现代计算机

XIANDAI JISUANJI

第29卷第16期（总第784期）

半月刊（1984年创刊）

2023年8月25日出版

主管单位 中山大学
主办单位 广州中山大学出版社有限公司
出版单位 广东现代计算机杂志社有限公司
发　　行 广东省报刊发行局（全国公开发行）
印　　刷 广州一龙印刷有限公司
社　　长 黄少伟
主　　编 石玉珍
编　　委 邹岚萍　熊锡源　李　文　石玉珍　梁嘉璐
地　　址 广州市海珠区新港西路135号
　　中山大学内（510275）
电　　话 020-84112089（编辑部）
网　　址 www.moderncomputer.cn
电子邮箱 tougao@moderncomputer.cn

ISSN 1007-1423
CN 44-1415/TP

邮发代码：46-121
定价：30.00元

邮局订刊二维码　现代计算机官方网站二维码

ISSN 1007-1423 16>
9 771007 142239

现代计算机 第29卷 第16期（总第784期） 2023年8月

2023年8月 第29卷
第16期 （总第784期）

ISSN 1007-1423
CN 44-1415/TP

现代计算机

MODERN COMPUTER

中山大学出版社 主办

中国期刊数据库CNKI全文收录期刊
中国学术期刊（光盘版）收录期刊
中文科技期刊数据库全文收录期刊
中国核心期刊（遴选）数据库收录期刊
中国学术期刊综合评价数据库收录期刊

- 研究与开发：计算机发展和软、硬件开发的理论研究
- 图形图像：重点为与图形图像相关的理论及实践研究
- 开发案例：基于某方面的计算机开发案例研究与分析
- 实践与经验：计算机应用的实例及心得

现代计算机
XIANDAI JISUANJI
第29卷 第16期（总第784期）2023年8月25日出版

目 次

研究与开发

图形图像

实践与经验

开发案例

期刊基本参数：CN44-1415/TP*1984*S*16*120*zh*P*¥30.00*1200*22*2023-08

Modern Computer
(Vol. 29, No. 16; Aug. 25, 2023)

CONTENTS

Research and Development

Graphic and Image

Practice and Experience

Development Solution

研究与开发

文章编号: 1007-1423（2023）16-0001-08 DOI: 10.3969/j.issn.1007-1423.2023.16.001

基于YOLOv5-seg的多模型电石检测分割系统

郝俊峰[1]，李玉涛[2*]，来博文[1]

（1. 河钢数字技术股份有限公司人工智能工作室，石家庄 050035；
2. 河钢数字技术股份有限公司技术中心本部，石家庄 050035）

摘要：针对实际的电石业务场景，基于YOLO系列模型设计了电石检测分割的方法，并开发了相应的系统。首先，通过采集大量的图像样本构建了较大规模的数据集并进行标注；为了节省算力和提高准确率，将检测和分割分为两个模块，模型分别为YOLOv5-l和YOLOv5-seg-m。在实验中，对模型的内部结构进行了一系列的改进，提高了算法的运算速度；并在此基础上对模型进行了参数量化，进一步节省了带宽、降低了模型的存储和运算所需的空间。相比于MaskR-CNN算法，在分割效果接近的情况下，速度得到了极大的提升。很好地完成了检测和分割电石的任务，为后续的生产自动化和无人化打下了基础，填补了相关检测的空白。之后将会针对排除光照影响、提高分割精度和评判电石质量模型继续研究。

关键词：电石；YOLOv5-seg；CSPNet；参数量化

0 引言

由于我国能源结构中石油和天然气的消耗远远高于其他能源，这导致我国长期处于“贫油少气”的能源结构状态。为了满足能源需求，我国煤化工领域和能源领域中的电石作为工业制作乙炔的关键原材料得到了广泛应用。作为世界上最大的电石生产国，我国的电石产能和产量均占据了世界总量的90%以上，充分体现了我国在电石领域的强大实力。据统计，2021年我国电石的产量高达2825万吨，证明了电石在我国能源经济中的重要地位。

电石的主要成分是碳化钙，它与水会产生剧烈的化学反应，从而产生乙炔气体。虽然电石在工业制造中的作用非常重要，但是我们也应该注意到它的潜在危险性。碳化钙的生产被列入了世界卫生组织国际癌症研究机构公布的致癌物清单中的三类致癌物清单，这表明我们需要在电石生产中引入安全措施来避免危险情况的发生。同时在生产过程中熔化了的碳化钙从炉底取出后冷却，这个环节被称作电石打样，需要人为参与等待冷却后破碎。冷却时间不宜过短或过长，所以需要人员长期看守。

为了确保电石生产的安全和高效，我们需要引入自动化、无人化的生产方式。这就需要电石的目标检测和实例分割技术的应用，因为它们是实现生产自动化、无人化的必要前提。实际上，电石的目标检测和实例分割都是比较困难的技术挑战，因为电石的状态具有多种多样的变化，比如空状态、电石棒、电石三种状态。其中，电石棒和电石较为相似，而且还会出现电石棒上有电石残留的情况。因此，在检测和分割中还会受到光照、炉火光、电石未冷却时的红光和紫光的干扰。传统视觉方法难以满足当前需求，因此我们需要应用先进的人工智能和计算机视觉技术来实现电石的目标检测和分割，通过目标检测和实例分割技术，可以实现对电石准确快速的识别和定位，同时提高生产效率，降低生产成本，确保生产安全；以

收稿日期：2023-04-28 **修稿日期：**2023-05-23

作者简介：郝俊峰（1991—），男，河北石家庄人，硕士学历，主要研究方向为工业机器视觉技术与应用；*通信作者：李玉涛（1984—），男，河北衡水人，高级工程师，主要研究方向为智能控制、工业互联网技术，E-mail：hbliyutao@hbisco.com；来博文（1991—），河北邯郸人，工程师，主要研究方向为工业人工智能技术

填补检测和分割技术在电石生产领域的空白。

随着深度学习和计算机视觉的发展，深度学习在目标检测和实例分割方向上已成为最优解。在应用于图像算法基础模型CNN的基础上发展出了很多研究分支，如区域卷积神经网络(R-CNN)系列、R-CNN[1]、Fast R-CNN[2]、Faster R-CNN[3]和Mask R-CNN[4]等的两阶段算法和以YOLO系列为代表的单阶段算法，近几年YOLO系列的迭代更新非常快。

YOLO(you only look once)算法是一种深度学习算法，由Redmon等[5]于2016年提出，它可以对图像中的物体进行实时检测和分类。这个算法以其快速和准确的检测能力而闻名，成为计算机视觉领域的一项重要技术。在过去的几年中，YOLO算法已经经历了很多的发展和改进，YOLOv1是第一版YOLO算法，它是基于卷积神经网络的物体检测方法。YOLOv1将图像分成$S \times S$个格子，每个格子预测B个边界框以及每个框的置信度和类别概率。在训练时，YOLOv1使用均方误差作为损失函数，同时也考虑了置信度的加权。YOLOv1相比于其他算法具有较高的检测速度，但是其检测精度略低于一些其他算法。YOLOv2在YOLOv1的基础上进行了改进，提高了检测精度和速度。YOLOv2采用了一系列的技术来提高其性能，例如利用BN(batch normalization)层加速收敛和提高准确性，使用多尺度训练和预测来适应不同尺度的物体，引入Anchor Boxes来更好地适应不同大小的物体，以及在训练时使用目标形状信息来帮助检测物体[6]。YOLOv3是YOLO算法的第三版，它在YOLOv2的基础上进行了改进，进一步提高了检测精度和速度。YOLOv3采用了一系列的技术来提高其性能，例如使用FPN(feature pyramid networks)来提取多尺度特征，引入多个输出层来检测不同大小的物体，使用更高效的Darknet-53作为主干网络，使用上采样来提高精度等[7]。YOLOv4在YOLOv3的基础上进行了改进，进一步提高了检测精度和速度。YOLOv4采用了一系列的技术来提高其性能，例如使用SPP(spatial pyramid pooling)结构来提取更全面的特征信息，引入CSP(cross stage partial)结构来提高网络的表达能力，使用YOLOv3-tiny网络结构来提高小物体的检测精度，引入Swish激活函数来提高准确率和泛化能力，使用DropBlock正则化来提高泛化能力等[8]。与YOLOv3相比，YOLOv4的检测精度得到了进一步提高，同时也保持了很高的检测速度。YOLOv5是由ultralytics公司提出的YOLO的一个独立分支，基于YOLOv4进行改进，使用更轻量级的网络结构，同时也保持较高的检测精度和速度[9]。再到最新的YOLOv5-seg和YOLOv8添加了实例分割的分支，实现了实例分割的功能，同时在性能上也有所提升。YOLO系列应用的领域也非常广泛，张慧春等[10]基于YOLOv5实现了对植物叶绿素含量的估测；王玲敏等[11]引入注意力机制用YOLOv5实现了安全帽佩戴的检测；尹靖涵等[12]使用YOLOv5实现了雾霾天气下交通标志的识别。

1 YOLOv5-seg的结构及改进

YOLOv5的网络主要包括Backbone(主干网络)、Neck(特征融合)、Head(预测)这三大主要部分。Backbone部分主要由Focus、CSPNet、SPP(空间金字塔池化)三种模块构成，其主要的功能是特征提取。Neck部分为FPN+PAN结构，通过自顶向下和自下而上的方式提取特征图。Head部分对Neck传进来的特征图进行结果的预测[13-15]。其结构如图1所示。

本次研究所使用的YOLOv5-seg算法主要是在YOLOv5的基础上添加了一条分割网络，并在其基础上进行了轻量化的改进，其结构如图2所示。

通过图1与图2的对比可以发现，在Backbone结构中Focus模块被替换为一个6 × 6的卷积。Focus模块将特征图进行分块切片操作，然后再将结果拼接使通道扩充，可以在信息不丢失的情况下提高计算力。而将其替换为一个6 × 6的卷积可达到相同的效果，虽然两者的计算量是等价的，但对于一些GPU设备卷积的计算更易进行。

而SPP模块则被改进为SPPF[16]，SPP的作用是将任意大小的特征图通过不同尺寸的最大池化层转换成固定大小的特征向量，其结构如图3(a)所示。SPPF的结构如图3(b)所示，其效果与SPP相同，但运算时间更少。

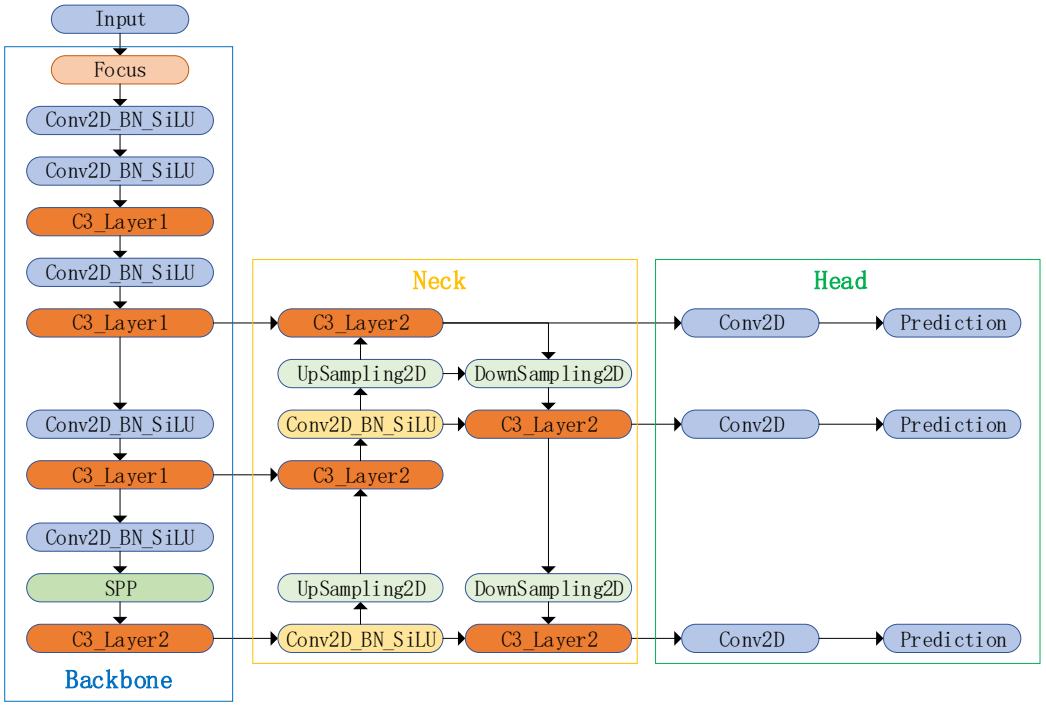


图 1　YOLOv5 网络结构

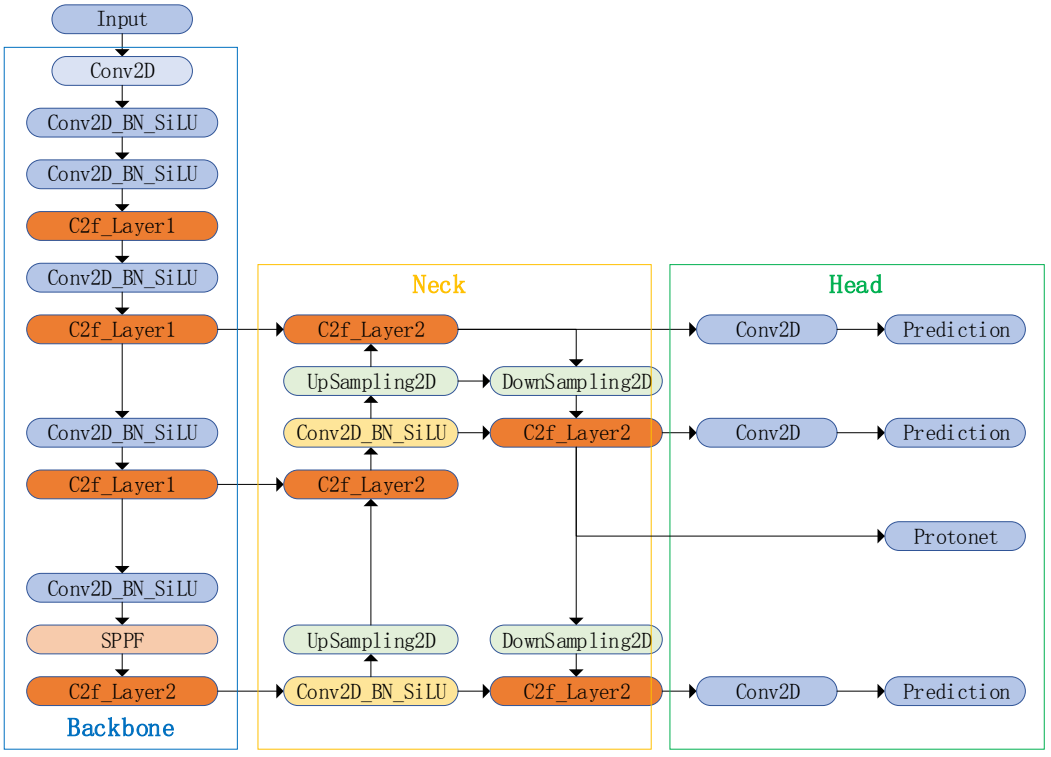
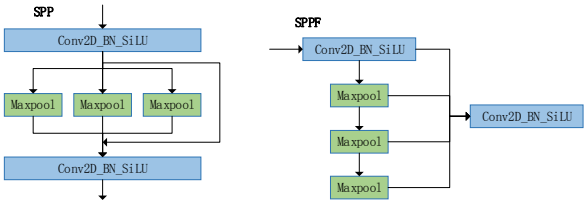


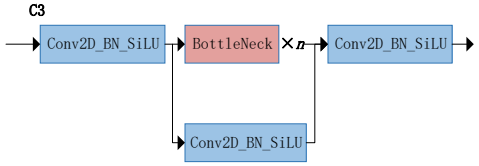
图 2　YOLOv5-seg 网络结构



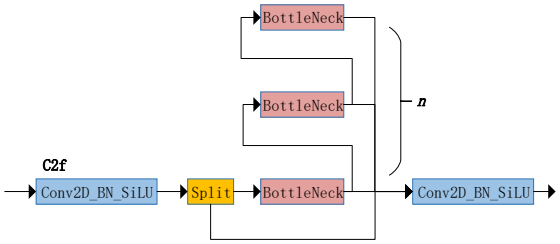
(a) SPP (b) SPPF

图 3 模块结构

YOLOv5 中的 C3 模块是基于 CSPNet 提取分流的思想[17]，同时结合 Bottlenneck 残差结构的思想而设计的，其结构如图 4(a)所示。为了进一步轻量化，并且丰富梯度流信息，在 C3 的基础上结合 ELAN[18]的思想设计了 C2f 模块，其结构如图 4(b)所示。



(a) C3



(b) C2f

图 4 模块结构

在 Head 部分，增加了一条 Protonet 作为分割的输出。算法的损失函数共有四个，分别为分类损失 cls_loss、边界框损失 box_loss、置信度损失 obj_loss 和分割损失 seg_loss。分类损失 cls_loss 用于判断模型是否能够识别出图像中的对象，并将其分类到正确的类别中；边界框损失 box_loss 用于衡量模型预测的边界框与真实边界框之间的差异，这有助于确保模型能够准确地定位对象；置信度损失 obj_loss 用于衡量模型预测的框(即包含对象的矩形)与真实框之间的差异[19-21]；分割损失 seg_loss 用于衡量模型预测的 mask 与真实的 mask 的差异。其中分类损失、定位损失和分割损失都是使用二元交叉熵损失函数[22]，其中 x 为样本，y 为标签分数，a 为预测输出分数，n 为样本总量，函数表达式如下：

$$C = -\frac{1}{n}\sum_{x}\left[y\ln a + (1-y)\ln(1-a)\right] \quad (1)$$

而置信度损失 obj_loss 则是用 Generalized Intersection over Union (GIoU)[23]来进行评估：

$$L_{\mathrm{GIoU}} = 1 - \frac{\left|B \cap B^{gt}\right|}{\left|B \cup B^{gt}\right|} + \frac{\left|C - B \cup B^{gt}\right|}{C} \quad (2)$$

其中：$B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$ 表示真实的回归框，$B = (x, y, w, h)$ 表示预测回归框，C 则为覆盖 B 和 B^{gt}，这样在不重叠的情况下，预测框也会向目标框回归。

2 电石识别分割系统

2.1 检测分割系统结构

相较于两阶段视觉算法，YOLO 模型具有相对轻量型的结构和较高的效率，基于 YOLOv5 算法我们设计了电石检测分割的方法，并开发了相应的系统。方法实现部分的主要步骤如图 5 所示。

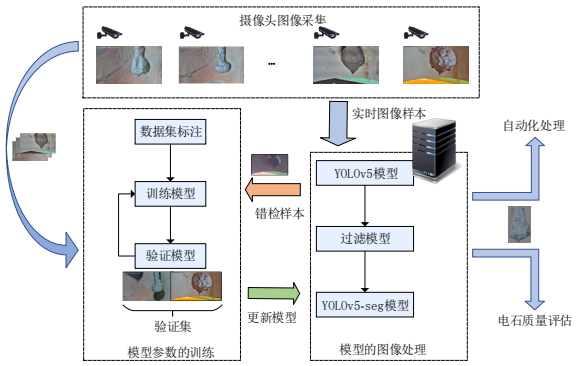
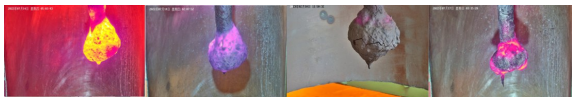


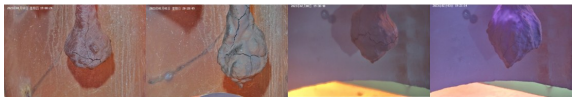
图 5 电石检测分割的主要步骤

由于我们只需要电石的分割输出，而检测算法的运算速度明显要高于分割算法，在训练集标注方面检测算法的标注成本也明显低于分割算法，所以我们先使用检测模型将电石检测出来，再传入分割模型进行分割，以达到节省算力的目的。同时针对电石棒与电石易混淆的问题，为了增加系统的容错率，在检测模型和分割模型之间添加了一个图像的二分类模型，对电石和电石棒进行区分。



（a）未冷却的电石（紫光、红光）

（b）受强日光影响的电石



（c）受不同炉光影响的电石（紫光、红光、蓝光）

（d）电石棒

图6　电石情况

使用的是YOLOv5-l模型进行目标检测，主要的检测目的是区分当前检测区域内是否有电石存在。根据实际场景的情况可以分为“空状态”“电石棒”和“电石”三种。为了减少模型的误检情况，在实际的实时检测中，系统设定连续15次“空状态”则判定为空，连续5次“电石”则判定为电石。系统将判定为电石的区域沿检测框裁剪出来，发送到分割模型进行下一步操作。由于在工业场景中干扰因素较多，经常会出现背景板污染、强光干扰、电石未冷却等因素，影响分割后图片的进一步使用。所以需要提前将这些情况排除，将误检或者效果不佳的样本收集起来，重新标注并训练和更新模型。

分割模型使用的是YOLOv5-seg-m模型，直接将检测模型传过来的电石图片进行分割，并将效果不佳的样本收集起来，进行标注并训练和更新模型，此模型的更新侧重于负例的更新，以便有效地排除受到干扰的情况。

2.2　模型的压缩

为了减少参数储存与内存占用空间，将模型进行参数量化操作。将32位浮点（FP32）网络参数量化到16位浮点数（FP16）或者8位整型（INT8），从而节省带宽加速运算，降低设备能耗。

而转换数值的方法则会显著影响预测的精度。以FP32到INT8为例，量化的映射方式主要可分为对称量化和非对称量化两种[24]。

对称量化是将FP32参数的最大值的正负区间投影在INT8的正负区间内$[-128,127]$，$[-\max_{F32},\max_{F32}]\rightarrow[-\max_q,\max_q]$，$x_f$为FP32下的参数，$x_q$为转换在INT8下的新参数，其转换关系如下：

$$x_q = round\left(x_f\frac{2^{n-1}-1}{\max|x_f|}\right) \tag{3}$$

非对称量化[25]是将FP32参数的最大值到最小值的区间投影在INT8的区间内$[0,255]$，$[\min_{F32},\max_{F32}]\rightarrow[\min_q,\max_q]$，$x_f$为FP32下的参数，$x_q$为转换在INT8下的新参数，其转换关系如下：

$$x_q = round\left(\left(x_f-\min_{x_f}\right)\frac{2^{n-1}-1}{\max_{x_f}-\min_{x_f}}\right) \tag{4}$$

3　实验方案和结果

本次实验主要是针对分割算法（YOLOv5-seg）的横向和纵向的比较。训练所使用的数据集是自制的电石数据集，使用labelme标注样本数为10000，样本中有摄像头采集的3840 × 2160原图，也有经过YOLOv5分割后的不定大小的样本。为了后期方便部署，YOLOv5-seg使用的权重是.onnx格式的。

实验中常采用召回率（Recall）、精确率（Precision）、mAP等评价指标[26]。首先表1给出二分类性能指标TP、FP、FN、TN的定义。

表1　TP、FP、FN、TN的定义

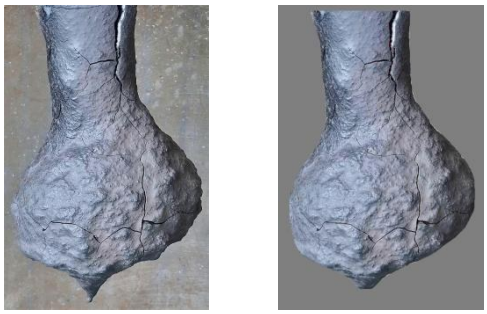
	预测1	预测0
实际1（P）	TP	FN
实际0（N）	FP	TN

T/*F*表示预测是否正确，*P*/*N*表示预测的标签。召回率的表达式为$Recall=\frac{TP}{TP+FN}$，精确率的表达式为$Precision=\frac{TP}{TP+FP}$。算法以均值平均精度（mean average precision，mAP）作为评价指标，P_{mAP}的计算表达式为：

$$P_{\text{mAP}} = \sum_{n=1}^{N} P(n)\Delta P(n) \quad (5)$$

YOLOv5-seg的不同模型和不同输入尺寸进行对比，输入尺寸分别为640 × 640和960 × 960；三种模型YOLOv5-seg_l、YOLOv5-seg_m和YOLOv5-seg_s，推理时使用的硬件配置GPU为1050 T，训练轮数为40轮，最佳参数的轮次集中在15～25轮。其中评估的参数为Recall、Precision、mAP、Predict_t、cls_loss、box_loss、obj_loss和seg_loss，表2为评估参数。

这些参数中参考价值最大的是mAP，可以较好地衡量出分割的效果；box_loss和obj_loss是描述检测框的效果的参数；综合参考mAP、损失函数、Recall、Precision和推理时间，可以看出640 × 640输入的模型的函数收敛情况不如960 × 960的收敛情况，但差距并不明显，从mAP、Recall、Precision上来看640 × 640输入的模型的效果明显要好于960 × 960输入的模型，所以我们选择640 × 640输入的模型。主要考虑到分割效果和推理时间，我们最后选择m_640模型。



（a）分割前 （b）分割后

图7 对比结果

表2 YOLOv5-seg不同模型和尺寸的评估参数

	Precision	recall	mAP	box_loss	seg_loss	obj_loss	cls_loss	Predict_t/ms
l_960	0.9324	0.9356	0.9575	0.01618	0.00638	0.00755	0.00016	255
m_960	0.9958	0.7898	0.9108	0.01840	0.00698	0.00826	0.00025	150
s_960	0.9869	0.7864	0.9216	0.01557	0.00640	0.00792	0.00011	77
l_640	0.9929	0.9488	0.9890	0.01797	0.00729	0.00791	0.00017	125
m_640	0.9825	0.9536	0.9905	0.01776	0.00736	0.00804	0.00016	75
s_640	0.9326	0.8203	0.9604	0.01705	0.00685	0.00765	0.00010	45

YOLOv5-seg与Mask R-CNN算法进行对比，使用相同的训练集和验证集进行训练。硬件配置GPU为NVIDIA T4，训练轮次为60轮。首先，在训练过程中，YOLOv5-seg在单轮的训练时间要比Mask R-CNN更短，收敛速度也更快；但从函数的收敛情况来看，Mask R-CNN的收敛更加稳定。从模型效果来看，二者效果基本达到一致，但在推理时间上，Mask R-CNN所需时间接近1 s，而YOLOv5-seg只需要45 ms。

4 结语

电石的检测和分割具有重要的研究与应用价值。但是，电石检测和分割存在着业务场景复杂、电石与电石棒易混淆、检测和分割精度要求高等问题。本文设计了面向实际业务场景的电石检测分割的方案与方法，在实际场景数据采集、电石状态与干扰状态的定义、数据标注等方面做出了重要贡献。提出的基于YOLOv5-l和YOLOv5-seg-m网络的电石检测分割方法在检测精度与速度上取得了较好的结果，对于电石的检测和分割的判识基本达到了应用的需求，且检测和分割的速度远超同类型的算法。之后将针对排除光照影响、提高分割精度和后续评判电石质量模型的方向继续深入研究。

参考文献：

[1] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[2] GIRSHICK R. Fast R-CNN[C]//Proceedings of

the 2015 IEEE International Conference on Computer Vision(ICCV),Santiago,Chile,2015:1440-1448.

[3] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: an evaluation of the state of the art [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012:34(4):743-761.

[4] 钟伟镇,刘鑫磊,杨坤龙,等. 基于 Mask-RCNN 的复杂背景下多目标叶片的分割和识别[J]. 浙江农业学报,2020,32(11):2059-2066.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016:779-788.

[6] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),Honolulu,HI,2017:6517-6525.

[7] WANG W S, WANG S, GUO Y C, et al. Obstacle detection method of unmanned electric locomotive in coal mine based on YOLOv3-4L [J]. Journal of Electronic Imaging,2022,31(2):23032.

[8] ZHENG Z Y, ZHANG T, LIU Z Y, et al. Arbitrarily oriented object detection in remote sensing images based on improved YOLOv4-CSP [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2022,15(1):9355-9368.

[9] 邵延华,张铎,楚红雨,等. 基于深度学习的YOLO目标检测综述[J]. 电子与信息学报,2022,44(10):3697-3708.

[10] 张慧春,张萌,边黎明,等. 基于YOLOv5的植物叶绿素含量估测与可视化技术[J/OL].农业机械学报:1-14[2023-05-23]. http://kns.cnki.net/kcms/detail/11.1964.S.20220224.0912.002.html.

[11] 王玲敏,段军,辛立伟. 引入注意力机制的YOLOv5安全帽佩戴检测方法[J]. 计算机工程与应用,2022,58(9):303-312.

[12] 尹靖涵,瞿绍军,姚泽楷,等. 基于YOLOv5的雾霾天气下交通标志识别模型[J]. 计算机应用,2022,42(9):2876-2884.

[13] YANG X, YAN J C. Arbitrary-oriented object detection with circular smooth label[C]//Proceedings of the European Conference on Computer Vision (ECCV),2020:677-694.

[14] 徐凤如,张昆明,张武,等. 一种基于改进YOLOv4算法的茶树芽叶采摘点识别及定位方法[J]. 复旦学报(自然科学版),2022,61(4):460-471.

[15] 孙泽强,陈炳才,崔晓博,等. 融合频域注意力机制和解耦头的YOLOv5带钢表面缺陷检测[J]. 计算机应用,2023,43(1):242-249.

[16] HADDAD Y, MARRACHE M, MELLOUL S, et al. Session peering provisioning framework analysis [C] //Proceedings of the International Conference on Software Science, Technology, and Engineering (SwSTE2014),2014.

[17] 任小康,刘行行. 基于改进的YOLOv3口罩佩戴检测和识别[J]. 计算机工程与科学,2022,44(10):1812-1821.

[18] WITTENBURG P, BRUGMAN H, RUSSEL A, et al. ELAN: a professional framework for multimodality research [C] //Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006),2006.

[19] 王一旭,肖小玲,王鹏飞,等. 改进YOLOv5s的小目标烟雾火焰检测算法[J]. 计算机工程与应用,2023,59(1):72-81.

[20] 马琳琳,马建新,韩佳芳,等. 基于YOLOv5s目标检测算法的研究[J]. 电脑知识与技术,2021,17(23):100-103.

[21] ZHANG R, WEN C B. SOD-YOLO: a small target defect detection algorithm for wind turbine blades based on improved YOLOv5 [J]. Advanced Theory and Simulations,2022,5(7):2100631.

[22] 杨宁,金生. 使用深度学习网络U-Net进行道路提取的研究与讨论[J]. 黑龙江大学工程学报,2020,11(4):1-8.

[23] 侯志强,刘晓义,余旺盛,等. 使用GIoU改进非极大值抑制的目标检测算法[J]. 电子学报,2021,49(4):696-705.

[24] 孙浩然,王伟,陈海宝. 基于参数量化的轻量级图像压缩神经网络研究[J]. 信息技术,2020,44(10):87-91.

[25] 杨颖振,赵志彪,赵宝新,等. 用于神经网络的压缩和推断加速的非对称量化:CN112085154A[P]. 2020.

[26] 徐小茹. 基于YOLO目标分割的船舶识别算法研究[J]. 梧州学院学报,2019,29(6):1-9.

(下转第14页)

文章编号: 1007-1423（2023）16-0008-07　　DOI: 10.3969/j.issn.1007-1423.2023.16.002

基于改进U-Net模型的近海水产养殖池塘信息提取

陈宇杨[1]，张　丽[2,3,4*]，陈博伟[2,3,4]，邱玉宝[2,3,4]

（1. 桂林电子科技大学信息与通信学院，桂林 541004；
2. 中国科学院空天信息创新研究院，中国科学院数字地球重点实验室，北京 100094；
3. 中国-东盟地球大数据区域创新中心，南宁 530022；4. 可持续发展大数据国际研究中心，北京 100094）

摘要：针对传统卷积神经网络算法在复杂环境背景下泛化能力弱、提取精度低的问题，提出了一种快速、准确的近海养殖池塘自动提取模型。该模型在U-Net模型的基础上进行改进，用新提出的多尺度特征提取结构(DC)取代了U-Net模型的传统卷积层。其中，DC结构融合了Inception模块和空洞残差模块，增强了模型提取养殖池塘特征的能力，降低模型训练过拟合风险，有效减少图像背景信息干扰。实验结果表明，改进U-Net模型的精确率、召回率、交并比、F1分数分别为91.73%、90.47%、91.12%、89.91%，优于其他的对比模型。该研究能够有效提高养殖池塘提取精度，实现养殖池塘快速、准确的监测。

关键词：养殖池塘；DC结构；Inception模块；改进U-Net模型

0　引言

据联合国粮食及农业组织(FAO)统计数据显示，全球水产养殖业为超过33亿人提供了人均动物蛋白摄入量的20%。2018年，全球捕捞渔业产量达到创记录的9640万吨，比前三年的平均值增加5.4%[1]。因此，水产养殖作为人类所需蛋白质的重要来源之一，在世界粮食安全中发挥着关键作用。从20世纪中叶开始，中国沿海水产养殖区迅速扩张，导致周围土地利用竞争激烈，在产生巨大经济效益的同时也对沿岸生态环境造成了一定的破坏，如近岸湿地的萎缩或消失、自然栖息地的破坏、生物多样性的丧失以及海岸带地区生态修复能力衰退[2-4]。因此，准确、快速地了解近岸地区养殖池塘的空间分布，对水产养殖的空间布局优化、自然资源的科学管理以及生态环境的保护至关重要。

卫星遥感技术的快速发展，为实现近岸养殖池塘准确、高效的提取提供了可靠的数据支持。它能够克服传统野外调查的不足，是监测和研究海岸带生态环境的重要手段[5-6]。目前，基于遥感影像提取水产养殖区的方法主要有基于像素、面向对象的图像分析、机器学习算法以及深度学习算法。其中，基于传统的基于像素、面向对象的图像分析，在某一特定区域中表现出了良好的识别能力，如马艳娟等[7]构建水体指数和波段运算函数用于近海水产养殖区的提取；吕巷艳等[8]采用归一化差异水体指数结合阈值的方法提取金沙县养殖水体信息；裴亮等[9]将归一化差异池塘指数与面向对象方法相结合，对天津海岸养殖区进行提取。但在面对复杂水域，出现“同谱异物”现象时，仅依靠光谱特征差异性容易导致错分，且其存在无法克服的“椒盐”噪声[10]。而面向对象的分类方法，虽然可以有效抑制“椒盐”噪声，但它

收稿日期： 2023-05-04　**修稿日期：** 2023-07-31
基金项目： 广西创新驱动发展专项资金项目(AA20302022)
作者简介： 陈宇杨(1998—)，男，广西贵港人，硕士研究生，研究方向为海岸带遥感技术与应用；*通信作者：张丽(1975—)，女，新疆伊犁人，博士，研究员，主要从事植被生态和海岸带遥感研究，E-mail：zhangli@aircas.ac.cn；陈博伟(1990—)，男，陕西汉中人，博士，助理研究员，主要从事新型传感器遥感机理和应用方面的研究；邱玉宝(1978—)，男，江西兴国人，博士，研究员，主要从事被动微波遥感、冰雪及北极环境遥感及数据科学研究

的分割参数选取往往取决于研究人员的经验知识且需要反复实验[11]。

机器学习算法虽然已被研究人员广泛应用于遥感图像识别与分类领域并取得不错的成效，但其存在特征选择上需要研究人员具备专业知识、算法结构较浅难以提高分类精度、自动化程度难以满足需求等问题[12]。近年来，深度学习技术在计算机视觉领域取得了巨大成功，相关遥感学者也开始将其应用于遥感图像语义分割领域，为实现近海水产养殖区识别开辟了另一条途径。如Cheng等[13]通过扩大感受野来增强网络模型学习高层特征的能力，提高了笼式养殖区和筏式养殖区的提取精度；Fu等[14]采用层次级联结构并引入注意力机制来优化特征空间，对不同种类的海水养殖区进行区分；苟杰松等[15]基于Deeplabv3+构建水产养殖水体语义分割模型；Sui等[16]采用ASPP结构，能够有效捕捉到海上筏式养殖区的多尺度信息，提高浮筏养殖区的提取精度。此类方法能够克服传统方法出现的“同谱异物”现象，改善了水产养殖区提取的边缘模糊问题。

基于深度学习方法提取水产养殖区具有准确、高效等特点。然而，上述的研究中大多是对海上的浮筏养殖区进行提取，相对于养殖池塘的自然地理环境背景更为单一，浅层的卷积神经网络模型在复杂背景情况下提取养殖池塘容易产生冗余信息，网络模型的泛化能力和鲁棒性表现不足。为提升遥感语义分割模型对养殖池塘的识别率，本文提出了一种基于U-Net改进的近岸养殖池塘自动提取模型，并与FCN[17]、E-Net[18]和U-Net[19]等经典语义分割模型进行比较，验证改进U-Net模型的有效性，以期为该领域相关研究的进一步发展提供参考价值。

1　研究区和数据源

本文的研究区为广西北部湾的官寨海和安铺港两个典型的规模化集聚型养殖区。如图1所示，地理坐标范围为109°49′～109°56′E，21°25′～21°30′N。研究区内养殖池塘主要位于海涂等低洼地区，一般是封闭的水体，形状近似矩形。研究区域内河流、水田、建设用地等多种不同的地物类型广泛分布，对养殖池塘提取任务造成一定挑战。

数据源采用欧空局提供的Sentinel-2多光谱扫描成像(multispectral scan imaging)遥感数据。Sentinel-2任务包含了2A、2B两颗同时运作的相同卫星，重访周期为5天，扫描幅宽为290 km，空间分辨率为10 m。首先在Google Earth Engine平台上完成影像的数据预处理工作，得到覆盖研究区域2020年6月的Sentinel-2 MSI(Level-2A)影像数据，包含红、绿、蓝、近红外4个波段，像素大小为7696×7059。然后利用ArcGIS软件对经过预处理的影像进行养殖池塘的标注，将其Id字段值设置为255。为防止计算机内存溢出，采用随机裁剪方法对影像进行裁剪，最终得到2100张像素大小为256×256的图像，按照8∶2比例制作为训练集和验证集。

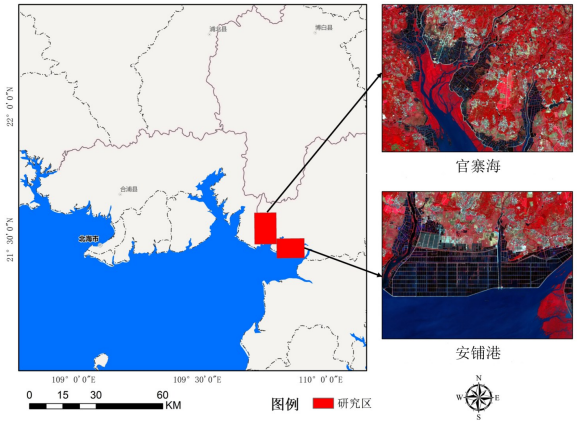


图1　研究区

2　研究方法

由于遥感图像的复杂环境背景干扰，目前的一些算法难以对养殖池塘进行准确的分割，因此本文对U-Net模型进行改进，得到更加适用于养殖塘分割的新算法。

2.1　改进U-Net模型

U-Net是一种基于FCN的改进型模型。如图2所示，它是一个对称的网络结构，包含了左侧编码器、右侧解码器以及中间的跳跃连接。其中，编码部分用于图像的特征提取，解码部分则逐步将编码得到的特征图恢复到原始图像大小。因此，整个网络的性能直接受编码端的特征提取能力和解码端的图像恢复能力的影响。

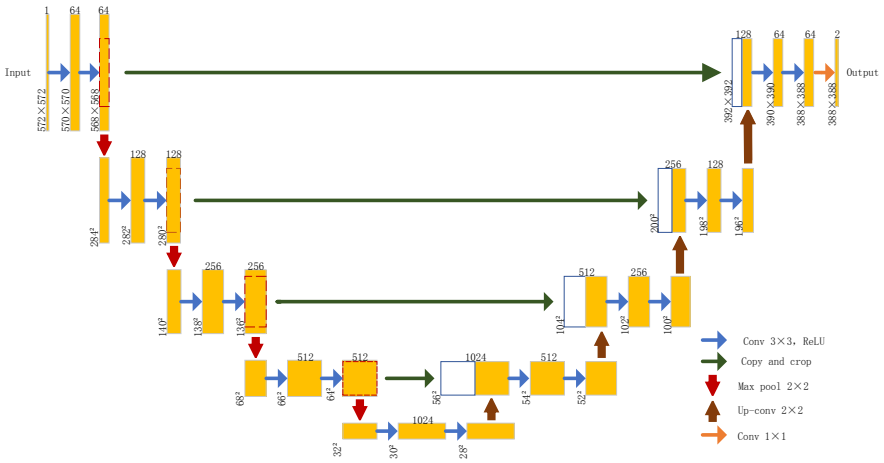


图 2　U-Net模型

在本研究中，U-Net模型是一个重要的结构和关键技术，它能够有效地进行图像语义分割任务，提高图像处理的准确性和效率。

本文提出的改进U-Net模型主要是在U-Net模型基础上进行改进，将提出的DC结构代替U-Net的传统卷积层，使模型具有更强的特征提取能力。改进U-Net模型包括编码部分、解码部分以及跳跃连接，采用的操作包括卷积操作（DC结构）、上采样（Up-Sampling）、最大池化（Max-Pooling）、跳跃连接（skip connection），其结构如图3所示。

原始U-Net网络的编码部分：编码部分可划分为5级，每级都是由2个3×3卷积、1个ReLU激活函数、1个最大池化层组成。而改进U-Net模型中，编码部分也分为5级，每级均由DC结构组成。DC结构主要是由Inception模块、空洞残差结构以及跳跃连接构成，它使得模型能够更好地对图像特征进行多尺度提取，保证网络模型对特征的稳定表达。并且，在每一个卷积层后面用ReLU层加快模型的收敛速度以及使用BN层缓解网络的梯度消失问题。最后，在DC结构的后面通过最大池化操作进行降采样。

原始U-Net网络的解码部分：解码部分也可划分为5级，前4级都是由上采样、3×3卷积组成，第5级为输出层。在改进U-Net模型中，解码部分全部使用DC结构代替原来的卷积层，同时，与原来的U-Net网络一样，仍然使用1×1卷积将最后一层的输出特征向量映射为一个类标签。

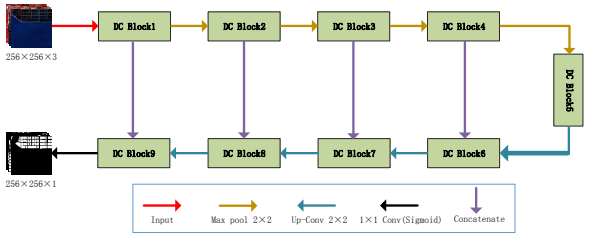


图 3　改进U-Net模型

2.2　DC结构

为了更准确地识别遥感影像中大小不同的养殖池塘，本文提出了一种多尺度特征提取结构（DC结构）。该结构由Inception[20]模块、空洞残差模块和跳跃连接组成，其结构如图4所示。DC结构能够有效提取养殖池塘的特征信息，加强特征表达能力，提升遥感影像中养殖池塘部分的识别精度和效率。

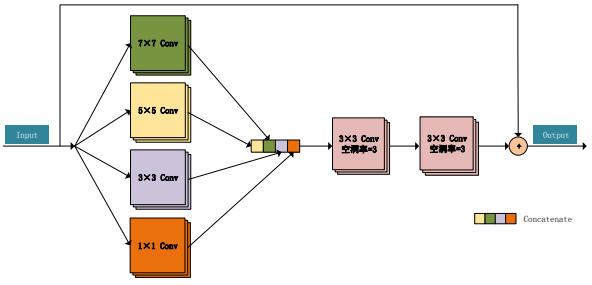


图 4　DC结构

如图4所示，首先，输入特征经过Inception模块的四个分支，每个分支分别进行1×1卷积、

3×3卷积、5×5卷积、7×7卷积，捕捉到不同尺度的图像特征，然后通过Concatenate拼接操作进行特征融合并输入到空洞率为3的空洞残差模块中，进一步对融合后的信息进行提取，最后通过一个长跳跃连接将输入和输出进行融合。DC结构能够提取不同尺度的特征信息，增强了模型对特征的复用和表达的准确性，降低模型训练过拟合风险，有效地提升了识别精度。

3　实验与分析

3.1　实验设置

该实验使用的仿真平台为PyCharm，使用keras及其TensorFlow端口，运用Python语言编程。计算机配置为Intel(R)CoreTM i7-10700F CPU@2.90 GHz，64.0 GB内存，NVIDIA GeForce GTX 3060，采用64位操作系统Windows10，Python3.6。为了客观地分析本文中养殖池塘提取模型的性能，实验中模型统一使用了Adam优化器、初始学习率为0.0001、批量大小设置为8、迭代周期为80次。

3.2　评价指标

为综合评估改进U-Net模型的有效性，实验使用了四种精度评估指标，包括精确率(Precision)、召回率(Recall)、交并比(Intersection-over-Union，IoU)和F1分数(F1-Score)。其中，精确率表示模型预测为正例的样本中，实际为正例的样本所占的比例；Recall表示实际为正例的样本中，模型正确预测为正例的样本所占的比例；IoU表示模型预测的边界框与真实边界框的重叠程度；F1-score是精确率和召回率的调和平均数，用于综合评估模型的性能。以上评价指标的计算公式如式(1)～(4)所示。

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

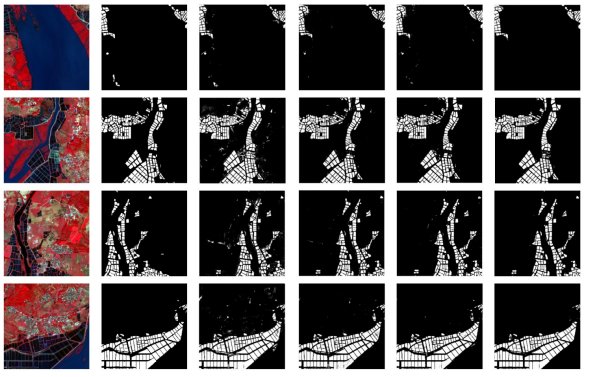
$$IoU = \frac{TP}{TP + FN + FP} \tag{3}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

其中，真正例(TP)表示模型预测结果和实际情况都为正例；假正例(FP)表示实际情况为反例，但是预测结果为正例；假反例(FN)表示实际情况为正例，但是预测结果为反例；真反例(TN)表示实际情况和预测结果都为反例。

3.3　消融实验

为了评估本文提出的改进后的U-Net模型的优越性，并验证该模型中每个模块的有效性，本文进行了消融实验。该实验都在同一台机器上进行，使用相同的数据集进行训练和验证，并采用相同的参数设置，以确保消融实验的结果可靠。其中，①U-Net：是没有任何改进的原始U-Net模型；②U-Net_1：在U-Net基础上加入空洞残差模块；③U-Net_2：在U-Net基础上加入Inception模块；④U-Net_3：本文所提出的改进U-Net模型，它加入了本文所提出的DC结构。消融实验结果和精度对比如图5和表1所示。



(a)图像　(b)标签　(c)U-Net模型　(d)U-Net_1模型 (e)U-Net_2模型 (f)U-Net_3模型

图5　消融实验结果

表1　消融实验精度对比

模型	Precision/%	Recall/%	IoU/%	F1-Score/%
U-Net	87.38	85.58	84.04	87.31
U-Net_1	90.78	89.01	87.82	89.62
U-Net_2	90.44	88.96	89.73	90.96
U-Net_3	91.73	90.47	91.12	89.91

在消融实验中，U-Net_1模型采用了空洞残差模块，相比原始的U-Net模型，它在精确率、召回率、交并比以及F1分数上分别提高了3.4、3.43、3.78和2.31个百分点。这是因为空洞残差模块可以增加模型的感受野，从而捕捉更大范围内的特征信息。U-Net_2模型采用了Inception模块，它比U-Net模型的精确率、召回率和交并比以及F1分数上分别提高了3.06、3.38、5.69

和3.65个百分点，这是因为Inception模块能够增加模型的宽度，有效提取到养殖池塘的多尺度特征信息。因此，采用空洞残差模块的U-Net_1模型和采用Inception模块的U-Net_2模型获得更高的精度。

与U-Net_1模型和U-Net_2模型相比，具有DC结构的U-Net_3模型在精确率、召回率以及交并比上得到了进一步的提升，说明了DC结构结合Inception模块的多分支结构以及空洞残差模块，既能增加网络模型的宽度，又能增大卷积提取的感受野，同时两者的结合减少了空洞卷积在卷积过程中带来的稀疏性。所以，DC结构能够提取到更多有效的不同尺度的特征信息，从而更好地辨别养殖池塘的轮廓。

3.4 不同方法的实验比较

上述的消融实验结果表明，改进U-Net模型在养殖池塘信息的提取上优于原始U-Net模型。为了进一步验证改进U-Net模型的分割性能，将改进U-Net模型与FCN、E-Net、U-Net模型进行实验对比。其不同模型的分割结果和精度对比分别如图6和表2所示。

从图6的分割结果中可以看出，采用FCN模型的分割结果明显差于其他几种模型，分割结果噪声较多，并不能清晰地识别出养殖池塘，这是因为FCN模型在特征提取过程中采用了连续下采样操作，通过不断降低特征图的分辨率来扩大感受野。然而，这种下采样操作会导致分辨率的下降，从而造成一定的信息损失。同时FCN模型在上采样造成的锯齿状边缘也会影响分割结果的质量。因此，在分割结果中可能会出现一些细节缺失。E-Net模型在大型堤防上的分割性能有所提升，是因为其采用了多尺度特征提取的方法，通过引入不同大小的卷积核和不同步长的卷积操作，对输入图像进行多尺度的特征提取，从而提高了模型对图像的感知能力和分割效果，但是仍然存在部分细小堤防与养殖池塘直接相连的问题。相比之下，U-Net模型能够有效消除堤防的影响，提取的养殖池塘边缘更加清晰，这得益于U-Net模型在上采样部分与跳跃连接相结合，逐层融合高层特征和低层特征，填充了一些局部细节，但在复杂场景下，与养殖池塘光谱特征相似的海水、湖泊、水库、河流等仍被错误归类，这是因为养殖池塘的形状和大小不一，而且存在大量的细节信息。U-Net模型在处理这些复杂目标时会出现过度拟合或欠拟合的问题，导致分割精度下降。而本文所提出的改进U-Net模型结合了DC结构，降低模型训练中的过拟合风险，充分考虑了养殖池塘的多尺度信息，能够更好地捕捉图像中的细节特征，增强了图像有用信息的表达。从表2也能明显地看到，改进U-Net模型的精确率、召回率、交并比和F1分数分别为91.73%、90.47%、91.12%和89.91%，要明显优于FCN、E-Net、U-Net三种模型。综上，本文的方法能够解决其他三种模型在提取养殖池塘过程中存在的问题，实现养殖池塘快速、准确的提取。

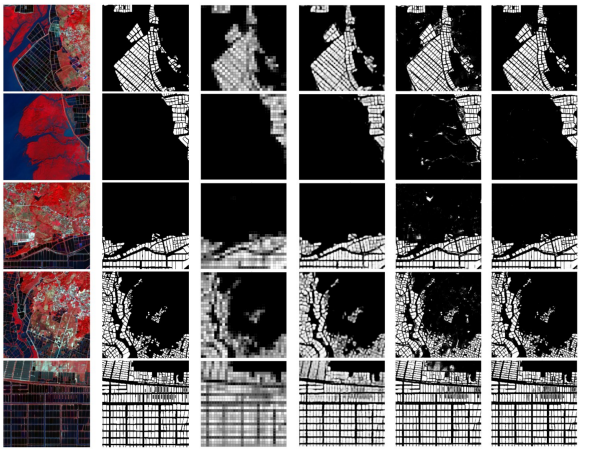


图6 不同模型的养殖池塘提取结果

表2 不同模型的提取结果精度对比

模型	Precision/%	Recall/%	IoU/%	F1-Score/%
FCN	67.13	63.49	60.77	55.97
E-Net	80.66	78.64	76.65	74.19
U-Net	87.38	85.58	84.04	87.31
改进U-Net	91.73	90.47	91.12	89.91

4 结语

本文基于U-Net模型改进提出了一个近海水产养殖池塘自动提取模型，并用本文提出的新DC结构替代了U-Net模型的卷积层，它能够增强模型对图像特征的表达能力，降低冗余信息的干扰。实验结果表明，改进U-Net模型在精确

率、召回率、交并比和F1分数上都优于FCN、E-Net、U-Net这三个模型，其中，精确率、召回率、交并比都达到90%以上，并且改进U-Net模型提取的养殖池塘边缘更为清晰。因此，本文的方法能够有效提高养殖池塘提取精度，为遥感图像自动分割提供了参考意义。

今后，我们将在以下两个重要方向继续开展研究：①将探索该模型在光谱信息极弱的遥感图像上对养殖池塘的提取效果；②利用不同来源、不同分辨率的遥感数据集，研究模型对不同数据集的有效性。

参考文献：

[1] FAO. The state of world fisheries and aquaculture 2020: sustainability in action [M]. Rome, Italy: [s. n.], 2020.

