 基于深度学习的中文 OCR 算法与系统实现[D]: [硕士学位论文]. 北京: 中国科学院大学(中国科学院深圳先进技术研究院), 2019.

[21] 尚果超. 基于深度卷积模型的手写中文文本识别[D]: [硕士学位论文]. 秦皇岛: 燕山大学, 2019.

[22] 刘小波, 徐波, 宋爱国, 等. 基于的变电站巡检机器人数字仪表识别算法[C]//江西省电机工程学会. 2019 年江西省电机工程学会年会论文集, 2019.