文章编号：2095-2163(2023)06-0097-06 中图分类号：TP391 文献标志码：A

面向多尺度折线图数据解析的两阶段方法

王赵宇，周国帆，舒 挺
（浙江理工大学 计算机科学与技术学院，杭州 310018）

摘 要：折线图数据解析在医疗、学术等领域具有重要价值，传统的基于图形学的方法需要对每个像素归类再进行后续识别从而影响了算法的效率，而基于神经网络的方法由于需要依赖显著的图表特征，在背景与线条颜色相近、折线较细等情况下存在效果不佳等问题。基于此，本文融合图形学方法和神经网络方法，提出了一种两阶段折线图数据解析方法。首先，基于色彩筛选线条上的像素对轴标签、边界、角点、网格等图表基础信息进行检测；其次，利用神经网络获取初步分割结果，结合传统图形学方法进一步完善分割效果并进行数值映射；最后，在多尺度折线图数据抽样数据集上评估本文算法，在轴标签检测中 $F1-Score$ 值达到 0.851，在线条分割中 $mIoU$ 值达到 0.921，在数值映射中平均欧式距离达到 0.087。实验结果表明该方法可以有效解决线条与背景色颜色接近、图表分辨率不高等场景下的折线图数据解析问题。

关键词：色彩筛选；折线图；数据解析

Two-stage method for multi-scale line chart data parsing

WANG Zhaoyu, ZHOU Guofan, SHU Ting
(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 Line chart data analysis is of great value in medical and academic fields. Traditional graphics-based methods need to classify each pixel and conduct subsequent recognition, which affects the analyzing efficiency. Neural network-based methods also have some disadvantages such as poor performance in the case of similar background and line colors and thin broken lines since the models mainly rely on prominent visual features. To solve these problems, this paper proposes a two-stage line graph data parsing method by integrating graphics and neural network methods. This method is based on color screening of the pixels on the line. Firstly, the basic information of the chart such as axis label, boundary, corner and grid are detected. Secondly, the neural network is used to obtain the initial segmentation results, and the traditional graphics method is incorporated to improve the segmentation results and carry out numerical mapping. The proposed algorithm is evaluated on the multi-scale line graph data sampling dataset. The $F1-score$ value of proposed method reaches 0.851 in axis label detection, the $mIoU$ value reaches 0.921 in line segmentation, and average Euclidean distance reaches 0.087 in numerical mapping. Experimental results show that this method can effectively solve the problem of line graph data parsing in the scene where the line color is close to the background color and the graph resolution is limited.

【Key words】 color screening; line chart; data parsing

0 引 言

可视化可以将不可见的数据现象转化为可见的图形符号，折线图是其主要形式之一，有广泛的应用场景。在医疗领域，听力图是一种特殊的折线图，表示各种频率下的听力水平，可以诊断听力损失，为患者选择合适的助听器；在科学文献中，大量的实验数据通常也可以以折线图的形式展现，而没有提供实验原始数据，从而制约了对实验的复现和对数据的分析复用。因此，抽取图表的底层数据在包括医疗、科研诸多领域具有重要应用价值。

折线图数据解析是指从缺乏底层数据的图表图像中，按照图形比例、位置等信息逆向提取折线数值的过程。图表自动化提取相比于人工估算，速度更快，可靠性更高。解决折线图数据解析问题主要面临以下 3 个挑战：

作者简介：王赵宇(1999-)，男，本科生，主要研究方向：计算机视觉；周国帆(2001-)，男，本科生，主要研究方向：计算机视觉；舒 挺(1979-)，男，博士，副教授，主要研究方向：计算机网络协议、软件测试、移动电子商务系统开发等。
通讯作者：舒 挺 Email：shuting@zstu.edu.cn

收稿日期：2022-07-08

(1)由于原始图片分辨率不高,导致轴标签检测的精确度下降;

(2)由于轴标签字体多样化,导致轴标签的识别难度增加;