[2] 陈一波，宋国宝，赵文星，等. 中国海水养殖污染负荷估算[J]. 海洋环境科学，2016，35(1)：1-6，12.

[3] CAO L, WANG W, YANG Y, et al. Environmental impact of aquaculture and countermeasures to aquaculture pollution in China[J]. Environmental Science and Pollution Research International, 2007, 14(7): 452-462.

[4] 熊超，袁俊雄. 广西北部湾近岸海域保护应坚持陆海统筹[J]. 环境经济，2020(19)：58-63.

[5] 文可，姚焕玫，黄以，等. 基于GEE的广西北部湾沿海水产养殖池塘遥感提取[J]. 农业工程学报，2021，37(12)：280-288.

[6] LIU W, LIU S, ZHAO J, et al. A remote sensing data management system for sea area usage management in China[J]. Ocean & Coastal Management, 2018, 152: 163-174.

[7] 马艳娟，赵冬玲，王瑞梅，等. 基于ASTER数据的近海水产养殖区提取方法[J]. 农业工程学报，2010，26(S2)：120-124.

[8] 吕巷艳，任金铜，欧阳力剑，等. 基于GF-1的喀斯特山区养殖水体信息提取：以金沙县为例[J]. 贵州工程应用技术学院学报，2020，38(3)：114-118.

[9] 裴亮，王金鑫，屈慧慧，等. 基于ONDPI的海岸养殖池塘遥感影像提取研究[J]. 海洋测绘，2020，40(5)：40-44.

[10] 王芳，夏丽华，陈智斌，等. 基于关联规则面向对象的海岸带海水养殖模式遥感识别[J]. 农业工程学报，2018，34(12)：210-217.

[11] 朱程. 基于中等分辨率遥感图像的海岸带养殖池提取算法[D]. 大连：大连海事大学，2020.

[12] 邵振峰，孙悦鸣，席江波，等. 智能优化学习的高空间分辨率遥感影像语义分割[J]. 武汉大学学报(信息科学版)，2022，47(2)：234-241.

[13] CHENG B, LIANG C, LIU X, et al. Research on a novel extraction method using Deep Learning based on GF-2 images for aquaculture areas [J]. International Journal of Remote Sensing, 2020, 41(9): 3575-3591.

[14] FU Y, YE Z, DENG J, et al. Finer resolution mapping of marine aquaculture areas using worldView-2 imagery and a hierarchical cascade convolutional neural network[J]. Remote Sensing, 2019, 11(14): 1678.

[15] 苟杰松，蒋怡，李宗南，等. 基于Deeplabv3+模型的成都平原水产养殖水体信息提取[J]. 中国农机化学报，2021，42(3)：105-112.

[16] SUI B, JIANG T, ZHANG Z, et al. A modeling method for automatic extraction of offshore aquaculture zones based on semantic segmentation [J]. ISPRS International Journal of Geo-Information, 2020, 9(3): 145.

[17] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[18] PASZKE A, CHAURASIA A, KIM S, et al. ENet: a deep neural network architecture for real-time semantic segmentation[EB/OL]. arXiv: 1606.02147, 2016.

[19] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation [C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 2015: 234-241.

[20] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818-2826.

Information extraction from offshore aquaculture ponds based on improved U-Net model

Chen Yuyang[1], Zhang Li[2,3,4*], Chen Bowei[2,3,4], Qiu Yubao[2,3,-4]

(1. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China;
2. Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;
3. China-ASEAN Regional Innovation Center for Big Earth Data, Nanning 530022, China;
4. International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China)

Abstract: A fast and accurate model for automatic extraction of offshore aquaculture ponds is proposed to address the problems of weak generalisation and low extraction accuracy of traditional convolutional neural network algorithms in complex environmental contexts. The model improves on the U-Net model by replacing the traditional convolutional layers of the U-Net model with a newly proposed multi-scale feature extraction structure(DC). The DC structure incorporates the Inception module and the null residual module to enhance the ability of the model to extract farm pond features, reduce the risk of overfitting the model training and effectively reduce the interference of image background information. The experimental results show that the accuracy, recall, cross-merge ratio and F1 score of the improved U-Net model are 91.73%, 90.47%, 91.12% and 89.91% respectively, which are better than the other comparison models. This study can effectively improve the extraction accuracy of farming ponds and achieve fast and accurate monitoring of farming ponds.

Keywords: aquaculture ponds; DC structure; Inception module; improved U-Net model

(上接第7页)

Multi-model calcium carbide detection and segmentation system based on YOLOv5-seg

Hao Junfeng[1], Li Yutao[2*], Lai Bowen[1]

(1. Artificial Intelligence Studio, HBIS Digital Technology Co., Ltd., Shijiazhuang 050035, China;
2. Technical Center Headquarter, HBIS Digital Technology Co., Ltd., Shijiazhuang 050035, China)

Abstract: A method for detecting and segmenting calcium carbide using YOLO series models was designed and a corresponding system was developed for practical carbide industry scenarios. First, a large-scale dataset was constructed by collecting a large number of image samples and annotating them, and to save computing power and improve accuracy, detection and segmentation were separated into two modules using the YOLOv5-l and YOLOV5-seg-m models respectively. In experiments, the internal structure of the models was improved to increase the computational speed, and the model was parameterized to further save bandwidth, storage, and computation space. Compared to the Mask R-CNN algorithm, the speed was greatly improved with similar segmentation results. The task of detecting and segmenting calcium carbide was successfully completed, laying the foundation for subsequent production automation and unmanned operation. Further research will be conducted to eliminate the influence of lighting, improve segmentation accuracy, and evaluate carbide quality models.

Keywords: calcium carbide; YOLOv5-seg; CSPNet; parameter quantification

文章编号: 1007-1423(2023)16-0015-07 DOI: 10.3969/j.issn.1007-1423.2023.16.003

基于频域Transformer的对抗生成网络去运动模糊算法

顾军华[1,2]，李 岩[1]，陈 晨[1]，牛炳鑫[1*]，李春杰[3]

（1. 河北工业大学人工智能与数据科学学院，天津 300401；
2. 河北省大数据计算重点实验室(河北工业大学)，天津 300401；
3. 河北高速公路集团有限公司，石家庄 050000）

摘要：对于图像不均匀的运动模糊问题，模型感受野的大小相当重要，基于自注意力机制的Transformer的感受野远大于传统CNN模型。并且引入了高效的频域处理模块，以往结合频域处理模糊的方法忽略了运动模糊在频域图中呈现出来的方向性特征，根据这种特征将其分解处理可以高效地恢复图像。实验中与目前较经典的几个方法进行了对比，模型在GoPro数据集上PSNR得分为32.24 dB，在RealBlur-J数据集上的得分为29.38 dB，效果相较于其他方法在两个数据集上分别提高了0.72 dB和0.83 dB。

关键词：图像去模糊；傅里叶变换；频域；Transformer；生成对抗网络

0 引言

图像去模糊方法一般是尝试通过去除模糊伪影来将模糊图片恢复成清晰图像。图像中的模糊现象受到许多因素的影响，如快速运动的物体、失焦、抖动等。由于图像去模糊是一种不适定问题，从模糊图像恢复到清晰图像是一个比较具有挑战性的任务。基于模糊核估计的模型[1,2]一般是同时估计模糊图像的模糊核及其清晰图像。但模糊核估计方法对噪声比较敏感且目前的模糊核估计方法适应情况少。现实世界中大多数运动模糊情况都属于不均匀运动模糊，如十字路口一类的动态场景，会有多个物体以不同幅度向不同方向运动。在图像去模糊任务上，还有一类是端到端的方法，这一类方法无需估计模糊核。目前端到端的方法主要分有对抗和无对抗两大类，其中大部分模型均使用卷积神经网络(convolutional neural networks, CNN)进行设计。

目前基于CNN的去模糊模型研究也都是尽量增大感受野以提高模型的恢复效果。虽然可以通过堆叠卷积层来提升感受野范围，但是堆叠过多的卷积层会增加参数量和计算量，并且更容易出现噪声。为了解决以上问题，本文提出了频域Transformer图像去运动模糊模型。Transformer[3]的自注意力机制可以对自身及其他所有位置的像素来建立依赖关系，相当于拥有全局感受野。对于去模糊任务而言，越大的感受野越容易处理剧烈的运动模糊情况。Dosovitskiy等[4]提出的视觉Transformer(vision transformer, ViT)模型证明了Transformer在图像领域的可行性，并且相对于CNN网络而言，Transformer没有归纳偏差问题。Transformer虽然具有强大的表示能力，但是仅从空域层面处理图像还是不够的，因此提出了空频处理模块(spatial frequency processing module, SFPM)。通过傅里叶变换将数

收稿日期：2023-05-05 **修稿日期：**2023-07-23

基金项目：河北省省级科技计划资助(20310801D)；河北省智慧交通大数据关键技术研究与应用示范(20310802D)

作者简介：顾军华(1966—)，男，河北石家庄人，教授，博士，CCF会员，研究方向为智能信息处理、数据挖掘；李岩(1994—)，男，天津人，硕士研究生，研究方向为图像处理、图像去模糊；陈晨(1998—)，女，河北邢台人，硕士研究生，研究方向为智能交通、交通流预测；*通信作者：牛炳鑫(1988—)，男，河北邯郸人，讲师，博士，CCF会员，研究方向为智能交通、物联网，E-mail:niubingxin666@163.com；李春杰(1975—)，男，河北石家庄人，正高级工程师，博士，研究方向为智能交通

据从空域转换到频域，利用频率成分和空域图像间的对应关系，一些在空域难以表述的问题在频域就能迎刃而解。采用傅里叶变换转换到频域后通过其中的平行条纹很容易观察到运动模糊的方向和大小，因此对频域图进行分解操作可以充分利用运动模糊在频域图中所体现的方向性特征，从而更有效地恢复运动模糊。由于频域是有限的频率表示图像，仅从频域处理可能会损失某些信息，因此空域和频域一同处理可以达到互补的目的。

1 频域Transformer图像去运动模糊模型

端到端模型的目的是仅将模糊图像I_b作为输入来恢复清晰图像，不需要模糊核。在训练阶段，有两个网络需要训练，一个是鉴别器，用于区分生成的结果和真实的清晰图像；另一个是生成器，对其输入模糊图片，并输出一个足以骗过鉴别器的结果。

1.1 模型架构

如图1所示，本文提出的网络基于生成对抗网络，由生成器和鉴别器两部分组成，以对抗方式进行训练。

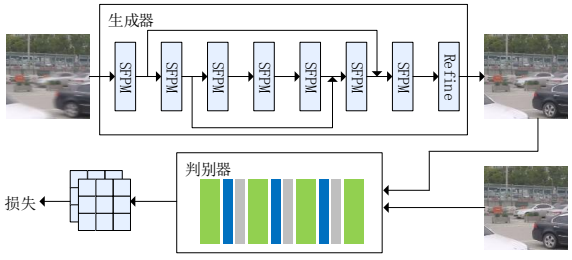


图1 模型整体架构

训练过程中向生成器输入模糊图像，而不是像原生的生成对抗网络(generative adversarial network, GAN)[5]一样输入一个随机噪声向量。然后通过判别器比较生成器的输出与清晰图像之间的差异。其中生成器使用频域Transformer，判别器使用PatchGAN来进行对抗训练。

1.2 频域Transformer生成器

生成器架构如图2所示，生成器整体采用U形结构并在编码器与解码器间加入了残差连接。给定一个模糊图片$I_b = \mathbf{R}^{H\times W\times 3}$，其中$H$表示图像的高度，$W$表示图像的宽度，3表示图像的RGB通道，模型首先使用卷积层将输入映射到高维空间来得到特征图。然后将特征图传递到后面的SFPM模块，每个SFPM模块包括多个Transformer模块和一个频域处理模块，它们以并行的方式处理输入图像。映射后的特征图通过空频处理模块(spatial frequency processing module, SFPM)向下传递多层，每个空频处理模块包含多个Transformer块以及一个频域处理模块，两者以并行的方式进行工作。每经过一个空频处理模块进行一次下采样，下采样通过卷积实现。带有下采样部分的左半部分是生成器的编码器，右半部分则是解码器。解码器中首先通过PixelShuffle操作在不丢失信息的前提下进行上采样，然后输入到空频处理模块进行学习。最后通过一个精炼模块来对输出作最后的调整。

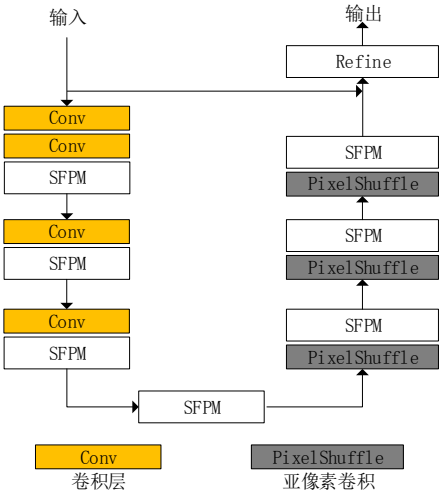


图2 生成器架构

Transformer的自注意力机制能够学习全局之间的关系，但是直接用它来处理高分辨率图像，会使得训练速度变得很慢，并且对于硬件要求也更高。为了满足处理大分辨率图片的需求，引入了局部自注意力机制，通过设置一个窗口大小来控制自注意力的计算范围，有效减少了计算量。同时结合U形网络，在浅层进行局部注意力计算，提取细节特征；在深层进行全局注意力计算，得到全局信息。编码器部分的每个SFPM模块都包含卷积下采样操作，输入每通过编码器其中的一个SFPM，尺寸就会进行

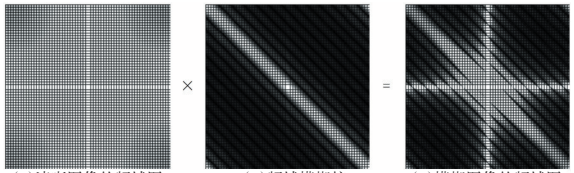
缩减，相应地，特征就会增加。当输入变得足够小时，就可以将自注意力应用于整个特征图，从而学习到全局信息。通过这种方法，既可以获得局部细节，又可以获得全局信息。在解码器部分使用的则是亚像素卷积上采样层，可以表示为如下公式：

$$PS(T)_{x,y,c} = T_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, c \cdot r \cdot mod(y,r) + c \cdot mod(x,r)} \tag{1}$$

1.3　空频处理模块

图像的退化过程可以理解为是将算子H作用在一个输入图片$f(x,y)$上来生成退化图像$g(x,y)$。如果算子H已知，那么可以很容易地将退化图像恢复成原始图像。但在很多情况下并不能得到这样一种准确的理想的算子，只能尽量近似它。根据卷积定理，空域中的卷积操作可以转换为频域乘法，因此一些直接在空间域表述非常困难，甚至不可能的任务在频域中变得非常普通。

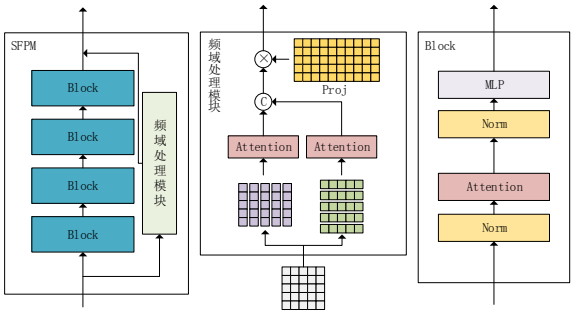
空域中的卷积操作转化到频域可以简化为乘积操作，图3中清晰图像的频域图(a)与频域模糊核(b)进行乘积可得模糊图像的频域图。同样地，模糊图像的频域图除以频域模糊核即可得到清晰图像的频域图。因此模型只需要拟合出图3(b)的倒数再乘上输入的模糊图像的频域图即可得到清晰图像。



(a)清晰图像的频域图　(b)频域模糊核　(c)模糊图像的频域图

图3　频域乘积示例

在现实世界中模糊图像与清晰图像在其频域图上仍然有明显的区别。场景中各种物体的边缘在频域图中相互叠加，频域中可见的条纹由整幅图像的抖动而来，图像的空域信息和频域信息可以相互补充，充分利用这两部分信息可以更好地抑制图像中的模糊伪影。空频处理模块设计如图4(a)所示，左侧为Transformer分支，用于处理空域信息；右侧分支为频域处理分支，用于对频域图像进行处理。空域信息和频域信息并行处理，最后进行融合得到SFPM模块的输出。随着U形网络的加深，SFPM模块所处理的特征维度就越高，且越注重全局信息，结合U形网络的特点，频域处理模块更是可以关注不同层次大小的频域图信息。



(a)空频处理模块(SFPM)　(b)频域处理模块　(c)Transformer块

图4　空频处理模块内部结构

为了在频域对图像进行处理需要先将图像转换到频域，通过离散傅里叶变换(discrete fourier transform, DFT)转换而来的频域图可以清晰地反映运动模糊的方向与大小。再将频域图沿水平垂直方向进行分解，然后计算各自的注意力，可以充分利用运动模糊在频域中体现的特征，有效地恢复频域图。DFT是采样后的傅里叶变换，它不包含组成图像的所有频率，可能会丢失一些细节信息，因此仅在频域处理是不够的，还要结合空域一起处理才能起到互补的作用。

由于模糊图像的频域图包含模糊图像中的运动方向及幅度，因此频域处理模块将频域图沿横纵轴进行分解，将不同方向的模糊频率分解为水平和垂直方向，然后对分解后的向量分别进行注意力计算。垂直分解的向量表示为$\boldsymbol{X}_i^v$，$i \in \{1,\cdots,W\}$，其中W表示图像宽度；水平分解的向量表示为$\boldsymbol{X}_i^h$，$j \in \{1,\cdots,H\}$，其中H表示图像高度。在分解后的向量进行注意力计算之前，首先对其特征维度使用1×1的卷积层扩充至其自身的两倍，由于频域图中包含丰富的不同方向和幅度的频率，通过扩充分解向量的特征数有助于提高模块的表达能力，不会丢失关键信息。该步骤如下所示：

$$\boldsymbol{X}_i^v, \boldsymbol{X}_j^h = DEC\left(DFT(\boldsymbol{X})\right) \tag{2}$$

$$\boldsymbol{X}_i^{v'}, \boldsymbol{X}_j^{h'} = Conv(\boldsymbol{X}_i^v, \boldsymbol{X}_j^h) \tag{3}$$

其中：DFT表示离散傅里叶变换，DEC表示向

量分解，$\boldsymbol{X}_i^{v'} \in \mathbf{R}^{W \times 2C}, \boldsymbol{X}_j^{h'} \in \mathbf{R}^{H \times 2C}$，分解向量在特征扩充后进行注意力计算。首先生成映射矩阵 $\boldsymbol{P}_n^Q, \boldsymbol{P}_n^K, \boldsymbol{P}_n^V \in \mathbf{R}^{2C \times (2C/N)}$，其中 N 表示多头注意力的头数，$n \in \{1, \cdots, N\}$。以水平分解向量为例使用下式计算其对应的查询与键值 $\boldsymbol{Q}_{\text{in}}^v, \boldsymbol{K}_{\text{in}}^v, \boldsymbol{V}_{\text{in}}^v$：

$$\left(\boldsymbol{Q}_{\text{in}}^v, \boldsymbol{K}_{\text{in}}^v, \boldsymbol{V}_{\text{in}}^v\right) = \left(\boldsymbol{X}_i^{v'}\boldsymbol{P}_n^Q, \boldsymbol{X}_i^{v'}\boldsymbol{P}_n^K, \boldsymbol{X}_i^{v'}\boldsymbol{P}_n^V\right) \tag{4}$$

注意力计算表示为

$$\boldsymbol{O}_{\text{in}}^v = \text{Soft}\max\left(\frac{\boldsymbol{Q}_{\text{in}}^v\left(\boldsymbol{K}_{\text{in}}^v\right)^{\text{T}}}{\sqrt{2C/N}}\right)\boldsymbol{V}_{\text{in}}^v \tag{5}$$

合并所有垂直分解向量的注意力输出即可得到垂直注意力输出 O^v。以同样的方式对水平分解向量进行计算可得 O^h。然后将垂直和水平的注意力输出合并起来并通过矩阵 $\boldsymbol{K} \in \mathbf{R}^{2W \times W \times 2C}$ 将输出映射为输入尺寸，并再次使用 1×1 的卷积层对特征维度进行聚合，表示如下：

$$O = iDFT\left(Conv\left(\boldsymbol{K} * Concat\left(O^v, O^h\right)\right)\right) \tag{6}$$

其中：$O \in \mathbf{R}^{H \times W \times C}$ 表示模块的输出，$iDFT$ 表示离散傅里叶变换的逆运算。傅里叶变换中，低频主要决定图像在平滑区域中总体灰度级的显示，而高频决定图像细节部分，如边缘和噪声。模糊问题也可以理解为原本高频锐利的边缘经过模糊之后变得更平滑，因此去模糊任务就是从比较平滑的低频区域将原本应是高频边缘的部分恢复出来。

1.4 损失函数

原生GAN的损失函数在两个数据分布完全不重合的时候，JS散度恒为 $\log 2$，所以模型也无法更新。因此损失函数选用WGAN-GP[6]进行训练，公式如下：

$$\begin{aligned} Loss_{\text{gan}} = {} & E_{x \sim P_G}\left[D(x)\right] - E_{x \sim P_{\text{data}}}\left[D(x)\right] + \\ & \lambda E_{x \sim P_{\text{penalty}}}\left[\max\left(0, \left\|\nabla_x D(x)\right\| - 1\right)\right] \end{aligned} \tag{7}$$

其中：惩罚项对分布 P_{penalty} 进行采样，分布 P_{penalty} 介于 P_{data} 与 P_G 之间，只需要保证在该分布的采样的梯度范数小于等于1即可。此外还增加了一个感知损失函数，感知损失认为生成图像是从内容图变换而来，通过计算内容损失不断迭代生成图片，使其越来越接近内容图。在去模糊任务中，生成图像指的是模糊图像，而内容图则是清晰图像。感知损失函数如下所示：

$$Loss_{\text{perceptual}}^{\phi, j} = \frac{1}{C_j H_j W_j}\left\|\phi_j\left(\hat{y}\right) - \phi_j\left(y\right)\right\|_2^2 \tag{8}$$

其中：j 表示VGG网络的第 j 层，$C_j H_j W_j$ 表示第 j 层特征图的尺寸。将对抗损失函数和感知损失函数组合起来就是模型最终的损失函数，其中 λ_1 和 λ_2 是两个超参数：

$$Loss = \lambda_1 Loss_{\text{gan}} + \lambda_2 Loss_{\text{perceptual}} \tag{9}$$

2 实验部分

2.1 数据集

实验使用GoPro[7]数据集作为训练集和测试集。它由3214幅模糊图像和对应的清晰图像组成，分辨率为 1280×720，其中2103对用于训练，1111对用于测试。它使用GoPro相机以视频的方式捕捉帧，然后对前后连续短曝光的帧进行平均，生成模糊图像。

另一个用于测试的RealBlur[8]数据集使用两个相同型号的相机来捕捉图像，其中一个是以高快门速度进行拍照，另一个是以低快门速度拍照。由研究人员手持两台相机构成的拍摄系统在街道中进行随机拍摄，该数据集采集到的模糊图像更为真实。它们都包含4738对图像，其中980对用于测试，其余用于训练。

2.2 实验设置

实验使用PyTorch来实现本文的模型，该模型无需预训练。训练是在带有AMD 5950X CPU以及单块Nvidia GeForce RTX 3090Ti GPU的服务器上进行的。实验将图像裁剪为几个大小为 256×256 的图像块，批量训练大小设置为1。在训练阶段，使用ADAM[9]优化器对模型进行训练。学习率设置为 10^{-6}，使用余弦退火策略逐步降低学习率。实验使用上述配置在GoPro数据集上训练了300轮，然后在模糊图像数据集GoPro和RealBlur上对所提出的去模糊模型分别进行了测试。

2.3 实验结果

实验在GoPro和RealBlur数据集上进行了测试。图5展示了各模型的效果对比。

图 5　去模糊方法对比

表1显示了模型在峰值信噪比(peak signal to noise ratio, PSNR)和结构相似度(structural similarity, SSIM)指标上优于其他方法。

表 1　模型测试结果

模型	GoPro		RealBlur-J	
	PSNR	SSIM	PSNR	SSIM
DeblurGAN[10]	28.70	0.85	26.97	0.80
DeblurGANv2[11]	29.55	0.93	28.10	0.82
PyNAS[12]	30.62	0.94	28.31	0.85
UHDVD[13]	31.33	0.92	27.96	0.85
SimpleNet[14]	31.52	0.94	28.55	0.86
频域Transformer	32.24	0.95	29.38	0.87

该模型在GoPro和RealBlur两个数据集上与几个经典的去模糊模型进行了比较，其中DeblurGAN和DeblurGANv2是同一个团队提出的，使用的是CNN构成的GAN网络。结果表明本文提出的模型在两个数据集上分别比其他方法至少提高了0.72 dB和0.83 dB，表现出较强的泛化能力。此外，图6展示了一些去模糊的结果图与频域图。

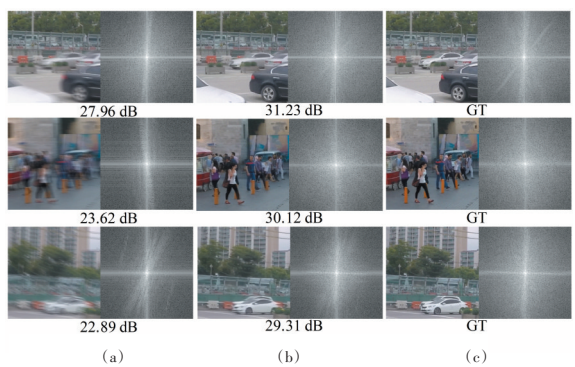


图 6　GoPro数据集的训练结果

由图中频域图像也可看出，模型对图像的恢复体现在频域中的效果很明显，成功滤除了造成模糊的频率条纹。

2.4　消融实验

本节详细分析了模型中组件的作用，并在GoPro和RealBlur两个数据集上进行对比实验，实验结果见表2。U形网络结构的性能优于简单的对生成器的输入和输出应用残差连接，如表2中的(A)和(B)所示。它将每个下采样阶段的特征图与对应的上采样阶段使用残差连接结合起来。在U形网络结构中，网络越深感受野越大，网络将更加关注更抽象的全局相关性。相比之下，当网络深度较浅时，它会更加关注纹理细节。

表 2　消融实验结果

生成器	GoPro		RealBlur-J	
	PSNR	SSIM	PSNR	SSIM
(A)single RC	30.84	0.92	27.54	0.83
(B)Unet	31.33	0.92	27.43	0.83
(C)Unet+频域(SFPM)	31.69	0.93	27.92	0.85
(D)Unet+频域(Block)	31.97	0.92	28.34	0.85
(E)Unet+Freq(MLP)	32.02	0.94	28.58	0.86
(F)Unet+Freq(Proj)	32.24	0.95	29.38	0.87

本节尝试将频域处理模块置于不同位置，结果显示频域处理模块的位置对于去模糊效果是有影响的。(C)尝试将频域处理模块置于与U形网络并行的位置，使其与Transformer的U形网络共享输入再合并输出。与之相比，(D)中将频域处理模块放入Transformer块或(E)中将频域处理模块放入MLP中的性能要更好。当频域处理模块与U形网络并行时，直接对图像进行处理，将输入转换到频域后对输入进行卷积运算。它只能用有限的频率在频域表示图像，这意味着它会丢失一些信息。但将其放入与Transformer块并行的位置，它就可以处理多个不同大小的特征映射，并可以获得全局和局部的相关性。关于表2中的(E)，是将频域处理模块部署在MLP模块旁边。虽然最后的分数有一定的提升，但计算量比(D)提高了3倍左右。在(F)中，再把它放回Transformer块的旁边，同时改变其内部结构。在(C)和(D)中，只在频域对图像进

行横纵向分解后的注意力计算，而(F)中，增加了一个投影矩阵，横纵向分解向量进行注意力计算后通过该投影矩阵可以自动学习关键信息，结果也表明该方法对于去模糊的效果有一定的改善。

3 结语

本文提出一种端到端的基于GAN的频域Transformer的去模糊模型，无须估计模糊核，主要面向不均匀运动模糊问题。文中对Transformer的内部结构进行了重新设计，加入了频域处理模块，使得模型能够同时在空域和频域上对图像进行恢复，充分利用了运动模糊在频域图中的特征，与Transformer处理空域互补，取得了更好的恢复效果。同时生成器U形网络的架构，使得模型能够处理高分辨率图片，并且可以学习到模糊图片的局部细节以及全局信息。得益于Transformer的特性，模型对于数据增广的收益要大于基于CNN的模型，数据量越大对模型效果越有利。模型受益于Transformer的自注意力机制，但同样受制于自注意力机制，因为对于高分辨率图片构成的输入序列已经属于超长序列，自注意力计算的代价很大，造成了模型参数量大、计算速度慢的问题，后续工作需在此基础上压缩模型参数量，提高模型运行效率，便于部署在计算资源有限的嵌入式终端，如交通监控摄像头等设备。

参考文献：

[1] CHAKRABARTI A. A neural approach to blind motion deblurring[C]//Proceedings of the European Conference on Computer Vision, Berlin, German, 2016:221-235.

[2] REN D, ZHANG K, WANG Q, et al. Neural blind deconvolution using deep priors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems, La Jolla, California, 2017.

[4] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//Proceedings of the International Conference on Learning Representations, 2021.

[5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27:1-9.

[6] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-11.

[7] NAH S, HYUN KIM T, MU LEE K. Deep multi-scale convolutional neural network for dynamic scene deblurring[C]//Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, 2017.

[8] RIM J, LEE H, WON J, et al. Real-world blur dataset for learning and benchmarking deblurring algorithms[C]//Proceedings of the European Conference on Computer Vision, Berlin, German, 2020.

[9] KINGMA D P, BA J. ADAM: a method for stochastic optimization[J]. Computer Science, 2014:1-15.

[10] KUPYN O, BUDZAN V, MYKHAILYCH M, et al. Deblurgan: blind motion deblurring using conditional adversarial networks[C]//Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, 2018.

[11] KUPYN O, MARTYNIUK T, WU J, et al. Deblurgan-v2: deblurring (orders-of-magnitude) faster and better[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Piscataway, NJ, 2019.

[12] HU X, REN W, YU K, et al. Pyramid architecture search for real-time image deblurring[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Piscataway, NJ, 2021.

[13] DENG S, REN W, YAN Y, et al. Multi-scale separable network for ultra-high-definition video deblurring[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Piscataway, NJ, 2021.

[14] LI J, TAN W, YAN B. Perceptual variousness motion deblurring with light global context refinement[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Piscataway, NJ, 2021.

（下转第59页）

文章编号: 1007-1423（2023）16-0021-06 DOI: 10.3969/j.issn.1007-1423.2023.16.004

面向电力领域自然语言理解的数据增强研究与实现

施俊威，宋 晖*

（东华大学计算机科学与技术学院，上海 201620）

摘要：探究面向领域智能问答中自然语言理解的数据增强问题。由于应用缺乏历史数据，且人工标注成本高，无法满足大规模训练自然语言理解模型的需求。因此，对传统数据增强的方法进行研究，提出使用基于对比搜索的关键字文本生成模型，以此生成了具有句式表达多样性的的数据集。实验结果表明，相比传统的集中搜索算法，使用对比搜索作为模型的解码策略能够生成更加准确和合理的电力领域问题文本，有效地降低了生成文本的词重复率。利用这些样本数据，成功地训练了一个高效准确的自然语言理解模型，提高了用户意图识别的准确率。这一研究对于智能问答领域的实际应用具有一定的参考价值。

关键词：自然语言理解；文本生成；对比搜索；数据增强；智能问答

0 引言

随着人工智能技术的不断发展和应用，电力领域也逐渐迎来了智能化转型的浪潮。在电力行业中，用户往往通过人工客服和电力公司网站等传统方式获取与用电相关的各种信息，然而，这些方式存在着诸多问题，如人工客服效率低、需要等待时间长。因此，在人工智能系统与电力系统的融合中，智能问答成为一种重要的应用方式和技术手段，将自然语言处理技术应用于电力系统中，提高沟通效率和工作效率，节省了人力资源[1]，使电力系统向自动化、智能化方向发展[2]。

在电力领域的智能问答中正确地理解用户的问题至关重要，目前通常采用自然语言理解（NLU）模型来实现[2]。在面向电力营销指标问答的应用时，由于业务场景很多，需要识别的槽也很多，直接标注企业提供的少量问题，构建NLU模型，识别准确率只能达到60%左右，离实际应用要求相去甚远。研究表明，大量的训练数据能够显著提高NLU模型的准确性[3]。在实际应用中，用户和企业提供的领域问题数据并不能满足模型训练要求，并且人工数据标注的成本大，所以需要使用数据增强技术生成大量的电力领域样本数据，以此满足NLU模型训练需求。

传统的文本数据增强方法包含词汇替换、文本表面转换以及随机噪声注入[4]。但是在面向电力领域NLU任务里，通过使用这些方法生成的文本不能保证文本语义的连贯性和文本多样性，这会导致NLU模型性能下降[5]，因此考虑采用生成式模型生成问题样本，生成模型能够基于给出的关键词生成问题句。我们首先通过槽值替换方法获得部分样本，再利用这些样本一起训练生成式模型。

目前生成式模型的解码方法一般是基于集集中搜索[6]，其生成的文本会出现重复词或者多个同类型词连续出现，导致生成的文本语义不正确。本文提出了一种基于对比搜索[7]关键词文本生成模型（neural text generaion with contrastive search，CSTG）。在文本生成过程中，生成的输

收稿日期：2023-04-26 **修稿日期：**2023-06-09

作者简介：施俊威（1998—），男，湖北天门人，硕士研究生，研究方向为自然语言处理；*通信作者：宋晖（1971—），女，上海人，教授，研究方向为自然语言处理，E-mail：songhui@dhu.edu.cn

出应该从模型预测的最有可能的候选词集合中选择，生成的输出应该具有充分的区分性，以便于与前文上下文的关系进行区分。通过这种方式，在电力领域样本数据生成任务中，模型所生成的问题文本既能够更好地保持语义连贯性，同时避免模型退化和在生成的文本中出现连续的重复词。实验结果表明，该模型降低了生成电力领域问题文本的重复率，从而使得训练后的NLU模型能够更加准确地理解用户的问题，提升电力指标检索系统的准确率。

1 基于对比搜索关键词文本生成模型

1.1 模型框架

在电力领域的指标问答应用中，NLU模型需要理解用户问句对应的业务领域、指标问法，以及关键槽值。如领域domain为配变异常，指标意图intent为query_total, 槽和对应的槽值slots: {“org”：“南因供电所”，“time”：“2022年1月”，“detail”：“情况”}。

我们首先使用传统的槽替换、值替换等方法，基于用户提供的典型和应用系统数据，获得的样本数据作为训练数据集，用于训练问题生成模型：模型将问题句的各种槽值(关键字)作为输入，输出一段完整的文本，该段文本应尽可能包含这些给定的领域关键词。输入领域关键词包含售电量、石家庄、3月和排名，经过模型的文本生成，生成完整的文本：“今年3月份在石家庄的售电量排名是多少?”