(3)由于折线与背景颜色差异不明显、网格粗细和颜色多样使得分割准确折线更加困难。

当前图表解析的主流方法分为两类:基于图形学的方法和基于神经网络的方法。在基于图形学的方法中,Huang[1]等利用连通区域分析分离文本和图中的数字信息,使用边映射和特定规则提取图形元素;Jayant[2]及 Falk[3]等开展了文本方向的相关工作,有助于文本识别;Yan Ping Zhou[4]等使用边界追踪和霍夫变换识别柱状图中的柱形。上述基于图形学的方法能够实现较高精准度,使得抽样数据更加可靠,但效率不高,且大多基于特定的图表规则,对于多变的折线图效果欠佳。神经网络具有自学习能力,通常更高效。Mathieu[5]等采用基于神经网络的目标检测模型,对目标物体定位,获取数值信息,此方法仅对散点图有效;Noah[6]等和 Jorge[7]等提出了定位数字和分析内容的方法,为多类图表数据提取提供了解决思路;Li[8]等提出了一种能够自动提取完整听力图信息的可行模型,但仅能够提取部分带有标记的数据,难以完整提取折线数据。基于神经网络的方法应用范围广,操作效率高,但精准度不如图形学方法。

针对线条与背景色颜色接近、图表分辨率不高等情况,本文提出了基于神经网络和图形学方法结合的两阶段折线图解析算法,利用神经网络解决图形学方法逐像素遍历效率不高的问题,利用图形学方法的高准确性解决神经网络由于线条边缘大量过渡颜色的像素导致数值提取不准的问题。本文使用神经网络高效获取初步提取结果,使用图形学方法基于色彩进一步完善线条分割效果,提出了一种折线图数据解析方法,使用擅长小目标检测的 Faster-RCNN 模型进行轴标签的目标检测,使用线性回归过滤掉因字号微小或字体多样导致识别错误的轴标签,使用双色彩筛选算法修正 Unet++语义分割模型的分割图提高折线图数值提取的准确度。折线图数据解析的两阶段方法如图 1 所示。

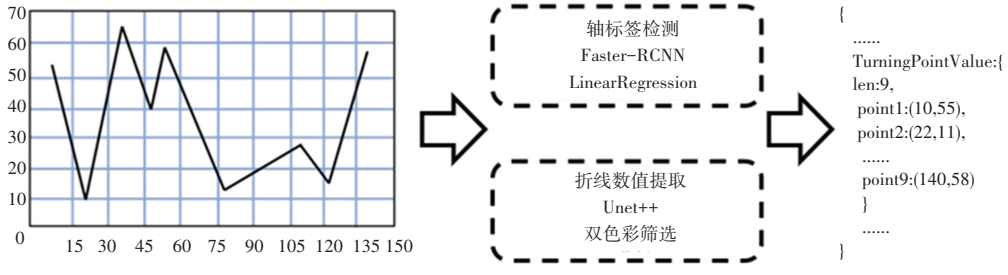


图 1 折线图数据解析的两阶段方法

Fig. 1 A two-stage approach on data analysis of line graphs

1 本文方法

本文提出的折线图数据解析算法如下:

(1)使用目标检测网络对原始图表进行目标检测,得到每个标签的信息:$Label_i = \{box_i: [b_{left}, b_{up}, b_{right}, b_{down}], value_i: v\}$。其中,$box_i$ 为标签的位置信息,依次为左、上、右、下边界,$value_i$ 为标签所代表的真实数值。

(2)借助 box_i 进行掩膜,使用 Canny 边缘检测和 FAST(Feature from Accelerated Segment Test)角点检测算法检测表格边界和表格角点以及表格网格。

(3)使用语义分割网络进行线条分割配合色彩筛选算法,得到折线 $Line = [x_1: [y_1 \cdots y_n] \cdots x_n: [y_1 \cdots y_n]]$。