本文设计了CSTG模型来实现生成过程，该模型是基于注意力机制的LSTM文本生成模型[8]，如图1所示。根据输入的电力领域关键词W生成与电力领域关键词相关的问题文本。

CSTG模型由编码和解码两部分组成，在编码部分，主要包含对关键词的编码以及设置关键词表达的覆盖向量，覆盖向量是控制关键词在文本生成中被表达的权重值向量，使得关键词尽可能地在文本中被表达出来。在解码部分，包含注意力模块和解码策略，注意力模块让解码器基于自身历史输出，来计算输出。解码策略则基于所给候选词选择最优项作为输出。

1.1.1 编码

在模型的编码阶段，首先模型将输入的关键词通过一个词嵌入层转换成向量表示$\boldsymbol{W}$，然后，通过引入注意力参数μ，我们可以将每个关键词表示为T_t，并将每个关键词的语义通过注意力机制传递到生成的单词中，该注意力参数定义为一个向量：μ_1、μ_2、…、μ_k，其中μ_i表示领域关键词词汇i的分数。在时间步t生成文本时，T_t的计算公式如下：

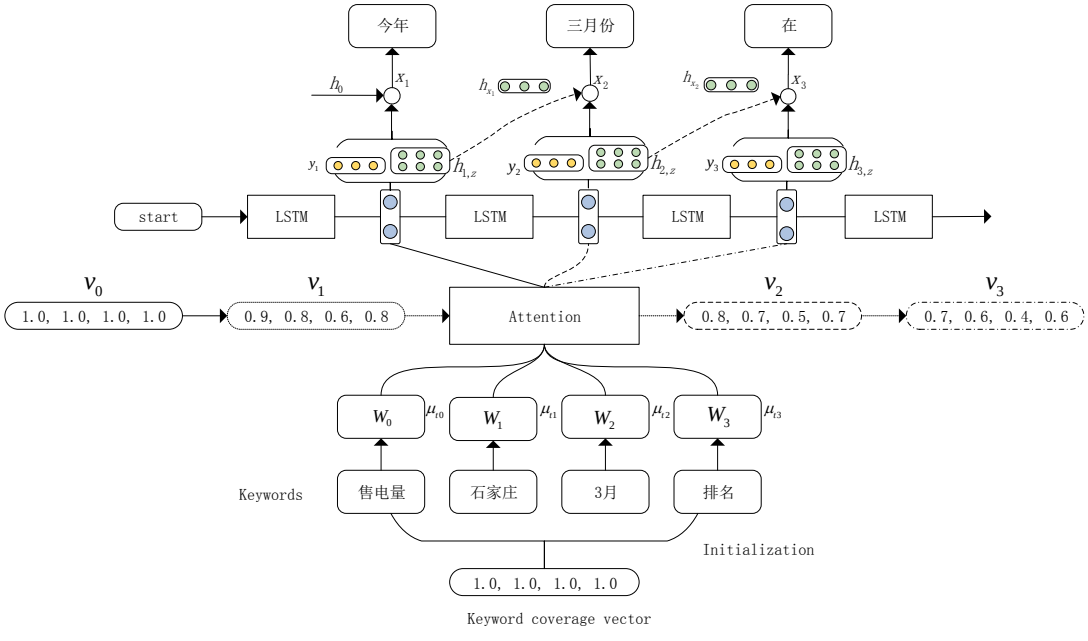


图1 模型结构

$$T_t = \sum_{j=1}^{k} \mu_{tj} \boldsymbol{W}_j \tag{1}$$

其中，$\boldsymbol{W}_j$表示第j个关键词的编码张量，μ_{tj}的计算方式如下:

$$\mu_{tj} = \frac{\exp(q_{tj})}{\sum_{i=1}^{k} \exp(q_{ti})} \tag{2}$$

并且q_{tj}的计算公式表示为

$$q_{tj} = V_{t-1,j} \boldsymbol{v}_a^{\mathrm{T}} \tanh(\boldsymbol{X}_a h_{t-1} + \boldsymbol{Y}_a \boldsymbol{W}_j) \tag{3}$$

其中$\boldsymbol{v}_a$, $\boldsymbol{X}_a$和$\boldsymbol{Y}_a$是三个矩阵，在模型训练期间需要进行优化。

1.1.2　关键词覆盖向量

模型设置一个关于所给关键词的覆盖向量$\boldsymbol{v}$，每一维度代表该关键词在将来的文本生成中被表达的程度，关键词覆盖向量通过一个参数φ_j进行更新，这个参数可以被视为关键词W_j的话语级重要性权重。该参数用于调整生成过程中领域关键词的表达程度，从而影响模型对关键词$\boldsymbol{W}_j$的关注程度，使得生成的文本能够更好地符合关键词$\boldsymbol{W}_j$的内容要求。在模型的生成过程中，模型会根据覆盖向量v和注意力机制调整生成策略，使得模型能够让未被表达的关键词在生成过程中表达，从而提升生成文本的可读性和完整性。在当前时间步t生成新的词时，第j个关键词$V_{t,j}$的计算方式如下:

$$V_{t,j} = V_{t-1,j} - \frac{1}{\varphi_j} \mu_{t,j} \tag{4}$$

其中$\mu_{t,j}$表示第j个关键词在时间步t的注意力权重。$\varphi_j = N \cdot \sigma\left(U_f \left[T_1, T_2, \cdots, T_k\right]\right), U_f \in \mathbf{R}^{kd_w}$。

在模型的训练过程中，可以利用已有的电力领域问题和关键词来训练模型的参数和关键词覆盖向量。在模型训练的过程中，可以采用基于梯度下降的方法对模型的参数和关键词覆盖向量进行更新。综合上述，基于关键词的文本生成模型可以帮助解决电力领域NLU任务中数据数量不足的问题，提高训练数据的数量和质量，从而提高模型的性能和准确率。同时，利用关键词覆盖向量可以让模型更好地理解关键词，并根据关键词生成符合要求的文本。

1.2　模型解码

1.2.1　文本生成

模型在生成文本阶段，是基于上一个生成的词预测下一个词，下一个预测词的概率根据下式得出。

$$P\left(y_t | y_{t-1}, T_t, V_t\right) = \mathrm{soft\,max}\left(q(h_t)\right) \tag{5}$$

上式中模型在时间步t的隐层状态h_t的公式如下所示:

$$h_t = f\left(h_{t-1}, y_{t-1}, T_t\right) \tag{6}$$

一般生成式模型，在得到预测词的概率后会选择搜索算法，从候选词中选取较为合适的词作为生成文本。本文选用对比搜索作为模型的解码策略，生成语义正确的文本数据。

1.2.2　对比搜索

基于关键词的文本生成模型普遍的解码方法包括集中搜索、贪心搜索和随机采样，虽然这些方法在一般情况下可以保证生成文本的语义正确性，但是在电力领域中，领域词之间存在相似性，使用这些方法容易导致生成的文本出现重复词或连续生成同类词，从而影响生成文本的质量。究其原因，是因为解码算法没有对语言模型的偏差合理规避[9]。为此，在基于关键词的文本生成模型的基础上，可以采用对比搜索的解码方法来解决这一问题。对比搜索的方法会基于候选词的选中概率，在每个解码步骤中选择模型预测的最可能候选词集合，在选取最优候选词时，生成的输出与前一个上下文有足够的区别性，以避免模型的退化。这种方法可以更好地保持生成文本与前缀的语义连贯性，并避免生成的文本质量下降。

对比搜索是一种用于解码的策略。在每个解码步骤中，该策略会从模型预测的最有可能的候选集合中选择输出，以确保生成的文本与人类编写的前缀之间的语义一致，并保持生成文本的词汇相似度矩阵的稀疏性，从而避免模型退化。一般解码方法会选择概率最高的候选项作为输出，而候选集合通常包含模型预测的前k个最高概率的选项，k的取值通常为3到10。对比搜索引入了一项惩罚机制，用于评估候选项与先前上下文的可区分性，从而保证生成的

文本具有足够的多样性。在时间步骤t，给定先前上下文$x_{<t}$，选择输出x_t的过程如下：

$$x_t = \arg\max_{z \in Z^{(k)}} \left\{ (1-\theta) \times P(y_t | y_{t-1}, T_t, V_t) - \theta \times \left(\max \left\{ s(h_{t,z}, h_{x_{t-1}}) \right\} \right) \right\} \tag{7}$$

在上式中$Z^{(k)}$是模型概率分布中前k个最高预测的候选集合，通常设置为3到10。第一项模型置信度是模型预测的候选z的概率。第二项衰变惩罚度量候选z相对于先前上下文$x_{<t}$的可区分性。s在公式中被定义为z的表示和$x_{<t}$中所有tokens的表示之间的相似度。这里$h_{t,z}$是表示在当前时间步候选词在模型中的隐层状态。$h_{x_{t-1}}$表示时间步$t-1$，选择的最优候选词的隐层状态，其中z的更大的退化惩罚意味着它更类似于上下文，因此更容易导致模型的退化。超参数$\theta \in [0,1]$调节这两个组件的重要性。当$\theta = 0$时，对比搜索退化为贪心搜索方法。

2 实验

2.1 数据集与实验设置

在当前的电力领域文本生成研究中，构建面向电力领域指标检索的问题数据集是非常关键的。样本数据集是JSON的文件格式，每一条数据包含问题文本、领域、意图、槽及槽对应的槽值。为了能够更好地训练和评估电力领域的文本生成模型，选择了基于问题模板生成的电力营销领域样本数据，并从企业真实数据中获取了用电量、售电收入、投诉数量、地市排名等多个问题句类型的数据。

该数据集包含了20920条样本数据，其中训练集包含17500条文本数据，测试集包含3420条文本数据。为了更好地模拟真实情况，数据集的关键词是选自于NLU样本问题的槽数据，能够更好地体现电力领域的语境和特点。通过这样的数据集构建，可以为电力领域的文本生成研究提供更加真实、具有代表性的数据，有助于研究者们更加深入地探究电力营销领域文本生成模型的性能和应用场景。

2.2 实验设置

在模型的实现上，该模型是基于LSTM的文本生成模型，其中使用到注意力机制，模型的超参数设置为：batch_size为32，候选词的数量k为3，训练轮次为45，最大关键词词数为5，embedding_dim为100，hidden_dim为512，最大句子长度为128，学习率为1e-7，ReLU作为激活函数。

2.3 评估指标

为了评估模型的性能，我们采用以下三个指标：

（1）Bleu指标通过比较机器翻译结果与参考翻译之间的n-gram重叠率来进行评估。

（2）MAUVE是一种衡量生成的文本和人类编写的文本之间的标记分布紧密度的指标。更高的MAUVE分数意味着该模型生成更多类似人类的文本。

（3）rep-n这个指标以生成文本中重复n-gram的比例来衡量序列级别的重复，其中，n表示连续词语或字符的数量。该重复率越低代表所生成的文本出现重复词越少。对于所生成的文本$\hat{t}$，n-gram级别的重复的定义如下：

$$\text{rep-}n = 100 \times \left(1.0 - \frac{\left| unique\ n-grams(\hat{t}) \right|}{\left| total\ n-gram(\hat{t}) \right|} \right) \tag{8}$$

（4）Diversity指标考虑到不同的n-gram级别生成的重复，可以视为模型退化的整体评估，较低的多样性意味着模型退化得更加严重。其公式如下：

$$diversity = \prod_{n=2}^{4} \left(1.0 - \frac{rep-n}{100} \right) \tag{9}$$

3 实验结果比较与分析

实验1通过在文本生成模型中使用不同解码方法，根据模型所生成的文本质量判断模型整体的性能。该实验中使用对比搜索和集中搜索两种解码方法，使用两种解码方法的模型在测试集中，实验1的结果见表1，使用对比搜索的模型相较于使用集中搜索提高了9%，说明在电力领域使用对比搜索相较于传统的集中搜索可以提高所生成文本的质量，模型的性能也会更好。

表 1　模型性能实验结果

解码方法	训练集	测试集	Bleu
对比搜索	17500	3420	**0.36**
集中搜索	17500	3420	0.27

在实验2中，根据训练数据集已有的领域关键词利用已训练的模型生成4000条文本，分别计算基于不同的解码方法所生成文本的文本重复率，以此评估生成的文本质量。实验结果见表2，不同模型所生成的文本的重复率使用不同的解码方法有了显著的下降，其中对比搜索相较于集中搜索，在Rep-2、Rep-3、Rep-4三个指标上都有较为明显的下降，分别降低10.88、5.27、2.61，并且Diversity与Mauve指标有一定提高，对比搜索相较于贪心搜索和Nucleus Sampling两种解码方法也有很大提升，表示对比搜索在该模型中能够显著降低生成文本的重复率，更好地保持模型生成文本与人类编写的文本之间的语义一致性以及保留生成文本的词汇相似度矩阵的稀疏性。

表 2　不同解码方法对文本重复率的影响

解码方法	数据量	Rep-2	Rep-3	Rep-4	Diversity	Mauve
对比搜索	4000	**1.62**	**0.52**	**0.24**	**0.98**	**0.678**
集束搜索	4000	12.5	5.79	2.85	0.80	0.671
贪心搜索	4000	12.99	7.42	3.88	0.77	0.551
Nucleus Sampling	4000	13.94	8.06	4.4	0.76	0.464

此外，通过对实验2中所生成的文本进行分析，在模型训练数据集中，根据不同文本长度的文本数据对应的领域关键词生成文本，分析不同的文本长度对文本重复率是否产生影响。如表3所示。

表 3　文本长度对重复率的影响

解码方法	数据量	平均文本长度	Rep-2	Rep-3	Rep-4	Diversity
对比搜索	1800	21.37	**2.80**	**1.12**	**0.54**	**0.96**
集中搜索	1800	21.37	12.05	5.85	3.12	0.80
对比搜索	1800	28.11	**0.67**	**0.00**	**0.00**	**0.99**
集中搜索	1800	28.11	13.72	6.13	2.82	0.79

在电力领域数据集，平均文本长度分别为21.37和28.11，每种文本长度选取基于不同领域关键词生成的1800条文本数据进行分析，其对比搜索的重复率指标Rep-2、Rep-3和Rep-4相较于集中搜索都有一定幅度的降低，整体重复率指标Diversity有一定提高。在分析不同长度的文本数据中，使用对比搜索的解码方法相较于集中搜索取得更好的效果，基于对比搜索的模型所生成的文本，其重复率对于集中搜索有明显的下降。

4　结语

本文针对智能问答中电力指标检索系统的NLU问题，训练NLU模型，其需要大量的样本问题满足训练模型需要，但企业搜集的数据较少，人工标注成本较大，不能满足需要，因此使用数据增强技术生成大量样本数据。我们构建了一个面向电力领域的样本问题数据集，提出并使用CSTG生成了与电力领域相关的问题样本。对比搜索的核心思想是在每个解码步骤中，从模型预测的最可能候选集中选择输出，保持生成文本与给定前缀之间的语义一致性和词汇的区分性，并避免模型的退化。实验结果表明，相对于传统的集中搜索，对比搜索能够生成更加准确和合理的电力领域问题，且能够有效地解决电力领域生成文本出现重复词的问题，一定程度上减少高频率文本的重复率[10]，使得模型在各项评估指标都得到了提升，NLU模型在使用模型所生成的文本数据后，模型识别用户意图的准确率有所提高，这也提升智能问答的电力指标检索的准确率。

参考文献：

[1] 田丽，洪福斌. 自然语言处理在电力智能问答领域的应用研究[J]. 科技与创新，2021，176(8)：5-8.

[2] 卢致亮，匡先进，曾赵锦. 基于自然语言处理的电力客服系统研究[J]. 中国新通信，2021，23(17)：129-130.

[3] BAYER M，KAUFHOLD M A，BUCHHOLD B，et al. Data augmentation in natural language processing：a novel text generation approach for long and short

text classifiers [J]. International Journal of Machine Learning and Cybernetics, 2023, 14(1): 135-150.
[4] 赵逸飞. 面向意图识别的数据增强方法研究[D]. 哈尔滨:哈尔滨工业大学, 2021.
[5] 万小军. 智能文本生成:进展与挑战[J/OL]. 大数据:1-14[2023-03-15]. http://kns.cnki.net/kcms/detail/10.1321.G2.20230214.1127.002.html.
[6] 许中卫, 李炜, 吴建国. 集中搜索算法的候选选取方法研究[J]. 计算机工 程 , 2007, 272(4) : 223-224, 227.
[7] SU Y, LAN T, WANG Y, et al. A contrastive framework for neural text generation [EB/OL]. arXiv: 2202.06417, 2022.
[8] FENG X, LIU M, LIU J, et al. Topic-to-essay generation with neural networks [C] //Proceeding of the Twenty-Seventh International Joint Conference on Artificial Intelligence(IJCAI), 2018: 4078-4084.
[9] 刘哲君. 面向开放式文本生成的智能算法研究与实现[D]. 成都:电子科技大学, 2021.
[10] 张兴信, 郭志刚, 陈刚. 受控文本生成技术研究综述[J]. 信息工程大学学 报, 2022, 23(2): 230-238.

Research and implementation of data enhancement for natural language understanding in the electric power field

Shi Junwei, Song Hui*

(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: Explore the data enhancement of Natural-language understanding in domain oriented intelligent question answering. Due to the lack of historical data and the high cost of manual labeling, it cannot meet the needs of large-scale training of natural language understanding models. Therefore, the traditional data enhancement methods are studied, and a keyword text Generative model based on comparative search is proposed to generate a dataset with diverse sentence expressions. Experimental results show that compared with the traditional beam search algorithm, the decoding strategy using comparative search as the model can generate more accurate and reasonable problem texts in the power field, and effectively reduce the word repetition rate of the generated texts. Using these sample data, an efficient and accurate natural language understanding model was successfully trained, which improved the accuracy of user intent recognition. This research has certain reference value for the practical application in the field of intelligent question answering.

Keywords: natural language understanding; text generation; contrastive search; data augmentation; intelligent question-answering

文章编号: 1007-1423（2023）16-0027-06　　DOI: 10.3969/j.issn.1007-1423.2023.16.005

基于图神经网络的学习推荐算法研究

李　月*，李　琳，陈　丽，王槐彬

（广东交通职业技术学院信息学院，广州 510650）

摘要：随着学习者个性化需求的增加和个体差异性的扩大，学习需要解决适应千人千面的个性化学习推荐问题。传统推荐算法主要面向欧式空间数据建模，忽略了现实中的图结构数据。从图神经网络角度出发，应用图卷积方式分别提取用户社交图结构下的用户特征数据，以及知识图谱结构下的知识特征数据，从而形成更加有效的学习推荐内容，实验数据也证明了算法的有效性。

关键词：人工智能；图神经网络；知识图谱；学习推荐

0　引言

随着大数据和人工智能技术的发展，科学技术应用于教育教学也变得越来越广泛。由于当前教育的入口呈现多样化趋势，学习者的差异性也越来越大。对于大学生学习者而言，就存在全国夏季统一高考、单独自主招生、中高职衔接、三二分段等多种多样的入口方式，这也导致同一届学生中的学情存在较大差异。不仅同一届学生的学情呈现多样性，同一门课程中的学生也会因为个人基础、兴趣爱好、接收方式的不同，导致同一批学生对同一门课程的学习产生不同的需求。

学生广泛存在的学情差异，也决定了教育教学资源不能是千人一面发布的静态内容。采用大数据及人工智能技术构建教育资源，可以充分利用技术手段提升学习者对于学习资源的利用效率，同时解决信息过载情况下的知识爆炸问题。同时，信息技术和新媒体的快速发展，也使得学生在信息选择、信息接纳、信息偏好上呈现出“千人千面”，学生所处的学习情境不同，也会对学习资源产生不同的需要。如，当学生处于线下教学时，他们对于学习资源的需求可能更多的是文字性资源，用以临时了解不清晰的知识点；而当学生处于线上教学时，他们可能更需要的是直观的视频和语音资源，同时需要更具真实感、体验感的交流互动方式。这些个性化学习需求也要求教学资源随之做出适应性变革。

如何解决大学生学习者的个性化学习需求，以及解决学习者在海量数据中快速获得有效资源的问题，已存在相关研究，其中推荐系统就是解决方式之一。推荐系统是海量数据时代兴起的一类系统，主要应用于各类电商网站、社交媒体、新闻资讯等系统中，用以解决海量信息中信息过载及长尾物品等问题。推荐系统通过对用户数据、物品数据、交互行为等多方面内容的计算和筛选，形成满足用户个性化需求

收稿日期：2023-04-18　**修稿日期：**2023-05-03

基金项目：广东省教育厅 2021 年教育教学改革研究与实践项目（GDJG2021111）；中国职业技术教育学会 2022 年度科研规划课题（ZJ2022B23）；中国交通教育研究会 2022—2024 年度教育科学研究课题（JT2022YB444）；广东交通职业技术学院 2021 年教研教改项目（GDCP-ZX-2021-005-N3）

作者简介：*通信作者：李月（1979—），女，湖北荆门人，讲师，硕士，研究方向为人工智能，E-mail：liyue@gdcp.edu.cn；李琳（1983—），女，陕西渭南人，硕士，研究方向为网络技术；陈丽（1978—），女，湖南邵东人，副教授，硕士，研究方向为图形图像与大数据；王槐彬（1980—），男，广东潮州人，硕士，讲师，研究方向为人工智能

的结果向用户进行推荐，以提升用户的使用感受、增加用户粘性、提升系统使用效率。

1 相关工作

推荐系统的发展经历了一个从传统的推荐算法到基于深度学习技术的算法的发展过程。传统的推荐算法主要有基于内容的推荐、协同过滤推荐和混合推荐三种模式。基于内容的推荐方法需要提取用户的偏好特征和物品的特征数据，推荐结果极大地依赖于特征信息的选择。协同过滤方法利用用户之间的相似性关系来发现不同用户间可能存在的潜在偏好相似性，但是模型好坏取决于用户评分数据的准确性和完整性，容易遭遇数据稀疏和冷启动问题。混合推荐方法通过融合不同的推荐模型进行推荐，包括特征层面的融合、算法层面的融合以及结果层面的融合多种形式[1]。

随着人工智能的发展，深度学习已经广泛应用于各个领域，并且取得了良好的应用成果。在图像识别、机器翻译、阅读理解、语音合成等方面，深度学习技术的引入都极大地提升了系统性能。基于深度学习的推荐系统一般通过输入层、模型层、输出层的三层架构实现从系统原始数据到输出结果的转换。通过多层感知机、卷积神经网络、自编码网络、循环神经网络等技术实现基于内容的推荐、协同过滤推荐和混合推荐等。总体而言，基于深度学习技术的推荐模型能够有效地融合多源异构数据，结果不依赖于人工选择特征，能实现从多源异构数据到预测的端对端训练，最大限度地发挥了用户的显性数据和隐性数据的价值，学习到数据的非线性的多层次抽象表达，从而有效提升推荐性能。

不管是传统的线性模型还是神经网络模型，所处理的数据主要都是针对欧式空间数据，然而在现实世界中，很多数据都是从非欧式空间数据产生的，例如分子结构的表达、社交网络关系、交通流量网络、人体骨骼结构等，都是具有明显图结构特征的数据类型。在推荐系统中，就存在用户和用户之间的社交关系网络、用户对物品的评价数据网络、物品间的层次网络数据等多种网络数据融合而成的复杂图状网络结构。对于图状结构的数据，因为数据之间不再存在固定不变的关系和位置，并且数据节点的结构不统一，因此目前常见的神经网络模型在处理这部分数据时并不适用。

但是由于图网络中的边信息、节点间的图结构信息等关联信息对于捕获节点之间的隐藏依赖关系、挖掘节点的特征值具有重大作用，因此通过对图结构数据直接进行计算能获得更为优质的推荐结果。

为解决学习中的学习推荐问题，同时结合图神经网络强大的数据建模和特征提取能力，我们提出了一种基于图神经网络的学习推荐模型。

图卷积神经网络是基于深度学习的卷积神经网络在图结构上的推广，它能同时对节点特征信息与结构信息进行端对端学习。图卷积神经网络适用性广，适用于任意拓扑结构的节点与图。在节点分类与边预测等任务上，在公开数据集上的效果要远远优于其他方法[2]。因此，将图卷积神经网络的模型应用于推荐系统，可以取得更好的推荐效果。

和传统的卷积神经网络相比，图卷积网络具有卷积网络的相同性质[3]：

（1）图卷积网络中局部参数共享，卷积算子是适用于每个节点的，算子在不同节点上处处共享。

（2）模型感受野正比于层数，最开始的时候，每个节点包含了直接邻居的信息，再计算第二层卷积时就能把二阶邻居的信息包含进来，这样参与运算的信息就更加充分。模型的感受野与卷积层数成正比，卷积层数越多，参与运算的信息就更多。

同时，图卷积网络同样具备深度学习的性质：

（1）图卷积网络具有层级结构，特征一层一层抽取，一层比一层更抽象，更高级。

（2）非线性变换，增加模型的表达能力。

（3）可实现端对端的训练，不需要定义任何规则，只需要给图的节点一个标记，让模型自己学习融合特征信息和结构信息。

图卷积网络同样具备深度学习的特性。比如，图卷积网络具有层级结构，特征一层一层

抽取，一层比一层更抽象、更高级。

基于图神经网络的特征提取能力，提出的基于图神经网络的学习推荐系统模型如图1所示。首先，采用图卷积方式对用户特征进行提取，然后再采用图神经网络对学习中的知识内容进行图卷积，获得面向知识内容的知识特征，最后根据获得的用户特征向量和知识特征向量进行学习预测，并以此向学习者进行学习推荐。

1.1　学习者建模

与其他系统中的数据不同，推荐系统中的一类重要数据是用户数据信息，而中华文化中对于用户有一句谚语是“物以类聚，人以群分”，也就是说由人构成的数据天然具有群体属性。因此，在推荐系统中，用户具有群体聚集的属性，就像现实世界中，熟悉的人总是更接近一样，他们之间的联系也会越紧密，这种紧密联系会形成图像上的聚集，导致群体的密度升高。同样的道理，学习者形成的社会网络中，也会存在这种群体相似性，也就是通过用户的好友可以刻画当前用户。比如一个热衷于点赞美食标签地点的用户，其好友也可能是喜欢打卡美食地点的人。将用户社交关系图中与当前用户节点有关联的好友用户进行图卷积聚合计算，就可以得到当前用户的隐层信息。虽然图卷积神经网络可以通过堆叠多层卷积层来获得更远用户的信息，但是依据社交理论，距离太远的用户其实与当前用户之间并不相似，因此模型中只需要聚合三跳以内的邻居信息即可。

用户具有非常多的属性特征，如原始属性包含性别、年龄、籍贯、学历等，隐含特征包含短期兴趣、长期兴趣、行为动作、活动轨迹，等等。要想获得学习者的用户特征向量表达，首先需要构建用户和用户之间的社交关系矩阵。并且基于图神经网络提取用户的隐层信息。一层图卷积可以获得当前中心节点一阶领居的节点信息，通过多层图卷积，就可以获得用户的二阶邻居和三阶邻居等信息。根据三度影响力理论，以及图卷积网络堆叠过多层会引起过平滑现象的特性，堆叠至三层即可。计算公式如下：

$$e_u^l = LeakyRelu\sigma AGG\left(\frac{1}{\sqrt{|N_u|}}\sum_{n\in N_u} W_1^l e_n^{l-1}\right) \quad (1)$$

其中，N_u表示用户u在社交图层中的所有一阶邻居，e_u^0是用户社交图层的原始输入，W_1^l是用户社交图层中第l层卷积的权重参数，e_n^{l-1}是用户u在用户社交图层中第$l-1$层卷积的邻居节点。

1.2　知识建模

知识推荐属于推荐系统的一种，Burke[4]认为知识推荐是一种利用用户和产品的知识，采用基于知识的方法生成推荐，推理哪些知识可能满足用户需求的系统。与一般意义上的推荐系统不同，面向知识内容的推荐在推荐内容、推荐目的上都有自己的特性。首先，基于知识的推荐产生的推荐内容是领域知识，这些领域知识之间存在较强的逻辑联系，同时知识背后隐含的信息也非常丰富，这些隐含内容也与知识推荐结果存在关联。同时，面向知识的推荐目的也与一般推荐存在区别。常见的电子商务推荐，一般推荐目的是希望提升用户的点击率和商品的曝光率，从而提升商品销量；社交媒

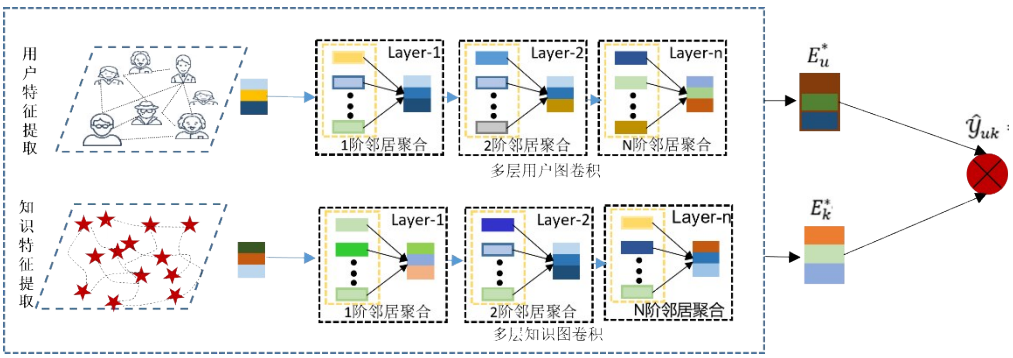


图1　系统模型

体推荐，其推荐目的是命中用户感兴趣的内容，从而提升用户对平台的体验感和忠诚度。与以上目的都不同，面向知识的推荐既不是希望用户购买知识，也不是为了迎合用户的喜好，而是通过感知用户情景，向用户推荐对其当前情境下真正有帮助的知识内容。这些差异都决定了知识推荐的推荐过程和结果与一般推荐系统存在区别。当前知识推荐的方法主要有基于用户情境信息的推荐方法、基于关联规则的推荐方法、基于知识图谱的推荐方法以及基于社会网络信息的推荐方法等[5]。

随着人工智能技术的发展和应用，知识图谱现已被广泛应用于智能搜索、个性化推荐等领域。知识图谱通常以三元组的形式存储实体及其关系，其将现实世界中的知识建模成(k_h, r, k_t)三元组的形式，其中k_h和k_t分别表示头实体和尾实体，r表示实体之间的关系。三元组不仅可以帮助我们理解知识实体之间的关系，也可以存储知识实体的属性。知识图谱中项目之间丰富的语义关联有助于探索它们之间的潜在联系，提高推荐结果的准确性，同时知识图谱中的各种关系有助于合理地扩大用户的兴趣，增加推荐项目的多样性，将知识图谱引入到推荐系统不仅有利于信息的挖掘和推荐结果的发散，还可以增强推荐的可解释性[6]。

知识图谱表示学习不仅要考虑图的结构特征，还需要考虑结点和边的语义类型信息。虽然TransE和DIstmult等知识图谱模型在一定程度上也能捕捉图的结构信息，但是图神经网络对于图结构特征信号考虑更为充分，因此可以利用图神经网络帮助知识图谱表示学习算法更好地捕捉图谱中的结构信息。对于学习系统中的知识特征提取，可采用以下公式计算：

$$e_k^l = LeakyRelu\sigma AGG\left(\frac{1}{\sqrt{|N_k|}}\sum_{n\in N_k} W_1^l e_n^{l-1}\right) \quad (2)$$

其中，N_k表示知识K在知识图谱中的所有一阶邻居，e_k^0是知识层的原始输入，也就是各个知识的知识图谱表示，W_1^l是第l层卷积的权重参数，e_n^{l-1}是知识在知识图谱网络层中第$l-1$层卷积的邻居节点。

模型最后用以上图卷积中得到的e_k^l和e_u^l进行预测，预测用户u对于知识k的评分，计算公式如下

$$\hat{Y}_{uk} = (E_u)^T E_k \quad (3)$$

将用户对于知识的评分进行降序排序，将前n个知识形成知识推荐列表推荐给用户即可。

模型在参数求解时的损失函数采用BPR损失函数，其公式定义为

$$Loss = \sum_{(u,i,j)\in D_s} \ln\sigma(\hat{y}_{ui} - \hat{y}_{uj}) - \lambda_\Theta \|\Theta\|_2^2 \quad (4)$$

其中D_s定义为

$$D_s = \{(u,i,j) | (u,i)\in R^+, (u,j)\in R^-\} \quad (5)$$

其中，R^+表示用户u对知识i的评价高于只是j的数据，R^-表示用户对只是i的评价低于只是点j的数据。$\hat{y}_u$表示用户对知识感兴趣程度的预测得分，λ_Θ表示控制L_2正则化强度以防止过度拟合的参数，Θ表示模型中所有可训练的参数。

2 实验与分析

为检验提出算法的有效性，选取了泛雅学习平台上广东交通职业技术学院《软件测试技术》课程中的学习数据作为实验数据。泛雅平台是超星公司“一平三端”智慧教学系统中的一环，“一平三端”是以在线教学平台为中心，涵盖课前、课中、课后的日常教学全过程，融合教室端、移动端、管理端各类教学应用于一体的信息化教学整体解决方案，即泛雅网络教学平台+手机端(学习通)+教室端(x.chaoxing.com投屏)+管理端(运行数据监控)。泛雅网络教学平台和学习通数据互通共享，已构成一个相对完善的学习网络，学习者之间的关联关系包含好友关注、点赞、讨论、聊天、学习者自行组建学习小组、学习者共读书籍、共同加入同一直播或在线公开课等多种社交行为[7]。以平台上广东交通职业技术建设的《软件测试技术》课程为例，可采集到学生数据867人，该部分学生数据可通过平台提供的身份信息，以及结合广东交通职业技术教务部门的数据信息，获得较为完整的用户属性信息，包括学生年龄、性别、籍贯、所在院系、过往学习成绩等数据，该部分统计信息可以较好地刻画学生的固有属性特征。同时，平台通过数据采集，还可以获得867人参与学习的行为数据和社交数据等。

对于该门课程下的知识内容，可以提取知识点433个，依据三元组和知识本义构建基于知识图谱的知识表达，实验中用到的数据情况见表1。

表 1　实验数据内容

类别	内容
用户数量	867
知识数量	433
用户固有属性	年龄、性别、省份来源、所属院系、年级、政治面貌、民族、培养类型、学生类别
用户社交行为	点赞、转发、提问、回答、学习时长、视频观看次数和时长、作业提交、章节测验、参与讨论、抢答

实验通过观测加入学习推荐前后的数据对比验证学习推荐对于学习者学习质量提升的作用。实验中随机抽取学习者100名，再随机分为两组进行对照。其中A组采用传统的资源学习方式，B组通过推荐算法对学习者提供知识推荐。考虑到学习后进行测验的数据并不能完全真实客观地反映学习者的学习效果，实验中还同步观测了学习者在是否产生学习推荐下的学习过程数据，包括学习总时长，同一知识点学习是否产生回看等。对于A/B对照组学生，从433个知识中随机抽取10个知识进行学习，考察这10个知识点的学习情况，得到的数据见表2。

表 2　实验结果

知识	知识	1	2	3	4	5	6	7	8	9	10
学习时长	A组	7.3	10.5	6.8	9.4	13.6	6.6	15.7	24.3	4.3	17.5
	B组	6.5	10.1	5.1	8.3	10.2	4.6	10.6	15.5	4.5	15.8
回看次数	A组	1	0	0	2	2	1	3	3	0	2
	B组	0	0	0	0	1	0	1	2	0	1
关联知识学习数量	A组	3	4	1	2	3	1	2	3	0	2
	B组	1	2	1	1	1	0	0	1	0	1
成绩	A组	87.4									
	B组	92.3									

通过以上数据对比可知，有推荐组B组学习者在绝大部分的知识学习上的平均学习时长明显低于无推荐组A组，同时可以看到，在知识点复杂的情况下，也就是学习时长较长的情况下，有推荐组的学习时长显著短于无推荐组，表明对于复杂的知识内容，学习推荐可明显缩短学习者对复杂知识的学习时间，显著提升学习效率。同时，有推荐组的学习者因为产生学习困惑而反复学习同一知识点的情况也明显好于无推荐组，表明学习推荐内容可以较好地帮助学习者理解知识，从而减少回看次数和学习时间。从最后的同一测验考察情况来看，有推荐组的学习者在学习效果上也明显高于无推荐组的平均考核成绩，表明学习推荐算法确实能提升学习者的学习效率和内化程度，从而提升学习质量[8]。

3　结语

学习推荐是解决学习者在海量学习数据中产生知识迷航问题的重要手段。在学习中加入学习推荐，可以有效提升学习者的学习效率，提升学习质量。基于人工智能技术在知识学习中产生有效的学习推荐，既可以提升资源利用率，又能有效缩短学习时长。对于学习型社交网络中的推荐问题，除可以采用图神经网络进行数据建模和提取外，如何更加有效地采用其他类型的图结构方法，以及如何更为有效地融合知识谱图技术，都是可以进一步深入研究的方向。

参考文献：

[1] 王一镝.融合卷积神经网络与协同过滤的个性化推荐系统[D].大连:东北财经大学,2021.

[2] 孙水发,李小龙,李伟生,等.图神经网络应用于知识图谱推理的研究综述[J].计算机科学与探索,2023,17(1):27-52.

[3] 吴静,谢辉,姜火文.图神经网络推荐系统综述[J].计算机科学与探索,2022,16(10):2249-2263.

[4] BURKE R. Knowledge-based recommender systems [EB/OL],2000. https://www.researchgate.net/publication/2378325_Knowledge-Based_Recommender Systems.

[5] 程开原,姚俊萍,李晓军,等.时态网络中知识图谱推荐:关键技术与研究进展[J].中国电子科学研究院学报,2021,16(2):174-183,188.

[6] 顾军华,佘士耀,樊帅,等.基于用户长短期兴趣与知识图卷积网络的推荐[J].计算机工程与科学,2021,43(3):511-517.

[7] 赵小伟,裴志朵."一平三端"智慧教学系统在工学一体化课程中的应用[J].电脑与信息技术,2020,28(3):88-90.

[8] 李月,王槐彬.基于改进位置社交的近邻知识推荐算法[J].计算机与数字工程,2020,48(1):34-38.

Research on learning recommendation algorithm based on graph neural network

Li Yue*, Li Lin, Chen Li, Wang Huaibin

(School of Information, Guangdong Communication Polytechnic, Guangzhou 510650, China)

Abstract: With the increase of learners' personalized needs and the expansion of individual differences, learning needs to address the problem of adapting personalized learning recommendations. Traditional recommendation algorithms are mainly oriented to Euclidean spatial data modeling, ignoring the realistic graph structure data. Applying graph neural networks, graph convolution is used to extract user feature data under user social graph structure and knowledge feature data under knowledge graph structure, respectively, so as to form more effective learning recommendation contents, and experimental data also prove the effectiveness of the algorithm.

Keywords: artificial intelligence; graph neural network; knowledge graph; learning recommendation

文章编号: 1007-1423（2023）16-0033-05 DOI: 10.3969/j.issn.1007-1423.2023.16.006

基于改进Cascade RCNN的集成电路板瑕疵检测算法

王代涛，李文杰，谢 波，俞文静*

（广州软件学院网络技术系，广州 510990）

摘要： 针对集成电路板视觉瑕疵检测中存在背景复杂多样、检测目标较小、瑕疵难定位的问题，以及工业生产中满足安全性和高精度的检测要求，提出了一种改进Cascade RCNN算法，并设计了网络模型。该模型以Cascade RCNN为基础检测网络，选用特征提取效果较好的ResNet50作为其骨干网络，并针对集成电路板表面缺陷尺寸变化较大的问题融入了FPN思想，将ResNet50输出的具有大量位置信息的浅层特征层和具有丰富语义信息的深层特征层进行均衡化融合，搭建了一种Cascade RCNN-ResNet-FPN模型架构。实验结果表明，针对集成电路板瑕疵这种小目标检测，改进Cascade RCNN算法相较于基本Cascade RCNN模型、SSD模型以及YOLO模型，安全性以及mAP均有较大的提升，适用于集成电路板瑕疵自动检测。

关键词： 集成电路板；瑕疵检测；Cascade RCNN；小目标；算法

0 引言

瑕疵检测通常是指对物体表面疵点的检测[1-2]。在集成电路板批量生产或者经年使用中，由于各方面因素的影响，会产生诸如漏孔、鼠咬、开路、短路、杂散、杂铜等瑕疵。迄今为止，对集成电路板的瑕疵检测主要还是依赖人工进行实现，存在着效率低、人工成本高和检测标准难以统一等问题。因此，急需通过一些高效、智能化的方案去解决这一问题。计算机视觉拥有高效、稳定、可长时间工作等特性，以计算机视觉替代人工进行瑕疵检测获得越来越多的认可。

近年来，物体表面缺陷检测已经成为深度学习领域的热门研究课题。然而，在集成电路板视觉检测中，由于主板图像背景复杂多样，检测目标小、瑕疵难定位，再加上工业生产中要求满足安全性和高精度检测，现有的很多算法模型并不适用于此。集成电路板行业需要既能满足检测精度和安全要求，又能在复杂背景下进行瑕疵检测的高效算法模型。

针对此问题，本文选用Cascade RCNN为基础检测网络，并选用特征提取效果较好的ResNet50作为其骨干网络，针对集成电路板检测目标较小的问题加入了FPN网络，将ResNet50输出的具有大量位置信息的浅层特征层和具有丰富语义信息的深层特征层进行均衡化融合，以提高检测质量与效率。最后得到的模型体积约为68.94 M，在实验环境下，其FPS为16.1、mAP和mAR高达93.8%与98.8%。可安全、有效地实现对集成电路板的瑕疵检测。

本文克服了Cascade RCNN算法对小目标和重叠度高的缺陷识别率低的问题，提出了一种基于Cascade RCNN的改进算法，使用ResNet50作为Cascade RCNN算法的骨干网络，并融入基于FPN网络改进的多尺度特征融合均衡化网络，搭建了Cascade RCNN-ResNet50-FPN网络模型，以下称为改进Cascade RCNN，整个算法模型检测精度、召回率较高，既精确又安全，可满足

收稿日期： 2023-04-07 **修稿日期：** 2023-07-24

基金项目： 2021年国家级大学生创新创业训练计划项目（DCXM2021077）；2022广东省科技创新战略专项基金攀登计划项目（pdjh2022b0653）

作者简介： 王代涛（2000—），男，湖南长沙人，本科生，主要研究领域为人工智能、机器学习、计算机视觉技术等；李文杰（2002—），男，广东湛江人，本科生，主要研究领域计算机视觉、数据分析等；谢波（2002—），男，广东河源人，本科生，主要研究领域计算机视觉、数据分析、系统开发等；*通信作者：俞文静（1982—），女，湖南湘潭人，硕士，副教授，主要研究领域为计算机图形图像处理、算法优化研究等，E-mail：judily0612@qq.com

工业集成电路板表面缺陷检测的需求。

1 相关基础理论

1.1 残差神经网络

ResNet50是经典的运用残差单元结构组成的网络，是在2016年由He等[3]共同提出，旨在解决因卷积神经网络不断迭代而可能导致的梯度消失和网络退化的问题。其网络层次如表1所示。

表1 ResNet50网络结构

Layer name	Output size	50-Layer
Conv1	112 × 112	7 × 7，64，stride=2
Conv2_x	56 × 56	3 × 3，max pool，stride=2 $\begin{bmatrix}1\times1, & 64\\3\times3, & 64\\1\times1, & 256\end{bmatrix}\times3$
Conv3_x	28 × 28	$\begin{bmatrix}1\times1, & 128\\3\times3, & 128\\1\times1, & 512\end{bmatrix}\times4$
Conv4_x	14 × 14	$\begin{bmatrix}1\times1, & 256\\3\times3, & 256\\1\times1, & 1024\end{bmatrix}\times6$
Conv5_x	7 × 7	$\begin{bmatrix}1\times1, & 512\\3\times3, & 512\\1\times1, & 2048\end{bmatrix}\times3$
	1 × 1	Average pool，1000-d fe，softmax
FLOPs		3.8×10^9

ResNet50包括一个步长为2的7 × 7卷积层、四个残差单元级联卷积层和一个全连接层。其中残差单元是由一个3 × 3的卷积核、两个1 × 1的卷积核和一个残差连接分支组成，其结构如图1所示。

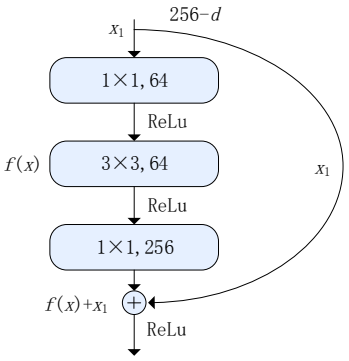


图1 ResNet50网络的残差单元结构

其中：x_1为输入，其期望输出为$h(x)$，但加入一个恒等映射x_1后原始学习特征变成$h(x)=f(x)+x_1$，因此残差是指$f(x)=h(x)-x_1$，使得网络拟合更容易，直接映射的加入也使得网络的下一层产生比前一层的图像信息多[2-3]。ResNet50计算量适中，能提取更多小目标特征信息。

1.2 多尺度特征融合均衡化网络

当前，FPN[4]是目标检测算法中最常用的特征融合网络，具体实现方式如图所示。FPN是在特征提取网络生成的多尺度特征图的基础上，将最后一层特征图做2倍上采样后与前一层特征图融合生成新的特征图，然后将新生成的特征图以同样的方式与前一层特征图进行融合，逐层重复以上操作，形成特征金字塔。但是特征金字塔网络长距离信息流动会导致缺陷信息的流失和更多的关注相邻分辨率，这种融合方式会产生缺陷特征信息不平衡问题。因此，本文基于FPN提出了一种多尺度特征融合均衡化网络，利用相同深度融合的平衡语义特征来增强多层次的特征。首先将通过FPN网络形成的每层特征图通过上采样操作变成相同尺度，然后进行像素级的聚合均值，再将聚合均值化的特征图通过与之前相反的采样操作形成不同尺度的特征图，从而得到缺陷特征信息更平衡的多尺度特征图，实现缺陷特征的多层复用和融合均衡化，提高小目标缺陷敏感度[5]。

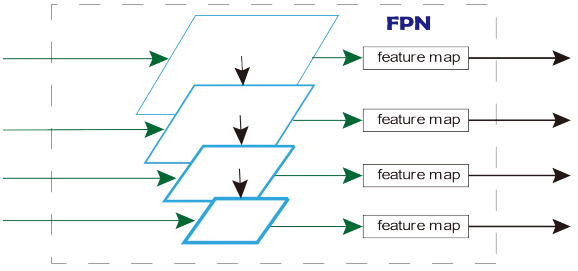


图2 多尺度特征融合均衡化网络结构

1.3 Cascade RCNN

Cascade RCNN是一种提高目标检测效果的方法，其关键在于采用4次框回归。其通过将RPN[4]输出的检测框输入回归模块（如图4所示），并将其输出作为下一层的输入，重复该过程以提高IoU阈值，从而提升预测框的定位效果。框回归模块通过生成d_x、d_y来调整矩形框

的位置，以及生成 d_w、d_h 来调整矩形框的长宽，即 Cascade RCNN 通过级联多个框回归模块来增强线性回归的能力，从而改善检测效果。

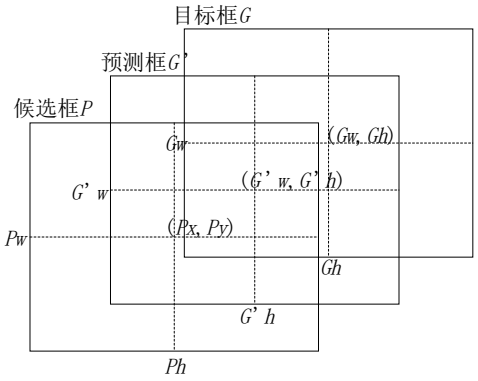


图 3　边框回归

2　本文方法

2.1　改进的 Cascade RCNN 检测算法

针对 Cascade RCNN 算法在进行缺陷检测时仍存在对小目标缺陷漏检的问题，提出了一种改进的 Cascade RCNN 算法。改进方式主要有：使用分类效果更好的 ResNet50 作为 Cascade RCNN 骨干网络，可提取更多的缺陷特征信息；借鉴 FPN 的思想，融合浅层特征的位置信息和深层特征的语义信息，可提高小目标的识别率。改进后的 Cascade RCNN 网络结构如图 4 所示。

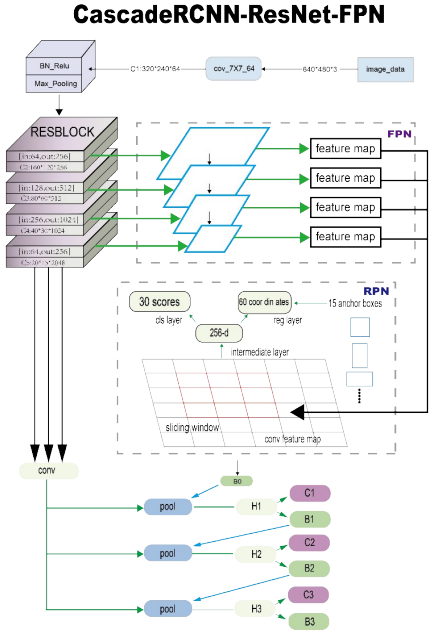


图 4　改进后的 Cascade RCNN 网络结构

2.2　Cascade RCNN-ResNet50-FPN 组合

本文网络模型的基本检测流程是通过 ResNet50 提取特征信息，再使用 FPN 融合各尺度特征图，生成语义与全局信息更多的特征图组，RPN 按尺度在对应的特征图上生成候选框，最终由多个 ROI Align 对候选框进行多次的框回归。

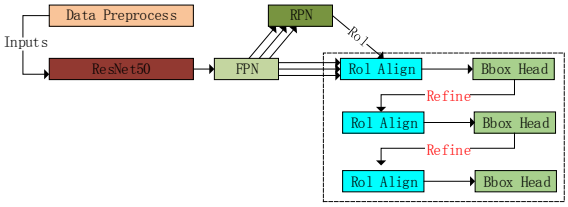


图 5　Cascade RCNN-ResNet50-FPN 结构

骨干网络部分采用 ResNet50，赋予模型强大的特征提取能力。ResNet50 最初作用是用来做分类网络的，要使其运用在目标检测中充当骨干网络，只需要将最后一层的全连接层删除，并将四个大块的特征图输入 FPN 中，最后从 RPN 中将预测框与 FPN 得到的特征图传入 Cascade RCNN 即可完成目标检测网络的搭建。

改进 Cascade RCNN 这一检测方法的优点是通过级联多个渐增阈值的框回归模块使其可以将原本低质量的检测框优化为高质量的检测框，对小缺陷目标更为敏感，提升了检测的准确率和召回率。

3　数据集

本文的集成电路板数据集来源于北京大学智能机器人开放实验室[5]，其中包含 1386 张缺陷图像，以 8∶2 的比例分为训练集与验证集，其主要由六种缺陷构成：漏孔、鼠咬、开路、短路、杂散、杂铜。各类缺陷尺寸相对较小，常用算法难以在该数据集上取得良好结果。

针对本文的模型，模型训练时对数据集图像进行了如下的处理：

（1）随机裁剪：对图片进行随机裁剪以提高模型鲁棒性；

（2）随机翻转：图片以 0.5 的概率进行水平、上下翻转；

（3）标准化：将图像数据进行标准化处理，既减去均值[123.675, 116.28, 103.53]，再除以标

准差[58.395, 57.12, 57.375]（如公式1所示）。

$$X_{std} = \frac{X - \mu}{\sigma} \quad (1)$$

式中：X为原始数据集中的样本特征，X_{std}为标准化后的样本特征，μ为该特征的均值，σ为改特征的标准差。

4 实验

4.1 参数细节

实验采用PyTorch框架[6]，主要CPU、GPU运算设备分别为Xeon（R）Platinum 8255C、Nvidia TeslaV100 * 8，采用SGD[7-8]作为优化器，初始学习率为0.0025，动量与权重衰减系数设置为0.0001，共50轮训练，其中前5轮为预热阶段，到达40与45轮时，学习率下降到之前的1/10。

本文采用ResNet-50作为骨干网络，颈部网络使用FPN，RPN头部与RoI头部均使用交叉熵损失函数与平滑L1损失函数作为分类损失函数与回归损失函数。此外，本文还使用相同参数训练了基础Cascade RCNN、SSD、YOLO，与我们的模型进行对比实验。

4.2 实验步骤

在进行缺陷检测的实验时，对Cascade RCNN缺陷检测算法进行改进，由训练至测试得到检测结果，其具体实施步骤为：

（1）将训练集作为网络输入，利用骨干网络ResNet50提取特征，其4个卷积层的输出作为多尺度特征图；

（2）在均衡化网络中对多尺度特征图进行特征融合以输出具有更加丰富和平衡信息的特征；

（3）将所述多尺度特征图送入共享RPN网络中生成建议框；

（4）将生成的缺陷建议区域和ResNet50网络的最后输出特征图结合，经池化运算，获取每类缺陷候选区域特征图；

（5）输出特征图输入到Cascade RCNN Head中后，获得每一类缺陷的分类信息及位置信息，经过非极大值抑制，选择最佳预测框；

（6）重复上述步骤至模型收敛后完成训练；

（7）对缺陷检测模型权重参数进行提取；

（8）对所述测试集使用检测模型进行检测，获取每类缺陷的检测结果以及定位结果。

5 结果分析

将验证集带入模型进行计算，结果如图6所示，相关实验运算结果见表2。

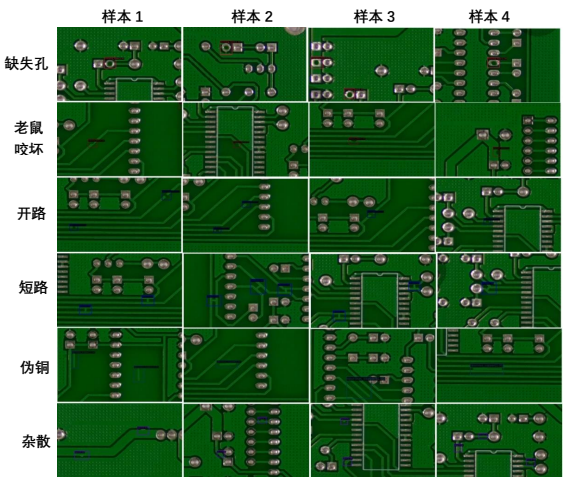


图6 检测结果

表2 实验数据指标对比

Model	#Param./M	FLOPs/G	FPS	mAP/%	mAR/%
Basic Cascade_RCNN	57.63	75.69	17.2	79.0	83.6
our	68.94	80.17	16.1	93.8	98.8
SSD	3.10	2.82	69.9	58.3	60.3
YOLO	3.67	6.58	48.1	49.3	53.6

通过分析可以得出以下结论：①本文模型对CascadeRCNN的改造是非常有效的，其参数量与计算量在增加不大的情况下，mAP从79.0%提升至93.8%，mAR从83.6%提升至98.8%，使得网络更加精准与安全；②本文模型对小缺陷目标检测相对SSD、YOLO与基础Cascade RCNN更为优越，能够较为精准地检测出数据集中非常小的缺陷目标，非常适合于PCB行业的工业检测。

6 结论与未来工作

本文针对集成电路板中小目标缺陷检测，以及基本Cascade RCNN检测算法对小目标检测能力弱的问题，提出了一种改进Cascade RCNN，通过ResNet50取代原网络的骨干网络VGG16，来提高网络模型的特征提取能力；借鉴FPN思想，提高多尺度特征融合均衡化能力，提高了

对小目标检测对象的敏感度；采用多个框回归模块级联来增强线性回归的能力。采用北京大学智能机器人开放实验室提供的集成电路板数据集进行实验研究，实验结果表明，本文所提出的改进Cascade RCNN算法模型对小目标检测对象有较高的检测效果和检测效率，且模型参数规模较小，适合部署与运行在高端FPGA设备上，可满足集成电路板表面缺陷检测的需求。

未来工作将在介绍改进后的Cascade RCNN检测算法，并将其应用在高端FPGA设备的基础上进行铝片表面工业缺陷检测。首先，对Cascade RCNN算法的两个改进部分的改进目的及改进方式进行详细的分析；然后，介绍该模型训练到测试获取检测结果的具体实现步骤等。

参考文献：

[1] 刘瑞明，孙帅成，黄佳炜，等. 基于图像纹理分析的布匹瑕疵检测综述[J]. 江苏海洋大学学报(自然科学版)，2020，29(2)：86-93.

[2] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition，2014：580-587.

[3] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)，2016：770-778.

[4] LIN T-Y，DOLLAR P，GIRSHICK R，HE K M，et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，2017：2117-2125.

[5] 郭同建，徐洋，陈慧敏，等. 基于深度学习的机织印花布疵点实时检测方法研究[J]. 东华大学学报(自然科学版)，2023，49(1)：33-38.

[6] PASZKE A，GROSS S，MASSA F，et al. PyTorch：an imperative style，high-performance deep learning library[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS)，2019：8024-8035.

[7] 戴林辉. PKU-Market-PCB：a benchmark dataset for PCB defect detection[DS]. HRI实验室PKU-Market系列数据集，2021. https://robotics.pkusz.edu.cn/resources/dataset/.

[8] REN S Q，HE K M，GIRSHICK R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[C]//Proceedings of the Advances in Neural Information Processing Systems (NIPS)，2015：91-99.

A PCB defect detection algorithm based on improved Cascade RCNN

Wang Daitao, Li Wenjie, Xie Bo, Yu Wenjing*

(Department of Network Technology, Software Engineering Institute of Guangzhou, Guangzhou 510900, China)

Abstract: Addressing the problems of complex and diverse backgrounds, small detection targets, and difficult defect localization in integrated circuit board visual defect detection, as well as the requirements for safety and high accuracy in industrial production, this paper proposes an improved Cascade RCNN algorithm and designs a network model. The model is based on the Cascade RCNN detection network and uses ResNet50 as the backbone network, which has better feature extraction performance. To address the problem of large variation in defect sizes on the surface of integrated circuit boards, FPN ideas are integrated, and the shallow feature layers with a large amount of position information and the deep feature layers with rich semantic information are balanced and fused. A Cascade RCNN-ResNet-FPN model architecture is constructed. Experimental results show that for small target detection of integrated circuit board defects, the improved Cascade RCNN algorithm proposed in this paper has significant improvements in safety and mAP compared with the basic Cascade RCNN model, SSD model, and Yolo model, and is suitable for automatic detection of integrated circuit board defects.

Keywords: PCB(printed circuit board); defect detection; Cascade RCNN; small target; algorithm

文章编号: 1007-1423（2023）16-0038-05　　DOI: 10.3969/j.issn.1007-1423.2023.16.007

基于多目标优化的相邻交叉口协调控制模型研究

汪　红*，何怡玲，代秋香

（湖北大学计算机与信息工程学院，武汉 430000）

摘要：针对道路车辆不断增加而导致交通拥堵的问题，提出一种基于多目标优化的相邻交叉口协调控制模型。首先根据车流的离散性，引入相位差协调控制，构建相邻交叉口多目标优化问题。接着，提出修正内点法进行求解得到目标优化后的配时策略。最后，根据实际的交通数据进行仿真。仿真结果表明，与使用遗传算法相比，采用修正内点法能得到Pareto最优解。并且所提的配时方案，能够降低平均延误和减少停车次数，提高了通行效率并缓解了交通压力。

关键词：协调控制；修正内点法；相邻交叉口

0　引言

城市道路上车辆的增加，使交通拥堵的状况越来越严重，其中交叉口是城市交通拥堵的多发地，采取合理的信号灯配时能够缓解这种状况。

近年来，国内外很多专家学者对信号灯配时进行深入研究。由于城市主干道交通流量大，为了提高干线的通行效率，信号协调控制应用在交通系统中。文献[1]提出一种基于车辆轨迹采样的干线协调控制模型，选择所有采样轨迹在主干道上的延误之和为优化目标，并使用粒子群算法进行求解。文献[2]提出一种基于双层优化的交通信号控制策略，以车辆排放、延误最少和交通流均匀分布为目标，实现对主干道的协调控制，并采用遗传算求解。文献[3]提出一种基于V2X的多交叉口协调算法，通过交通信息分配绿灯时间，以最大优化干道上的绿波段，进而实现车辆流畅通过交叉口。然而，干线协调控制有效提高了主干路的交通效率，但是忽略了支路的通行效率。文献[4]考虑支路的绿波和行人过街时间，提出一种基于改进的遗传算法的综合绿波协调控制模型。上述文献求解多目标优化问题大多采用智能算法，在保证解的最优性方面有一定局限性。文献[5]从不同交通流的角度对车辆延误进行研究，提出一种基于双层规划的协调优化模型，以车辆延误和行人延误为优化目标。文献[6]考虑关联交叉口间车流、路段距离、信号配时方案等因素，研究了一种基于稳态理论的“相位差-延误”模型，定量分析交叉口相位差与延误之间的关系。文献[7]基于V2X技术，以交叉口负荷、行程时间和舒适性为优化目标，提出一种基于NSGA-II算法的相邻交叉口多目标协同优化模型。这些研究发现，协调控制能够实现缓解交通拥堵的目的，但是很少考虑交叉口间车流的离散特性。

综上，本文研究在低饱和度交通状态下，考虑车流的离散性，引入相位差协调控制，提出一种相邻交叉口多目标协调控制模型，以车辆延误和停车率为评价指标来衡量交通效果，然后利用修正内点法对模型进行求解。

1　多目标优化模型的建立

考虑相邻两个交叉口的车流特点，将相邻两个交叉口标定为外部进口道和内部进口道进行研究[8]。其中，内部进口道优化模型引入相

收稿日期：2023-05-08　　**修稿日期：**2023-05-19

作者简介：*通信作者：汪红（1998—），女，湖北荆州人，硕士研究生，主要研究方向为交通信号控制，E-mail：2013457187@qq.com；何怡玲（1998—），女，湖南永州人，硕士研究生，主要研究方向为MIMO无线通信、协作通信、导频污染等；代秋香（1996—），女，湖北随州人，硕士研究生，主要研究方向为智能反射面辅助通信

位差协调控制，构建车辆延误模型和停车率模型，外部进口道优化模型采用车辆延误和停车次数作为评价指标，进而构建相邻交叉口多目标协调控制模型。

1.1　内部进口道优化模型

受上游交叉口信号灯控制影响，离散车流到达下游交叉口遇到红灯排队而产生延误，从而导致车队尾部受阻或者头部受阻。其中，为了便于协调控制，将两个相邻交叉口的周期和绿信比设置为相同。考虑两个相邻的交叉口信号对车流离散行为的影响，采用Roberston车辆离散模型[9]，来预测计算车流从上游抵达下游交叉口的到达率。

1.1.1　车队尾部受阻

上游交叉口的车辆以饱和流率S_1驶离停车线，在交叉口间发生离散行为，当车队在下游交叉口正好赶上绿灯时，车队以到达率q_2驶离交叉口，车队尾部部分车辆被截断，需要排队等待下一个绿灯时间。累积到达车辆数与延误时间的关系如图1所示。

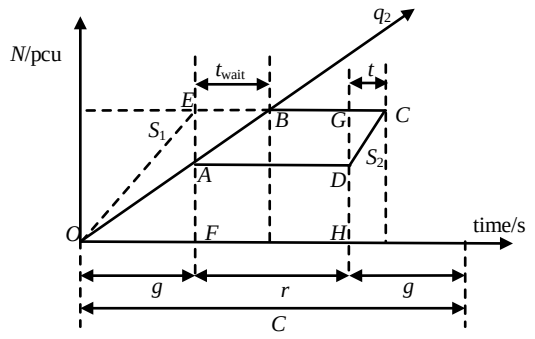


图 1　车队尾部受阻延误

其中：t_{wait}表示车队尾部车辆等待红灯的时间，s；t表示车队尾部车辆驶离交叉口所耗用的时间，s；S_1表示上游交叉口的驶离率，pcu/s；q_2表示下游交叉口的到达率，pcu/s；S_2表示下游交叉口的驶离率，pcu/s；r表示有效绿灯间隔时间，s；g表示有效绿灯时间，s；C表示周期时长，s。

两交叉口之间的协调相位差为

$$\frac{x}{v_f}-t_{\text{wait}}=\varnothing_{\text{up, down}} \tag{1}$$

其中：x表示相邻交叉口的间距，m；v_f表示路段间的平均行驶速度，m/s；t_{wait}表示车队尾部车辆等待红灯的时间，s；$\varnothing_{\text{up, down}}$表示上游到下游的相位差，s。

根据车辆到达和驶离过程，$S_{\square ABCD}$的面积为一个周期内车队尾部受阻的总延误：

$$\begin{aligned}D_1&=S_{\square ABCD}=\frac{1}{2}(BG+AD)AE+\frac{1}{2}tAE\\&=g(C-g)(S_1-q_2)-\frac{g^2(S_1-q_2)^2}{2q_2}+\frac{g^2(S_1-q_2)^2}{2S_2}\end{aligned} \tag{2}$$

一个周期内车队尾部受阻的平均延误：

$$\begin{aligned}\overline{D_1}&=\frac{D_1}{g(S_1-q_2)}\\&=C-g-\frac{g(S_1-q_2)}{2q_2}+\frac{g(S_1-q_2)}{2S_2}\end{aligned} \tag{3}$$

根据式(2)可以得到一个周期内车队尾部受阻的停车率：

$$h_1=\frac{AE}{EF}=\frac{g(S_1-q_2)}{q_3C} \tag{4}$$

1.1.2　车队头部受阻

车队在下游交叉口遇到红灯排队，当绿灯亮起时，排队累积的车辆以饱和流率S_2驶离交叉口。累积到达车辆数与延误时间的关系如图2所示。

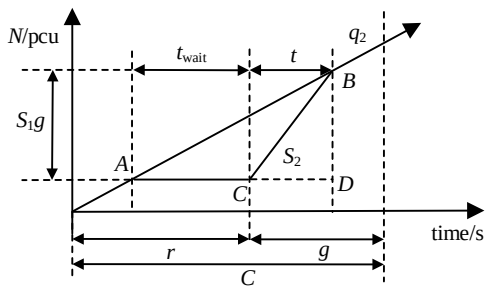


图 2　车队头部受阻延误

其中：t_{wait}表示车队头部车辆等待红灯的时间，s；t表示车队头部车辆驶离交叉口所耗用的时间，s。其它各参数的意义同上。

两交叉口之间的协调相位差为

$$\frac{x}{v_f}+t_{\text{wait}}=\varnothing_{\text{up, down}} \tag{5}$$

根据车辆到达和驶离过程，$S_{\triangle ABC}$的面积为一个周期内车队头部受阻的总延误：

$$D_2 = S_{\triangle ABC} = \frac{1}{2} AC \times BD = \frac{g^2 {S_1}^2 (S_2 - q_2)}{2S_2 q_2} \quad (6)$$

一个周期内车队头部受阻的平均延误：

$$\overline{D_2} = \frac{D_2}{gS_1} = \frac{gS_1(S_2 - q_2)}{2S_2 q_2} \quad (7)$$

根据式(6)可以得到一个周期内车队头部受阻的停车率：

$$h_2 = \frac{S_1 g}{q_3 C} \quad (8)$$

1.1.3 车辆随机到达受阻

考虑车辆随机到达产生的延误，采用*Webster*延误模型中的随机延误公式，得到车辆随机到达的平均延误为

$$\overline{D_3} = \frac{x^2}{2q(1 - x)} \quad (9)$$

其中：x表示下游交叉口的饱和度；q表示下游交叉口的到达率，单位为pcu/s。

通过上述研究分析，得到车队均匀到达时，$x/v_f \geqslant \varnothing_{up,down}$，车队尾部受阻；$x/v_f < \varnothing_{up,down}$，车队头部受阻。

$$\gamma = \begin{cases} 1, & \dfrac{x}{v_f} \geqslant \varnothing_{up,down} \\ 0, & \dfrac{x}{v_f} < \varnothing_{up,down} \end{cases} \quad (10)$$

因此相邻交叉口内部进口道的平均延误为

$$\overline{d_{in}} = \gamma \overline{D_1} + (1 - \gamma) \overline{D_2} + \overline{D_3} \quad (11)$$

相邻交叉口内部进口道的停车率为

$$\overline{h_{in}} = \gamma h_1 + (1 - \gamma) h_2 \quad (12)$$

1.2 外部进口道优化模型

本文是在低饱和交通状态下进行研究，延误模型采用Webster延误模型[10]：

$$\overline{d_{out}} = d_{ij} = \frac{C(1 - \lambda_{ij})^2}{2(1 - \lambda_{ij} x_{ij})} + \frac{{x_{ij}}^2}{2q_{ij}(1 - x_{ij})} \quad (13)$$

其中：$\overline{d_{out}}$表示外部进口道的平均延误，s·pcu^{-1}；d_{ij}表示i相位j进口道每辆车的平均延误时间，s·pcu^{-1}；C表示交叉口的一个信号周期，s；λ_{ij}表示i相位j进口道的绿信比；x_{ij}表示i相位j进口道的饱和度；q_{ij}表示i相位j进口道的车流量，pcu/s。

根据Webster公式，可以得出一个信号周期内交叉口车辆的停车率[11]为

$$\overline{h_{out}} = h_{ij} = \frac{1 - \lambda_{ij}}{1 - y_{ij}} \quad (14)$$

其中：$\overline{h_{out}}$表示外部进口道的停车率；h_{ij}表示i相位j进口道的停车率；C表示交叉口的一个信号周期；λ_{ij}表示i相位j进口道的绿信比；y_{ij}表示i相位j进口道的流量比。

通过对相邻交叉口多目标优化问题进行研究，得到相邻交叉口的平均延误模型和停车率模型，如式(15)和(16)所示。

$$\bar{D} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\overline{d_{in}} q_{in} + \overline{d_{out}} q_{out})}{\sum_{i=1}^{n}\sum_{j=1}^{m}(q_{in} + q_{out})} \quad (15)$$

$$\bar{H} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\overline{h_{in}} q_{in} + \overline{h_{out}} q_{out})}{\sum_{i=1}^{n}\sum_{j=1}^{m}(q_{in} + q_{out})} \quad (16)$$

其中：$\bar{D}$、$\bar{H}$分别表示相邻交叉口的平均延误时间、停车率；$\overline{d_{in}}$表示内部进口道的平均延误，s·pcu^{-1}；$i = 1, 2, \cdots, n$表示交叉口的相位数；$j = 1, 2, \cdots, m$表示进口道的方向；$\overline{d_{out}}$表示外部进口道的平均延误，s·pcu^{-1}；q_{in}、q_{out}分别表示内部进口道、外部进口道对应的车流量，pcu；$\overline{h_{in}}$表示内部进口道的停车率；$\overline{h_{out}}$表示外部进口道的停车率。

因此多目标优化模型为

$$\min f = \left(\bar{D}, \bar{H}\right)^T$$
$$\text{s.t} \begin{cases} g_{min} \leqslant g_i \leqslant g_{max} \\ C_{min} \leqslant C \leqslant C_{max} \\ \sum_{i=1}^{n} g_i = C - L \end{cases} \quad (17)$$

其中：g_i表示i相位的有效绿灯时间；g_{min}、g_{max}分别表示有效绿灯时间的最小值和最大值；C_{min}、C_{max}分别表示周期的最小值和最大值；n表示交叉口的相位数；L表示一个周期的损失时间。

2 多目标优化模型的求解

为了使提出的多目标优化模型达到最优，采用最短距离评价函数来构建单目标优化问题[12]，并采用归一化法对目标函数进行无量纲化处理，具体如式(18)所示。

$$\min f = \left(\frac{\bar{D} - \overline{D_{min}}}{\overline{D_{max}} - \overline{D_{min}}}\right)^2 + \left(\frac{\bar{H} - \overline{H_{min}}}{\overline{H_{max}} - \overline{H_{min}}}\right)^2 \quad (18)$$

式(18)是一个多变量约束的非线性规划问题，采用对数障碍函数内点法进行求解。

3　数据仿真分析

3.1　仿真数据

为了验证提出的多目标优化模型的有效性，以文献[13]相邻两个交叉口的交通数据为例，信号控制相位均为四相位，其中第一相位是东西方向直行和右转，第二相位是东西方向左转，第三相位是南北方向直行和右转，第四相位是南北方向右转。

3.2　模型验证分析

本文利用Matlab仿真平台分别实现修正内点法和遗传算法[14]对多目标优化模型的仿真分析。结合实例数据，周期取60～150 s，各个相位的有效绿灯时间最大为60 s，由于各个路口车流量不同，第一相位和第三相位的有效绿灯时间最小为30 s，第二相位和第四相位的有效绿灯时间最小为20 s。同时经过多次实验分析，确定遗传算法的种群大小为500，交叉概率为0.4，变异概率为0.1。

为了验证模型的有效性，将本文提出的模型作为模型一，与模型二：由式(13)、(14)建立的不考虑相位差的多目标优化模型进行比较。利用修正内点法，分别对模型一和模型二进行模拟仿真，两个模型的优化效果如图3所示，两个模型的仿真结果见表1。

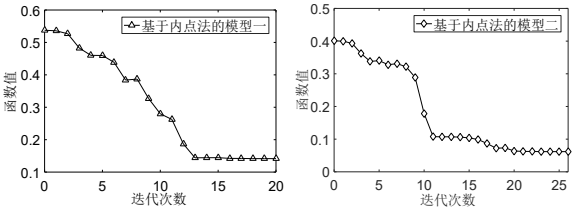


图3　两个模型的优化效果

表1　两个模型的仿真结果

模型	平均延误/s	停车率
模型一	39.7535	0.7930
模型二	44.8504	0.8836

从图3和表1可以看出，通过修正内点法进行优化，两个模型的优化结果都达到收敛。其中，采用本文提出的第二种模型时，车辆平均延误时间和停车率分别为44.8504 s和0.8836。采用本文提出的第一种模型，相应的结果分别为39.7535 s和0.7930，车辆平均延误降低了11.4%，停车率减少了10.3%。由此可以得出，本文提出的基于相位差协调的多目标优化模型是合理的。

3.3　算法验证分析

本次实验用于验证相比遗传算法，修正内点法对基于相位差协调的多目标优化模型的控制效果更优。两种算法对多目标模型的优化效果如图4所示，两种算法的优化结果见表2。

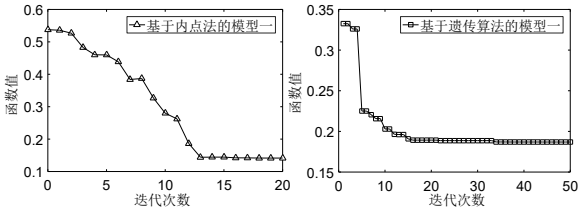


图4　两种算法对多目标模型的优化效果

表2　两种算法的优化结果

配时方法	平均延误/s	停车率
内点法	39.7535	0.7930
遗传算法	40.4246	0.7940

从图4和表2可以看出，修正内点法和遗传算法分别对基于相位差协调的多目标优化模型进行优化，结果都具有收敛性。与采用遗传算法相比，采用修正内点法得到的优化结果更理想，优化后相邻交叉口的平均延误时间降低1.7%，停车率减少0.1%。由此可以得出，相比遗传算法，修正内点法在进行相邻交叉口信号多目标优化配时时能够得到相对更优的控制效果。

4　结语

本文根据相邻交叉口车流的动态特点，将相邻两个交叉口标定为外部进口道和内部进口道。以平均延误和停车率为优化目标，提出外部进口道多目标优化模型。引入相位差协调控制，构建内部进口道的平均延误模型和停车率模型，进而提出一种基于多目标优化的相邻交叉口协调控制模型，并且通过修正内点法仿真验证相邻交叉口多目标优化模型和优化算法的有效性。实验结果证明，与遗传算法相比，采

用修正内点法进行优化，相邻交叉口的车辆平均延误降低、停车率减少，一定程度上提高了交叉口的通行效益。本文基于相邻交叉口进行多目标优化协调控制研究，后续研究将考虑区域路网构建相应的多目标优化模型并仿真验证模型的性能。

参考文献：

[1] YAO J R, TAN C P, TANG K S. An optimization model for arterial coordination control based on sampled vehicle trajectories: the STREAM model[J]. Transportation Research Part C: Emerging Technologies, 2019, 109: 211-232.

[2] ZHOU S P, LI H R, ZHANG Q Y, et al. Traffic signal coordination control optimization considering vehicle emissions on urban arterial road[J]. Journal of Computational Methods in Sciences and Engineering, 2019, 21(1): 1-7.

[3] SHI Y J, LI J J, HAN Q M, et al. A coordination algorithm for signalized multi-intersection to maximize green wave band in V2X network[J]. IEEE Access, 2020, 8: 213706-213717.

[4] ZHANG J, SHANG H Y, LI X X, et al. An integrated arterial coordinated control model considering green wave on branch roads and pedestrian crossing time at intersections[J]. Journal of Management Science and Engineering, 2020, 5(4): 303-317.

[5] LI D W, SONG Y C, CHEN Q. Bilevel programming for traffic signal coordinated control considering pedestrian crossing[J]. Journal of Advanced Transportation, 2020, 2020: 1-18.

[6] 刘庆广，顾玉牧，林培群. 关联交叉口延误计算模型研究[J]. 重庆交通大学学报(自然科学版)，2021，40(3)：22-26.

[7] 赵红专，吴浩，卢宁宁，等. V2X环境下主干路相邻交叉口多目标协同优化模型研究[J]. 计算机应用研究，2022，39(10)：3003-3007.

[8] CAO H W, LUO J. Coordinated optimization control of regional traffic signals based on vehicle average delay model[C]//Proc. of the 2nd International Conference of Intelligent Robotic and Control Engineering, Los Alamitos, CA: IEEE Computer Society, 2019: 72-77.

[9] 严丽平，张默可，郭成源，等. 基于量子粒子群算法的实时多交叉口信号控制[J]. 计算机仿真，2021，38(10)：180-184.

[10] LI Y, QIN Z, ZHU C M. Optimal design of transportation signal control at the intersection based on Webster signal timing method[J]. Journal of Physics: Conference Series, 2021, 1972(1): 012130.

[11] 陈廷伟，高研. 面向交通信号优化改进快速非支配排序遗传算法研究[J]. 计算机应用研究，2018，35(5)：1320-1324.

[12] 邓兰梅，黄天民. 一种融合相对有效性的两阶段理想点法[J]. 运筹与模糊学，2017，7(4)：110-137.

[13] 张兰，雷秀娟，马千知. 基于粒子群优化算法的多交叉口信号配时[J]. 计算机应用研究，2010，27(4)：1252-1254.

[14] 张丽莉，高雪溢. 考虑交叉口延误的交通微循环路网优化研究[J]. 重庆交通大学学报(自然科学版)，2021，40(12)：27-32.

Coordinated control model of adjacent intersection with multi-objective optimization

Wang Hong*, He Yiling, Dai Qiuxiang

(School of Computer Science and Information Engineering, Hubei University, Wuhan 430000, China)

Abstract: The increasing number of vehicles has led to traffic congestion on the road. In this paper, we propose a coordinated control model of adjacent intersection with multi-objective optimization at to solve the traffic congestion. Firstly, the phase difference coordination control is introduced to construct a multi-objective optimization problem of adjacent intersections according to the discrete nature of the traffic flow. And then, the modified interior point method is also presented to obtain a target-optimized timing strategy. Finally, simulation based on actual traffic data. The simulation results show that the modified interior point method can obtain a Pareto optimal solution compare to the genetic algorithm. Furthermore, the proposed timing scheme can reduce the average delay and the parking times. Therefore, the proposed scheme will improve the traffic efficiency and relieve the traffic pressure.

Keywords: coordinated control; modified internal point method; adjacent intersection

图形图像

文章编号: 1007-1423(2023)16-0043-06 DOI: 10.3969/j.issn.1007-1423.2023.16.008

基于YOLOv5的工具表面缺陷检测系统

焦俊祥，金若男，李慧姝，方 武*

（苏州经贸职业技术学院信息技术学院，苏州 215009）

摘要：在工业生产过程中，产品质量极易受到现有生产技术等客观条件的影响，因此需要对产品进行质量检验，其中，表面缺陷是产品质量合格的重要指标之一。现如今在表面缺陷检测方面的常见技术有渗透探伤、超声波检测、机器视觉等。利用基于深度学习的YOLOv5算法通过机器视觉识别工具表面不同缺陷种类，为工具产品的质量检验提供便利。

关键词：目标检测；卷积神经网络；工具表面缺陷检测；YOLOv5

0 引言

目前，已有学者开展了工具表面缺陷检测的相关研究，研究中针对不同的应用场景使用了不同的技术，例如：磁粉探伤、渗透探伤、涡流检测、超声波检测和基于深度学习算法的机器视觉等技术。近年来，随着深度学习各种算法的发展，很多新的目标检测模型被提出，其中一些模型由于独特的设计和出色的性能被广泛应用在各个领域，例如 Faster R-CNN 算法、R-FCN 算法和 SSD 算法等，但这些算法的性能提升主要依赖于大型数据集和前期样本标注，对于工具表面的缺陷识别不太友好。本文使用的是基于深度学习的机器视觉方法，该方法在工业应用中包括三个主要环节：数据标注、模型训练与模型推断，使用的算法为YOLOv5网络算法，YOLOv5单阶段目标检测算法，该算法的优势主要有两点：①在输入端完成了数据增强，因此对训练数据集的要求不会太过苛刻；②该算法模型在硬件上部署方便，对于表面缺陷检测有较强的适应性与实用性。通过使用公开的数据集NEU-DET数据集与YOLOv5训练出来的模型，可以准确识别表面的各种缺陷，如划痕、凹坑等，如图1所示，该实验结果充分证明了基于深度学习的缺陷检测方法在工业中的可行性。

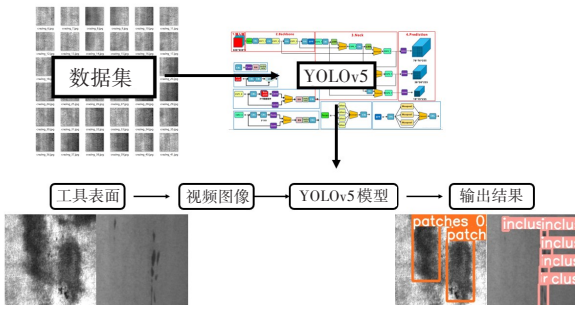


图1 使用NEU-DET数据集和YOLOv5网络算法对工具表面缺陷进行识别

1 研究进展

目前，国内针对不同场景下的表面缺陷检测的方法有许多种，例如：王攀峰等[1]提出的Q355B钢板截面磁粉探伤；王赫等[2]通过荧光渗透探伤测量裂纹长度；周庆祥等[3]使用以常规涡流检测为基础的多通道涡流检测技术，该技

收稿日期：2023-04-16 **修稿日期：**2023-07-23

作者简介：焦俊祥(2002—)，男，江苏泰州人，学生，专业方向为物联网技术；金若男(2003—)，女，安徽芜湖人，学生，专业方向为物联网技术；李慧姝(1990—)，女，江苏扬州人，博士，讲师，研究方向为机器视觉、计算机应用技术；*通信作者：方武(1981—)，男，湖北仙桃人，博士，副教授，研究方向为机器视觉、深度学习与物联网技术，E-mail：swmyang@126.com

术凭借检测灵敏度高、检测范围大等优势广泛应用于金属构件的检测；孙贺斌等[4]通过超声波检测结合铝合金焊缝声学特征验证了超声波检测技术在铝合金焊缝缺陷方面的实用价值；崔国宁等[5]通过卷积神经网络使用机器视觉技术实现了对输油管道缺陷尺寸的智能识别；路生亮等[6]通过改进Faster R-CNN算法，解决了对热轧钢板表面小目标识别精度差的问题，使其检测速度与识别精度有了很好的提升；耿朝晖等[7]改进Faster R-CNN算法，提高了PCB板缺陷检测的效率，在增强后的数据集上面取得了96.65%的mAP值；楚雅璐等[8]改进SSD算法，解决了因形变引起的服装图像识别准确率低的问题，与原SSD模型相比其平均预测精度mAP值提升可达3.63%；杨耀[9]认为可以采用机器视觉技术来完成裂纹磁痕判别，通过改进Mobile Net V3模型实现对裂纹磁痕的识别，查全率96.5%，查准率91.7%，平均单图识别速度1.7 s；易文渊[10]使用一种改进的Canny边缘检测算法来提取缺陷图像，并设计了一种基于UNET和ResNet的二阶段深度学习网络算法来检测轴承部件外壁缺陷；葛钊明等[11]改进YOLOv5的特征检测算法，引入卷积注意力模块并改进了损失函数，在检测速度不变的情况下，改进后的模型在测试集上的mAP提高了1.2%，召回率提高1.5%；李衍照等[12]提出一种Mosaic+Mixup的数据集增强方法，引入轻量型的GhostNet网络代替YOLOv5主干网络中CSP1模块中的残差模块，改进的YOLOv5的平均精度均值(mAP)达到96.88%，单张图片检测时间不超过50毫秒。高慧芳等[13]提出一种基于DCGAN与YOLOv5s的缺陷识别方法，设计了深度卷积生成对抗网络(DCGAN)扩充缺陷数据集，并优化了YOLOv5s网络的C3模块，优化后YOLOv5s算法的mAP值、精确率、召回率分别提高到85.4%、85.2%、83%。

2 YOLOv5

2.1 数据集

YOLOv5的数据集包含有六种典型工具的表面缺陷，如图2所示，即轧制氧化皮(RS)、斑块(Pa)、开裂(Cr)、点蚀表面(PS)、内含物(In)和划痕(Sc)。该数据库包括1800个灰度图像：六种不同类型的典型工具表面缺陷，每一类缺陷包含300个样本。对于缺陷检测任务，数据集包含了注释，指示每个图像中缺陷的类别和位置。如图3所示，对于每个缺陷，黄色框是指示其位置的边框，绿色标签是类别分数。

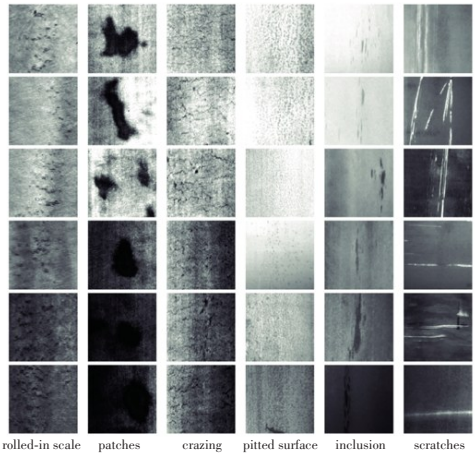


图2 数据集中的六种典型工具表面缺陷

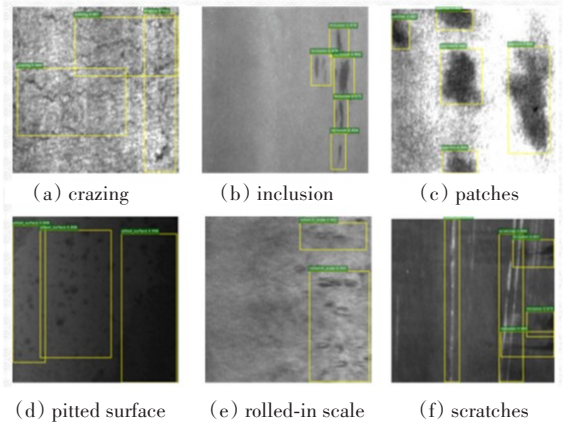


图3 数据集对六种典型缺陷的标注

2.2 YOLOv5单阶段目标检测算法

现阶段目标识别的主流算法有：R-CNN、YOLO、SDD等，其中YOLO算法主要应用于快速目标检测，如自动驾驶、无人车等这类需要快速对所获得的图像和视频进行识别并且反馈，以供控制系统做出迅速响应的场景。YOLO采用一步到位的方式，即输入整个图片，直接输出为目标识别的区域和目标类型。为了降低训练好的预测模型在硬件上的部署难度，并且考虑实际应用中环境因素的多样性，这里采用了YOLOv5算法，使用的预训练模型为由YOLOv5官方提供的YOLOv5s、YOLOv5m、YOLOv5l、

YOLOv5x，这四种网络结构在COCO数据集下的性能各有优劣：YOLOv5s网络最小，速度最快，AP精度也最低。另外三种网络算法比YOLOv5s网络结构复杂，AP精度提升，但计算速度也越来越慢，如图4所示。

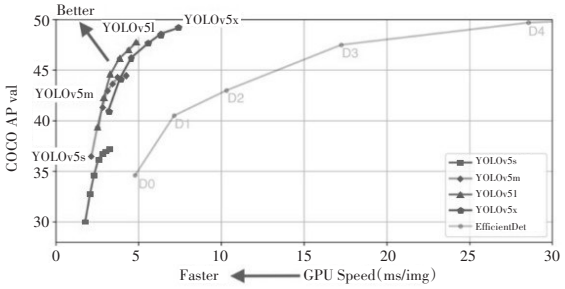


图4　YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5m预训练模型在COCO数据集的性能表现

YOLOv5的网络结构主要分为四个部分：输入端、Backbone（主干网络）、Neck网络和Prediction（输出端）。

（1）YOLOv5在输入端进行了Mosaic数据增强、自适应锚框计算、自适应图片缩放。

（2）YOLOv5在Backbone里面有两个结构，分别为Focus结构与CSP结构，其中Focus里最重要的是切片操作，而CSP结构相较于YOLOv4版本主要是设计了两种CSP结构，以YOLOv5s网络为例，CSP1_X结构应用于Backbone主干网络，另一种CSP2_X结构则应用于Neck中。

（3）YOLOv5在Neck里采用了FPN+PAN结构，相较于YOLOv4的Neck中采用的都是普通的卷积操作，YOLOv5的Neck结构中采用借鉴CSPNet设计的CSP2_X结构，加强网络特征融合的能力。

（4）YOLOv5在Prediction（输出端）进行的操作有：Bounding box损失函数、nms非极大值抑制。其中，nms非极大值抑制是在原有的CIOU_Loss基础上添加了DIOU_nms，主要作用是对一定区域内多个被重叠遮挡的目标进行检测，CIOU_Loss作为目标Bounding box损失函数，计算过程如式（1）、式（2）所示：

$$CIOU_{Loss} = 1 - CIOU = \left(IOU - \frac{Distance_{2^2}}{Distance_{C^2}} - \frac{v^2}{(1 - IOU) + v} \right) \tag{1}$$

$$v = \frac{4}{\Pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w^{p}}{h^{p}} \right)^2 \tag{2}$$

2.3　建立数据集并使用YOLOv5进行模型训练

数据标准标注软件使用labelimg库对图片中的目标位置进行人工标注，可以得到一个包含有图片的标签、标注框的中心坐标、标注框的相对宽与高信息的标签文件：“.txt”（如图5所示）。将原始图片数据和人工标注的标签共同作为原始数据集，包含训练集、验证集和测试集。通过官网下载预训练模型YOLOv5s.pt对本地数据集进行训练，训练中通过更改Weights、cfg、data、epochs、batch-size等一系列变量针对不同训练集和验证集中的目标类别数目、识别目标名称进行训练，最终得到两个模型文件：best.pt与last.pt，获得了模型文件的可视化数据：识别目标的类别和数目、模型的深度、模型的宽度、anchors、Backbone、head。

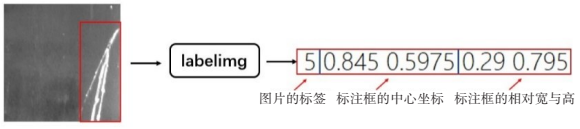


图5　对目标缺陷在labelimg库中使用的人工标注方式示意图

3　模型性能评估

首先计算了训练后模型的混淆矩阵，如图6所示，这个矩阵可以方便地看出机器是否将两个不同的类混淆了（比如说把一个类错当成了另一个）。

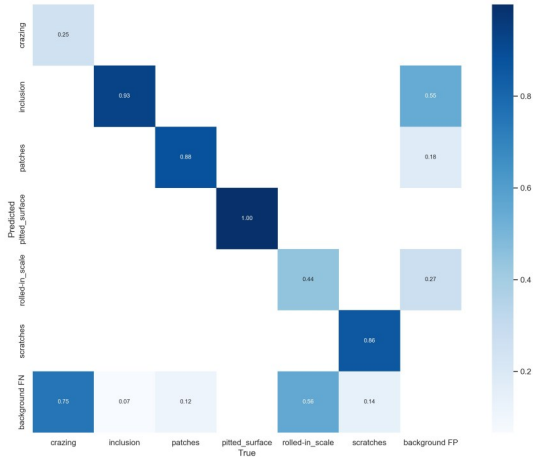


图6　模型的混淆矩阵

进一步地，对训练后的模型预测单一类的准确率P_curve进行了计算，如图7(a)所示。当confidence越大的时候，类别检测越准确。对召回率(查全率)和置信度之间的关系R_curve进行了计算，结果显示置信度越小时，类别检测得越全面，如图7(c)所示。还计算了训练后模型的均值平均精度PR_curve，结果显示精度越高，召回率越低，如图7(b)所示。所以只有曲线尽可能的地接近(1，1)点，才可以在准确率很高的前提下，尽可能检测到全部的类别。本文还计算了模型的F1_curve，F1-score是分类问题的一个衡量指标。F1-Score的值从0到1，1是最好，0是最差。模型评估结果如图7(d)所示。

接着分析了模型的数据指标，如图8所示，其中，train/box_loss：预测框与真实框的损失函数，值越小预测框的准确率越高；train/obj_loss：目标检测损失，值越小目标检测越准；train/cls_loss：分类损失均值，值越小分类越准；metrics/precision：精度(找对的正类/所有找到的正类)；metrics/recall：召回率，本应该被找对的正类；val/box_loss：验证集边缘损失；val/obj_loss：验证集目标检测损失均值；val/cls_loss：验证集分类损失均值；metrics/mAP@0.5:0.95(mAP@[0.5:0.95])表示在不同IoU阈值(从0.5到0.95，步长0.05)(0.5、0.55、0.6、0.65、0.7、0.75、0.8、0.85、0.9、0.95)上的平均mAP；metrics/mAP@0.5：表示阈值大于0.5的平均mAP，该指标为PR曲线的面积，代表模型的平均准确率，mAP越高越好。

在训练时设置的*batch-size*=4，即每4张图片拼成一张照片送进YOLOv5中训练。这里的图片一共12张，如图9所示。使用训练好的模型进行了测试，如图10所示，val_batchX_labels为验证集第x轮的实际标签，即：由人工标注过的图片信息。val_batchX_pred为模型对val验证集第*x*轮里的图片类别的标签的预测。

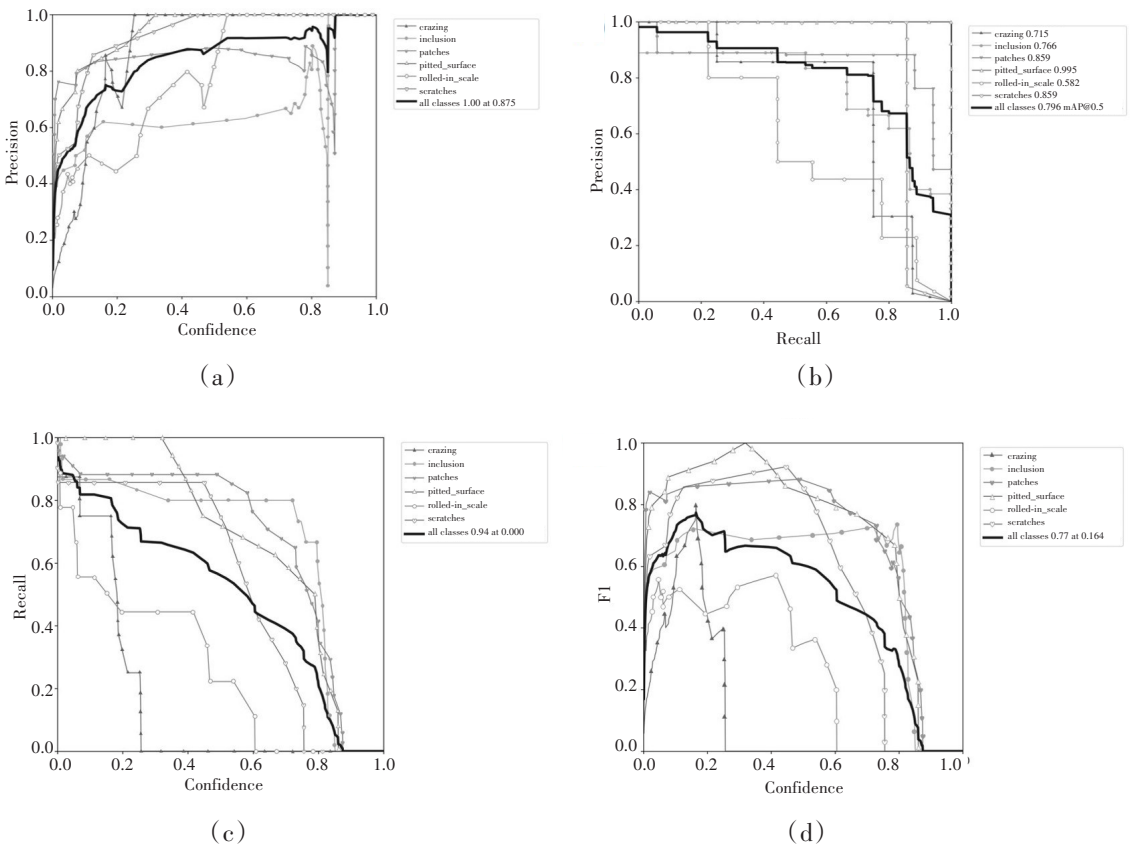


图7 模型的P曲线、R曲线、PR曲线和F1分数曲线

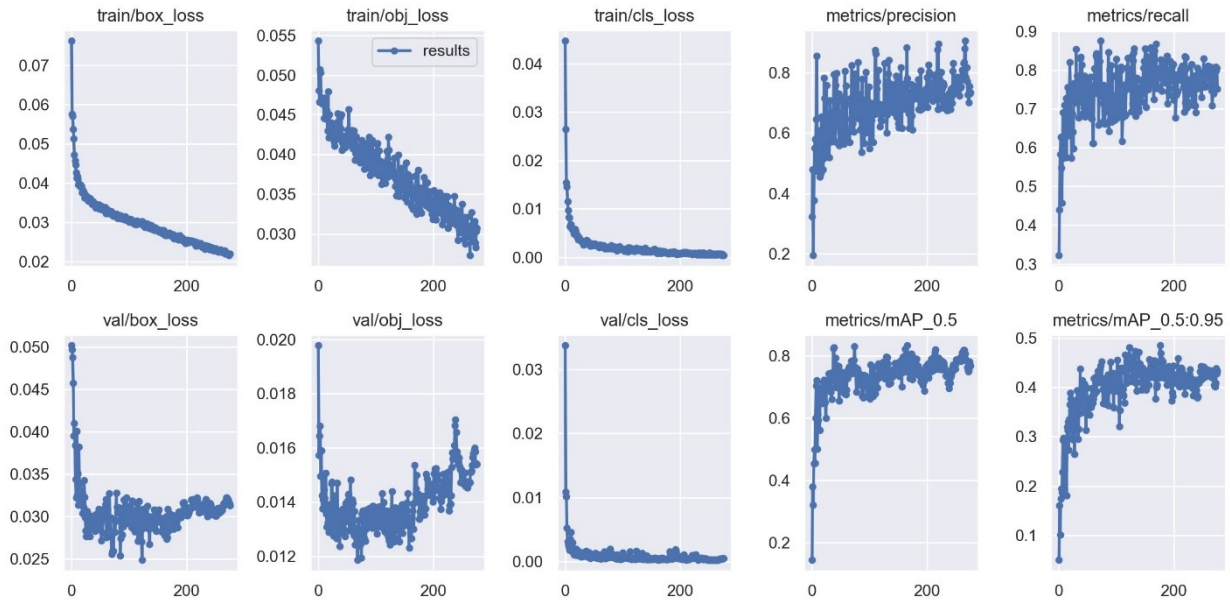


图 8　模型的性能指标

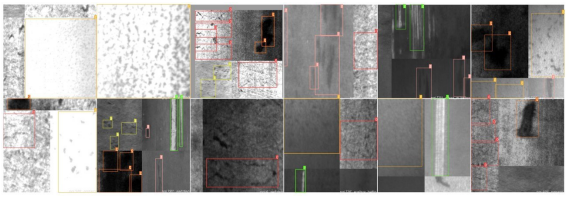


图 9　模型训练

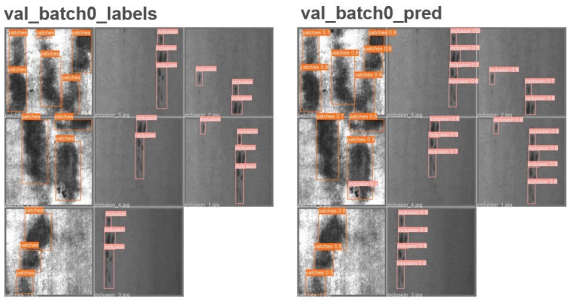


图 10　模型测试

4　模型在硬件下的使用

选用的硬件平台是jetson nano 4 GB版本，首先在jetson官网下载它的操作系统并将其写入SD卡中，插入jetson，启动开机。配置环境方面，需要更新CMake、PyTorch库、torchvision库以及一些YOLOv5运行时的依赖库，再将训练好的YOLOv5模型加载过来，进入detect.py文件修改parse_opt函数的weights与source，分别为生成的模型文件的路径与推理方式，在YOLOv5里面通过了三种模型的推理方式，分别为图片(default=文件路径)、视频(default=文件路径)、摄像头(default=0)。然后运行detect.py就可以在jetson平台上使用模型了。这里选择的是摄像头识别的方式，分别对两种操作工作进行缺陷检测，结果显示均能识别出表面缺陷种类：scratches(划痕)，如图11所示。

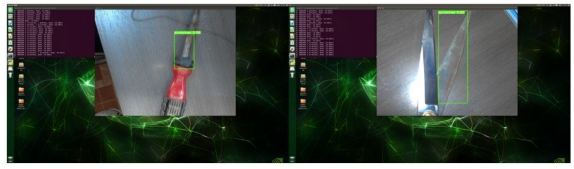


图 11　模型对两种不同工具进行缺陷检测结果

5　结语

相较于其他的工具表面缺陷检测技术，本文提出了一种基于YOLOv5的工具表面缺陷检测算法。实验证明，通过YOLOv5的深度学习算法对数据集进行训练得出的预测模型，搭载在硬件上可用于操作工具的表面缺陷检测。

参考文献：

[1] 王攀峰，李明，康学勤. Q355B钢板截面磁粉探伤显示磁痕的原因[J]. 上海金属，2022，44(6)：101-107.