(4)根据轴标签值和标签像素坐标进行映射计算。

折线图数据解析的两阶段方法主体流程如图 2 所示。

1.1 轴标签检测

Faster R-CNN 能够更好地检测小目标,对于本任务中的检测轴标签具有天然优势。

轴标签检测在图片中找到标签位置,并对标签进行识别,算法如下:

首先,用 Faster R-CNN 模型对图表中的数字进行目标检测,得到标签的 $box_i: [b_{left}, b_{up}, b_{right}, b_{down}]$:加载预训练权重模型,用多尺度折线图数据抽样数据集进行微调训练,生成新的网络模型。利用 box_i 边框信息将轴标签裁剪,使用灰度图进行边界调整,如果边界像素灰度值大于 $255 * 0.95$,判定为数字部分,从而得到更加精准的数字框。

将数字框覆盖的图像内容输入位数分类网络：经过卷积层、激活函数、最大池化层对图像进行特征提取，全连接层进行分类，最终分为 3 类（本文假设轴标签位数最大为 3 位，可根据实际情况调整），并得到预测结果；根据数字位数对数字进行等分裁切，将单个数字框输入数字识别卷积神经网络，在全连接层将数字分为十个类别，再进行数字组合，从而得到轴标签数值。

最后，通过线性回归过滤错误值。轴标签一般是符合线性递增排列的，可以对一列标签数值和像素坐标根据最小二乘法进行线性拟合，从而过滤错误值，算法具体描述见表 1。

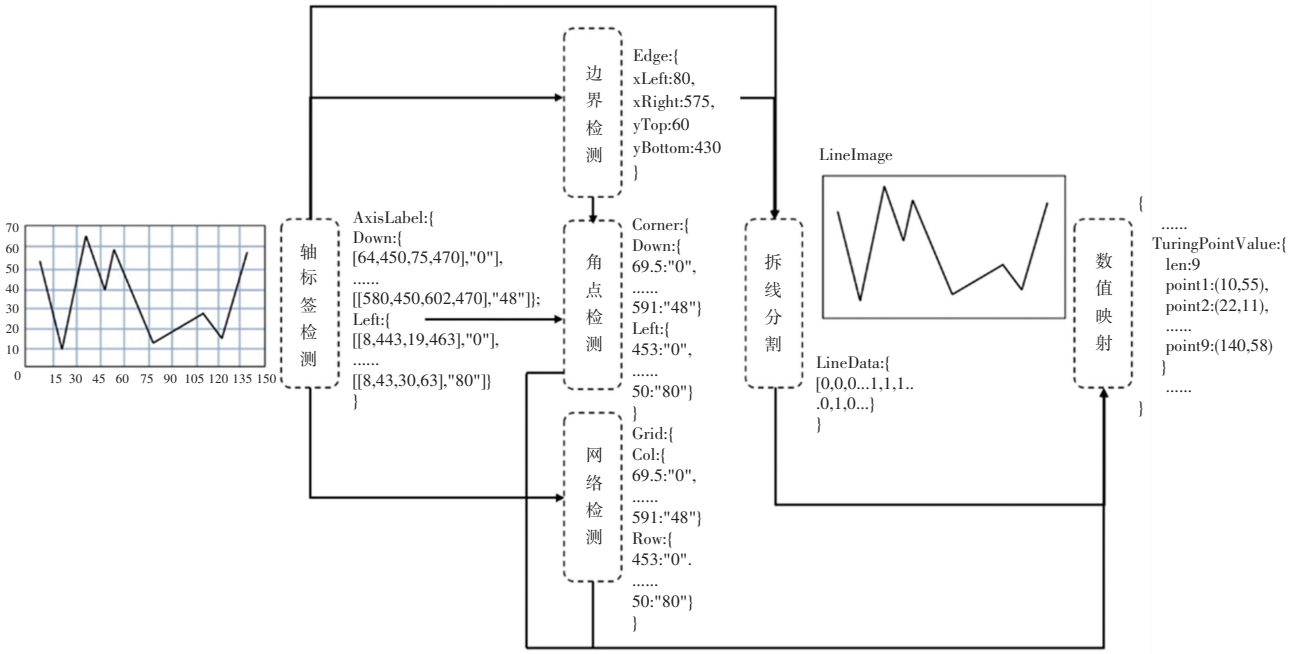


图 2　折线图数据解析的两阶段方法主体流程

Fig. 2　The main process of two - stage method for data analysis of line graph

表 1　线性回归过滤错误值算法

Tab. 1　Linear regression error filtering algorithm

算法　线性回归过滤错误值算法
输入　一列标签列表 $p_i = (t_{xi}, t_{yi}, v_i)$
输出　修正后的标签列表 $np = (t_{xi}, t_{yi}, v_i)$
(1)针对中心坐标 (t_{xi}, t_{yi}) 利用最小二乘法进行线性拟合；
(2)计算最大偏差值，如果大于 0.01，判定存在错误值，将该值删除；否则判定无错误值；
(3)对新的标签列表重新进行线性拟合，重复(1)、(2)阶段，直至算法结束。

1.2　折线图数值提取