[2] 王赫，白素平，闫钰锋. 荧光渗透探伤裂纹长度测量技术研究[J]. 长春理工大学学报(自然科学版)，2021，44(6)：22-27.

[3] 周庆祥，李刚卿，侯凯，等. 铝合金部件裂纹缺陷的多通道涡流检测仿真与试验研究[J]. 电子测试，2022，36(17)：34-38.

[4] 孙贺斌，周治伊，李军，等. 铝合金焊缝超声波检测准确率影响因素研究[J]. 焊接技术，2021，50(1)：99-103.

[5] 崔国宁，杨理践，耿浩，等. 基于卷积神经网络的管道缺陷量化识别方法[J]. 仪表技术与传感器，2022(10)：99-103.

[6] 路生亮，马驰，胡辉，等. 基于Faster R-CNN的钢板表面缺陷识别研究[J]. 电脑编程技巧与维护，2021(10)：110-113.

[7] 耿朝晖，龚涛. 基于改进Faster R-CNN的PCB板表面缺陷检测[J]. 现代计算机，2021，27(19)：89-93.

[8] 楚雅璐，顾梅花，刘杰，等. 改进SSD的服装图像识别方法[J]. 纺织高校基础科学学报，2022，35(4)：95-102.

[9] 杨耀. 轴承套圈荧光磁粉探伤图像识别与定量方法研究[D]. 上海：东华大学，2022.

[10] 易文渊. 基于机器视觉的轴承部件磁粉探伤系统软件设计与实现[D]. 成都：电子科技大学，2021.

[11] 葛钊明，胡跃明. 基于改进YOLOv5的锂电池极片缺陷检测[J]. 激光杂志，2023，44(2)：25-29.

[12] 李衍照，于镭，田金文. 基于改进YOLOv5的金属焊缝缺陷检测[J]. 电子测量技术，2022，45(19)：70-75.

[13] 高慧芳，金永，柴国强，等. 基于DCGAN与YOLOv5s的火箭弹非金属壳体缺陷识别方法研究[J]. 固体火箭技术，2022，45(6)：949-955.

Tool surface defect detection system based on YOLOv5

Jiao Junxiang, Jin Ruonan, Li Huishu, Fang Wu*

(Information Department, Suzhou Institute of Trade & Commerce, Suzhou 215009, China)

Abstract: In the process of industrial production, product quality is easily affected by the existing production technology and other objective conditions, so it is necessary to carry out quality inspection of products, among which, surface defects is one of the important indicators of product quality. Nowadays, the common technologies in surface defect detection include penetration detection, ultrasonic detection, machine vision and so on. YOLOv5 algorithm based on deep learning was used to identify different types of defects on the surface of tools through machine vision, which provided convenience for quality inspection of tool products.

Keywords: object detection; deep convolutional neural network; tool surface defect detection; YOLOv5

实践与经验

文章编号: 1007-1423(2023)16-0049-05 DOI: 10.3969/j.issn.1007-1423.2023.16.009

基于FPGA的超宽带低频信号发生器设计

谭晓刚*，蔡昌恒，高 峯

（贵州航天计量测试技术研究所，贵阳 550009）

摘要：针对电磁兼容性试验中超宽带、高精度、多功能激励源的需求，采用了一种基于“FPGA+PLL+DDS”的频率合成技术，实现了一定频率范围、调制方式灵活可控的信号发生器。设计的信号发生器中，FPGA负责解析并执行上位机下发的指令，同时计算相关DDS控制字并通过并行SPI更新DDS芯片配置，最终输出满足需求的调制信号。测试结果表明设计的信号发生器功能、指标均达到了电磁兼容性试验要求，现已应用于某电磁兼容性试验测试设备中。

关键词：电磁兼容性试验；FPGA；PLL；DDS；信号发生器

0 引言

电磁兼容性(electromagnetic compatibility, EMC)试验在科研装备生产中是一项重要的考量因素，电磁传导是设备电磁兼容性优劣的主要判断标准，EMC标准中对相关试验CS114、CS115、CS116测试要求进行了描述，在此过程中需要信号发生器产生干扰信号[1-2]。以CS114为例，测试设备需要产生一定频率范围的干扰信号，经功率放大器放大后注入受试电缆，通过监测受试电缆的感应电流变化，测试电缆束传导敏感度[3-4]。本文设计的信号发生器控制核心选用深圳国微SMQ2V3000，支持并行SPI配置的直接数字频率合成(direct digital synthesis, DDS)芯片，大大提升了配置速度，通过上位机设置可输出不同频率、不同幅值的连续波，还可以输出不同调制频率、调制度可变的幅度(amplitude, AM)调制和占空比可变的脉冲(pulse, PU)调制信号，其性能满足了EMC标准中电磁传导敏感度测试的需求。

1 设计方案

本文设计的信号发生器采用FPGA+锁相环频率合成(phase locked loop, PLL)+ DDS技术实现，信号发生器设计方案如图1所示，信号发生

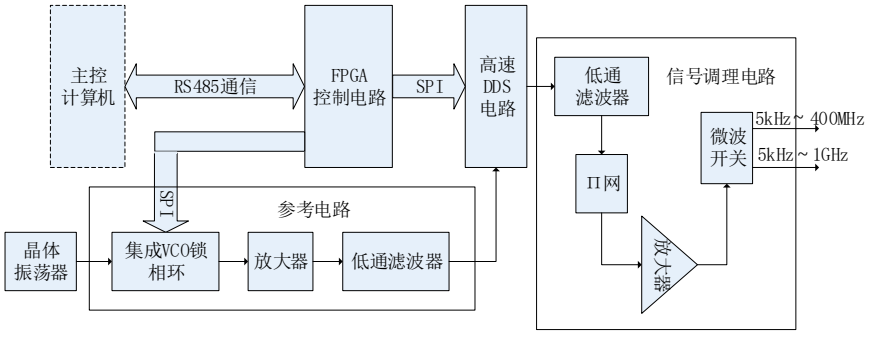


图 1 信号发生器设计方案

收稿日期：2023-01-13 **修稿日期：**2023-03-08

作者简介：*通信作者:谭晓刚(1996—),男,贵州印江人,硕士,助理工程师,研究方向为嵌入式软件开发,E-mail:794354126@qq.com;蔡昌恒(1996—),男,贵州石阡人,本科,助理工程师,研究方向为微波电路设计;高峯(1989—),男,贵州贵阳人,硕士,高级工程师,研究方向为微波电路设计

器采用硬件系统和控制软件系统协同的方式完成，设计方案中PLL和DDS芯片输出由FPGA配置，提高了使用灵活性，在不同的应用场景下，可通过FPGA修改PLL和DDS芯片配置满足使用需求。PLL+DDS的组合可以提高输出信号相噪、杂散等指标。FPGA作为控制核心，晶振信号输入PLL芯片产生参考信号，DDS芯片输出目标信号，最后通过信号调理得到满足条件的输出信号[5]。

2 硬件电路设计

2.1 参考电路

为保证时钟信号的频谱纯度，参考电路基于PLL频率合成的方式产生高频、高纯度参考信号，为满足输出信号带宽要求，参考电路需为DDS电路提供4 GHz的时钟信号。晶体振荡器产生100 MHz信号作为集成VCO锁相环芯片参考，通过FPGA控制该锁相环芯片产生4 GHz信号，经放大满足时钟信号输入电平要求，再经滤波处理后作为DDS电路的参考时钟信号输入[6]。PLL芯片输出参考信号计算公式如下：

$$f_{PLL} = (f_{REF}/R)*(N_{int} + N_{frac}/2^m) \quad (1)$$

其中：f_{REF}为参考输入；R为预分频比；N_{int}为整数分频比；N_{frac}为小数分频比；m为小数分频数值范围。

2.2 频率合成电路

DDS电路为专用单片集成电路，以参考电路输出的4 GHz时钟信号为参考，DDS电路根据FPGA控制电路产生的控制字(幅度、频率、相位)实现频率、幅度、相位等参数的控制，输出1 Hz小步进射频信号。DDS输出频率计算公式如下：

$$f_{DDS} = f_{REF}* FTW/2^N \quad (2)$$

其中：f_{DDS}为DDS输出频率，f_{REF}为参考信号频率，FTW为频率控制字，N为频率控制字精度。

2.3 信号调理电路

信号调理电路由低通滤波器、π网、放大器以及微波开关组成，其中低通滤波器用于滤除泄露的参考时钟信号，避免参考时钟信号进入后级电路造成输出信号失真；π网用于调节微波放大器输入信号功率以保证其工作在线性区；微波放大器用于将DDS输出信号功率放大至所需的量级；微波开关将信号从频段上分为两个支路输出，其中低频段支路进入后级电路单元，高频支路直接输出到外部，作为频率扩展输出。

3 FPGA控制软件设计

控制软件数据流图如图2所示，根据功能将控制软件系统分为：通信模块、PLL模块、协议解析模块、AM调制模块、PU调制模块、DDS模块。其中通信模块负责与上位机交互，PLL模块在每次上电时或初始化时进行配置，其它主要模块的功能及具体实现在下文介绍[7]。

3.1 协议解析模块

协议解析模块是信号发生器核心，根据指令指挥系统运行。该模块主要通过通信模块收

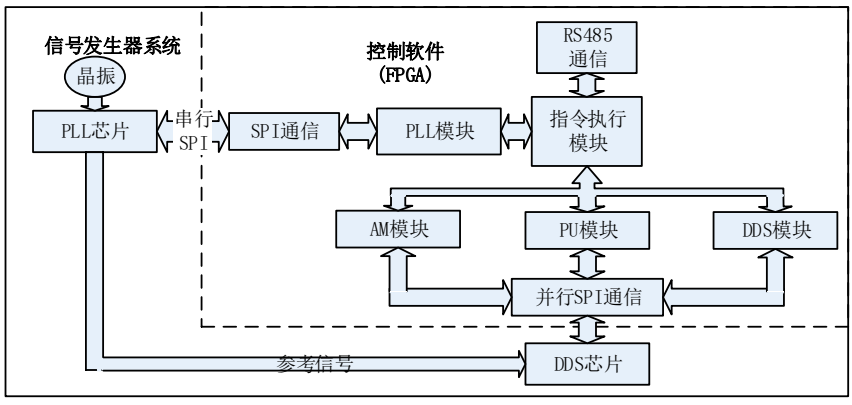


图2 控制软件数据流图

发指令，完成指令解析和执行。本次设计中，上位机可发送6个字节数据，首字节为指令字，末字节为校验字，中间4个字节为数据字。上位机通过6个字节控制信号发生器初始化、输出开关、输出连续波的功率和频率，也可设置调制功能并配置调制度和占空比、调制频率。

3.2　DDS模块

DDS模块根据协议解析模块发出的使能执行初始化信号发生器、配置频率和功率功能。初始化功能需要配置DDS芯片的通用寄存器、功能寄存器以及初始化频率相位寄存器。输出信号的频率和功率配置需要根据数据计算频率控制字，通过并行SPI配置频率控制寄存器完成频率切换，功率配置与频率配置基本相同。

3.3　调制模块

调制模块包括了AM调制和PU调制两个模块，在调制过程中可通过上位机发送指令修改载波频率、调制频率、调制度和占空比以及峰值，也可以关闭输出、关闭调制、切换调制方式。

3.3.1　AM调制模块

为产生精度更高的正弦波，AM调制模块使用FPGA中的DDS IP产生正弦波，根据正弦波采样幅值、调制度以及公式(2)计算DDS幅度控制字，循环计算刷新DDS寄存器输出AM调制信号。

AM调制流程如图3所示，设计中将FPGA IP核配置为相位增量值输入30 bit、输出幅度值21 bit，根据调制波频率和IP核手册计算相位增量参数，产生满足条件的正弦波。由于产生的正弦波是补码表示，因此需要将补码平移并恢复为原码。最后根据调制度和平移后的正弦波采样幅值计算最终正弦信号的幅值，结合DDS芯片的幅值计算公式计算得到幅值控制字，使用SPI写入DDS芯片完成信号输出。从IP核产生正弦波到完成DDS芯片更新，需要至少6个时钟周期，为提高输出精度，在写入幅度控制字时，开始计算下一幅度控制字，依次循环不间断配置输出AM调制波。

3.3.2　PU调制模块

设计使用的DDS芯片拥有8对频率、相位、幅值寄存器组，可通过外部开关切换寄存器组，PU调制根据此特点完成设计。即选用一个寄存器组作为信号输出寄存器组，另一个寄存器组则配置输出为无输出的控制字，根据调制波频率在两个寄存器组之间切换完成PU调制。

PU调制模块流程如图4所示，PU调制模块运行时根据占空比和调制波频率计算高电平周期和完整周期，采用两个周期分别计数高电平周期和完整周期。计时器开启调制后，使能DDS引脚切换到有信号输出的寄存器组，高电平计时器和周期计时器开始计时，当高电平周期完成后，切换到无输出的寄存器组。

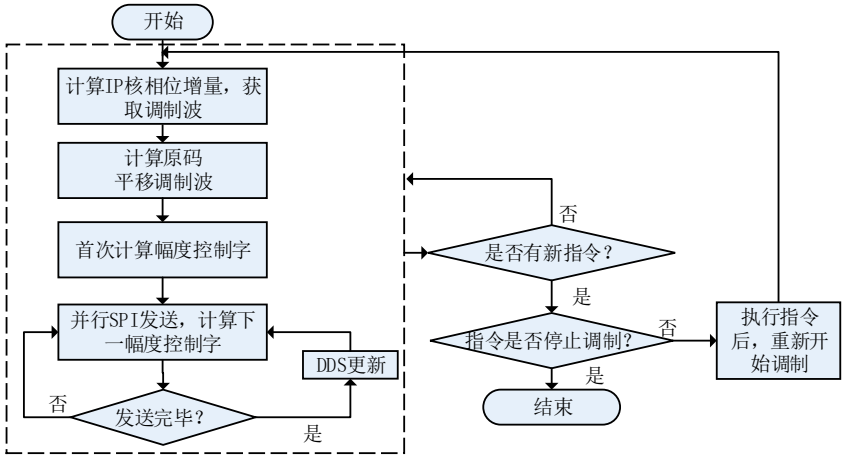


图3　AM调制流程

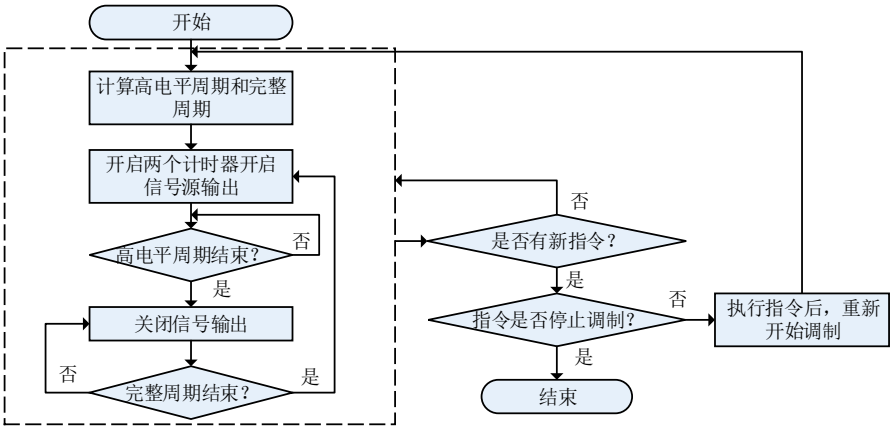


图 4　PU调制流程

4　结果测试

根据设计方案，经制版加工、焊接后的信号发生器印制电路板如图5所示。

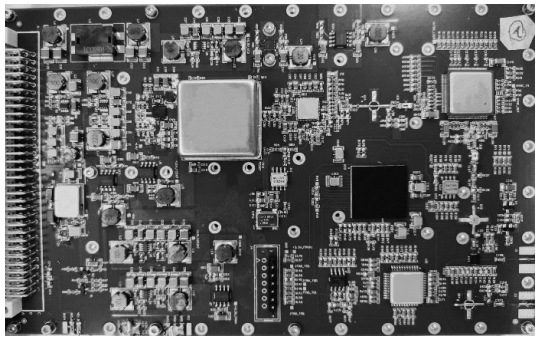


图 5　信号发生器印制电路板

图6所示为1 GHz频点输出频谱仪显示结果。

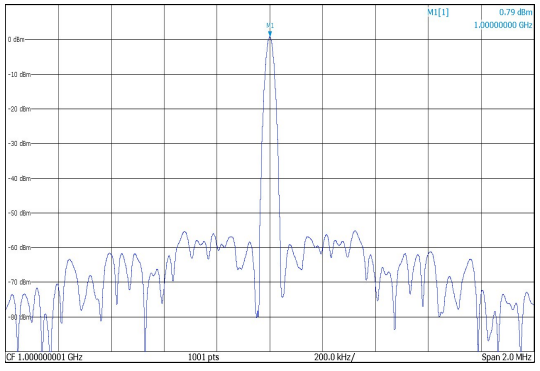
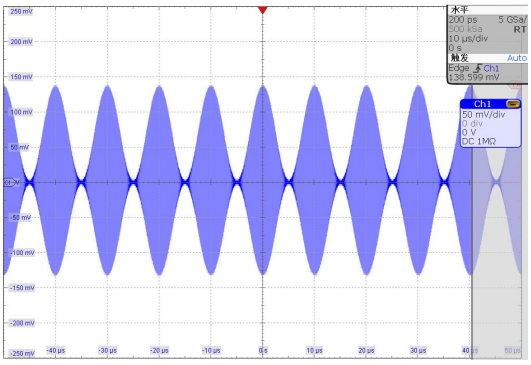
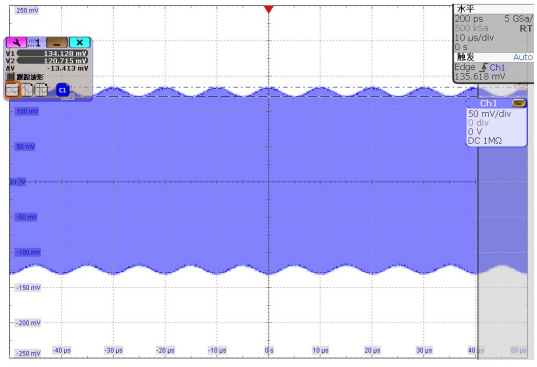


图6　输出端1 GHz、1 dBm输出

图7、图8所示为载波100 MHz时不同调制状态的AM调制和PU调制波形。结果表明设计的信号发生器AM调制波频率最高可达100 kHz，调制度设置范围10%～100%；PU调制波频率最高可达200 kHz，占空比设置范围10%～99%。

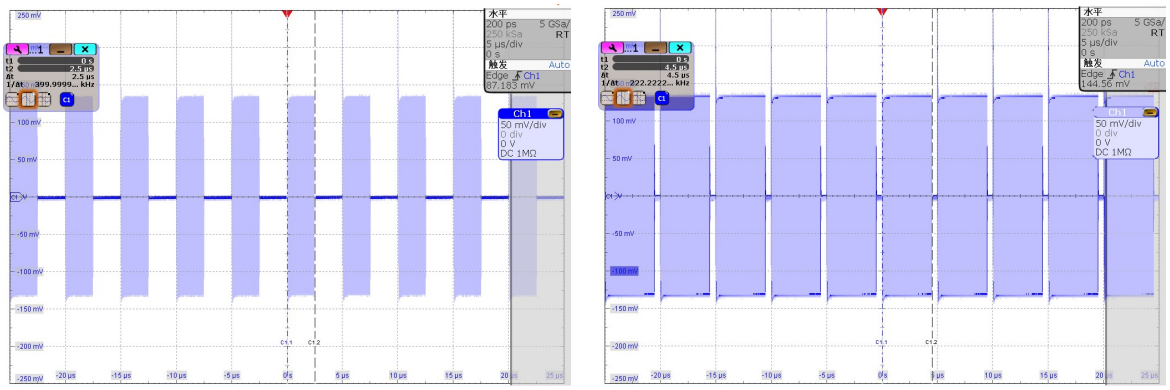


（a）100%调制度



（b）10%调制度

图 7　AM调制波形



（a）50%占空比　　（b）90%占空比

图8　PU调制波形

5　结语

本文设计了一款基于“FPGA+PLL+DDS”技术的国产器件信号发生器，使用FPGA作为核心控制器，通过PLL+DDS的技术得到超带宽、高精度、多功能的激励源，输出频率范围5 kHz～1 GHz，频率分辨率达到了1 Hz，实现了一定调制频率范围的AM调制和PU调制。试验结果表明，该技术设计的信号发生器各项性能指标满足了EMC试验的需求，目前该信号发生器已应用到研制的EMC测试设备中并通过项目验收。

参考文献：

[1] 徐亮，胥朋. GJB151B传导敏感度测试要求的分析[J]. 微波学报，2016，32(S2)：441-443.

[2] 陈世钢. GJB151B-2013解析[J]. 安全与电磁兼容，2014(2)：15-24，69.

[3] 王月焜. CS114电缆束注入传导敏感度试验方法探究[J]. 现代制造技术与装备，2018(5)：54-55，57.

[4] 张孙虎. CS114测试原理和主要问题解析[J]. 电子技术与软件工程，2018(1)：90-91.

[5] 张坤，陈义，张子才. 基于锁相环的频率合成器的设计[J]. 现代电子技术，2007(19)：110-111，114.

[6] 李映晟，严厚伟，杜睿. 模块化中频信号源设计[J]. 火力与指挥控制，2021，46(6)：155-160.

[7] 卫冠荣，文丰，贾兴中. 基于FPGA的多通道高精度信号源设计[J]. 单片机与嵌入式系统应用，2022，22(4)：83-87.

Design of ultra-wideband low-frequency signal generator based on FPGA

Tan Xiaogang*, Cai Changheng, Gao Feng

(Guizhou Aerospace Institute of Measuring and Testing Technology, Guiyang 550009, China)

Abstract: In response to requirements of ultra - wideband, high precision, and multi-functional excitation sources in Electromagnetic Compatibility tests, this paper adopts a frequency synthesis technology based on “FPGA+PLL+DDS” to achieve a signal generator with a certain frequency range, flexible and controllable modulation mode. In this signal generator, FPGA is responsible for parsing and executing the instructions issued by upper computer, calculating relevant DDS control words, and updating DDS configuration through parallel SPI, and finally outputting the modulation signals that meet the requirements. The test results show that functions and indicators of the designed signal generator meet requirements of EMC test. It has been applied to an EMC test equipment.

Keywords: electromagnetic compatibility(EMC); FPGA; PLL; DDS; signal generator

文章编号: 1007-1423(2023)16-0054-06 DOI: 10.3969/j.issn.1007-1423.2023.16.010

基于改进蝴蝶优化算法的工程应用

储 敏*，李 宣，陈学财

（贵州师范大学教育学院，贵阳 550025）

摘要：为解决经典蝴蝶优化算法易陷入局部最优、收敛速度慢的缺点，提出一种融合折射反向学习和黄金正弦分割指引机制的改进蝴蝶优化算法。融入黄金正弦指引机制来引导粒子飞行方向，可以增强算法的局部探索能力，帮助蝴蝶更快地找到食物；引入折射反向学习，提高种群多样性并实现对更多区域的勘探，再对参数自适应化后，算法的全局和局部平衡效果也有所改善。本研究通过对10个不同特征的测试函数以及三杆工程设计问题(Three-Bar Truss)进行实验和测试，测试结果均表明改进后的算法在寻优中具有更好的收敛速度和寻优精度，还进一步验证了改进方法的功能性。

关键词：蝴蝶优化算法；折射反向学习；自适应化；结构优化；黄金正弦算法

0 引言

近些年来，人们希望通过模拟大自然生物种群为了生存而相互配合协作的特性，各种群智能优化算法(intelligence optimization algorithm)被设计出来。例如粒子群优化算法(PSO)[1]、灰狼算法(GWO)[2]、海洋捕食者算法(MPA)[3]等。蝴蝶优化算法(butterfly optimization algorithm，BOA)[4]由 Arora 等[5]于 2017 年提出，其具有参数少、结构简单等优点，并且已经解决了无线传感器网络族首选择、图像分割[6]等问题。为了提高算法对高维优化问题的处理能力，李彦苍等[7]融合最优领域扰动和反向学习策略的蝴蝶优化算法，但其精度也有待提升；Arora 等[8]将莱维飞行策略引入，但却降低了算法的群体多样性，使其易陷入局部最优。

上述等人的改进虽然在某些方面改进了算法性能，但是在求解高维问题时依旧存在很大的提升空间。本文提出了一种融合折射反向学习和黄金正弦指引机制的蝴蝶优化算法，加入折射反向学习，可以提升算法的收敛速度和全局探索能力，再对蝴蝶个体感官系数自适应化使算法的局部开发能力进一步增强。随后使用10个典型测试函数和一个工程算例进行仿真实验，结果表明，改进的蝴蝶优化算法在大多数函数上取得更好解的同时，收敛速度也得到了显著提升。

1 蝴蝶优化算法

蝴蝶优化算法是模拟蝴蝶觅食和交配行为的优化方法，其中每只蝴蝶都有属于自身的香味，散发出来以便其它蝴蝶能够嗅到，起到一种相互吸引的作用。香味的浓度可用如下公式表示：

$$f_i = cI^a \tag{1}$$

式中：f_i为蝴蝶的香味感知量；c为感官模态，理论上可以取$[0,\infty)$；I为刺激强度；a为基于香味吸收程度的幂指数，通常取$[0,1]$。

然后算法将会使每只蝴蝶有概率性地进入到两个关键步骤，即全局搜索阶段和局部搜索阶段，搜索过程可分别用式(2)和(3)表示：

收稿日期：2023-05-15 **修稿日期：**2023-07-19

作者简介：*通信作者：储敏(1997—)，女，安徽安庆人，在读硕士生，主要研究方向为深度学习算法，E-mail：1610546421@qq.com；李宣(1997—)，女，贵州遵义人，在读硕士生，主要研究方向为信息化教学；陈学财(1998—)，男，江西赣州人，在读硕士生，主要研究方向为信息化教学

$$x_i^{t+1} = x_i^t + \left(r^2 \times g^* - x_i^t\right) \times f_i \tag{2}$$

$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) \times f_i \tag{3}$$

式(2)和(3)中的x_i^t表示第i个蝴蝶个体在第t代的位置，x_j^t表示第j个蝴蝶个体在第t代的位置，x_k^t表示第k个蝴蝶个体在第t代的位置。

2　改进的蝴蝶优化算法

2.1　折射反向学习方法

折射反向学习(refracted opposition-based learning, ROBL)[9]是在反向学习的基础上对其反向过程进行的一种改进，具有很好的全局搜索能力，可以避免种群在算法前期迅速聚集而导致其种群多样性降低的问题。其二维原理如图1所示。

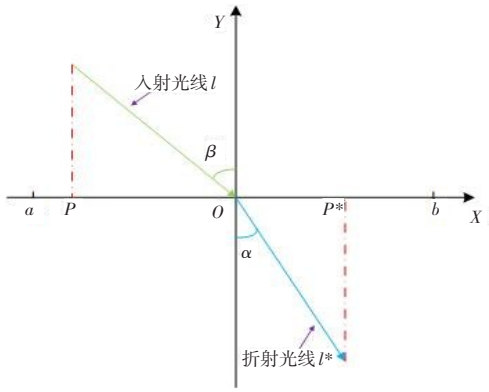


图1　折射反向学习示意图

在图1中，X轴P的搜索空间是$[a,b]$，对应一个维度的解空间，l,l^*是点P在搜索空间上的入射光线和折射光线，其对应的入射角和折射角分别为α,β。由此可知折射率n即为$\sin\alpha/\cos\beta$，O点是搜索区间$[a,b]$的中点，只需要设定透镜的缩放系数$k = l/l^*$，根据折射原理就可得到P点与其折射反向学习后的P^*点的关系式：

$$P^* = \frac{a+b}{2} + \frac{a+b}{2kn} - \frac{P}{kn} \tag{4}$$

当我们所研究的问题维度增加时，可将式(4)推广到下式来计算粒子折射反向学习过后的新位置：

$$x_{i,j}^* = \frac{a_j+b_j}{2} + \frac{a_j+b_j}{2k} - \frac{x_{i,j}}{k} \tag{5}$$

式中：$x_{i,j}$表示第i个粒子在第j维的值，$x_{i,j}^*$表示$x_{i,j}$经过折射反向学习后形成的解，a_j、b_j分别表示当前种群第j维的最大值和最小值。

2.2　黄金正弦指引机制

黄金正弦算法(golden sine algorithm, Gold-SA)是Tanyildizi提出的新算法，主要依据单位圆和正弦函数的定义逻辑，可以用单位圆上的正弦值与搜索代理空间融合进行寻优，具有参数少、易实现等特点。其原理如图2所示。

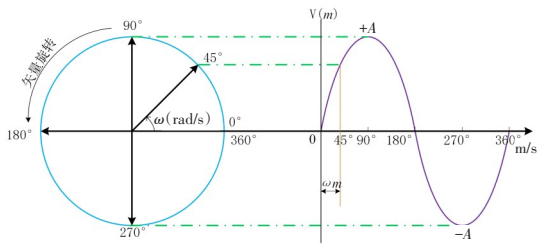


图2　正弦三角函数与单位圆的关系原理图

其中：A是振幅，ω是角频率，单位为rad/s，m为时间，单位为s。正弦的角频率ω和幅值A随迭代次数的变化而变化。

Gold-SA与其它算法融合可以为算法提供更快的收敛速度。其目的是缩小搜索范围，加强算法的局部搜索能力，其黄金分割系数如下式：

$$\begin{cases} c_1 = a_1(1-h) + b_1 h \\ c_2 = a_1 h + b_1(1-h) \end{cases} \tag{6}$$

式中：a_1,b_1为黄金分割搜索初始值，一般取$a_1 = \pi, b_1 = -\pi$；h为常数，通常取$h = (\sqrt{5}-1)/2$。

蝴蝶优化算法后期局部迭代能力变差，难以跳出局部最优解，将局部迭代更新后蝴蝶位置按照式(7)指引到新的位置，并与之前该蝴蝶的适应度值进行比较，留下较好的位置和解

$$x_i^{t+1} = x_i^t \left|\sin(r_3)\right| - r_4 \sin(r_3) \left|c_1 D^t - c_2 x_i^t\right| \tag{7}$$

式中：D^t是第t次迭代中第j个蝴蝶的位置，r_3,r_4为随机数，范围可取：$r_3 \in [0,2\pi], r_4 \in [0,\pi]$。

2.3　自适应度变化

感官形态系数c可以取$[0,\infty]$范围内的任意值。但是，为了与改进后的算法相适应，经多次实验后发现，将其改为随迭代次数减小更能平衡全局和局部搜索。感觉模态c可表示为

$$C_{t+1} = C_t + \left[\frac{0.5}{\left(c_t \cdot T_{\max}\right)}\right] \tag{8}$$

其中：C_{t+1}是$t+1$代时的感觉模态，C_t是t代时的感觉模态，T_{max}是最大迭代次数。

2.4 ORGABOA实现流程

综上所述，改进的蝴蝶优化算法(ORGABOA)的具体优化流程如图3所示。

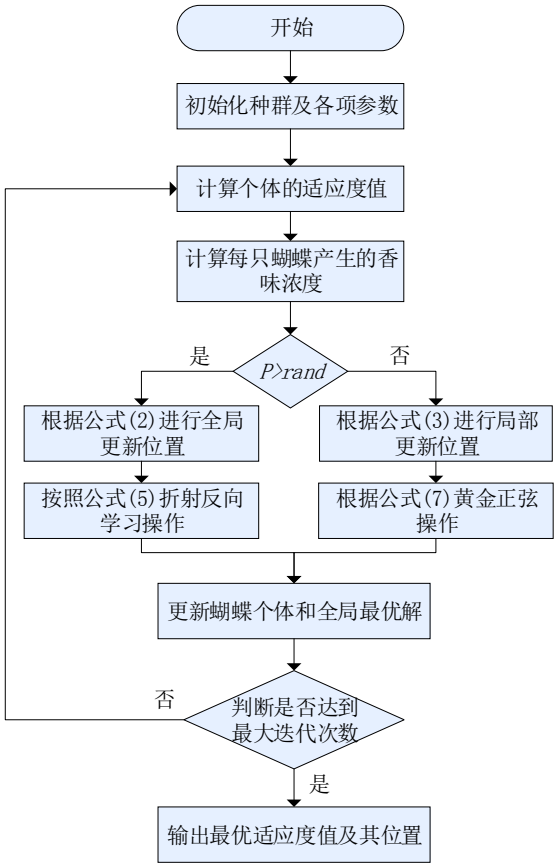


图3 改进的蝴蝶优化算法流程

3 仿真实验与分析

3.1 基准测试函数

为验证本文ORGABOA的有效性，选择了三种算法和只加黄金分割的蝴蝶优化算法进行比较分析，分别是：BOA、GWO、MPA、GABOA。折射反向学习策略中$k=10000$，其它算法中的各项参数设置均与原论文一致。此外，为了保证公平性，实验中种群大小统一为30，最大迭代次数为500。同时为了降低算法的随机性和偶然性，算法在每个测试函数上都进行30次独立实验。取四个基准测试函数对改进算法进行验证，测试函数见表1。

表1 基准测试函数

函数编号	测试函数	维数dim	搜索范围	最优值
F1	Sphere $^{IBOA}_{BOA}$	50	[-100，100]	0
F2	Schwefel1.2	50	[-100，100]	0
F3	Sumpower	50	[-100，100]	0
F4	Rastrigin	50	[-5.12，5.12]	0
F5	Ackley	50	[-20，20]	0
F6	Alpine	50	[-10，10]	0

3.2 实验结果的分析

由表2可以看出，本文提出的ORGABOA对所选的单峰测试函数均达到了函数的理论最优值，尤其是标准差体现出来的稳定性非常好，说明改进算法在求解单峰的问题上有很好的寻优能力，求解精度也很理想。此外，ORGAMPA对复杂的多峰函数也展示出非常好的寻优性能，尤其是F4，F6达到理论最优解的同时又拥有非常好的平均值和标准差，说明改进算法的局部开发能力有了极大的提升，而F5的结果也可以看出算法的性能提升也是比较明显的。从平均值和标准差的结果也能看出ORGAMPA相比其他算法具有更好的稳定性和鲁棒性。

根据实验数据，绘出函数迭代图像，由图4可以更加形象地看出，ORGAMPA收敛速度和寻优精度比其他算法都要好，曲线下降的速度非常快，同时也保证了算法的抗早熟能力。

4 IBOA在三杆桁架设计问题中的应用

三杆桁架设计问题[10]目的是使三杆的体积最轻，并且也要使每根构件的应力满足要求，其模型如图5所示，该问题的数学表达式如下：

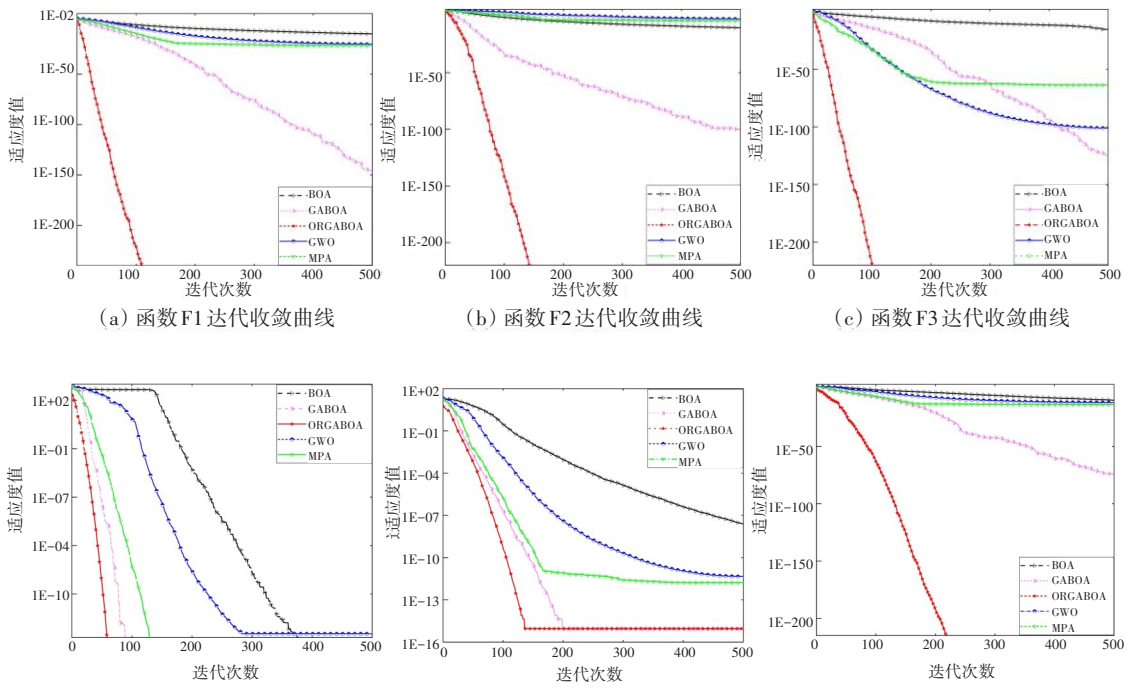
$$\text{s.t.}\begin{cases} g_1=\dfrac{\sqrt{2}A_1+A_2}{\sqrt{2}A_1^2+2A_1A_2}P-\sigma\leqslant 0 \\ g_2=\dfrac{A_2}{\sqrt{2}A_1^2+2A_1A_2}P-\sigma\leqslant 0 \\ g_3=\dfrac{1}{A_1+\sqrt{2}A_2}P-\sigma\leqslant 0 \\ \min f(A_1,A_2)=\left(2\sqrt{2}A_1+A_2\right)\cdot L \end{cases} \tag{9}$$

式中：$0\leqslant A_1, A_2\leqslant 1$, $L=100\ \text{cm}$, $P=2\ \text{kN/cm}^2$, $\sigma=2\ \text{kN/cm}^2$

ORGABOA和原算法对该问题求解的最优解

表 2　基准测试函数实验结果

统计量	算法	F1	F2	F3	F4	F5	F6
最优值	BOA	8.46E−11	6.29E−11	6.82E−14	**0.00E+00**	2.68E−08	1.77E−10
	GABOA	2.69E−128	2.87E−93	6.31E−116	**0.00E+00**	**8.88E−16**	1.34E−74
	ORGABOA	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**8.88E−16**	**0.00E+00**
	GWO	4.08E−20	0.0827626	3.85E−93	5.58122	2.93E−11	0.005418
	MPA	1.02E−21	0.0434093	5.85E−60	**0.00E+00**	4.98E−12	2.97E−13
平均值	BOA	8.33E−11	6.64E−11	8.71E−14	59.07353	2.91E−08	5.84E−10
	GABOA	2.6E−126	3.95E−67	4.2E−104	**0.00E+00**	**8.88E−16**	3.29E−65
	ORGABOA	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**8.88E−16**	**0.00E+00**
	GWO	6.82E−20	0.678642	1.62E−84	3.679921	2.59E−11	0.001041
	MPA	3.73E−21	0.036556	4.48E−59	0.00E+00	5.79E−12	3.57E−13
标准差	BOA	7.96E−12	7.46E−12	4.8E−14	132.3865	2.57E−09	3.24E−10
	GABOA	9.8E−126	2.12E−66	2.3E−103	**0.00E+00**	1.97E−31	1.23E−64
	ORGABOA	**0.00E+00**	**0.00E+00**	**0.00E+00**	**0.00E+00**	**9.86E−32**	**0.00E+00**
	GWO	6.61E−20	2.016873	8.72E−84	3.826088	1.25E−11	0.001115
	MPA	5.35E−21	0.092594	2.3E−58	**0.00E+00**	2.91E−12	3.53E−13



（a）函数F1达代收敛曲线　（b）函数F2达代收敛曲线　（c）函数F3达代收敛曲线

（d）函数F4达代收敛曲线　（e）函数F5达代收敛曲线　（f）函数F6达代收敛曲线

图 4　函数迭代收敛曲线

见表3，结果表明，ORGABOA比原算法的结果更好，最优变量为$\vec{A}$=[0.78868428 0.40822241]，相应的最优体积为263.89584301。两种算法对三杆桁架的优化迭代收敛曲线见图6。

表3 三杆桁架设计问题的比较结果

优化算法	优化设计变量		优化结果
	A_1	A_2	
ORGABOA	0.788677621	0.40824524	263.8958433
BOA	0.7882447	0.4094669	263.8959796

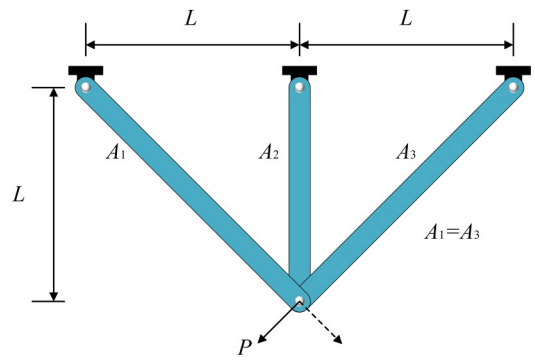


图5 三杆桁架模型

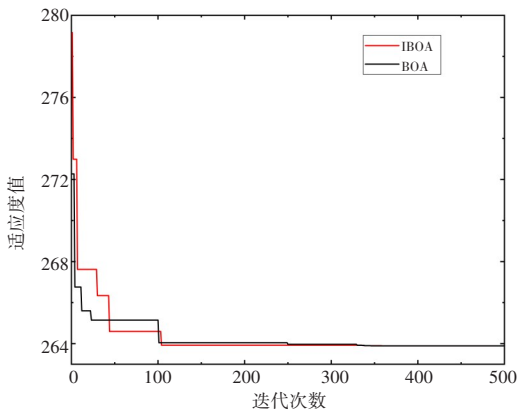


图6 三杆桁架体积的收敛曲线

5 结语

本文对传统的蝴蝶优化算法易造成收敛速度慢、种群多样性较差等问题，提出了一种基于折射反向学习以及自适应化的改进蝴蝶优化算法。通过在算法中引入折射反向学习，再依据蝴蝶个体感官系数自适应化策略，平衡算法的局部开发和全局搜索能力的同时也有利于提高算法的种群多样性。同时实验测试结果表明改进后的算法具有更好的收敛速度和寻优精度，对典型的工程设计问题的测试进一步验证了改进方法的功能性。接下来的研究重点是希望将该改进算法应用在实际的大型工程和社会实践中。

参考文献：

[1] KENNEDY J, EBERHART R. Particle swarm optimization [C] //Proceedings of ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 1995: 5263228.

[2] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer [J]. Advances in Engineering Software, 2014, 69: 46-61.

[3] FARAMARZI A, HEIDARINEJAD M, MIRJALILI S, et al. Marine predators algorithm: a nature-inspired metaheuristic [J]. Expert Systems with Applications, 2020, 152: 113377.

[4] ARORA S, SINGH S. Butterfly optimization algorithm: a novel approach for global optimization [J]. Soft Computing, 2019, 23(3): 715-734.

[5] ARORA S, SINGH S. Node localization in wireless sensor networks using butterfly optimization algorithm [J]. Arabian Journal for Science and Engineering, 2017, 42(8)3325-3335.

[6] SOWJANYA K, INJETI S K. Investigation of butterfly optimization and gases Brownian motion optimization algorithms for optimal multilevel image thresholding [J]. Expert Systems with Applications, 2021, 182.

[7] 李彦苍，卜英乔，朱海涛，等. 融合最优邻域扰动和反向学习策略的蝴蝶优化算法[J]. 中国科技论文，2021，16(11)1181-1188.

[8] ARORA S, SINGH S. Butterfly algorithm with levy flights for global optimization [C] //Proceedings of the 2015 International Conference on Signal Processing, Computing and Control(ISPCC), 2015.

[9] TIZHOOSH H R. Opposition-based reinforcement learning [J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2006, 10.

[10] COELLO C A C. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art [J]. Computer Methods in Applied Mechanics and Engineering, 2002, 191(11/12): 1245-1287.

Engineering application based on improved butterfly optimization algorithm

Chu Min*, Li Xuan, Chen Xuecai

(School of Education, Guizhou Normal University, Guiyang 550025, China)

Abstract: In order to solve the shortcomings of the classical butterfly optimization algorithm, which is easy to fall into local optimum and slow to converge, so an improved butterfly optimization algorithm combining refraction reverse learning and self-adaptation is proposed. After introducing refraction reverse learning to improve population diversity and explore more regions, and adapting parameters, the global and local equilibrium effects of the algorithm are also improved. In this study, four test functions with different characteristics and three-bar Truss were tested. The test results show that the improved algorithm has better convergence speed and search accuracy in the search, and further verify the functionality of the improved method.

Keywords: butterfly optimization algorithm; refraction reverse learning; adaptive strategy; structure optimization; golden sine algorithm

(上接第20页)

Generative adversarial network for motion deblurring algorithm based on frequency transformer

Gu Junhua[1,2], Li Yan[1], Chen Chen[1], Niu Bingxin[1*], Li Chunjie[3]

(1. School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 300401, China;
2. Hebei Province Key Laboratory of Big Data Computing(Hebei University of Technology), Tianjin 300401, China;
3. Hebei Expressway Group Limited, Shijiazhuang 050000, China)

Abstract: For the problem of non-uniform motion blur, the size of the receptive field of the model is quite important. The receptive field of the Transformer based on self-attention mechanism is much larger than that of the traditional CNN model. An efficient processing module in frequency domain is introduced. The directional feature of motion blur presented in frequency domain image is ignored in the previous processing methods combined with frequency domain blur. The PSNR score of the model on GoPro dataset is 32.24 dB, and the score on RealBlur-J dataset is 29.38 dB. Compared with other methods, the effect is improved by 0.72 dB and 0.83 dB on the two datasets, respectively.