在折线图数值提取之前，需要进行边界检测、角点检测、网格检测等预处理。使用标签定位与识别的结果掩膜，进行 Canny 边缘检测，根据色彩阶跃变化对包含角点的外边界进行调整；使用 FAST 特征点检测算法进行角点检测；使用与边界检测类似的方法进行网格检测。

采用 Unet++语义分割模型对折线图进行语义分割，得到一阶段的分割图。

折线图往往线条很细，且包含网格背景等无用信息，本文设计了双色彩筛选算法得到第二阶段分割图，具体步骤如下：

(1)对第一阶段分割图进行像素色彩统计，选择数量最多的像素值进行第一次色彩筛选；

(2)从图表边界向内延伸进行色彩统计，进行第二次色彩筛选，生成第二阶段分割图。

最后，在第二阶段分割图中仍存在边缘不清晰、跨度大、内部残缺的问题，需要进行填充修补。从左边界向右扫描，将最高点和最低点之间进行填充，得到饱满的折线；依照目标色彩从原图中向四周探寻，将第二阶段分割图向外延伸，获得完整的折线。部分线条会由于网格、过渡像素等的干扰形成中断缺

失,使用左右两点的计算均值代替该列的像素值和位置。

1.3 数值映射计算

针对折线图数值映射计算,本文给出以下定义:

(1)像素坐标:目标像素点距图表上界的像素数量为 y_p,目标像素点距离作边界的像素数量差值为 x_p,像素坐标记为 (x_p,y_p);

(2)图表坐标:目标点经映射计算后在图表中所代表的真实横纵坐标值,记为 (x_r,y_r);

(3)由于线条存在宽度,故取分割图中每一列像素中点作为映射输入点。

获取到线条的分割图后,需要根据轴标签的中心位置和映射输入点进行计算。首先,将标签集合 K 中的每一个轴标签记为 $p_i=(t_{xi},t_{yi},v_i)$ $(i=1,2,3\cdots)$,其中,(t_{xi},t_{yi}) 为标签的中心像素坐标,v_i 为标签所代表的真实数值。

分别取相邻轴标签的轴向像素数量差值和真实数值差值,两两之间计算单位像素坐标值所代表的真实图表坐标值,取平均值,对于每一个待求图表坐标的映射输入点,与每一个轴标签 p_i 进行如公式(1)的计算,var 即所求坐标数值,即可完成对于折线图的数据解析。

$$var=\frac{\sum(v_i+(d_y-t_{yi})\times dist_{ave})}{len(p_i)} \tag{1}$$

其中,d_y 表示当前标签的纵向像素值;$dist_{ave}$ 表示图表单位像素的坐标值跨度范围;$len(p_i)$ 表示标签个数。

2 实验与分析

2.1 实验测试问题

本文实验设计主要测试算法在以下两个方面的效果:

(1)测试本文方法针对不同字号、不同字体样式的适应能力。在含有 50 种不同字体、随机字号的数据集上计算单字符、完整标签的识别精确率、召回率、$F1$ 值检测轴标签衡量算法对不同大小、字号轴标签定位和内容识别的准确程度,并比较线性回归策略对算法的提升效果;

(2)测试本文方法针对折线、网格、背景颜色接近的性能。在含有多种背景、线条、网格样式的数据集上,计算线条分割的平均交并比、准确度、精确度、平均欧式距离,衡量算法在不同的干扰背景下抽取线条和数值映射的准确程度,并比较双色彩筛选策略对算法的提升效果。

2.2 测试数据集

目前还没有公开的大规模折线图数据抽样数据集,为了验证本文方法的实用性和有效性,构建一个折线图数据抽样数据集,包含具有 50 种轴标签字体、随机轴标签位置、9 种背景颜色、随机网格颜色、3 种网格宽度、7 种网格数量、随机折线颜色、12 种折线宽度的 5 000 张折线图。针对轴标签检测实验,本文引入 ICDAR 2019 扫描收据数据集,该数据集具有 1 000 个完整的扫描收据图像,在此数据集的基础上引入本文设计的数据集共 6 000 张图片进行轴标签检测算法的性能评估。

2.3 轴标签检测实验

2.3.1 实验设置

将数据集按照比例 6∶3∶2 划分训练集、验证集和测试集。轴标签检测实验参数设置见表 2。

2.3.2 评价指标

本文使用标准的目标检测评价指标来衡量轴标签检测的性能,即单字符识别精确率、单字符识别召回率、完整标签识别精确率、完整标签识别召回率、$F1-Score$。精确率计算公式(2)、召回率计算公式(3)如下:

$$Precision=\frac{TP}{TP+FP} \tag{2}$$

$$Recall=\frac{TP}{TP+FN} \tag{3}$$

其中,TP 表示真正例,即被分类器正确地判定为正例的样本数;FP 表示假正例,即被分类器错误地判定为正例的样本数;FN 表示假负例,即被分类器错误地判定为负例的样本数;TN 表示真负例,即被分类器正确地判定为负例的样本数。

表 2 轴标签检测实验参数设置

Tab. 2 Parameter setting in axis label detection experiment

模型	输入大小	Batchsize	Epoch	学习率	损失函数	优化器
定位模型	900×900	4	50	0.005	Smooth L1 Loss	SGD
位数识别模型	28×28	16	50	0.001	CrossEntropyLoss	Adam
字识别模型	28×28	16	50	0.001	CrossEntropyLoss	Adam

2.3.3　实验结果

轴标签检测实验结果见表 3。轴标签检测实验结果示意图如图 3 所示。轴标签检测实验中采用的线性回归策略对结果优化对比结果如图 4 所示。

表 3　轴标签检测实验结果

Tab. 3　Axis label detection results

方法	单字符识别精确率	单字符识别召回率	完整标签识别精确率	完整标签识别召回率	完整标签 *F*1 值
未经过线性回归	0.907	0.888	0.836	0.834	0.835
经过线性回归	0.929	0.845	0.882	0.823	0.851

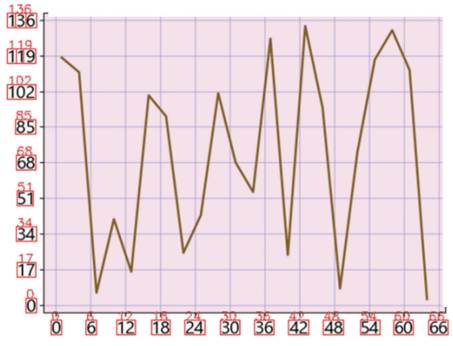


图 3　轴标签检测实验结果示意图

Fig. 3　Schematic diagram of axis label detection results

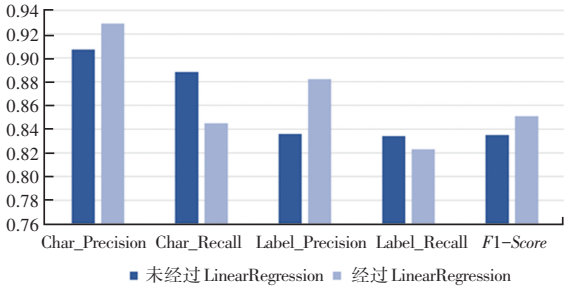


图 4　线性回归策略对结果优化对比图

Fig. 4　Comparison of results on linear regression strategy

实验测试了使用线性回归方法对实验结果的影响，实验结果表明该方法可以滤除错误的轴标签识别结果，即会提升精确率并降低召回率，但考虑到本文采用坐标值映射的方法是最多等值策略，故算法的精确率更为重要，即允许漏识别，避免误识别，因此经过线性回归后会过滤掉部分错误识别的结果，虽然会使得字符召回率和轴标签召回率下降，但对整体算法的准确度是有益的。同时对于少部分图片，由于字体过小导致个别字符识别错误或因字体样式（如部分字体的 9 和 0 过于相似）导致个别字符识别错误，线性回归能够大幅减弱轴数字检测错误带来的副作用，有利于提高后续数值提取的准确度。

2.4　折线图数值提取实验

2.4.1　实验设置

选用数据集中 4 000 张图片进行训练，1 000 张图片用于测试。参数设置见表 4。

表 4　折线图数值提取实验参数设置

Tab. 4　Parameter setting in line graph value extraction experiment

参数	参数值
输入大小	224×224
Batchsize	16
Epoch	100
初始学习率	0.001
学习率衰减策略	Poly
Poly 策略 power 值	0.9
损失函数	DiceLoss
优化器	Adam

2.4.2　评价指标

本文使用标准的语义分割评价指标来衡量线条分割的性能，指标为平均交并比（*mIoU*）、准确度、精确度。平均交并比计算，公式（4）：

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{4}$$

本文采用平均欧式距离作为评价指标评估预测值与真实值的偏离程度，将图标横向 5 等分，通过 5 个点与轴标签的中点之间的像素差值和轴便签的单位值（即两个邻接轴标签的数值差值）计算折线所表示数值，公式（5）：

$$d = \sqrt{\sum_{i=0}^{4}\left(\frac{y_i}{k} - \frac{y_i'}{k}\right)^2} \tag{5}$$

其中，y_i 表示真实值；y_i' 表示预测值；k 表示纵坐标最大标签值。

2.4.3　实验结果

折线图数值提取实验结果见表 5。折线图数值提取实验结果示意图如图 5 所示。折线图数值提取实验双色彩筛选策略对结果优化对比图如图 6 所示。本文算法计算的平均欧氏距离为 0.087。

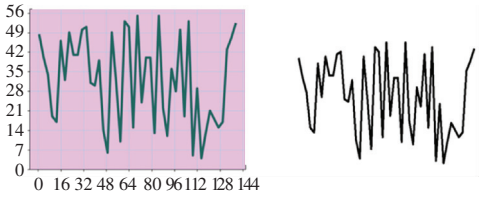
实验测试了双色彩筛选算法对实验结果的影响。实验结果表明，相比于只使用深度神经网络的方法，结合传统的图形学方法进行调整，大幅提高了线条分割的精准度，提升幅度为 53.2%，这主要是由于一些边缘像素导致神经网络进行语义分割的结果包含了大量无用信息，算法滤除了无用信息且修复了残缺的线条，从而提高了性能，但不可避免的引入

一些耗时操作。对于难以界定的边缘像素,无论将其归于线条类还是折线线条类都会导致分割线条不够平滑,不过由于坐标映射时会将该列像素值的中点作为映射输入,故只要对于上下两侧的边缘像素以同种策略进行筛选则不会影响结果。坐标值映射产生的误差的主要原因是网格颜色与线条颜色过于接近导致的映射纵坐标取值错误、边界定位出错导致的横坐标取点出错。虽然本文的方法是有效的,但仍然存在一些局限性,如训练不是端对端的,有些工作是重复的会降低操作效率。