Keywords: image deblurring; fourier transform; frequency domain; transformer; generative adversarial network

文章编号: 1007-1423（2023）16-0060-05 DOI: 10.3969/j.issn.1007-1423.2023.16.011

多维Star-QAM星座优化的SCMA码本

张 丽*

（西华大学理学院，成都 610039）

摘要：对一种多维正交振幅调制（QAM）优化的稀疏码多址接入（SCMA）码本进行研究。不同于四环的星形正交振幅调制（Star-QAM），设计的码本增加了一组内外环从而构造了六环的三组子星座，利用四环Star-QAM的证明结果，将三组内外环半径比设置为固定值，以最大化最小欧氏距离（MED）和最小乘积距离（MPD）为优化目标，在多目标粒子群优化算法（MOPSO）的帮助下，构造了性能较优的用户码本。通过数值仿真结果表明，与已有的码本相比，提出的码本在下行瑞利衰落信道下的误码率（BER）更小，性能更优。

关键词：稀疏码多址接入；星形正交振幅调制；粒子群算法；下行瑞利信道

0 引言

随着物联网（IoT）和大规模机式通信（mMTC）设备[1]应用的快速扩展，支持依赖于第四代（4G）长期演进（LTE）[2]继承的物理层解决方案的通信设备具有挑战性。由于第5代（5G）和第6代（6G）通信中频谱稀少且拥挤，正交多址（OMA）可能不适用于物联网应用的要求。为了实现更高层次的连通性和频谱效率[3]，非正交多址技术（NOMA）被认为是一种很有前途的技术。因此，提出了非正交多址技术之一的稀疏码多址（SCMA）[4]。

与通过预先分配的资源元素（REs）重复传输特定星座的低密度签名（LDS）方案相比，SCMA方案将信息位映射到在预先分配的REs[5]上的多维星座，以最大限度地提高码字对的最小欧氏距离。因此，SCMA具有更好的符号映射多样性增益[6]。同时，受益于组合调制和扩展过程[7]，SCMA以更高的系统复杂度为代价，获得了比其他的CD-NOMA技术更好的误差性能，因此SCMA码本的设计尤为重要。

如今，码本的设计仍然是一个开放性的问题，目前已经有多种次优码本的设计，但是并未有最优的码本。Liu等[8]提出了一种将上行无线通信的稀疏码多交调幅（QAM）划分为多个子集，实现优化后的稀疏码多址（SCMA）码本的优化设计。Mheich等[9]使用黄金角调制（GAM）点来构造上行和下行稀疏码多址（SCMA）系统的码本，在下行链路中，使用GAM的点集为SCMA码本构建一个多维母星座，而在上行链路中，GAM的点集被直接映射到用户码本。Zhang等[10]提出了一种构造下行稀疏码多址（SCMA）系统码本的方法，每个子信道的码本都是基于最常用的QAM调制的，同时用低误差概率准则来衡量信道在高斯信道和瑞利衰落信道中的SCMA码本性能。Li等[11]提出了一种新的SCMA码本，它展示了不同用户之间的功率不平衡的下行传输，基于Star-QAM母星座结构，并借助遗传算法，对所提出的码本的最小欧氏距离（MED）和最小乘积距离（MPD）进行了优化。Yu等[12]提出了一种基于星形正交振幅调制（Star-QAM）信号星座的SCMA码本设计方法。基于四环Star-QAM信号星座的MC设计方案，验证当内半径之比为0.63且内外半径之比为1/3时，码本性能最优。

本文在四环Star-QAM信号星座上，再增加

收稿日期：2023-04-27 **修稿日期：**2023-05-16

作者简介：*通信作者：张丽（1998—），女，四川绵阳人，硕士研究生，主要从事稀疏码的研究工作，E-mail:734792847@qq.com

了一组内外圆环，同时将三组内外半径比设置为1/3，外部半径加入旋转参数，三组内半径为尺寸优化参数，利用多目标粒子群优化算法(MOPSO)[13]，最大化最小欧氏距离(MED)和最小乘积距离(MPD)，从而求解出相关参数构造用户码本。仿真结果表明该码本较之前码本相比误码率得到一定控制。

1　SCMA系统介绍与码本设计

1.1　SCMA系统介绍

考虑一个下行链路SCMA系统，其中基站发射机同时与共享K个资源节点的J个用户进行通信，定义系统的过载因子$\lambda = (J/K) > 1$。不同于低密度签名(LDS)，映射器和扩展程序在SCMA中被合并在一起，一个SCMA编码器直接将输入的$\log_2 M$位比特流映射到一个大小为M的K维复数星座X，然后叠加J个用户的码字进行传输。在接收端接收到的信号向量$\boldsymbol{y} = (y_1, y_2, \cdots, y_K)^{\mathrm{T}}$可以表示为

$$\boldsymbol{y} = \mathrm{diag}(\boldsymbol{h})\sum_{j=1}^{J}\boldsymbol{x}_j + \boldsymbol{n} \tag{1}$$

其中：$\boldsymbol{h} = (h_1, h_2, \cdots, h_K)^{\mathrm{T}}$表示发射机到用户的信道向量，且满足$h_k \sim CN(0,1)$，diag(.)表示对角化向量，$\boldsymbol{x}_j = (x_{1,j}, x_{2,j}, \cdots, x_{K,j})^{\mathrm{T}}$表示从码本$X_j$中选择的用户$j$的$K$维复数码字，其中$1 \leqslant j \leqslant J$。$\boldsymbol{n} = (n_1, n_2, \cdots, n_K)^{\mathrm{T}}$表示均值为零和方差为$\delta^2 \boldsymbol{I}_K$加性高斯白噪声(AWGN)向量[14]。

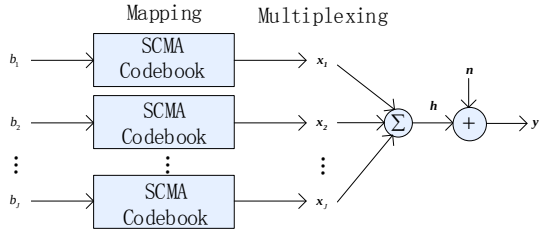


图1　SCMA下行链路模型图

1.2　码本设计

在SCMA信号的发送端，SCMA编码器执行将$q = \log_2 M$编码位的信息映射为大小为M的K维复数码本，该定义为

$$f: \mathbf{B}^{\log_2 M} \rightarrow X, \boldsymbol{x} = f(\boldsymbol{b}) \tag{2}$$

其中：$\boldsymbol{b}$表示输入信道编码比特，$\boldsymbol{x} \in X \subset \mathbf{C}^k$是一个$K$维复数码字的稀疏向量，其中$\|X\| = M$。

本文通过直接构造子星座来获得SCMA用户码本，具体步骤如下：

(1) 构造子星座S_1, S_2, S_3,大小均为$1 \times M$；

(2) 根据因子图矩阵$\boldsymbol{F}_{K\times J}$将$S_1, S_2, S_3$分配到因子图矩阵的非零元素位置，将大小与子星座相同的零向量分配到因子图的零元素位置，从而获得用户码本。

1.3　最小欧氏距离和最小乘积距离

在SCMA系统中，J个用户的码本在K个资源块上同时传输形成叠加码字，构成有M^J个叠加码字的叠加星座V。最小欧氏距离$E_{\min}$是M^J个叠加码字之间$M^J(M^J-1)$种距离中的最小值，它是高斯信道下误码率性能的关键指标，可表示为

$$E_{\min} = \min\left\{\left|x(p)-x(q)\right|, \forall x(p), x(q) \in V, x(p) \neq x(q)\right\} \tag{3}$$

其中第n个叠加码字$\boldsymbol{x}(n)$表示为

$$\boldsymbol{x}(n) = \sum_{j=1}^{J}\boldsymbol{x}_j(m_j), m_j = 1, 2, \cdots, M \tag{4}$$

在瑞利衰落信道下，最小乘积距离$P_{\min}$是主要的性能指标。计算包含所有J个用户码本的$P_{\min}$如下：

$$P_{\min} = \min\left\{\prod_{k=1, x_{j,p}^k \neq x_{j,q}^k}^{K}\left|x_{j,p}^k - x_{j,q}^k\right|, 1 \leqslant p < q \leqslant M\right\} \tag{5}$$

其中：$x_{j,m}^k$表示与用户j的码本相关联的第k个资源块的第m个码字[11]。

2　下行瑞利衰落信道下SCMA的码本设计及优化

2.1　SCMA码本设计

在本文中，首先参照文献[12]中的结论，将三组内外环半径比值设置为1/3，构造三个子向量$\boldsymbol{a}_1 = (-3*R_1, -R_1, R_1, 3*R_1)$，$\boldsymbol{a}_2 = (-3*R_2, -R_2, R_2, 3*R_2)$，$\boldsymbol{a}_3 = (-3, -1, 1, 3)$。其次对三个子向量的外环引入旋转参数，增加自由度，三个子向量表示为$\boldsymbol{A}_1 = (-3*R_1*b_0, -R_1, 3*R_1*b_0)$，$\boldsymbol{A}_2 = (-3*R_2*b_0, -R_2, R_2, 3*R_2*b_0)$，$\boldsymbol{A}_3 = (-3*b_0, -1, 1, 3*b_0)$，其中$b_0 = e^{i*2*\pi*\alpha}$。再对$\boldsymbol{A}_1$,

$\boldsymbol{A}_2, \boldsymbol{A}_3$各自进行整体的旋转操作形成三个子星座$S_1 = b_1 * \boldsymbol{A}_1$，$S_2 = b_2 * \boldsymbol{A}_2$，$S_3 = b_3 * \boldsymbol{A}_3$，其中$b_1 = e^{i*2*\pi*\theta_1}$，$b_2 = e^{i*2*\pi*\theta_2}$，$b_3 = e^{i*2*\pi*\theta_3}$。最后，根据因子图矩阵$\boldsymbol{F}$确定非零向量和零向量的位置，根据签名矩阵$\boldsymbol{S}$确定三个子星座的分配，本文的因子图矩阵$\boldsymbol{F}$、签名矩阵$\boldsymbol{S}$和文献[11]中的因子图矩阵、签名矩阵相同。为了降低码本的平均峰值功率比(PAPR)，对用户码本按照文献[15]的规则进行交织操作。本文的资源块数量为4，用户数量为6，用户码本大小为4 × 4，则用户1的码本可表示为

$$\boldsymbol{X}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -3*R_1*b_0*b_1 & -R_1*b_1 & R_1*b_1 & 3*R_1*b_0*b_1 \\ 0 & 0 & 0 & 0 \\ R_2*b_2 & -3*R_2*b_0*b_2 & 3*R_2*b_0*b_2 & -R_2*b_2 \end{bmatrix} \tag{6}$$

类似地，我们能获得其余的用户码本。

2.2 下行瑞利衰落信道下码本优化

在下行瑞利衰落信道下，以最大化最小欧氏距离$E_{\min}$和最小乘积距离$P_{\min}$为优化目标，优化函数如下：

$$\begin{cases} \max\{E_{\min}; P_{\min}\} \\ \text{s.t. } 1 < param(i) < 3, \forall i = 1, 2 \\ 0 < param(t) < 1, \forall t = 3, 4, 5, 6 \end{cases} \tag{7}$$

最终根据多目标的粒子群优化算法得出结果，如图2所示。

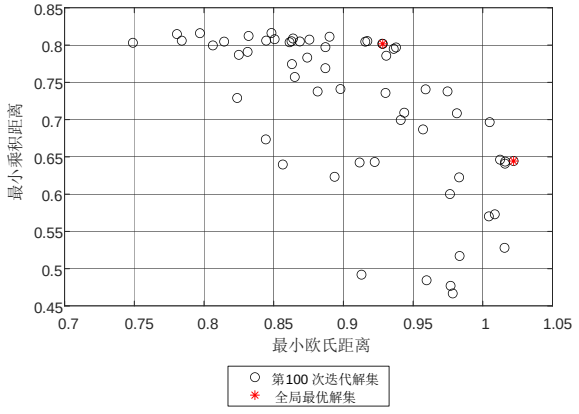


图2 下行瑞利信道下粒子群算法优化图

通过图2可知，经过100次迭代之后，精英库中有两组帕累托(Pareto)解[13]。这两组解则构成全局最优解集，最优解有相应的目标函数值。

3 仿真结果

本文对多维星形正交振幅调制星座优化的SCMA码本进行参数优化，将其命名为Prop码本。将Prop码本与Star-QAM码本[12]、DEMC码本[16]进行比较，为了便于比较，将三种码本的平均能量都定为1，具体情况见表1。考虑多维圆环半径设计标准，Pareto解2的半径参数太过于接近，不符合设计要求，因此舍去，最终选取Pareto解1($R_1 = 1.35, R_2 = 1.50, \theta_1 = 0.30, \theta_2 = 0.58, \theta_3 = 0.85, \alpha = 0.03$)作为目标解。

表1 不同码本比较

	码字平均能量	最小欧氏距离	最小乘积距离
Star-QAM码本	1	0.90	0.72
DEMC码本	1	0.58	0.50
Pareto解1	1	1.02	0.64
Pareto解2	1	0.93	0.80

在高斯信道和下行瑞利衰落信道下，对Prop码本与Star-QAM码本、DEMC码本进行BER性能优劣比较，相关的仿真参数见表2。根据图3的仿真结果所示，在高斯信道下，当$BER = 1e-4$时，Prop码本相比Star-QAM码本有1 db性能增益，相比于DEMC码本有3 db性能增益。在下行瑞利衰落信道下，当$BER = 1e-3$时，Prop码本相比Star-QAM码本、DEMC码本分别有0.5 db和5 db的性能增益。

表2 码本仿真参数

参数名称	参数值
用户数J	6
资源块数K	4
码本大小M	4
信息数量Num	1e5
过载系数λ	150%
迭代次数T	6
信道环境	高斯信道，下行瑞利信道

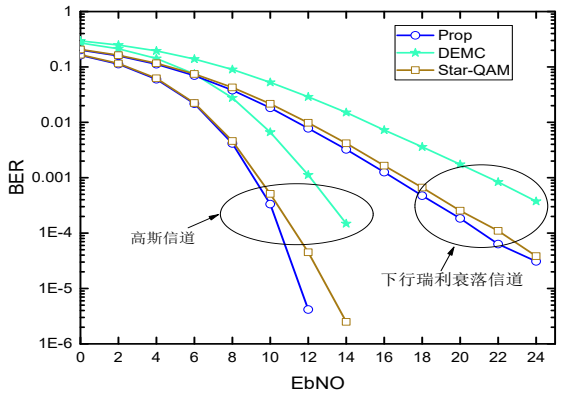


图3　码本误码率比较

4　结语

本文提出的多维Star-QAM星座SCMA码本设计，通过引入新的半径参数、外环旋转算子参数以及子向量的旋转算子参数，增加自由度。再以最大化最小欧式距离和最小乘积距离为目标函数，在多目标粒子群优化算法的帮助下，找出符合要求的目标解，从而求出用户码本。仿真结果显示，本文构造的码本在误码率性能上相比之前的码本有所提高。

参考文献：

[1] LIU P, AN K, LEI J, et al. SCMA-based multiaccess edge computing in IoT systems: an energy-efficiency and latency tradeoff[J]. IEEE Internet Things, 2022, 9(7): 4849-4862.

[2] YUAN Y, WANG S, WU Y, et al. NOMA for next-generation massive IoT: performance potential and technology directions[J]. IEEE Communications Magazine, 2021, 59(7): 115-121.

[3] CHEN Y-M, HSU Y-C, WU M-C, et al. On near-optimal codebook and receiver designs for MIMO-SCMA schemes [J]. IEEE Transactions on Wireless Communications, 2021: 115-121.

[4] NIKOPOUR H, BALIGH H. Sparse code multiple access[C]//Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications(PIMRC), London, UK, 2013: 332-336.

[5] VAMEGHESTAHBANATI M, MARSLAND I D, GOHARY R H, et al. Multidimensional constellations for uplink SCMA systems: a comparative study [J]. IEEE Communications Surveys & Tutorials, 2019, 21(3): 2169-2194.

[6] SAMRA H, DING Z, HAHN P M. Symbol mapping diversity design for multiple packet transmissions [J]. IEEE Transactions on Communications, 2005, 53(5): 810-817.

[7] LEI T, NI S, CHENG N, et al. SCMA-OFDM codebook design based on IRM[C]//Proceedings of the 2021 International Wireless Communications and Mobile Computing(IWCMC), Harbin, China, 2021: 735-739.

[8] LIU S, WANG J, BAO J, et al. Optimized SCMA codebook design by QAM constellation segmentation with maximized MED [J]. IEEE Access, 2018, 6: 63232-63242.

[9] MHEICH Z, WEN L, XIAO P, et al. Design of SCMA codebooks based on golden angle modulation [J]. IEEE Transactions on Vehicular Technology, 2019, 68(2): 1501-1509.

[10] ZHANG Y, LIU Z, XIAO P, et al. A novel SCMA codebook design method based on low error probability criteria[C]//Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing(ICSIP), Nanjing, China, 2021: 937-941.

[11] LI X, GAO Z, GUI Y M, et al. Design of power-imbalanced SCMA codebook[J]. IEEE Transactions on Vehicular Technology, 2022, 71(2): 2140-2145.

[12] YU L, LEI X, FAN P, et al. An optimized design of SCMA codebook based on star-QAM signaling constellations[C]//Proceedings of the 2015 International Conference on Wireless Communications & Signal Processing(WCSP), 2015: 1-5.

[13] COELLO C A C, PULIDO G T, LECHUGA M S. Handling multiple objectives with particle swarm optimization [J]. IEEE Transactions on Evolutionary Computation, 2004, 8(3): 256-279.

[14] CHEN Y M, CHEN J W. On the design of near-optimal sparse code multiple access codebooks [J]. IEEE Transactions on Communications, 2020, 68(5): 2950-2962.

[15] CAI D, FAN P, LEI X, et al. Multi-dimensional SCMA codebook design based on constellation rotation and interleaving[C]//Proceedings of the 2016 IEEE 83rd Vehicular Technology Conference(VTC Spring), Nanjing, China, 2016: 1-5.

[16] HOU Z, XIANG Z, REN P, et al. SCMA codebook design based on divided extended mother codebook [J]. IEEE Access, 2021(9): 71563-71576.

（下转第68页）

文章编号: 1007-1423(2023)16-0064-05 DOI: 10.3969/j.issn.1007-1423.2023.16.012

微博数据爬虫的检测方法研究

黄志高*

（泉州师范学院物理与信息工程学院，泉州 362000）

摘要： 针对常见的分布式网络爬虫提出了一种对策，研究了爬虫检测的方法，并分析了分布式爬虫如何绕过这些方法。通过关注网络流量遵循功率分配的属性来检测分布式爬虫。当我们按请求数量对网页进行排序时，大多数请求都集中在最常请求的网页上。此外，还会有一些普通用户通常不会要求的网页。但是爬虫会请求这些网页，因为它们的算法旨在通过解析网页来迭代请求，以收集爬虫遇到的每个项目。因此可以假设，如果某些 IP 地址频繁用于请求位于功率分配图长尾区域的网页，则这些 IP 地址可以归类为爬虫节点。网络流量数据的实验结果表明，该方法可以有效地识别出0.02%误报的分布式爬虫。

关键词： 分布式网络爬虫；长尾域值；爬虫检测

0 引言

网络爬虫在各个领域用于收集数据，即使目标站点禁止机器人爬虫，某些网络爬虫也会收集数据，某些 Web 服务尝试通过反爬虫程序方法检测爬虫活动并阻止爬虫程序访问网页，但某些恶意 Web 爬虫通过修改其标头值或分发源 IP 地址来伪装自己[1]，从而绕过检测方法，就好像它们是普通用户一样。

一些公司禁止网络爬虫访问他们的网页，原因如下：首先，网络爬虫可能会降低网络服务器的可用性；其次，网络服务器中的内容被视为公司的知识产权。竞争公司可以复制网络服务器中提供的全部数据，竞争公司可能会向客户提供类似的服务。本文研究了传统的反爬虫方法和各种回避技术，表明传统的反爬虫方法不能阻止分布式爬虫。然后，提出了一种新的反爬虫方法，即长尾阈值模型(LTM)方法，该方法逐渐将分布式爬虫的节点IP地址添加到阻止列表中。实验结果表明，该方法能够有效识别误报率为0.02%的分布式爬虫。在传统的基于频率的方法中[2]，当增加阈值以检测更多的爬虫节点时，误报也会相应增加。

1 传统反爬虫方法及其缺陷

1.1 使用HTTP标头信息进行过滤

基本爬虫程序发送请求而不修改其标头信息，Web 服务器可以通过检查请求标头来区分合法用户和爬虫程序，此标头检查方法是一种基本的反爬虫方法。但是，如果爬虫试图将自己伪装成合法用户，它将使用来自 Web 浏览器的标头信息或类似于浏览器的 HTTP 标头信息重置[3]。这使得 Web 服务器很难通过简单地检查请求标头来确定客户端是爬虫程序还是合法用户。

1.2 基于访问模式的反爬虫

基于访问模式的反爬虫方法根据客户端生成的请求模式将合法用户与爬虫程序进行分类。如果客户端仅连续请求特定网页，而不调用通

收稿日期： 2023-04-17 **修稿日期：** 2023-08-08

基金项目： 2018年福建省中青年教师教育科研项目(JT180381)

作者简介： *通信作者:黄志高(1982—),男,福建泉州人,讲师,硕士研究生,研究方向为网络通信技术、数字图像处理,E-mail:45684906@qq.com

常应请求的网页，则该客户端将被视为爬虫程序。执行主动爬虫的爬虫程序预定义了爬虫程序想要收集的核心网页，爬虫程序请求特定的网页而不请求不必要的网页。在这种情况下，Web 服务器可以识别客户端不是合法用户。通过分析客户端访问模式的Web服务，该服务可以根据预定义的普通用户的访问模式将爬虫程序与普通用户区分开来。虽然这种方法可以根据访问模式识别爬虫，但一些爬虫甚至通过分析网络日志来伪装他们的访问模式。

1.3　基于访问频率的反爬虫

基于访问频率的反爬虫方法通过访问频率阈值作为特定时间范围内的最大访问次数来确定客户端是爬虫还是合法用户。如果来自客户端的请求数在预定义的持续时间内超过某个阈值，则 Web 服务器会将客户端分类为爬虫程序。这种方法有两个众所周知的问题。首先，它对分布式爬虫有漏洞。如果攻击者使用分布式爬虫(如Crawlera)，则可以管理每个爬虫节点的访问速率保持在阈值以下[4]。其次，普通用户和爬虫程序共享单个公共 IP 地址，容易被误识别为爬虫程序。

2　阻止分布式爬虫

如上所述，分布式爬虫程序可以绕过传统的反爬虫方法。我们提出了一种新技术来检测和阻止传统反爬虫技术无法防御的分布式爬虫。

2.1　所需的爬虫节点数

为了使分布式爬虫收集网站的全部数据，必须满足以下条件：

$$C_n \geqslant \frac{U_m}{(T_d*30)} \tag{1}$$

其中:U_m是一个月内更新的项目数，T_d是每个 IP 地址的最大请求数，C_n是爬虫程序节点数(IP 地址)，30 是一个月的天数。T_d乘以 30 得到每月的请求数。

爬虫程序节点需要收集每月所有更新的数据。每月更新数据数除以每月最大请求数。例如，如果一个月内有 Web 服务更新30000个项目，并且该服务具有限制规则，即具有多个(例如100)请求的 IP 地址将被阻止，并且尝试从 Web 服务收集每个项目的攻击者将需要例如 10 个爬虫程序节点来避免限制。因此，随着U_m增加或T_d减少，应该增加C_n并以数字表示网站难以抓取的级别。

2.2　生成长尾区域

一个月内更新的项目数量不能随意增加。因此，防止分布式爬虫的一种简单方法是T_d减少，但这也会显着增加误报。在本文中，我们通过反转 Web 流量的一般特征并利用分布式爬虫尝试复制 Web 服务器的整个数据的事实来解决这个问题。如果项目按访问率排序，我们可以在图表中看到指数递减曲线，如图 1 所示。大多数网络流量集中在最常请求的项目上，并且有一个长尾区域具有较低的访问率。我们计算了此长尾区域的最大请求计数，并将此值设置为T_{d3}。

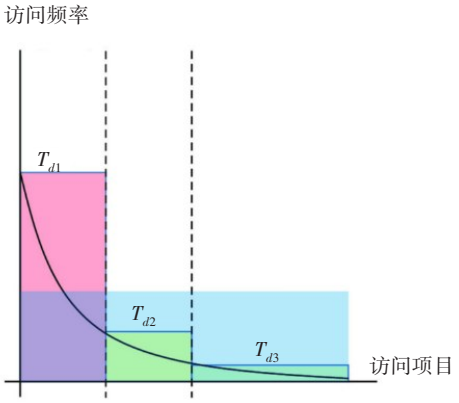


图 1　每个链接的访问频率

在信息论中，不太可能的事件比可能的事件更具信息量，而长尾地区的事件比其他事件更不可能。这意味着 Web 服务可以从长尾区域的请求中查找更多信息。因此，当客户端不断请求长尾区域中的项目时，Web 服务可以增加计数，直到达到T_{d3}，而不是达到T_d平均值。这意味着 Web 服务可以设置更敏感的阈值，而不会增加误报率。

2.3　长尾区域节点缩减

为了使攻击者从 Web 服务收集整个数据，攻击者还必须访问长尾区域中的项目。但是，

攻击者并不确切知道哪些项目属于长尾区域。利用这种信息不对称性，服务提供商可以轻松识别比其他IP地址更频繁地访问项目的IP地址。这些已识别的爬虫程序的IP地址将包含在阻止列表中，并且阻止列表中的IP地址数将为C_m。如果我们开始通过长尾间隔增加该C_m值，攻击者将使用较少数量的IP地址进行爬行，并且会在T_d间隔内增加C_m[5]。

2.4 虚拟项目

服务提供商可能会添加虚拟项目来检测爬虫程序，并且合法用户无法访问虚拟项目，因为虚拟项目没有用户界面或隐藏。生成虚拟项目的方法很少，它可能以HTML标签的形式存在，但属性设置不会显示在屏幕上，或者可能包含普通用户不感兴趣的垃圾信息。但是，对服务执行顺序访问的爬虫程序可能会访问虚拟项目。通过这一特性，虚拟项目可以作为长尾区域的延伸。在本文中，我们不会在实验中包含虚拟项目，以便与不包含任何虚拟项目的真实流量日志进行公平比较[6]。

3 实验测试

实验旨在评估爬虫检测模块对网络流量的分类性能。将LTM方法与基于正常访问频率的反爬虫方法在爬虫节点的最大数量和误报率上进行了比较[7]。在实验中使用这个数据集有两个因素，一个是用户数，另一个是网站中的项目数。用户数量很重要，因为如果用户数量较少，某些用户可能会偏向流量模式。为了实现这一目标，我们开发了一个基于Python的数据工具和一个模拟器。在数据预处理工具中，如图2所示对原始流量数据进行预处理，以计算单个URL的访问频率，并对属于长尾区域的集进行分类。每当发生新的访问时，模拟器根据预处理的数据确定访问节点是否为爬虫程序。

3.1 数据源

NASA在2005年7月共公布了2493425份访问日志。我们将这些日志解析为csv格式，该格式由四列组成，包括IP地址、日期、访问目标和访问结果。连接的IP地址总数为41958，项目数为21534。

3.2 数据预处理和流量分配

我们在实验的预处理阶段执行了三个步骤。第一步将日志拆分为两个数据集：训练集和测试集。在NASA访问日志中，前24天日志设置为训练集，最后一个日志设置为测试集。第二步筛选出一些访问日志，以计算更准确的访问计数。某些请求合并为单个请求，以防止重复计数。例如，当用户访问html文件时，他们还可以访问链接的图像文件。这可能会强制访问计数成倍增加。因此，我们删除了一些对图像文件的请求。此外，我们还从实验中排除了访问结果不成功的请求日志。

表1 预处理的网络流量数据

指标	数值
总项目数/个	6649
长尾区域项目数/个	5355
平均访问计数	184.76
长尾项目访问计数	1.88

实验中构建了一个预处理的流量数据集，该数据集由来自21649个原始数据的6649个项目组成，并且它有一个由5355个项目组成的长尾区域。总访问数平均值为184.76，长尾区的平均访问数为1.88。两个平均值之间的差异只是表明可以在爬虫检测算法中设置一个更灵敏的阈值。

实验还统计了按访问频率排序时最常访问的链接，到最不常访问的链接的特征[8]。比率是指按访问计数对所有项目进行排序时每个组所在的间隔。访问平均值是指属于每个组的每个项目的平均访问次数，最大访问量是指每个组中项目之间的最大访问计数。

3.3 虚拟仿真

在仿真实验中，检查LTM是否能够检测和禁用分布式爬虫程序IP地址组，将实际Web流量输入到LTM时检查误报率。整个爬虫程序检

测流程如图2所示。

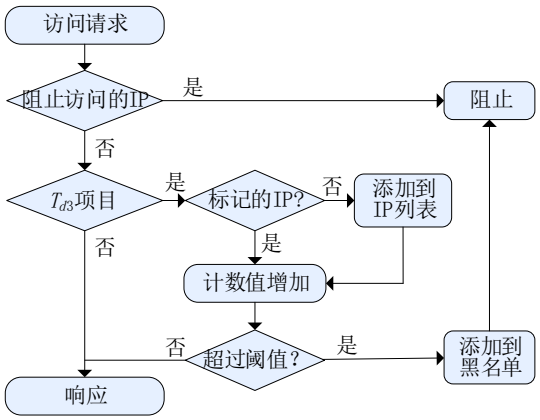


图2　爬虫检测流程

3.4　分布式爬虫检测模拟

从表2的实验数据，我们可以观察到爬虫IP地址集合逐渐减少，直到所有IP地址都被完全阻止。当第一个爬虫节点IP地址超过T_{d3}阈值IP被封禁，节点减少计数呈指数级增长。这是因为其他爬虫节点得到了更多的访问负荷，并且当爬虫节点被访问时，更多的项目必须访问被阻止。

表2　实验数据

	域值	最大节点数/个	错误率/%
LTM	10	426	0.1239
LTM	20	222	0.0275
LTM	35	128	0.0046
FBA	10	573	8.8064
FBA	20	299	2.8819
FBA	35	389	19.446

实验阈值设置为20，爬虫项目包含222节点。LTM检测到了整个爬虫项目所有节点，错误率约为0.0275%，远低于传统的基于频率的爬虫检测方法。在表2中，我们将LTM的结果与通用的基于频率的反爬虫(FBA)方法的结果进行了比较。LTM对分布式的爬虫检测性能非常依赖于项目的数量和长尾比率。考虑到这个限制，仿真实验是使用旧的NASA交通数据进行[9]，项目的总数比通常的现代网络项目要小得多。如果有10倍数据和类似访问的频率分布的项目，我们提出的方法可以从由2000个节点组成的网络项目中检测出分布式爬虫。

在实验中，LTM达到了最小的错误率为0.0046%，而FBA仅达到0.0367%，这意味着LTM的检测可靠性比经典FBA方法提高了500%。当我们将阈值设置为35，这时LTM达到最小误报率的值，FBA方法比LTM方法多了19.400%的误报率。

4　结语

本文介绍了长尾阈值模型(LTM)，并展示了LTM如何有效地检测分布式爬虫，相比之下先前的方法是脆弱的。通过模拟真实的网络流量数据，LTM有效地识别了分布式爬虫，并显示出极低的误报率。针对网络服务的非法网络爬取成为严重的安全威胁[10]。考虑到一些爬虫开发者将分布式爬虫代理服务用于非法目的，LTM可以提高网络服务的数据安全性。

参考文献：

[1] 孙握瑜. 基于Python的新浪微博爬虫程序设计与实现[J]. 科技资讯,2022,20(12):34-37.
[2] 陆莉莉. 基于爬虫的社交平台舆情用户追踪系统设计与实现[J]. 电脑知识与技术,2022,18(1):26-28.
[3] 鲜敏. 面向海量网络数据的收集方法研究与分析[J]. 电脑编程技巧与维护,2021(8):64-66.
[4] 陈肖润东. 网络爬虫行为反不正当竞争法规制[D]. 北京:北京交通大学,2021.
[5] 郑明丽. 基于深度学习的网络虚假信息识别研究[D]. 武汉:中南财经政法大学,2021.
[6] 黄红桃,杨玉翔. 微博网络水军数据获取研究[J]. 信息与电脑(理论版),2021,33(9):169-171.
[7] 褚祎. 基于微博舆情监测系统的研究与设计[D]. 西安:西北大学,2021.
[8] 王小兵. 基于长短期记忆网络的新冠疫情微博情感分析研究[D]. 合肥:安徽大学,2021.
[9] 方自建. 基于爬虫技术和文本聚类的COVID-19舆情分析[D]. 湘潭:湘潭大学,2021.
[10] 董少林,李钟慎. 采用Scrapy分布式爬虫技术的微博热点舆情信息获取与分析[J]. 电脑与信息技术,2020,28(5):23-26.

Research on detection method of Weibo data crawler

Huang Zhigao*

(School of Physics and Information Engineering, Quanzhou Normal University, Quanzhou 362000, China)

Abstract: This paper proposes a countermeasure against common distributed web crawlers, studies the methods of crawler detection, and analyzes how distributed crawlers bypass these methods. Detect distributed crawlers by focusing on the property that network traffic follows power distribution. When we sort web pages by number of requests, most requests are concentrated on the most frequently requested web pages. In addition, there will be pages that the average user would not normally request. But crawlers request these web pages because their algorithms are designed to iterate the request by parsing the web page to collect every item the crawler encounters. Therefore, we can assume that certain IP addresses can be classified as crawler nodes if they are frequently used to request web pages located in the long-tail region of the power distribution graph. Experimental results on network traffic data show that the method can effectively identify distributed crawlers with 0.02% false positives.

Keywords: distributed web crawler; long tail domain value; crawler detection

(上接第63页)

Multi-dimensional Star-QAM constellation optimized SCMA codebook

Zhang Li*

(College of Science, Xihua University, Chengdu 610039, China)

Abstract: Research on a Sparse Code Multiple Access(SCMA) Codebook Optimized for Multidimensional Orthogonal Amplitude Modulation(QAM). Different from the four-ring quadrature amplitude modulation (Star-QAM), the designed codebook constructs three subconstellations of six-ring by adding a set of inner and outer rings. Using the proof results of four-ring Star-QAM, the radius ratios of the inner and outer rings of the three groups are set to a fixed value, and the maximization of the minimum Euclidean distance (MED) and the minimum product distance (MPD) is taken as the optimization objective. With the help of multi-objective particle swarm optimization(MOPSO), a user codebook with better performance was constructed. Numerical simulation results show that compared with existing codebooks, the proposed codebook has lower bit error rate(BER) and better performance in downlink Rayleigh fading channels.

Keywords: sparse code multiple access; star quadrature amplitude modulation; particle swarm optimization; downlink rayleigh channel

文章编号: 1007-1423（2023）16-0069-04 DOI: 10.3969/j.issn.1007-1423.2023.16.013

采用CNN进行中文文本分类

火善栋*

（重庆三峡学院计算机科学与工程学院，万州 404100）

摘要：通过卷积神经网络，运用两种不同的方案对中文文本分类进行了对比实验，实验结果表明，对于中文文本分类问题，仅采用CNN中的全连接层并选择合适的特征词和激活函数就可以达到比较理想的分类效果。

关键词：卷积神经网络；中文文本分类；特征词；激活函数

0 引言

文本分类是自然语言处理领域最活跃的研究方向之一，从新闻分类、商品评论信息情感分类到微博信息打标签等辅助推荐系统都有着非常广泛的应用。随着深度学习的不断推广和发展，图像分类与识别也变得越来越成熟和简单，有很多人开始尝试采用深度学习进行文本分类，但是，文本有别于图像，文本分类的一个首要问题就是如何把文档转换成神经网络可以识别和计算的数字表达形式。对于文档的数值表示，现在比较流行的有独热（One-Hot）编码模型、基于词频统计的词袋模型以及Word2Vec[1]模型等。基于词频统计的词袋模型是文档分类中用的比较多的文档向量转换模型，词袋模型通过统计每个特征词的TF_IDF[2]值来将文本文档转换成对应的文档向量，再通过比较这些文档向量的相似度，实现文本分类问题，但这种模型存在维度高、特征稀疏等特点，导致在进行文本分类时存在计算量大、效率低的问题。为了有效解决这些问题，本文基于词袋模型，通过采用提取一定数量的特征词构建特征词典，来大大缩小文档向量的维度，然后通过CNN采用两种不同的方案进行训练和测试，实验结果表明，仅采用CNN中的全连接网络就可以比较好地实现中文文本分类。

本文采用两种方案对九类（环境、交通、教育、经济、军事、体育、医药、艺术、政治）共1995篇中文文档采用PyTorch构建CNN[3]进行训练和测试（其中训练文档为1815篇，测试文档为180篇）。

1 方案1：只采用CNN全连接层进行实验

其实验和测试具体步骤如下：

1.1 构建训练集文档的特征词典

（1）读取每一类文档中的所有文档，过滤掉其中的非中文符号，并进行中文分词；

（2）过滤掉分词中的单字词，统计每类文档中每个词出现的次数，按单词出现的次数从大到小进行排序；

（3）为了减少特征词的数量，只从每类文档中选取一定数量的词（本实验分别对每类排在前1000、1500、2000个词进行了提取）；

（4）对每类文档中抽取的特征词进行合并，去掉重复的词，构成训练文档特征词典并以npy文档的形式进行保存。

收稿日期：2023-04-30 **修稿日期：**2023-05-11

作者简介：*通信作者：火善栋（1974—），男，湖北孝感人，硕士，讲师，研究方向为智能信息系统，E-mail：463727715@qq.com

1.2 统计逆文档频率(IDF)

(1)以特征词典为基础，遍历所有的训练文档，统计特征词典中每一个特征词包含训练文档的数目并计算其IDF，其计算公式为

$$IDF = \log\frac{docNumber}{number + 1} \quad (1)$$

其中：*docNumber*为训练文档的文件总数，*number*为包含特征词的文件数目。

(2)将所有特征词的IDF值以npy文档的形式进行保存。

1.3 构建每一个训练文档的特征向量(TF_IDF)

(1)以特征词典为基础，统计每一个特征词在每一篇训练文档中的词频(TF)，词频等于特征词在对应文档中出现的次数除以该文档的总词条数。

(2)将每一个特征词的TF与IDF相乘得到特征词典中每一个特征词的TF_IDF值，从而构建所有训练文档的文档向量，为了便于训练，将每一个训练文档的文档向量以npy文档的形式进行保存。

1.4 采用CNN全连接网络层进行网络训练

为每一个训练文档向量添加对应的类别标签(1～9)并打乱次序，以16个训练文档为一个批次进行训练，其网络模型如图1所示。

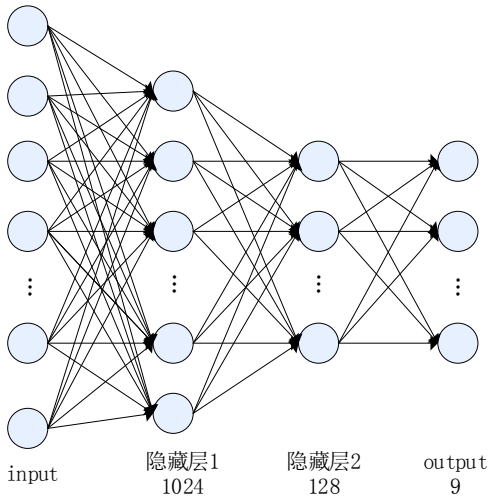


图1 CNN全连接网络的结构

本文对每类文档按词频进行降序排列，分别选取每类排在前1000、1500和2000的词条构建特征词典进行训练(其文档特征向量构建方式和激活函数如表1所示)。图2为选取每类排在前1500个词条构建特征词典所得到的训练迭代次数与损失误差结果图，从训练结果图可以看出，当迭代次数达到2300次左右时，其损失误差的最小值基本处于稳定状态，为了防止过拟合，训练时采用了dropout策略，发现训练损失达到一个稳定值后仍然出现一些细微的波动。

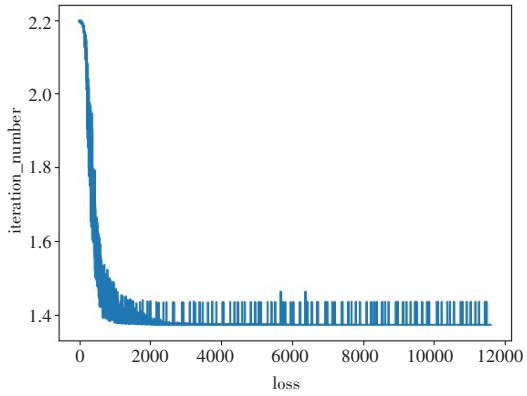


图2 全连接(tanh)训练结果

1.5 训练模型测试

导入训练过程中保存的特征词典文档和IDF文档得到训练集特征词典和每个特征词的IDF值，并以特征词典和IDF值为基准对180篇(每类分别为20)测试文档进行类别判断，

其测试结果如表1所示，从测试结果来看，分类的准确性都在96%以上。

表1 卷积网络全连接层实验结果

每类文档选取特征词数量	训练集文档特征词典词数量	激活函数	正确率/%	
			TF	TF_IDF
1000	5288	regu	97.2	96.1
		tanh	97.2	97.8
1500	7854	regu	96.7	97.8
		tanh	96.7	98.3
2000	10350	regu	96.1	97.8
		tanh	96.7	98.3

2 方案2:采用CNN进行实验

方案2的实验步骤与方案1的实验步骤基本相同，不同的是方案1的文档特征向量是一维的，而方案2的文档特征向量是二维的(模仿单

通道图片），在进行特征输入时，将一维文档特征向量转换成$M \times M$或者$(M-1) \times M$的二维文档特征向量，进行转换时后面位数不够的地方以0进行填充。其向量转换效果如图3所示。

$$[1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10] \Rightarrow \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 0 & 0 \end{bmatrix}$$

图3　一维向量转二维向量示意图

方案2网络结构如图4所示。

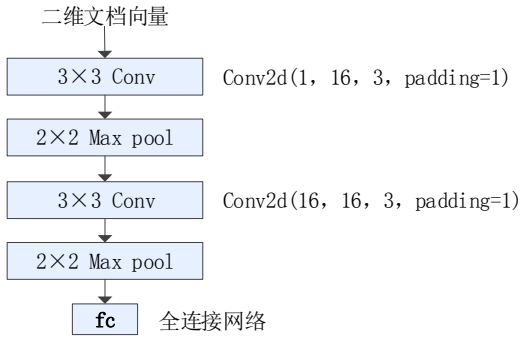


图4　方案2网络结构实验

从图4可以看出，方案2其实就在方案1全连接网络的基础上添加了两个卷积层和两个最大池化层，以每类文档取前1500特征词条为例，其网络训练迭代次数与损失误差结果如图5所示，从训练结果来看，当迭代次数达到2500时，其最小损失误差基本处于稳定状态，但是损失误差的波动范围比较大，其训练效果没有单一的全连接网络好。

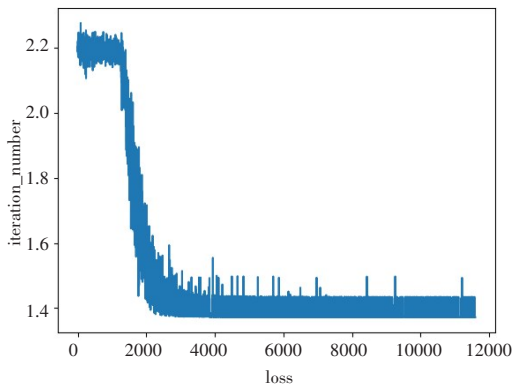


图5　卷积神经网络训练结果

与方案1类似，采用180篇文档对卷积网络的训练结果进行了测试，其实验结果见表2，从实验结果来看，采用卷积对文档特征向量进行降维和特征提取，虽然能够取得一定的准确率，但是其分类效果不如全连接网络。

表2　卷积网络实验结果

每类文档选取特征词数量	训练集文档特征词典词数量	激活函数	正确率/% TF_IDF
1000	5288	tanh	86.1
1500	7854	tanh	86.7
2000	10350	tanh	89.4

3　结语

通过以上实验结果可以看出，对于中文文本分类问题，当采用TF_IDF方法来构建文档特征向量时，为了避免文档特征向量的维度过大，可以先根据每个词条在每类文档中出现的次数，从中提取一定数量的特征词来构建特征词典。从表1的实验结果来看，当提取的特征词达到一定数量时，随着特征词数量的增多，对文档分类结果的影响非常小，因此，对于中文文本分类问题，可以通过这种方法来降低特征词典的长度，从而达到提高训练和分类速度的目的；另外，从表1的实验结果可以看出tanh激活函数略优于regu( )，TF_IDF统计方法总体上略优于TF统计方法；相比于卷积神经网络，全连接网络更适合做文本分类，而且准确率比较高，运算量比较小。

参考文献：

[1] TSHITOYAN V, DAGDELEN J, WESTON L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature[J]. Nature, 2019, 571: 95-98.

[2] 宗成庆. 统计自然语言处理[M].2版. 北京：清华大学出版社，2013.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks[C]//Proceedings of the Advances in Neural Information Processing Systems, 2012, 25: 1106-1114.