表 5 折线图数值提取实验结果

Tab. 5 Experimental results of numerical extraction of line graph results

	神经网络结果	本文算法结果
mIoU	0.601	0.921
accuracy	0.935	0.991
precision	0.588	0.987



(a) 折线图原图 (b) 折线图数值提取

图 5 折线图数值提取实验结果示意图

Fig. 5 Schematic diagram of experimental results of line graph value extraction

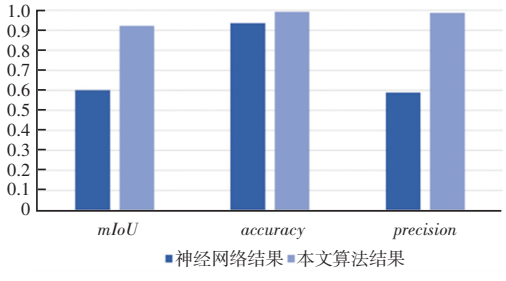


图 6 双色彩筛选策略对结果优化对比图

Fig. 6 Effect of dual color screening strategy on final results

3 结束语

为了重新抽取图表中的数据信息,本文提出了一整套折线图数据解析方法,使用擅长小目标检测的 Faster R-CNN 和 Unet++两大模型,分别对标签数字进行定位识别和对曲线进行语义分割;使用线性回归辅助解决字号微小、容易识别错误的问题;使用双色彩筛选算法修正语义分割模型的结果,提高对于细微线条提取的准确度。另外,本文方法还可以拓展,对于现实中更多复杂的表格类型,可以抽取相应的特征点,进一步地细化折线图解析方法。

参考文献

[1] HUANG W, TAN C L. A system for understanding imaged infographics and its applications[C]//Proceedings of the 2007 ACM symposium on Document engineering. 2007: 9-18.

[2] JAYANT C, RENZELMANN M, WEN D, et al. Automated tactile graphics translation: in the field[C]//Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility. 2007: 75-82.

[3] BÖSCHEN F, SCHERP A. Multi-oriented text extraction from information graphics [C]//Proceedings of the 2015 ACM symposium on document engineering. 2015: 35-38.

[4] ZHOU Y P, TAN C L. Hough technique for bar charts detection and recognition in document images [C]//Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101). IEEE, 2000: 605-608.

[5] CLICHE M, ROSENBERG D, MADEKA D, et al. Scatteract: Automated extraction of data from scatter plots [C]//Joint European conference on machine learning and knowledge discovery in databases. Springer, Cham, 2017: 135-150.

[6] SIEGEL N, HORVITZ Z, LEVIN R, et al. Figureseer: Parsing result-figures in research papers[C]//European Conference on Computer Vision. Springer, Cham, 2016: 664-680.

[7] POCO J, HEER J. Reverse-engineering visualizations: Recovering visual encodings from chart images [C]//Computer graphics forum. 2017: 353-363.

[8] LI S, LU C, LI L, et al. Interpreting audiograms with multi-stage neural networks[J]. arXiv preprint arXiv:2112.09357, 2021.

(上接第 96 页)

[15] ZHU F, ZHANG Y, LIN C, et al. A universal designated multi-verifier transitive signature scheme[C]//Information Security and Cryptology: 13th International Conference, Inscrypt 2017, Xi′an, China, November 3-5, 2017, Revised Selected Papers. Cham: Springer International Publishing, 2018: 180-195.

[16] LIN C, ZHU F, WU W, et al. A new transitive signature signature schemel[J]. Lecture Notes in Computer Science, 2016, 9955(8): 45-52.

[17] NOH G, JEONG I R. Transitive signature schemes for undirected graphs from lattices [J]. KSII Transactions on Internet and Information Systems (TIIS), 2019, 13(6): 3316-3332.

[18] NOH G, CHUN J Y. Identity-based transitive signature scheme from lattices [J]. Journal of the Korea Institute of Information Security & Cryptology, 2021, 31(3): 509-516.