（下转第80页）

开发案例

文章编号: 1007-1423（2023）16-0072-09 DOI: 10.3969/j.issn.1007-1423.2023.16.014

中成药多背景数据协作共享与可视化系统的设计与实现

阎玉婷*，李雨红，郑水飞

（江西中医药大学计算机学院，南昌 330004）

摘要：近年来，中成药的使用频率和认可度逐步提升，相关信息平台、医疗终端，以及互联网持续产生海量多背景中成药数据。这些数据来源众多、演化较快、不易跟踪利用，难以有效支撑中成药相关领域的管理、决策与服务。基于此，利用Python爬虫技术，采集中成药大数据，尝试整合多种中成药信息资源，利用PostgreSQL数据存储技术，构建包括中成药政策、技术与流通等多背景的数据库；其次使用Flask框架设计实现系统相关的功能与界面，为不同权限的用户提供不同程度的数据管理服务；最后借助ECharts技术对用户关心的一系列药品指标数据进行可视化展示。系统力图改善中成药多背景数据孤立现象，更好地为领域专家或普通用户提供中成药多背景数据的协作共享与可视化服务。

关键词：中成药；Python爬虫；数据协作共享；ECharts可视化；知识图谱

0 引言

中医药是我国重要的卫生、经济、科技、文化和生态资源，传承创新发展中医药是新时代中国特色社会主义事业的重要内容[1]。随着中成药使用规模的不断扩增，药品说明书不明确、用药不规范、安全性评价体系不完善等问题逐渐显现，且目前针对中成药多背景数据而建立的数据库和应用系统较少，很多中成药信息得不到及时的更新，不同数据之间的关系也难以得到很好的挖掘利用，这一定程度上给患者的健康安全带来隐患，导致用药不良事件的发生[2]。

基于此，本文构建中成药多背景数据协作共享与可视化系统，旨在通过中成药数据资源协作共享及可视化方式增强中成药数据管理，降低中成药不良反应发生率，实现信息交流与共享，积极为中成药使用规范化和中医药行业发展提供服务。

1 研究现状

中医药信息资源所涉及的范围与学科越来越广泛，海量的中医药数据通过互联网、各类医疗终端设备等渠道产生，围绕这些数据的采集、分析、应用而展开的研究也日益丰富。为了实现面向用户临床与科研需求，建立多学科、多类型的个性化、特色化中医药信息数据存储与研究利用，更好地服务于读者和用户[3]，中医药信息科研工作者们一直以来都想要寻找便捷、有效的方式来对这些信息资源进行管理，来实现数据资源的有效利用。

目前，我国已经建立了一些中医药科技文献数据库和中医药事实型数据库[4]，这些数据库是我国中医药信息数字化的成果，为中成药的应用和研究提供了有力支撑。但这些数据库也存在以下问题：一是数据内容单一化明显，具体表现为数据来源主要是医院信息科和相关研究机构，且这些数据大多是属于同一类型，

收稿日期：2023-05-05 **修稿日期：**2023-06-02

基金项目：国家级大学生创新创业训练计划项目（202210412026）；江西省卫生健康委科技计划项目（202211404）

作者简介：*通信作者：阎玉婷（2001—），女，山东青岛人，在读本科，研究方向为计算机科学与技术，E-mail：1033302674@qq.com；李雨红（1999—），女，江西抚州人，在读本科，研究方向为计算机科学与技术；郑水飞（1998—），男，江西上饶人，在读本科，研究方向为医学信息工程

如中成药本身具有的属性：药品名称、成分、主治等，缺乏流通信息，如广告招商等。二是数据获取和更新不及时，目前中成药数据信息的标准化水平不一，导致数据清洗复杂程度增加，工作量增大，进而降低了数据更新的及时性。三是数据展示直观性不好，数据的呈现方式主要是表格和列表，难以对数据进行直观的理解和分析。

在计算机分类科学中，利用人眼的感知能力对数据进行交互的可视化表达以增强认知的技术，称为可视化[5]。可视化不仅仅是对海量数据的简单描述，也能通过数据挖掘技术，整合各类资源，帮助人们挖掘到不可见或难以直接显示的数据关系，从而提高数据分析、决策的能力以及数据价值转化的效率。随着数据可视化在中医药领域的研究、探讨、应用日益增多，暴露出的问题也随之增多，主要为以下几个方面：①不同领域可视化的表示方法各不相同，难以提出规范性的可视化方法；②随着数据网的复杂程度越来越高，单一的可视化方法难以表现出数据的内在特性与价值；③中医药数据之间多存在强关联，不同的处理方式、显示程度会导致可视化结果不明显[6]。基于此，许多学者尝试对可视化进行研究和探索，例如：肖永飞[7]提出了医学数据三维可视化交互的解决方案；李正等[8]对中药制药过程中运用的数据集成、数据挖掘与可视化技术进行了具体的探讨。

在此基础上，本文论述了如何利用Python爬虫技术和PostgreSQL数据库，构建涵盖中成药基础信息、技术与标准、流通广告、行业政策等多背景的动态数据库，同时结合ECharts可视化技术探究数据价值、展示数据关系，以期改善中成药多背景信息孤立现象，为领域专家进一步研究提供数据服务支持。

2　系统设计

2.1　总体设计

基于前期的实践调研和文献查阅进行系统需求分析，利用爬虫技术从互联网获取中成药相关数据，对数据进行实体识别、属性提取、整理分析，存入对象关系型数据库PostgreSQL，构建中成药多背景动态数据库；同时采用数据挖掘等技术对数据关系进行多维度探讨，最后利用具有良好扩展性和兼容性的Flask框架和ECharts技术实现数据的增删改查、批量导入导出和可视化等功能。系统具体实现流程如图1所示。

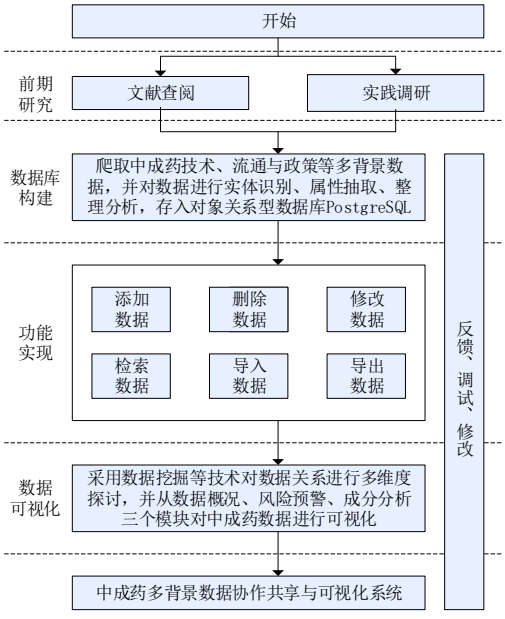


图1　系统具体实现流程

2.2　数据库设计

根据系统功能需求分析，将数据整理识别出12个实体，并为每个实体提取相应的属性，以期提高数据的标准化和规范化。系统主要实体间联系E-R图如图2所示，系统主要实体的

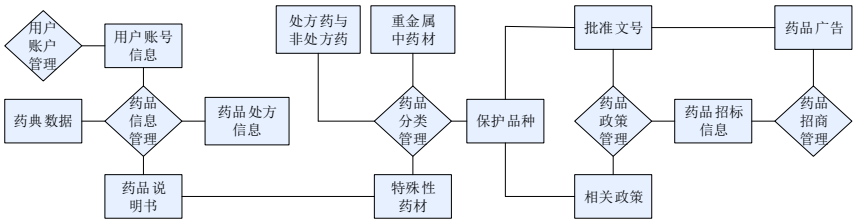


图2　系统主要实体间联系E-R图

属性见表1。

表 1 系统主要实体的属性表

实体	属性
用户账户	用户名、密码、用户权限
药品处方	药品名称、英文名称、处方、特征、功能主治、用法用量、注意事项、规格、存储、临床应用、来源
药典数据	药品名称、拼音名称、处方、治法、形状、鉴别、检查、含量测定、功能主治、用法用量、注意事项、规格、储存
药品说明书	标题、备注、正文
处方药与非处方药	药品名称、现用通用名、规格、甲乙分类、是否双跨、公告发布时间、公告链接
特殊性药材	药品名称、分类、毒性等级、毒性分类依据
重金属药材	药材名称、标准来源、配方
保护品种	药品名称、保护品种编号、剂型、药品批准文号、保护终止日、生产企业、企业链接
批准文号	药品名称、药品规格、生产单位、批准文号、同类别公司数、批准日期、医保类别
相关政策	发布日期、政策正文链接
药品招标	通用名、商品名、剂型、规格、包装转换比、单位、中标价、质量层次、生产企业、投标企业、中标省份、发布日期
药品广告	药品名称、生产企业、广告批准文号、处方分类、药品批准文号、广告生效日期、广告失效日期、详细信息、省份

系统在用户信息表中专门设置“用户权限”属性，只有拥有相应权限的账号，才能对系统进行操作，以期解决中成药数据库建设过程中出现的数据库完整性、安全性和隐私保护问题。用户权限与功能对应图如图3所示。

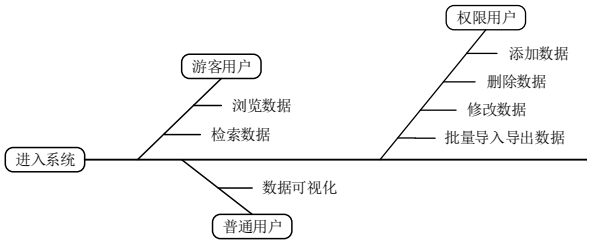


图 3 用户权限与功能对应图

2.3 系统功能设计

本文采用“前端页面+Flask框架后端+PostgreSQL数据库”的架构设计，系统具有较强的扩展性和可维护性。前端主要应用HTML、CSS、JavaScript设计Web前端页面，包括主页、登录/注册、中成药、药典数据、药品招标、相关政策等页面。中成药数据动态采集与可视化系统的实现分为前端和后端两个部分。后端则是使用Python语言连接数据库，与前端进行数据交互，审核和响应用户的操作，实现中成药数据的分页展示、增删改查、批量导入导出等功能。系统架构和交互流程如图4所示，系统页面结构设计如图5所示。

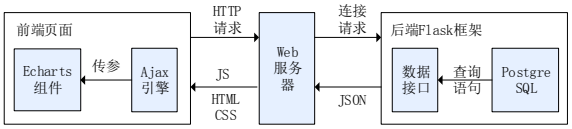


图 4 系统架构和交互流程

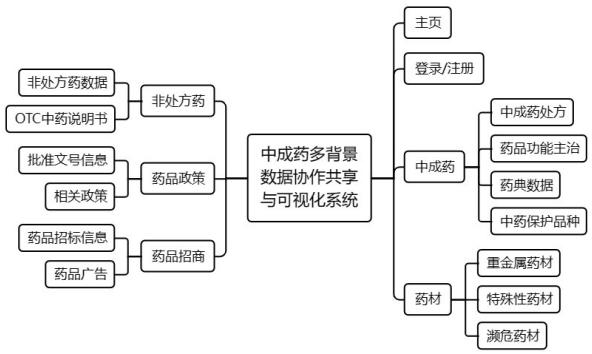


图 5 系统页面结构设计

2.3.1 数据增删改查模块设计

系统在每个中成药数据共享页面都提供了查询功能，用户既可以进行模糊查询，也可以针对某一信息进行精确查询，以便用户更快捷地获取信息。该系统每一个页面提供的查询字段不同，是通过相关性分析和热度分析筛选出对用户而言更有价值的字段，在保证数据有用性的同时，也使系统简洁美观，从而提高用户的使用体验。一般用户和科研工作者可以登录系统查询中成药的处方、功能主治、注意事项等中成药相关技术指标，中成药生产商则可以对药品广告、批准文号、招标信息等流通信息进行查询。

系统在为用户提供中成药多背景数据共享的同时，还允许用户之间相互协作，共同完善

数据库系统内的数据资源。但为了保障数据库系统内数据的安全性和系统运行时的稳定性，增加数据、删除数据、修改数据功能只对权限用户进行开放，游客用户和普通用户无权进行此类操作。

（1）添加数据功能：该功能可以帮助用户将来源不同的数据资源整合为可用的数据集，为用户提供更加便捷和高效的数据协作共享服务，用户在输入对应的字段后即可完成对单条数据的增加。

（2）删除数据功能：中成药数据有着较强的时效性，系统中存储的数据可能会随着时间的推移而变得过时或者出现重复存储的情况，用户可以使用删除数据功能来释放存储空间，降低数据冗余。当用户点击删除时，系统会弹出确认提示框，点击“确认”按钮后即可完成删除操作。

（3）修改数据功能：当用户发现某条数据有细微差错或流通信息有滞纳性时，可以使用该功能对该数据进行修改和更新。用户点击某条数据的“编辑”按钮后，系统将弹出该条数据内容编辑框，用户可以对相应字段进行更改，点击更新即可完成对数据的修改。

2.3.2　数据批量导入导出模块设计

为了提高对数据的处理效率，减少手动输入数据的错误率，系统提供了批量导入数据和批量导出数据的功能。用户点击“上传文件”按钮，在弹出的文件选择框中选择已根据字段整理好的Excel文件，再点击“批量导入”按钮即可完成大量数据的写入。这使得用户不必一条条添加数据，大大节省了时间，提升了协作效率。同时，用户也可以根据需求批量导出数据。用户可以对整个数据库进行下载，也可以先对数据进行条件检索，然后点击“导出数据”按钮将检索结果以Excel的格式保存到本地电脑，便于对数据进行分发和研究。同样因为考虑到数据安全和系统稳定性，该部分功能也只面向权限用户开放。

2.3.3　数据可视化与应用模块设计

数据可视化可以将数据之间潜在的关系以图表的形式形象直观地展示出来，帮助用户从不同的角度分析和研究数据，使其更好地理解和发现数据的规律和趋势。因此，该系统将爬取到的中成药大数据进行清洗与整合，将有价值的药品数据用不同的ECharts图表展示出来，为用户提供可视化服务。同时，对中成药成分和说明书进行挖掘，分析中成药中含有的有毒成分、用法用量、不良反应等数据，构建中成药知识图谱，实现对中成药的风险预警，帮助用户做出更安全的用药决策。

3　系统的实现与应用

3.1　数据增删改查

（1）查询数据：当用户在页面输入相应的检索条件时，系统通过request()方法获取前端表单传来的各字段内容，连接数据库，通过‘SELECT * FROM’查询语句实现对数据的检索，最后使用render_template()方法将数据传到前端页面完成数据的查询。

（2）添加数据：系统将通过request()方法获取用户在弹出的数据添加框中填写各字段信息，再在后端定义常量来接收这些数据，通过比对数据库，查询是否已存在该条数据，如果不存在，就将数据存入数据库并将其更新到系统页面，否则添加失败。系统添加数据功能页面如图6所示。

（3）删除数据：用户进行删除操作时，系统将会弹出一个确认框以供确认，有效地防止用户因操作不当而误删数据的情况。如果用户确认删除该条数据，系统后端将调用相应的remove()方法完成对数据的删除。

（4）修改数据：当用户需要对某条数据进行修改时，可以在弹出的数据修改框中修改数据内容并点击“更新数据”按钮，前端将数据传给后端相应的update()方法进行处理，再将处理结果存储到数据库，同时传回前端对系统页面进行更新，从而完成对数据的修改。

3.2　数据批量导入导出

批量导入数据功能主要使用Python语言的xlrd类库实现。当用户点击“上传文件”按钮把Excel文件信息传递给系统，后端将接收到文件并通过for循环完成对文档的扫描读取，最后使用insert into语句将数据存入对应的数据表中。系统批量导入数据功能页面如图7所示。

图 6 系统添加数据功能页面

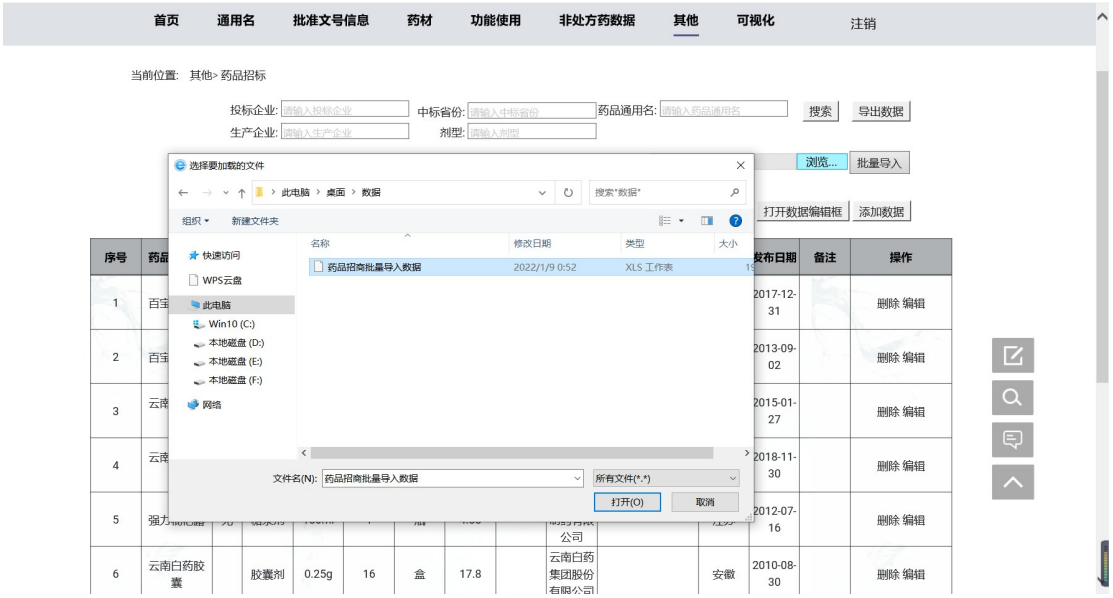


图 7 系统批量导入数据功能页面

批量导出数据功能主要使用Python语言的xlwt类库实现，当用户点击“导出数据”按钮后，响应头会将文件类型传递给浏览器，系统后端使用Workbook()方法创建Excel文件，最后通过for循环和write()方法将数据写入单元格，完成数据的批量导出。

3.3 数据可视化图表应用

ECharts是一款基于JavaScript的数据可视化图表库，提供直观生动、可交互、可个性化定制的数据可视化图表。它能够兼容绝大部分主流浏览器，可以跨平台流畅运行，并且具备千亿级数据可视化渲染能力[9]。同时ECharts还支持各种图表相互组合[10]，有助于解决因可视化方法单一而难以表现数据的内在特性与价值的问题。

本系统通过对数据关系的挖掘，选取了用

户感兴趣的部分，如特殊性药材信息、中成药保护品种信息、中成药广告信息、中成药热词进行可视化展示，页面详情如图8所示。

（1）特殊性药材毒性分析：采用饼状图展示不同毒性的中药材的数量和占比，用户可以点击“查看详情”，进一步查询不同级别含有毒性药材制药量的相关分析。

（2）中成药保护品种剂型：采用柱状图展示中成药保护品种的剂型分布。

（3）全国中成药广告分布：以中国地图为载体，以省份为统计单位，用深浅不同的颜色表示该省份的中成药广告数量，用户可以非常直观地了解中成药广告分布情况。

（4）中成药保护终止日期：将柱状图和折线图结合起来，从多维度展示到达终止日期的中成药保护品种的数量的逐年变化趋势。

（5）中成药热词词云：词语字体越大，索引热度就越高，可以进一步反映当下中成药研究热点。

3.4　中成药知识图谱的构建与应用

3.4.1　构建中成药知识图谱

知识图谱是在互联网大数据背景下产生的一种新颖的知识管理与服务模式，能够捕捉并呈现事物之间的关联关系，可以将琐碎零散的知识片段连接起来[11]。知识图谱能将多背景的领域概念整合起来，更好地反映中成药数据中尚未发现的关联性和规律性，为解决中医药领域的“知识孤岛”问题提供了理想的技术手段[12]。同时，知识图谱能够为用户提供全面、准确的中成药数据支持，帮助用户进行用药决策，提高用药决策的准确性和效率。

构建知识图谱最主要的一个步骤就是把数据从不同的数据源中抽取出来，然后按一定的规则加入到知识图谱中，即知识抽取。本文选取中成药名称作为实体结点，以成分、功能主治、用法用量、不良反应和禁忌为一级属性结点，并以其相应的内容作为二级属性结点，运用ECharts建立中成药知识图谱。以“惊风散”为例，通过知识图谱可以发现惊风散主治腹胀、积食、小儿急惊风等病症，会出现恶心、呕吐、便溏等不良反应。“惊风散”知识图谱如图9所示。同时，我们用不同的颜色将实体结点和一级属性结点予以区分，并且在知识图谱左上角设置了“自定义展示内容”的功能，以解决知识图谱关联过多结点而影响展示效果的问题[13]。

3.4.2　中成药知识图谱可视化应用

本系统以《中华人民共和国药典(2020版)》[14]收录的83种有毒中药(以半数致死量(LD50)分级法[15]为依据)建立“大毒”“有毒”“小毒”列表，将用户输入的中成药的成分与三个列表分别进行遍历对比，得到有毒成分，实现中成药

图8　系统可视化图表页面详情

成分安全性风险分析功能。同时，将中成药说明书中的注意事项从数据库中提取出来，结合前文构建的中成药知识图谱，将数据展示到页面，从而建立中成药风险预警模块。用户可以使用该模块来检索某中成药的有毒成分以及服用该药时的注意事项，还可以通过知识图谱的形式了解其成分、功能主治、用法用量等信息。“惊风散”的风险预警界面如图10所示。

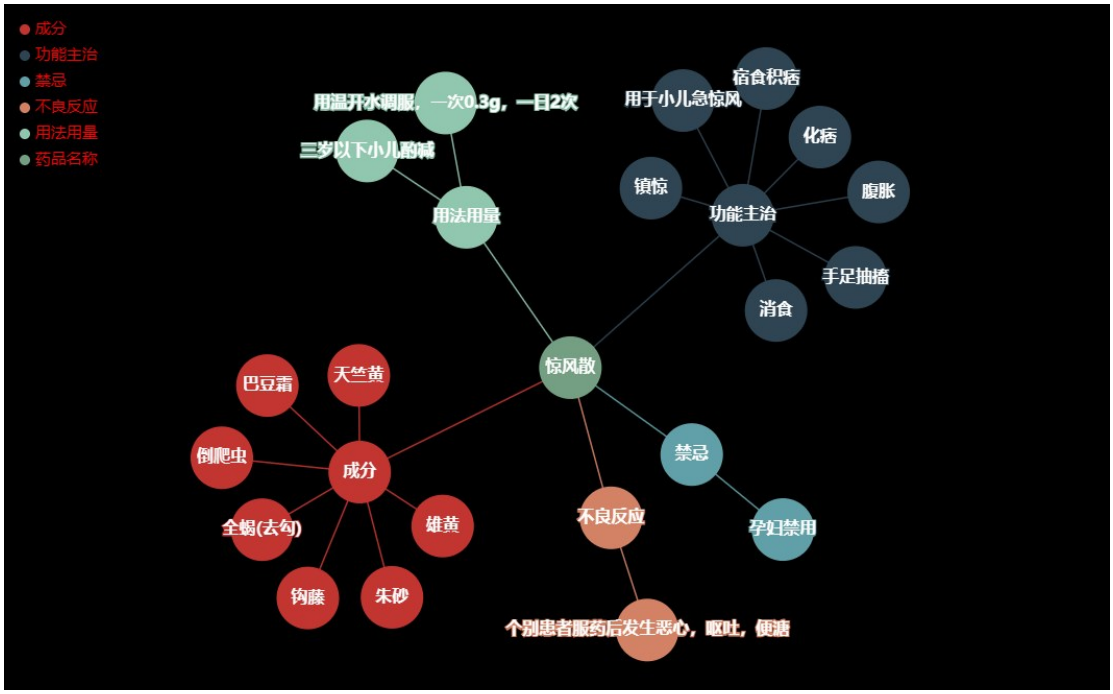


图 9 “惊风散”知识图谱

图 10 “惊风散”的风险预警界面

4　结语

本文讲述了中成药多背景数据协作共享与可视化系统的设计与实现，在调查和分析国内外研究现状的基础之上，运用Python爬虫技术采集中成药大数据，对数据进行实体识别、属性提取、整理分析，并将其存入PostgreSQL数据库，构建包括中成药政策、技术与流通等多背景的数据库。通过使用Python轻量级框架Flask设计实现系统增删改查等相关功能，应用HTML页面框架、CSS页面样式、JavaScript等技术进行前端页面设计。同时，为了深度挖掘探讨中成药数据之间的潜在映射关系，除了使用ECharts可视化图表对数据进行基础展示之外，还构建了风险预警知识图谱，使人们对中成药的成分、用法用量、服用禁忌有更清晰的了解，从而降低中成药不良反应事件的发生率。目前，本文设计实现的系统仍然存在着不足和缺陷，在后续研究中将继续从以下三个方面深入研究：

（1）增强系统数据覆盖面。目前本系统数据库中尚无中成药生产厂家、市场销售等相关数据，后续将通过爬虫技术从互联网中收集或依靠用户的协作共享来建立该部分数据的数据库，进一步提高数据库的完整性。

（2）优化数据库系统结构。随着数据量的增加，后续将严格按照数据规范化理论对数据库进行优化，控制数据冗余，避免异常操作，提高系统的稳定性、数据库的安全性，更好地确保数据库中数据的有效性和一致性。

（3）深入探究数据隐藏关联。目前系统风险预警模块尚且只能进行有毒成分的检测以及简单的知识图谱检索，未能深度分析有毒成分与中成药规格用量和不良反应之间的关系。后期将继续挖掘中成药相关数据，丰富优化风险预警知识图谱，不断提高中成药用药的安全性，促进中成药现代化、信息化发展。

参考文献：

[1] 国务院办公厅关于印发中医药振兴发展重大工程实施方案的通知[A]. 中华人民共和国国务院公报，2023(8)：24-35.

[2] 徐建华.《国家药品不良反应监测年度报告(2021年)》发布[N]. 中国质量报，2022-04-08(002).

[3] 亢力，高宏杰，李敬华，等. 新型科研模式下的中医药信息数据基础建设研究[J]. 中国医学创新，2014，11(6)：79-80.

[4] 黄玲玲. 我国中医药数据库建设存在的问题及对策[J]. 中华医学图书情报杂志，2012，21(2)：18-20.

[5] 吴英彬，郭福亮. 基于复杂数据处理的可视化技术研究[J]. 计算机与数字工程，2015，43(11)：1985-1989.

[6] 王艺，任淑霞. 医疗大数据可视化研究综述[J]. 计算机科学与探索，2017，11(5)：681-699.

[7] 肖永飞. 医学数据三维交互可视化方法的研究[D]. 哈尔滨：哈尔滨工业大学，2010.

[8] 李正，康立源，范骁辉. 中药制药过程数据集成、数据挖掘与可视化技术研究[J]. 中国中药杂志，2014，39(15)：2989-2992.

[9] 周洪斌，陈立平，刘连浩. 基于ECharts的数据可视化应用[J]. 沙洲职业工学院学报，2021，24(1)：3-9.

[10] 许梦雅. 基于Echarts技术的企业数据可视化的设计与开发[J]. 现代信息科技，2022，6(6)：90-92，96.

[11] 于彤，刘静，贾李蓉，等. 大型中医药知识图谱构建研究[J]. 中国数字医学，2015，10(3)：80-82.

[12] 于彤，李敬华，于琦，等. 中医养生知识图谱的构建与应用[J]. 中国数字医学，2017，12(12)：64-66.

[13] 邓宇，周卫强，张振铭，等. 基于名老中医医案的知识图谱构建[J]. 湖南中医杂志，2019，35(7)：186-187.

[14] 国家药典委员会. 中华人民共和国药典[M]. 北京：中国医药科技出版社，2020：1088.

[15] 樊青，关建红. 5种临床常用辛味中药饮片急性毒性初探：探讨毒性中药界定与分级标准的不足[J]. 世界中西医结合杂志，2021，16(8)：1443-1446，1452.

Design and implementation of multi-background data sharing and visualization system for Chinese patent medicine

Yan Yuting*, Li Yuhong, Zheng Shuifei

(College of Computer Science, Jiangxi University of Chinese Medicine, Nanchang 330004, China)

Abstract: In recent years, the use frequency and recognition of proprietary Chinese medicines have gradually increased. Relevant information platforms, medical terminals and the Internet continue to generate massive multi-background data of proprietary Chinese medicines. These data come from a large number of sources, evolve rapidly, and are not easy to track and utilize, which is difficult to support the management effectively, decision-making and services in the fields related to Chinese patent medicine. Based on this, Python crawler technology was used to collect big data of Chinese patent medicine, try to integrate a variety of Chinese patent medicine information resources, and use PostgreSQL data storage technology to build a multi-background database including Chinese patent medicine policy, technology and circulation. Secondly, the Flask framework is used to design and implement system-related functions and interfaces to provide different degrees of data management services for users with different permissions. Finally, a series of drug index data concerned by users were visualized by ECharts technology. The system tries to improve the isolation of Chinese patent medicine multi-background data and provide better collaborative sharing and visualization services for domain experts or ordinary users.

Keywords: Chinese patent medicine; Python crawler; data collaboration and sharing; ECharts visualization; knowledge graph

(上接第71页)

Chinese text classification using convolutional neural networks

Huo Shandong*

(School of Computer Science and Engineering, Chongqing Three Gorges University, Wanzhou 404100, China)

Abstract: Through the convolutional neural network, two different schemes are used to conduct comparative experiments on Chinese text classification. The experimental results show that for the Chinese text classification problem, only the fully connected layer in CNN is used and the appropriate feature words and activation functions are selected. A more ideal classification effect can be achieved.

Keywords: convolutional neural network; Chinese text classification; feature words; activation function

文章编号: 1007-1423（2023）16-0081-07 DOI: 10.3969/j.issn.1007-1423.2023.16.015

拦截与清除恶意捆绑软件系统的研究与开发

刘毓彬，叶传奇*，张少博，朱溪洋，秦梦雯，刘晓雨

（河南科技大学软件学院，洛阳 471003）

摘要：如今的互联网行业欣欣向荣，但也滋生出如垃圾信息轰炸、钓鱼网站盗取信息等问题。为解决最为突出的恶意捆绑软件下载的问题，设计了针对捆绑软件的拦截与清除系统，系统使用特征值匹配、云查杀、文本分析等技术实现了对新软件的下载前过滤、安装中检查、使用时监督的防护功能。本系统大小约为同类样本平均值的48.4%。在预防恶意捆绑软件下载功能，系统的页面检测准确性指标与资源占用指标分别高于同类软件平均得分的6%与13%。

关键词：计算机软件与理论；捆绑软件；特征值匹配；系统安全；文本分析

0 引言

互联网的快速发展给人们带来了巨大的便利，但伴随着互联网行业发展而滋生的恶意下载捆绑软件的问题也已困扰互联网用户十余年[1]。在2022年的315消费者晚会曝光多个P2P类型恶意捆绑软件下载器网站后，该问题成为了社会关注的焦点。

对于含有恶意代码软件的检查，传统的防护方案是使用Sandbox技术，正如张灿岩[2]所提出的：在含有恶意代码软件侵入用户计算机后，使用API Hook技术，防护软件将可能的恶意软件放入沙箱隔离运行，其方法具有普适性，但也存在系统侵入性过大，资源开销较大等缺点。本系统旨在将含有恶意代码的软件主要拦截在用户下载操作之前，并对少量逃避拦截的恶意下载捆绑软件使用特征值匹配、云查杀[3]等技术实现清除，以达到减轻系统负担与减小系统侵入性的目的。用户在应用本系统后有利于减轻信息的认知压力与电脑的使用压力。

1 主要技术

1.1 获取用户最新浏览记录功能

功能输入：当前用户计算机运行中所有浏览器的数据库路径

功能输出：用户浏览网址信息的数据库最新记录

Step1：全局监听，等待用户点击

```
                                    //加载全局监听实例
Logger.getLogger(GlobalScreen.class.getPackage().
  getName()).setLevel(Level.OFF);
  GlobalScreen.registerNativeHook();
  GlobalScreen.addNativeMouseListener(this);
```

Step2：获取浏览器进程

```
Runtime runtime = Runtime.getRuntime();
Process process = runtime.exec("tasklist");
        //使用runtime单例获取当前进程名称，pid，
                                        占用内存等
if(进程含有浏览器进程)执行step3;
else return
```

收稿日期：2023-04-06 **修稿日期：**2023-07-27

基金项目：河南科技大学2022年度大学生研究训练计划（SRTP）项目（2022257）

作者简介：刘毓彬（2001—），男，河南郑州人，在读本科，研究方向为软件工程；*通信作者：叶传奇（1969—），男，湖北武汉人，副教授，博士，主要研究方向为智能信息处理与模式识别、智能图像处理与图像融合研究，E-mail：287142898@qq.com；张少博（2002—），男，河南周口人，在读本科，研究方向为软件工程；朱溪洋（2002—），男，河南驻马店人，在读本科，研究方向为最优化理论、算法设计与分析；秦梦雯（2002—），女，河南安阳人，在读本科，研究方向为软件工程；刘晓雨（2001—），女，河南安阳人，在读本科，研究方向为软件工程

Step3：获取浏览器历史记录数据库文件

```
Runtime runtime = Runtime.getRuntime();
Process process = runtime.exec("whoami");
    //使用runtime单例获取当前用户名称，并将完整
            用户名称截取为当前用户特定用户名；
```

根据文件修改日期新旧确定用户文件存放位置，拼接数据库存放路径；

以火狐浏览器数据库路径为例：

```
String filesrc="C:\\Users\\"+用户名称+"\\
  AppData\\Roaming\\Mozilla\\Firefox\\Profiles";
hisdataURL= filesrc +最新修改日期的文件夹名称 +
  "\\places.sqlite";
```

根据数据库路径复制数据库文件至指定位置

```
File srcFile=new File(src);
      //防止出现浏览器写，本系统读的并发性错误
```

Step4：读取历史记录数据库数据

根据文件后缀名判断数据库类型，并建立相应连接

以主流浏览器使用的sqlite3数据库为例：

```
Class.forName("org.sqlite.JDBC");
Connection conn=DriverManager.getConnection("jdbc:
  sqlite:"+fileName);
```

根据不同浏览器选择不同的查询语句，获取到历史记录。

Step5：对数据库中历史记录判断新旧

```
If (! lastNewPagebyBrowser.containsKey(i)){
      //如果最后浏览记录为空，说明功能启动后，
                          用户首次点击
lastNewPagebyBrowser.put(i,hisdataarr.get(0));
                  //则记录当前最新一条浏览记录
                          删除复制的历史文件
}else{
                  //最后记录不为空，获取新页面数据
 int mapindex=lastNewPagebyBrowser.get(i).getId();
     //获取记录到的上一次对应浏览器名称标志的
                          历史记录最新标号
 int Previndex=hisdataarr.get(0).getId();
            //取出记录数组中最后一条记录的id值
                          删除复制的历史文件
 if(mapindex==Previndex){
   //上一次记录的最后标号等于最新记录的标号，
                  说明点击后没有新页面
   return null;
 }else {
   int section=Previndex-mapindex;
   for(int j=0; j<section;j++) {
                          新的历史记录加入数组
 }
     lastNewPagebyBrowser.replace(i, hisdataarr.get
       (0));
                          //更新最后浏览记录
     return新页面列表
   }
}
```

Step6：功能一次执行结束，等待用户再次点击激活

1.2 Web页面文本内容关键词分析

Step1：通过URL获取相应页面

例：Jsoup.connect(*HTMLUrl*).get();

Step2：使用正则表达式进行数据清洗

去除script的正则表达式匹配语句：

```
regEx_script = "<[\\s]*? script[^>]*? >[\\s\\S]*?
  <[\\s]*? \\/[\\s]*? script[\\s]*? >";
```

去除style样式的正则表达式匹配语句：

```
regEx_style = "<[\\s]*? style[^>]*? >[\\s\\S]*?
  <[\\s]*? \\/[\\s]*? style[\\s]*? >";
```

去除HTML标签的正则表达式匹配语句：

```
regEx_html = "<[^>]+>";
```

去除特殊字符的正则表达式匹配语句：

```
regEx_special = "\\&[a-zA-Z]{1,10};";
```

去除转义的正则表达式匹配语句：

```
regEx_transfer = "[\\u0000-\\u001f\b]";
```

将多个空格替换成1个空格：

```
Pattern.compile("\\s+").matcher(Str).replaceAll(" ");
```

得到以空格为分隔符的页面文本数据

Step3：使用基于mmseg算法的Jcseg API进行中文分词与词义分类[4]

```
SegmenterConfig config = new SegmenterConfig(true);
      //创建默认单例词库实现，并且按照config配置
                          加载词库
ADictionary dic = DictionaryFactory.createSingleton
  Dictionary(config);
```

```
        //依据给定的ADictionary和SegmenterConfig
                                来创建ISegment
ISegment seg = ISegment.COMPLEX.factory.create
  (config, dic);
以空格为分割符页面文本数据存入数组;
For(i=0; i<数组长度; i++){
将数组[i]字符串分词;
将数组[i]字符串词义分类;
seg.reset(new StringReader());
while ( (word = seg.next()) ! = null ) {
   将分词,分词义后的字符串的名词成分暂存;
   if(! wordMap.containsKey(word.getValue())) {
     wordMap.put(wordkey,1);
      }else {
        int wordvalue=wordMap.get(wordkey)+1;
        wordMap.put(wordkey,wordvalue);
      }
   }
}
```

Step4：将HashMap中的词语按词频排序

根据HashMap中值的大小，对HashMap使用快速排序得到词频由高到低的数组；

Step5：关键词得分

选取数组中出现次数大于(最高次数+最低次数)/2的词组作为关键词进行计分，公式如下：

$$\text{关键词得分} = \left[1 + \frac{\text{词频}}{\text{词组词频总数}} n + \left(\frac{\text{关键词词频排名次第}}{\text{关键词总个数}}\right)\right] \times \text{数据库反馈因子}$$

注：①“词组词频总数”：HashMap中所有词组出现次数之和；n为常数，是平衡因子，用于防止因词组词频总数(即页面文字内容总量)过多或过少导致的失衡；②“数据库反馈因子”：查询数据库，返回数据库中词组标记的基础得分，得分区间为[-1，1]。

Step6：页面内容总得分

$$\text{页面内容总得分} = \frac{\left(\sum_{1}^{n} \text{关键词得分}\right)}{n}$$

1.3　特征值匹配

哈希值特征码的构造主要由节的大小加上节的哈希值组成，程序执行时会主动扫描文件的节头，判断是不是PE文件，如果是，则打开文件，以文件头的字节流构造特征码，并与病毒库中的数据进行匹配，如果匹配成功则返回病毒名称，失败则返回空。同时，本功能的病毒库以特征码的头一个字符为标准划分为0～9、A～G十七个子病毒库，在病毒扫描时，以特征码头一个字符为准跳到对应的子病毒库中进行数据匹配，减少了扫描数据的数量，从而节约了病毒扫描所需的时间。

MD5信息摘要算法(MD5 Message-Digest Algorithm)，一种被广泛使用的密码散列函数，可以产生出一个128位(16字节)的散列值(hash value)，用于确保信息传输完整一致。理论上，任意两个文件、字符串不会有相同的MD5值[5]。因此我们建立了以MD5为基础的病毒特征码数据库，再对用户电脑全盘扫描使获取用户系统文件的MD5特征码与病毒特征码库比对，进行病毒文件的查杀。工作流程如图1所示。

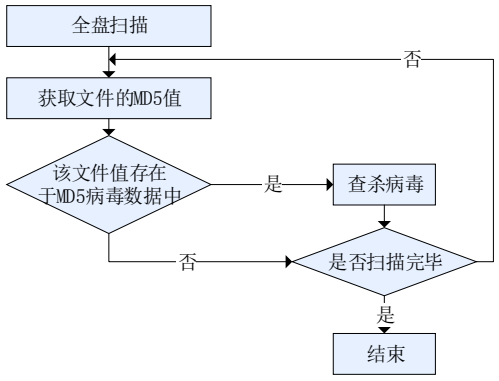


图1　MD5特征值匹配流程

1.4　云查杀技术

云查杀技术依赖于云计算技术，是将安全引擎和病毒木马特征库放在服务器端，将主要的计算、查询等资源开销较大的操作交由服务器处理，从而减轻用户个人计算机的资源占用负担，并获取更快的查杀速度、更短的反映时间以及更优秀的查杀效果[6]。

2　功能模块实现

2.1　页面拦截功能模块实现

功能需求：检查用户浏览的网站是否安全，阻止用户浏览包含恶意捆绑软件下载连接的网站。

功能核心：获取当前用户浏览的网页信息；数据库信息筛选；网页文本内容分析。

功能实现：首先，全局监听用户点击，用户每次点击后获取当前运行中的全部进程，筛选出所有浏览器进程，根据运行中的浏览器名称获取到其本地数据库路径。将数据库文件拷贝至本系统指定目录中，以防止浏览器的写与本系统的读并行执行造成的死锁问题。

然后，浏览器数据库记录新历史数据时具有‘只会将首次访问的页面记录到数据库中，若非首次访问，数据库则只更新相关记录的访问次数与最后访问时间’的特性。本功能利用该特性获取到当前用户首次浏览的页面。用户首次点击后，程序记录到用户上次使用时最后一条历史记录的ID值并将其作为标号，当用户第二次点击后，获取到的当前最新历史记录的ID值与标号之间的差额记录数列即为用户两次点击之间首次访问的页面，更新标号后循环往复，从而获取到用户的浏览信息。

再次，获取到新历史记录后，程序将对新记录逐条进行安全性初筛。程序根据用户选择的安全防护等级执行不同的初筛策略：用户选择低等级防护时，程序只进行黑名单域名匹配；用户选择中等级(默认等级)防护时，首先进行黑名单全域名匹配，若页面不为黑名单中页面，则采用了Jaro-Winkler Distance与LevenshteinDistance两种编辑距离[7]进行黑名单域名相似度匹配，同时进行页面名称特征词匹配，将各项匹配结果分别加权参与安全性评判，当值达到阈值时，认为该记录可能含有危险性需要进一步分析；用户选择高等级防护时，首先进行黑名单全域名匹配，若页面不为黑名单中页面，首先进行白名单特征匹配，特征值匹配结果为1(匹配结果范围为0～1)，即白名单有该条记录时返回页面安全，否则本历史记录与有相同域名的白名单列表逐条相似度匹配获取到该页面得到的最高分，并减去黑名单标题特征匹配的加权值后获取到最终值，若该值达到安全线值，则判断该页面安全，并添加该记录至白名单中，否则判断该页面需要下一步检测。页面拦截模块处理流程如图2所示。

最后，对于标记为需进一步检测的记录将进行更加细致的检测。为每一条记录开启一个新的线程，使用Jsoup通过URL获取HTML全部内容，若Jsoup获取网页内容失败，则使用本功能自编写的程序获取网页内容。使用正则表达式将页面中文本提取。使用Web页面文本内容关键词分析技术得到页面内容得分并加权；使用Jsoup获取所有页面中的超链接标签并逐个判别其链接是否安全，得到页面链接安全性得分

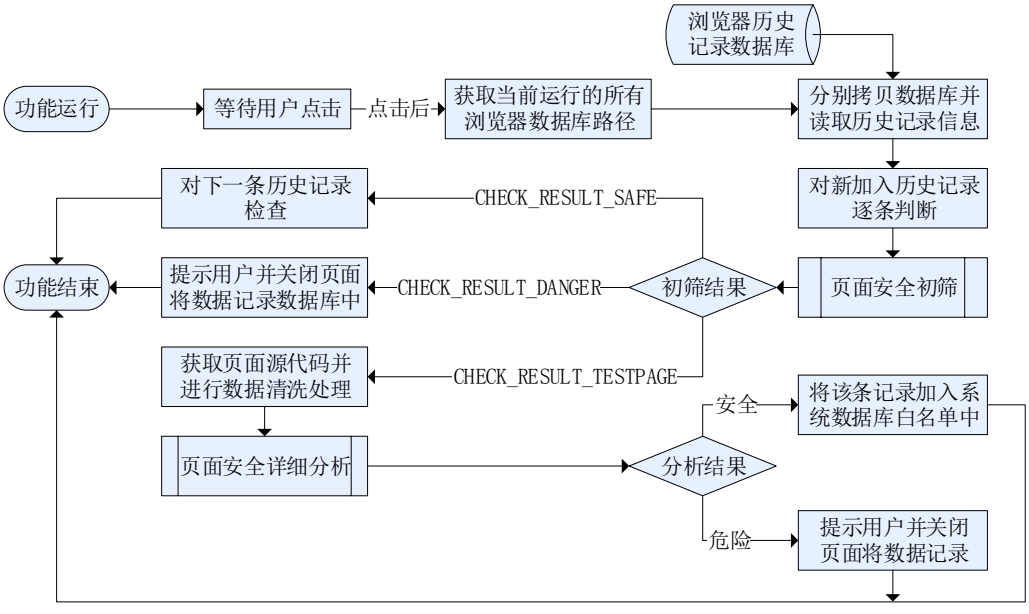


图2 页面拦截功能模块处理流程

并加权；获取网页ICP备案号，通过工业和信息化部政务服务平台的ICP/IP地址/域名信息备案管理系统(https://beian.miit.gov.cn)获取到页面备案信息，得到页面合法性得分并加权。最终通过综合得分判别该页面是否安全。若该页面安全则将其添加到系统白名单中，不安全则通知用户并关闭页面。

2.2　下载监控功能模块实现

功能需求：监控用户下载的文件并进行病毒查杀；自定义病毒扫描。

功能核心：监控文件的下载；常见格式压缩包的解压缩；病毒扫描。

功能实现：首先，文件系统中的文件夹均会被WatchService(基于本机操作系统实现监控)监控，在监控过程中，程序获取到新建文件的绝对路径，并根据新建文件的绝对路径进行病毒扫描。

需要说明的是，计算机系统生成新文件的途径除了用户下载，还可能为用户复制本地文件产生的备份文件，程序运行产生的缓存文件，等等，如果对所有新生成的文件都进行扫描，无疑加大了程序的负担。而本功能根据“在文件下载过程中系统会产生一个临时文件用于存放下载文件片段而在文件下载合并完成后临时文件被删除”的特性增加判断，从而获取到用户下载的文件。本功能仅对用户下载的文件进行病毒检测，提高了系统运行效率。下载监控功能模块处理流程如图3所示。

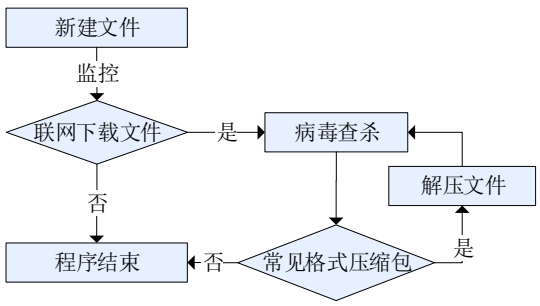


图3　下载监控功能模块处理流程

其次，在各种文件格式中，因为压缩包内部可以包含多个文件且杀毒程序无法直接检查未解压压缩包中的文件，所以压缩包格式的文件可能导致病毒文件特征值匹配的失败。针对压缩包格式的文件扫描问题，本功能先将压缩文件解压到指定的临时文件夹下，再对该文件夹下的所有文件进行病毒扫描，如果出现病毒就给予用户相应的提示信息。无论是否有病毒，系统在扫描完成之后都即刻清空该临时文件夹。然而，压缩文件的格式有多种，不同的压缩格式使用不同的编码方式，考虑到系统开发时效性与程序适用性，本功能只针对RAR、ZIP、7Z三种最常见压缩格式的未加密压缩文件进行解压与病毒扫描。

2.3　本地文件病毒检测功能模块实现

功能需求：检测用户计算机磁盘文件。用户上传文件进行检测并出具检测结果。

功能核心：扫描磁盘文件；利用特征值匹配检测文件；使用云查杀技术对用户提交的文件进行检查并返馈分析报告。

功能实现：功能运行后，用户电脑磁盘中的文件均会被FileScanService(基于本机系统实现扫描模块)扫描并计算出文件的MD5值，每个文件都与本系统的病毒库进行MD5特征值匹配[5]。用户还可以自定义区域进行病毒扫描。如果文件的MD5值匹配中病毒样本记录，系统将会提醒用户并建议用户对此文件进行删除或者对该软件进行卸载。但此种方法不能有效检测出混合病毒，因为病毒文件混入了其他文件后，其对应的MD5特征值也将改变；其次，此种方法对变形病毒的检测能力也较弱。

本系统得到了合作公司(北京云海协同科技有限公司)旗下VirScan网站的使用授权与技术支持。对于上述问题，系统将没有检测出病毒的文件上传至VirScan服务器进行云查杀并给予检测反馈[6]。

用户还可以单独调用云查杀功能。当用户将文件上传给系统后(文件大小不得超过50 Mb)，系统调用相关API将用户上传的文件提交给VirScan，VirScan将该文件置于一个虚拟环境，使用47款扫描引擎进行病毒扫描并逐个反馈结果。系统对反馈结果进行分析，得到最终的文件检测报告。下面以“virus.exe”为例，云查杀部分结果见表1所示。

表 1 云查杀部分结果展示

引擎	结果(显示为病毒类型或无检出)
AVG	Win32：Trojan-gen
Antiy	Trojan/Win32.AGeneric
JiangMin	Trojan.Generic.dngqp
JiangMin	Trojan(0054b5be1)
Qihu360	Win32/Ransom.Filecoder.HgIASOgA
QQ手机管家	无检出
McAfee	无检出

3 程序运行性能比较

本系统相比于主流的计算机防护杀毒软件具有体积更小的特点，在十余种主流统计样本中，本系统大小约为样本平均大小的48.4%。系统与部分主流计算机防护杀毒软件大小对比如图4所示。

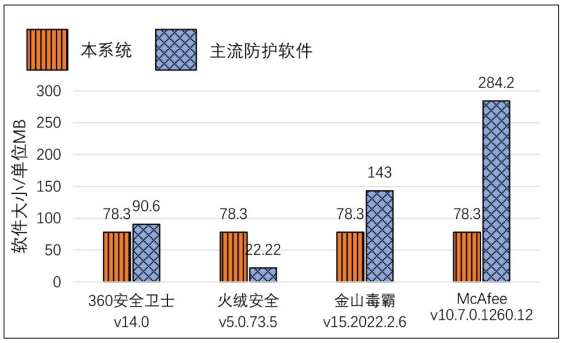


图 4 系统与部分主流计算机防护杀毒软件大小对比

另外在预防恶意捆绑软件下载至用户电脑功能上，相比于主流的计算机防护杀毒软件的平均防护水平，本系统的页面检测准确性指标与系统资源占用指标分别提升6%与13%。各项实现结果得分如图5所示。

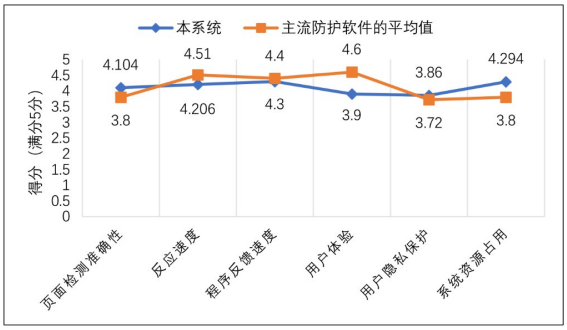


图 5 “预防恶意捆绑软件下载”功能中本系统得分与主流防护软件各项平均得分对比

系统在检测和清除恶意软件和病毒的软件程序仍有不足，如病毒库样本量过小，对于病毒变种的应对方法单一，扫描时间过长等问题。可借鉴张灿岩[2]、何国贤[3]等人使用的沙盒检测技术来改进上述问题。

4 结语

本文使用SSM开发框架[8]依据成熟的MVC设计模式[9]，实现了对恶意捆绑软件进行有效的拦截和清除的系统。本系统旨在将含有恶意代码的软件主要拦截在用户下载操作之前，并对少量逃避拦截的恶意下载捆绑软件使用特征值匹配、云查杀等技术实现清除，具有较小的系统开销，较快的运行速度等特点。用户应用本系统后可以保护其不受恶意软件的侵害，减轻了用户使用电脑时的认知负担，操作压力等。

参考文献：

[1] 刘庆平. 浏览器安全问题的研究与解决方案[D]. 上海:上海交通大学,2011.
[2] 张灿岩. 用于恶意代码检测的沙箱技术研究[D]. 哈尔滨:哈尔滨工程大学,2013.
[3] 何国贤. 基于沙盒技术的多平台恶意程序分析工具研究与实现[D]. 成都:电子科技大学,2013.
[4] 张中耀,葛万成,汪亮友,等. 基于MMSEG算法的中文分词技术的研究与设计[J]. 信息技术,2016(6):17-20.
[5] 魏晓玲. MD5加密算法的研究及应用[J]. 信息技术,2010,34(7):145-147,151.
[6] 许蓉,吴灏,张航. “云安全”检测技术安全性分析[J]. 计算机工程与设计,2012,33(9):3309-3312.
[7] CAHYONO S C. Comparison of document similarity measurements in scientific writing using Jaro-Winkler Distance method and Paragraph Vector method[J]. IOP Conference Series:Materials Science and Engineering,2019,662(5):052016.
[8] 李洋. SSM框架在Web应用开发中的设计与实现[J]. 计算机技术与发展,2016,26(12):190-194.
[9] 田娟,徐钊. 基于J2EE的MVC设计模式的分析与思考[J]. 计算机与现代化,2010(10):54-58.

(下转第98页)

文章编号: 1007-1423（2023）16-0087-06 DOI: 10.3969/j.issn.1007-1423.2023.16.016

基于Node.js的开源架构Electron赋能前端开发

邓杰海*，刘 薇，汤小燕

（江西中医药高等专科学校教育技术中心，抚州 344000）

摘要：长期从事Web前端开发的软件从业人员，要从事后端应用开发，需要学习面向对象编程和泛型编程，面对这些与前端JavaScript语言完全不同特性的语言，这对长期从事前端开发的人员来讲，学习技术曲线还是比较陡峭。为了前端开发人员付出较小的学习成本能快速地从事后端软件的开发，基于Node.js的开源架构Electron是一个不错的选择。通过编写实现一些功能的程序描述了使用JavaScript、Html、CSS等基于Electron架构进行后端开发的过程。事实证明使用Electron开源架构进行后端开发对于前端开发人员是非常友好的，并且也是可行的。

关键词：Node.js架构；Electron架构；JavaScript语言；跨平台；IPC通信

0 引言

软件开发可分为基于浏览器的前端Web开发和可以直接访问本地软硬件资源的后端开发。前端开发的软件架构比较丰富，有如jQuery、vue、react和angular等应用非常广泛的架构，通过HTML、CSS、JavaScript完成前端软件开发。基于浏览器的前端Web开发受浏览器沙箱操作的限制，前端程序访问本地资源受到很大的约束，而在网上流动的满足各行业的业务规则的数据，最终需要通过后台服务端程序将数据持久化到本地硬盘上，可以说前端应用与后端应用相辅相成、缺一不可。

常用的后端跨平台桌面应用开发的框架有Mono、QT等，要求具有C、C++、C#编程经验。长期从事Web前端开发的软件从业人员，要从事后端桌面应用开发，面对这些与前端JavaScript语言完全不同特性的语言，需要学习面向对象编程和泛型编程，这对前端开发人员来说，有一定难度。然而，在基于JavaScript引擎V8的Node.js横空出世之后，使得精通HTML、CSS和JavaScript语言的前端工程师，可以使用整合了Node.js和WebKit技术的Electron架构，在后端桌面服务端程序开发中大放异彩。

1 JavaScript语言运行环境Node.js

Node.js是Javascript的一个服务器端运行环境，它使得Javascript可以脱离浏览器的束缚而运行在服务器环境下，可以像PHP、Python等语言一样进行服务器端程序的开发[1-2]，它是一个为实时 Web（Real-time Web）应用开发而诞生的平台，它从诞生之初就充分考虑了在实时响应、超大规模数据要求下的架构的可扩展性[3]，这使得它摒弃了传统平台依靠多线程来实现高并发的设计思路，而采用了单线程、异步I/O、事件驱动式的程序设计模型[4]。

Node.js与Chrome浏览器相比，除了HTML、WebKit和显卡这些UI相关技术没有支持外，其他功能与Chrome浏览器十分相似[5]，它们的组成结构如图1所示。

收稿日期：2023-04-29 **修稿日期：**2023-05-12

基金项目：江西省教育厅科学技术研究项目重点项目（GJJ213301）

作者简介：*通信作者：邓杰海（1975—），男，江西抚州人，硕士，副教授，高级工程师，研究方向为全栈式应用开发，E-mail：358373740@qq.com；刘薇（1988—），女，江西鹰潭人，硕士，讲师，研究方向算法设计与分析；汤小燕（1984—），女，江西抚州人，硕士，讲师，研究方向信息化教育技术

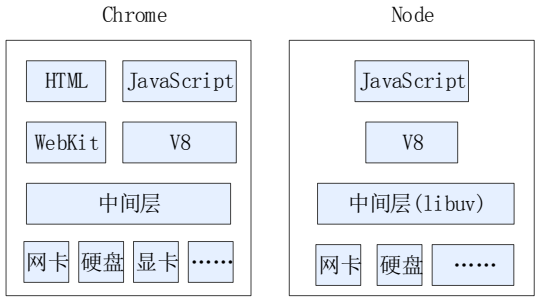


图 1 Chrome与Node.js结构图

从图1可看出，Node.js开发人员抽象出了Libuv层，使得Node.js的异步非阻塞模型可以兼容Windows和基于POSIX的操作系统等平台。在Windows操作系统平台下，采用IOCP机制实现，基于POSIX的操作系统平台下，则对Libev和Libeio库进行抽象封装来实现事件驱动和异步I/O[6]。由于Node.js平台抽象了中间层(libuv)，使得Node.js实现跨平台成为可能。

在windows10操作系统64位CPU架构安装Node.js v13.14.0软件过程如下：

(1)创建d:\nodejs目录；

(2)在d:\nodejs目录下创建node_global和node_cache子目录；

(3)打开https://nodejs.org/dist/v14.16.1/网页，选择node-v14.16.1-x64.msi下载并安装至d:\nodejs目录下；

(4)设置npm config set prefix “d:\nodejs\node_global”和npm config set cache “d: \nodejsnode_cache”；

(5)查看设置npm config get cache、npm config get prefix；

(6)在系统环境变量添加NODE_PATH，输入路径为：D:\nodejs\node_global(重启)；

(7)设置npm镜像源npm install-g cnpm--registry=https://registry.npm.taobao.org、npm config set registry https://registry.npm.taobao.org-global、npm config set disturl https://npm.taobao.org/dist-global；

(8)添加系统变量path的内容(重启)，因为cnpm会被安装到D:\nodejs\node_global下，而系统变量path并未包含该路径。

(9)在cmd命令状态输入node-v、npm-v、cnpm-v，如果都能正常显示版本，则表示安装成功。

2 跨平台桌面应用开发架构Electron

Electron是一个使用JavaScript、HTML和CSS构建桌面应用程序的框架。嵌入Chromium和Node.js到二进制的Electron允许操作者保持一个JavaScript代码并创建在Windows、macOS和Linux等平台上运行的跨平台应用——不需要本地开发经验[7]。

在Electron开发架构，有主进程和渲染进程的概念。主进程执行的文件就是模块描述文件package.json中main数据项指定的文件，也是开发的桌面应用程序的入口文件，它负责建渲染进程、管理应用程序的生命周期、处理本地文件系统和网络请求等任务。在Electron应用程序中，主进程由一个Node.js模块组成，可以通过它来访问底层操作系统的API和功能。主进程还可以通过进程间通信(IPC)机制与渲染进程通信，实现双方之间的数据传输和消息交互。每一个Web网页对应一个渲染进程，各个渲染进程之间是相互独立的，都是由主进程创建。主程序与渲染进程之间、渲染进程之间可以直接或间接通过消息发送或共享存储区等方法进行数据交互。

进行Electron跨平台桌面应用程序开发，可以使用Visual Studio Code、Sublime Text或Atom等编辑器，它们都是最常用和流行的跨平台代码编辑器，都支持Electron开发。选择好代码编辑器，就开始安装Electron跨平台桌面应用程序开发架构，可以使用Node.js带的包管理器npm进行安装，如果安装下载速度较慢，可以设置淘宝镜像使用cnpm进行安装。如果要在电脑上安装多个版本的Electron，则不需要全局安装，本文只介绍在Windows平台下使用npm进行全局安装Electron。在Windows平台下使用cmd命令打开一个命令输入窗口，在窗口中输入npm install-g electron命令进行全局安装，安装完成之后输入electron-v命令，如果能正常显示版本，则表示安装成功。

3 在Electron中实现闪烁托盘程序

以创建一个托盘程序为例，来说明使用Electron进行软件开发的过程，程序实现如下功能：

(1)程序初始界面，在渲染进程网页中显示

默认的托盘程序菜单项的说明信息、操作系统信息、Node.js和Electron版本信息；

（2）托盘图标每隔500毫秒显示两个图标中的一个，以达到动态效果；

（3）点击关闭按钮，程序界面隐藏，在任务栏的托盘位置显示闪烁托盘图标；

（4）对闪烁托盘图标点击鼠标右键，程序主界面显示相应的菜单项的说明信息、操作系统信息、Node.js和Electron版本信息；

（5）在Windows的任务管理器中不显示程序进程，以防备通过Ctrl+Alt+Del快捷键强制退出程序；

（6）定义全局快捷键Ctrl+Alt+k，弹出是否退出程序对话框，如果选择确定则程序退出，如果不知道这个快捷键则不能退出程序，除非系统重启。

为了实现上述功能，在项目一共创建了七个文件，文件清单见附件，项目文件结构如图2红色框所示，程序运行的初始界面如图3所示。

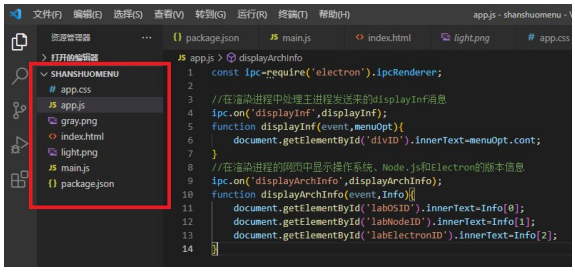


图2　文件结构

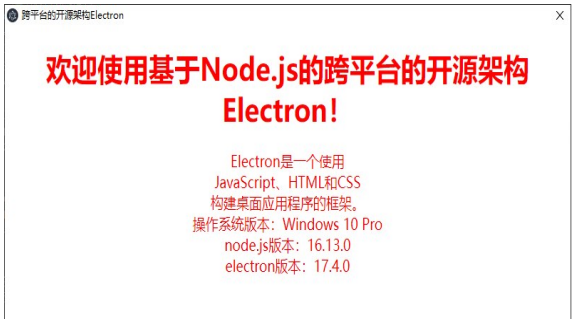


图3　初始界面

4　结语

Electron开发跨平台桌面应用程序的优点包括以下几点：

（1）跨平台：能够将一份代码打包成不同平台的应用，减少开发和维护成本；

（2）快速开发：使用Node.js和JavaScript等现有的开发技术和库，快速开发出高质量的桌面应用程序；

（3）易于维护：应用程序的维护和更新非常简单，能够快速修复bug、更新版本；

（4）大量插件：Electron拥有一个丰富的插件库，能够方便地集成第三方库和组件。

Electron开发跨平台桌面应用程序包括以下缺点：

（1）资源占用：由于Electron应用程序需要同时运行Chromium浏览器和Node.js运行环境，因此它的资源占用会比较高，可能会导致应用程序在老旧计算机上运行缓慢；

（2）安全性：由于Electron应用程序缺乏完善的安全措施，因此容易受到黑客攻击和恶意代码的侵入；

（3）大小：Electron应用程序的文件大小比较大，如果需要向用户提供快速的应用程序安装体验，需要考虑这一点；

（4）学习成本：如果不熟悉Electron，学习和使用它可能需要一些时间和精力。

总体来说，Electron开发跨平台桌面应用程序是一种很好的方式，但是在选择它之前需要综合考虑以上这些优缺点。

参考文献：

[1] 高原．服务器端Javascript技术研究[J]．信息与电脑，2012(01)：78-80

[2] TILKOV S，VINOSKI S. Node.js：using javascript to build high-performance network programs[J]. IEEE Internet Computing，2010，14(6)：80-83

[3] 邹向阳，邹和辉，刘戎．基于物联网与三维可视化的弹药库实时监测系统[J]．火力与指挥控制，2015(1)：30-33.

[4] BYVOID. Node.js开发指南[M]．北京：人民邮电出版社，2012：4-7.

[5] 朴灵．深入浅出node.js[M]．北京：人民邮电出版社，2017：2-3.

[6] 陈瑶．基于Node.js高并发web系统的研究与应用[D]．成都：电子科技大学，2014：15.

[7] OpenJS基金会和Electron贡献者们．Electron开发文档[EB/OL]，2021. https://www.electronjs.org/zh/docs/latest/.

附件：

```
                                    //package.json文件清单
{
 “name”：“shanshuomenu”,
 “version”：“1.0.0”,
 “description”：“创建一个闪烁的托盘程序”,
 “main”：“main.js”,
 “scripts”：{
  “test”：“echo \”Error：no test specified\“ && exit 1”
 },
 “author”：“Djh”,
 “license”：“MIT”
}
                                    //main.js文件清单
‘use strict’;
const electron=require(‘electron’);
const app=electron.app;
const BrowserWindow=electron.BrowserWindow;
const path=require(‘path’);
const { dialog } = require(‘electron’);
const os = require(‘os’);
                                //用于创建和管理菜单模块
const Menu=electron.Menu;
                            //用于创建系统托盘图标的模块
const Tray=electron.Tray;
                        //此处s字母小写,造成模块引用错误
const globalShortcut=electron.globalShortcut;
let mainWindow=null;
let appIcon=null;
                                //托盘显示菜单命令选项
const menuOpts=[
  {
    lab:‘Electron’,
    cont:‘Electron是一个使用\nJavaScript、HTML和
      CSS\n’+‘构建桌面应用程序的框架。’
  },
  {
    lab:‘Node.js’,
    cont:‘Node.js是Javascript的一个服务器端运行
      环境,\n’ +‘它使得Javascript可以脱离浏览器的
    束缚而运行在服务器环境’
  },
  {
    lab:‘论文名称’,
    cont:‘跨平台的开源架构Electron\n赋能国产软
      硬件环境桌面应用程序开发!’
  }
];
function displayOpt(menuOpt){
  //向渲染进程发送IPC消息,将需要在界面显示的
                                    内容发送至渲染进程
  mainWindow.webContents.send(‘displayInf’,menu-
Opt);
      //当主界面程序隐藏起来时,调用下面语句将
                                    主界面显示出来
  mainWindow.show();
}
                                //映射托盘菜单项与对应操作
function adddInfToMenu(menuOpt){
  return{
    label:menuOpt.lab,
    type:‘normal’,
                            //对菜单项左键单击的响应事件
    click:()=>{displayOpt(menuOpt);}
  };
}
app.on(‘window-all-closed’,()=>{
      // 判断表示当前运行的操作系统不是MacOS,
       // 在MacOS中,process.platform的值是darwin
  if(process.flatform！ ==‘darwin’){
    app.quit();
  }
});
app.on(‘ready’,()=>{
                                //用于创建系统托盘图标
  appIcon = new Tray(path.join(__dirname,‘light.
    png’));
                                //用于创建菜单的方法
  let cMenu=Menu.buildFromTemplate(menuOpts.map
(adddInfToMenu));
          //鼠标停在托盘程序图标显示的提示信息
  appIcon.setToolTip(‘托盘程序’);
   //设置对托盘图标点击鼠标右键菜单的内容和行为
  appIcon.setContextMenu(cMenu);
  mainWindow = new BrowserWindow({
                                //隐藏窗口的初始显示
    //这通常在应用程序启动时需要执行某些初始化
        操作,例如加载资源文件、进行身份验证等
  //在这些操作完成之后,窗口可以通过调用win.show()
                    或 win.maximize() 等方法来显示窗口
    show: false, //隐藏菜单栏,需要时按alt又显示
    autoHideMenuBar:true, //禁止改变主窗口尺寸
    resizable: false, //取消window自带的关闭最小化等
                                // frame: false,
```

```
                                    // 设置不显示最小化按钮
    minimizable: false,
                    //设置为true,不在任务管理器中显示
    skipTaskbar: true,
    webPreferences: {
      //启用nodeIntegration选项可以在渲染进程中
                                          使用Node.js模块
      nodeIntegration: true,
        //为 false 时,渲染进程中的 JavaScript 代码
                    可以直接访问 Node.js 的原生 API
      contextIsolation:false
    }
});
mainWindow.show();
              //加载渲染进程中要显示的HTML页面
mainWindow.loadURL(`file://${app.getAppPath()}/
  index.html`)
    //渲染器进程已经完成渲染并可以执行与DOM
                                          相关的操作
mainWindow.webContents.on('dom-ready',()=>{
          //向渲染进程发送默认显示的信息的消息
    displayOpt(menuOpts[0]);
       //将操作系统、Node.js和Electron架构的版本
                信息在渲染进程的网页中显示出来
    mainWindow.webContents.send('display
      ArchInfo', [os.version(), process.versions.node,
      process.versions.electron]);
});
  //注册全局CTRL+ALT+K组合键,确认是否退出程序
globalShortcut.register('CommandOrControl+Alt+
  K', () => {
    dialog.showMessageBox({
      type: 'question',
      buttons: ['取消', '确定'],
      defaultId: 1,
      title: '退出程序确认',
      message: '确定退出程序吗?',
      detail: ''
    }).then((result) => {
      if (result.response === 1) {
          clearInterval(timer);
          mainWindow=null;
          cQuit=true;
          app.quit();
      }
    }).catch((err) => {
      console.log(err);
    });
  })
  //当点击关闭按钮时,此值为false;当退出程序时,
          将此值设置为true,使得程序能正常退出
  let cQuit = false;
  mainWindow.on('close', (event) => {
 //当点击关闭按钮时,此值为假,把主界面隐藏只在
     托盘程序中显示程序图标,当按退出程序全局
  // 快捷键,此值为true,则不会拦截close消息事件,
                                    使得程序正常退出
  if (! cQuit) {
    mainWindow.hide();
      //阻止事件的默认行为,拦截close消息事件,
                                    使得程序不会退出
    event.preventDefault();
  }else{
    console.log('test2');
  }
  });
});
var count = 0,timer=null;
      //每隔500毫秒改变托盘图标以达到动态效果
timer=setInterval(()=>{
  if (count == 0) {
    appIcon.setImage(path.join(__dirname, 'light.
      png'));
    count=1;
  } else {
    appIcon.setImage(path.join(__dirname, 'gray.
      png'));
    count=0;
  }
 }
, 500);
          //会在所有窗口都已关闭并且没有其他的
                                  程序正在运行时被触发
app.on('will-quit',()=>{
                                          //注销全局快捷键
  globalShortcut.unregister('CommandOrControl+Alt+
    K');
});
                                  //index.html文件清单
<! DOCTYPE html>
<html>
  <head>
    <title>跨平台的开源架构Electron</title>
    <link rel="stylesheet" href="app.css" />
```

```
    <script>if (typeof module === 'object') {window.
      module = module; module = undefined;}</script>
    <script>if (window.module) module = window.
      module;</script>
  </head>
  <body>
    <div>
      <h1>欢迎使用基于Node.js的跨平台的开源
        架构Electron! </h1>
    </div>
    <div id="divID">
    </div>
    <div>
      <div>操作系统版本:<label id="labOSID">
          </label></div>
      <div>node.js版本:<label id="labNodeID">
          </label></div>
      <div>electron版本:<label id="labElectronID">
          </label></div>
    </div>
    </div>
  </body>
  <script>
    require('./app');
  </script>
</html>
```

```
                                    //app.js文件清单
const ipc=require('electron').ipcRenderer;
 //在渲染进程中处理主进程发送来的displayInf消息
ipc.on('displayInf',displayInf);
function displayInf(event,menuOpt){
  document.getElementById('divID').innerText=
    menuOpt.cont;
}
                //在渲染进程的网页中显示操作系统、
                        Node.js和Electron的版本信息
ipc.on('displayArchInfo',displayArchInfo);
function displayArchInfo(event,Info){
  document.getElementById('labOSID').innerText=
    Info[0];
  document.getElementById('labNodeID').innerText=
    Info[1];
  document.getElementById('labElectronID').
    innerText=Info[2];
}
                                    //app.css文件清单
div {
  font-size: large;
  color: red;
  text-align: center;
}
```

Open source architecture based on Node.js Electric enabling front-end development

Deng Jiehai*, Liu Wei, Tang Xiaoyan

(Education Technology Center of Jiangxi College of Traditional Chinese Medicine, Fuzhou 344000, China)

Abstract: Software practitioners who have been engaged in front-end development of the Web for a long time need to learn object-oriented programming and generic programming in order to engage in back-end application development. Facing these languages with completely different characteristics from the front-end JavaScript language, for those who have been engaged in front-end development for a long time, the learning curve is still steep. In order for front-end development to pay a small learning cost to quickly engage in back-end software development, the Node.js based open source architecture Electron is a good choice. The process of backend development based on Electron architecture using JavaScript, Html, CSS, etc. were described by writing programs that implement some functions. Facts have proved that the use of the Electron open source architecture for back-end development is very friendly and feasible for front-end development.

Keywords: Node.js architecture; electron architecture; JavaScript language; cross platform; IPC communication

文章编号: 1007-1423（2023）16-0093-06 DOI: 10.3969/j.issn.1007-1423.2023.16.017

医疗大数据隐私信息泄露途径分析及保护举措

李晓蕾，王　猛*，刘钰周

（济宁医学院医学信息工程学院，日照 276826）

摘要：通过分析医疗信息生命周期各阶段的用途，探讨信息泄露的风险和途径、医疗信息和个人隐私信息泄露的社会影响。在此基础上，从国家层面数据隐私保护法律法规的完善、用户层面的隐私保护意识的提升和医疗单位层面的信息安全防护技术的加强等方面提出相应的建议。

关键词：医疗大数据；隐私；隐私保护

0　引言

新型冠状病毒肺炎疫情让国内医疗信息化系统、远程诊断等互联网医疗健康相关领域再次激活，医疗数据量呈现“爆炸式”增长的趋势。据统计，2020 年全球的医疗数据量高达 35ZB[1]。医疗数据包括门诊住院记录、药物临床试验研究数据、人类相关基因组学、代谢、胚胎组学、疾病监测数据等[2]，涉及范围广而复杂，具有数据量巨大、数据结构复杂、动态化、隐私性等特点。医疗数据对人类医学发展研究工作至关重要，充分利用医疗信息资源会对人类医疗健康发展创造极大的价值。但频繁发生的医疗信息泄露事件警示公众，信息泄露的途径愈加复杂化，迅猛发展的黑客技术严重威胁到医疗隐私数据安全。例如美国三军健康管理处数据泄露事件[3]作为最大的数据泄露事件，造成数亿美元的损失。因此，本文从医疗隐私信息产生的各阶段生命周期分析泄露途径和社会影响，从法律法规完善、提升社会群体医疗隐私安全意识、加强医疗隐私信息安全防护技术等角度进行探讨和研究，并给出相关建议。

1　医疗大数据构成及其生命周期

医疗大数据主要包括：生物大数据、临床大数据、医疗机构大数据、健康大数据[4]。生物大数据主要包括基因组学、转录组学、蛋白质组学、代谢组学等生物学数据，其数据特点为：高价值、数据量大、数据类型多样等[5]。临床大数据主要产生于患者临床诊疗的过程中，包括但不限于：患者个人身体指标数据、姓名、年龄、就诊记录等信息，其数据具有隐私性的特点。医疗机构大数据主要产生于各医疗单位和组织，包括医院消费记录、药物、器材、医疗工作者等信息。健康大数据主要来源于现代化的医疗产品，例如：运动手环等可穿戴设备、健康检测 APP 等，它们可以实时监测用户的身体指标信息、用户的浏览记录，甚至是使用者的位置信息，等等。医疗大数据具有完整的生命周期，包括数据采集、存储、处理、共享、销毁五大阶段，具有阶段性的数据特征[4]。

1.1　数据采集

医疗大数据的采集环节主要分为主动采集和被动采集两种方式，信息主要来源于用户和

收稿日期：2023-05-06　**修稿日期：**2023-07-31

基金项目：山东省大学生创新创业训练计划项目（S202110443006）

作者简介：李晓蕾（2000—），女，山东德州人，在读本科，研究方向为网络安全与软件开发；*通信作者：王猛（1981—），男，山东济宁人，硕士，副教授，研究方向为计算机网络及信息安全，E-mail：apollo636@foxmail.com；刘钰周（2001—），男，山东济南人，在读本科，研究方向为网络安全与软件开发

医疗机构[6]。主动采集主要指用户通过互联网或者医疗APP进行的个人信息登记；或者用户在使用可穿戴医疗设备时，通过传感器采集的个人识别数据，如身高、体重、血压、心率等生理数据[7]。这些数据大都由设备供应商采集并存储，是隐私信息泄露的重灾区。被动采集主要是医疗机构在医疗工作过程中产生的大量医疗信息，包括医院临床诊断的患者信息、医保信息等各类信息。攻击者可以通过钓鱼攻击、篡改上传服务器地址、篡改数据定向等方式在数据采集过程中截获采集到的数据。

1.2 数据存储

存储量如此之大的数据，若保存不当很容易导致信息泄露、篡改等。而医疗机构往往关注的是医疗行业本身，在信息技术特别是隐私信息的保护方面有很大的欠缺。攻击者一般会采取直接破坏性攻击医疗信息系统，利用拒绝服务攻击或信息炸弹等方式使医疗系统瘫痪，或者通过植入病毒程序、后门程序等方式进行数据监听，以此获取大量医疗隐私数据[8]。另外，除了本地存储，有些医疗大数据还被存储在云端。云端提供给用户访问权限和资源管理的服务，由于身份认证、密钥管理、权限管理等访问控制措施方面的不完善，也会导致信息的泄露[9]。

1.3 数据处理

数据处理是指从海量数据中提取所需信息，运用高精度算法分析数据，为医学研究提供有价值的信息、进行数据加工和分析的过程。医疗数据的处理过程中，数据访问者的权限安排不合理或者数据脱敏处理不当，都会导致数据应用人员访问到原始数据，使得数据泄露风险增加。其次，通过高精度算法违规挖掘信息、违规备份信息，此类违规行为都会造成医疗数据泄露的严重后果。另外，在医疗数据应用处理的过程中，也极易因为工作人员的操作不当而导致数据泄露。此外，在数据处理过程中未能及时地进行数据检查、更新系统，也将导致医疗数据泄露危险系数升高[10]。

1.4 数据共享

存储在不同医疗机构的数据通过数据共享才会产生更大的价值。但在网络环境下，数据共享带来便利的同时也给患者带来了风险。在数据传输的过程中，攻击者往往采取数据监听的方式窃取数据。当患者的数据存储在云平台上，由于数据脱敏不完善又或者脱敏技术过于简单，也会有很高的数据泄露风险。另外在大数据的查询、访问过程中，不严格的权限访问、不合规的存储使用都将导致数据泄露。

1.5 数据销毁

数据一旦不再进行预期目的的分析、长期没有任何访问需求、超过生存时间戳以及存储冗余都应该进行数据销毁。数据销毁最彻底的方式是将存储数据的硬件存储介质通过粉碎或焚烧进行销毁，但造成了一定的浪费，所以基本上没有得到广泛应用。如果仅仅是将数据删除或者存储介质格式化，很容易被攻击者利用专业设备和技术对数据进行恢复和还原，从而造成隐私数据泄露。在云环境下，用户失去了对数据的物理存储介质的控制权，无法保证数据存储的副本同时也被删除，导致传统删除方法无法满足大数据安全的要求。如果存储在云端的数据删除不彻底，极有可能使敏感数据被违规恢复，从而导致隐私信息面临泄露的风险。

综合而言，医疗数据具有规模大、增长速度快、容量大、价值潜力不可估量的特点。而在医疗信息生命周期从数据产生到销毁之间，由于医疗信息安全的保护性不强及网络安全技术不到位，极易导致信息泄露。

2 医疗隐私泄露事件原因分析

2.1 医疗数据的高价值与法律保护之间的失衡

信息时代医疗领域机构之间的信息共享，已经极大地推动了医疗科学的发展，也为人们提供了更优质的医疗服务。但医疗数据因自身的高价值也成为不法分子的觊觎目标，窃取信息的第三方可以从中获得很大的利益。尤其是信息的二次拼接利用，即将海量的医疗数据通过一定算法进行数据筛选拼接，从而得出更有价值的信息[11]。患者的隐私信息已然成为最有价值的“商品”，使得医院信息系统成为“黑客”攻击的重要目标，造成大量医疗数据的泄

露。通过制定法律法规来保护隐私一直是世界各国关注的热点问题，与国外隐私保护立法相比，国内立法起步较晚，且立法推进工作存在明显的滞后性。目前我国对于隐私的法律保护散见于其他法律之中，缺乏对隐私保护的单独立法[12]。

2.2　医疗领域工作者信息安全保护意识欠缺

Verizon发布的一篇网络安全报告显示，在全世界范围内，医疗行业是唯一一个内部威胁高于外部威胁的行业，内部从业人员对医疗数据的泄漏达到了惊人的程度。医疗领域工作者在工作过程中可以直接接触患者的隐私信息，而由于其自身信息安全保护意识和法律意识欠缺，造成医疗隐私数据有意或无意的被泄露。如随手拍摄、转发涉及患者个人隐私的信息；随意收集、转卖患者信息；随意放置、丢弃信息登记本等纸质材料等[13]。另外，由于信息安全保护意识欠缺，使得医院内信息系统、PC机、自助服务机、无线网络等的防护措施不完善而容易受到黑客攻击，同时工作人员也可能会受到社会工程学攻击，导致医疗数据的泄露[14]。

2.3　公民个人信息防护意识的缺乏

根据调研分析，公民对于医疗信息泄露事件是担忧的，并且认为泄露的主要原因是商业利益的诱惑[15]。但是公民并不十分清楚如何从个人角度降低泄露医疗信息的风险。患者往往因缺乏辨识能力而轻信网上的诊疗机构，从而在网站或APP上留下身份证号、姓名等私人信息，或将更多的身体指标信息盲目投医。而这些不正当盈利机构会从网站或APP的存储云端，收集到用户上传的数据，并将各处收集的数据进行信息拼接、筛选，最终患者的大量信息被挖掘出来，造成患者“半透明”地暴露于互联网中。尤其是老年人群体的个人信息保护意识和防骗防诈意识更为欠缺，某些健康保健诈骗机构通过电话营销的方式骗取老年人信任，从而获取老年人的隐私信息并以此牟利。

2.4　医疗机构网络建设安全防范系数低

开放的网络环境和相对脆弱的安全防御，使医疗机构成为了入侵者的乐园。入侵者可以较轻易地进入医院网络，访问、盗取和破坏敏感数据[16]。开放或半开放的网络环境是医疗行业的典型特征，很少有行业像医疗行业一样是一个非安全的开放环境。如医生的工作电脑几乎处于人人可以接触的开放环境，大量不安全的自助服务设备使入侵者有机可乘，广泛使用的无线网络也为入侵者进入医院网络提供了便利。面对相对开放的网络环境，医疗机构的安全防御技术却相对脆弱。目前医疗机构内部使用最多的安全防御技术是防火墙和入侵检测系统，但这两项技术都是基于已知的病毒或攻击，难以抵御新型病毒的入侵和花样百出的黑客攻击。

2.5　新型电子医疗产品带来安全隐患

智能腕表、运动手环等新型可穿戴电子产品以及新兴的各类保健智能设备在为用户提供实时身体指标监控和优质医疗服务的同时，也增加了用户个人信息和医疗数据的泄露风险。这些设备不仅能收集用户的基本信息，而且还能采集实时的客观生命体信息和主观输入的事件信息。这些有关个人隐私的信息容易受到黑客的攻击，一旦用户隐私受到攻击，就有可能威胁到用户的生命健康。另外，这些设备采集的信息大都存储在设备厂商的云服务器上，而设备厂商的安全防护不足或存在安全漏洞将使得攻击者轻易盗取用户的云端信息，导致大量用户医疗数据的泄露。

2.6　APP过度索取敏感权限

当下手机APP越权过度收集手机用户信息，并导致侵权、信息泄露事件屡见不鲜。用户在安装完APP登录时，第一步往往需要签订同意该APP权限的协议，其中名词过于专业化、通过隐含信息过度索取权限。如“您提供的数据与资料将授予本运营商独家永久使用权”。用户往往不会详细阅读协议而盲目选择同意，将自己的许多个人信息泄露。或者APP在安装时大都会出现“是否允许访问您的位置”提示，这就是为了收集用户的活动轨迹，从而了解用户的行为习惯。又或者APP要求获取发送短信、访问通讯录、访问通话记录等与应用无关的权限，而这些正是用户要保护的隐私，如果APP获取该权限就可能造成隐私泄露。据统计，2020年3月中旬至今，工信部等相关部门前后

22次通报或下架违规收集个人信息的APP，涉及超1100款APP。

3 医疗隐私信息保护对策及建议

3.1 医疗隐私信息安全保护提供法律保障

目前国内关于个人健康医疗信息的保护在法律层面还未有专门立法，健康医疗领域对于个人信息的保护要求散见于法律、部委/地方性法规和规章当中。具体到医疗卫生领域相关法律中，《精神卫生法》《医师法》《护士管理办法》等法律和规章中都规定了医护工作者对于个人健康信息隐私保护的责任义务。但这些法律规范只是对医疗机构内部人员进行了一定的约束，对于泄露患者隐私的个人和机构进行一定力度的处罚，而对于外部的威胁不能起到法律的震慑作用。2021年11月1日开始实施的《个人信息保护法》是我国首部针对个人信息保护方面的专门法律，这部法律的出台，使得个人信息保护更加体系化，对个人信息的处理、跨境提供、个人的权利、处理者的义务、个人信息保护部门的职责以及法律责任进行了系统的规定。自此，我国形成了以《网络安全法》《数据安全法》《个人信息保护法》为核心的网络法律体系，为数据应用和个人信息权益保护提供了基础制度保障。但由于我国《个人信息保护法》实施时间不长，健康医疗领域尚未形成较为完善的隐私保护管理实践[17]。

3.2 提升公民安全保护意识及能力

近些年由于公民安全保护意识不到位、个人信息安全防护能力欠缺而导致信息泄露的事件时有发生。政府部门应当通过各种渠道加大对个人信息保护的法律知识宣传、对公民进行个人信息犯罪的警示教育。公民在日常生活中也要提高个人信息保护意识：在网站或APP上谨慎填写和上传个人信息、加强对自己的账号和密码的保护、定期对手机和电脑进行木马病毒查杀、不随意连接公共场所Wi-Fi、谨慎设置手机APP的权限等。

3.3 增强医疗相关行业工作人员信息安全意识

医疗行业要加强对从业人员法律意识的培养，严格控制数据的使用权限和遵循最小范围使用原则。医疗机构、疾控部门等要重视数据安全风险评估的工作，事前有效预防安全风险[18]。另外，信息安全培训计划应严格执行并落实到位。医药领域的互联网企业要切实提高数据安全意识，进行有效的数据安全管理，积极维护系统和更新升级，防止黑客恶意攻击。要有严格的数据分级机制，从信息产生、存储、传输、访问、销毁等诸多环节要有完备的安全运营体系，充分保证数据完整性和可靠性。还应依照相关法律法规，制定行业保密协议，防止内部人员泄漏和贩卖医疗电子数据。

3.4 加强信息安全防护技术

医疗大数据生命周期中的各个阶段，都存在隐私数据泄露的风险，需要采取相应的技术手段来应对。医疗大数据全生命周期保护技术如图1所示。在数据采集阶段通过匿名化技术来隐藏数据与个体之间的联系，通过差分隐私技术在保留统计学特征的前提下去除个体特征以保护用户隐私，通过同态加密技术对某些数据字段进行加密来防止敏感数据泄露[19]；在数据存储阶段主要使用加密存储技术来保证数据即使被偷窥也不泄露其中蕴含的信息，使用审计技术来验证数据完整性以确保数据不被篡改；在数据处理阶段，通过访问控制技术来确保用户对敏感数据具有“最小权限的原则”可以避免和防止大多数对数据库有意或无意的侵害，通过隐私数据挖掘保护技术可以在不侵犯数据所有者隐私的情况下发现数据中隐含的知识；在数据共享阶段主要通过加密技术、扰乱技术和访问控制技术来进行隐私数据的保护；在数据销毁阶段主要通过对数据存储介质的物理销毁和数据的多次覆写两种方式来完成。面对日益增加的数据量和大规模的数据融合应用需求，逐渐成熟的区块链、隐私计算等新型数据保护技术是未来个人隐私保护发展的新方向。

3.5 推进医疗健康行业信息安全监管

除国家部门进行相关的法律规定，政府及各监管部门都应协调各方，为保护公民医疗隐私信息进行整个行业的监管，打击窃取公民医疗隐私信息谋取暴利的违法行为。根据相关研究建议，为全民进行公民医疗隐私信息的保护，

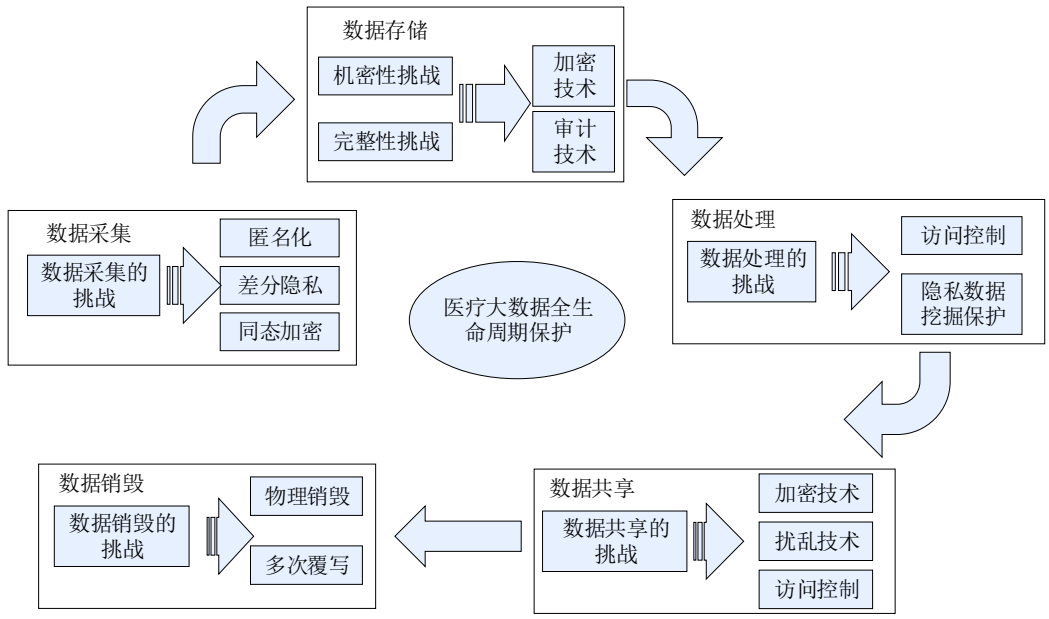


图 1　医疗大数据全生命周期保护模型

对健康大数据信息安全进行等级分类[20]，要求信息访问者为实名制。当发现信息泄露事件时，使查找访问者变得有迹可循，方便追寻信息泄露之处。同时整个医疗信息安全行业也应设置严格的惩罚制度，对进行不合法的信息泄露行为的个体和组织追究相应责任。同时培养相关医疗信息技术人才，各医疗组织机构设置相应信息管理部门，专门进行信息筛查工作，使每个医疗组织的医疗信息来源透明、去向安全，打击贩卖信息谋取利益的行为。

4　结语

在信息技术的迅猛发展下医疗健康领域将会发生重大的变革与发展，这既是机遇也是挑战。医疗健康大数据的价值是一把双刃剑，合理分析利用将极大促进人类医疗健康卫生领域的发展，造福人类。倘若医疗隐私信息安全性无法得到保障，将对社会各行各业带来很大的风险和冲击。面对医疗隐私信息安全问题，本文从医疗隐私数据的构成及数据生命周期进行分析，探讨信息泄露的风险和途径，并从多个层面探讨了医疗大数据隐私保护的方法和措施，以期更好地促进医疗大数据隐私保护。

参考文献：

[1] 王乐子,母健康,朱翀,等. 国外医疗信息化领域隐私数据保护现状及其启示[J]. 医学信息学杂志,2019,40(2):40-46.

[2] 宋扬,贾王平,韩珂,等. 健康医疗大数据的应用及其挑战[J]. 中国慢性病预防与控制,2021,29(3):220-223.

[3] 李亚子,尤斌,王晖,等. 医疗保险信息泄露案例分析及对我国安全隐私保护的借鉴[J]. 医学信息学杂志,2014,35(2):6-12.

[4] 张振,杨翠湄,徐静,等. 健康医疗大数据应用发展现状与数据治理[J]. 医学信息学杂志,2022,43(7):2-8.

[5] 王强芬. 大数据时代医疗隐私层次化控制的理性思考[J]. 医学与哲学(A),2016,37(5):5-8.

[6] 秦盼盼,李亚子,郭珉江,等. 生命周期视角下电子健康档案隐私保护研究[J]. 中国数学医学,2016,(3):2-3.

[7] 刘容京. 可穿戴健康监测系统隐私保护模型设计与实现[D]. 西安:西安电子科技大学,2018.

[8] 余妍霓,邓亚男,谢鑫. 浅谈黑客常用攻击手段及安全防护[J]. 科技与创新,2021,(4):74-75.

[9] 胡谷雨. 医疗隐私数据安全共享的机制研究[D]. 南宁:广西民族大学,2020.

[10] 刘桂锋,阮冰颖,包翔. 数据生命周期视角下高校科学数据安全内容框架构建[J]. 情报杂志,2021,40

(2):146-153.
[11] 殷明雪. 健康医疗大数据建设与隐私保护之冲突与协调[J]. 医学与社会,2023,36(2):125-131.
[12] 任晓勇. 我国互联网医疗中患者隐私权保护的法律完善[D]. 太原:山西财经大学,2021.
[13] 万琦. 互联网医疗中的患者隐私权保护研究[D]. 泸州:西南医科大学,2022.
[14] 冯基,王明芹,赵媛,等. 互联网医疗的隐私保护与信息安全探讨[J]. 科学与信息化,2019(19):53-54.
[15] 钱庆. 个人医疗保险信息隐私保护及信息共享认知的调查分析[J]. 中华医学图书情报杂志,2016,25(9):13-17.
[16] 吴超. "互联网+医疗"信息安全问题及对策[J]. 集成电路应用,2019,36(3):74-75.
[17] 张馨文. 大数据时代个人健康医疗信息的法律保护研究[D]. 郑州:河南财经政法大学,2022.
[18] 罗康. 医疗数据发布的隐私泄露风险评估系统设计与实现[D]. 贵阳:贵州大学,2022.
[19] 尚靖伟. 医疗大数据隐私泄露行为分析与建模[D]. 昆明:云南财经大学,2020.
[20] 胡雅婧. "互联网医疗"信息安全监管研究[J]. 中国卫生法制,2021,29(6):96-99.

Analysis and measures of medical bigdatprivacy information disclosure

Li Xiaolei, Wang Meng*, Liu Yuzhou

(School of Medical Information Engineering, Jining Medical University, Rizhao 276826, China)

Abstract: By analyzing the use of each stage of the life cycle of medical information, this paper discusses and analyzes the risks and ways of information disclosure, and explores the social impact of medical information and personal privacy information disclosure. On this basis, this paper puts forward corresponding effective measures and suggestions from three aspects: the improvement of data privacy protection laws and regulations at the national level, the improvement of privacy protection awareness at the user level and the strengthening of information security protection technology at the medical unit level.

Keywords: medical big data; privacy; privacy protection

(上接第86页)

Research and development to intercept and remove malicious bundled software systems

Liu Yubin, Ye Chuanqi*, Zhang Shaobo, Zhu Xiyang, Qin Mengwen, Liu Xiaoyu

(School of Software Institute, Henan University of Science and Technology, Luoyang 471003, China)

Abstract: Today's Internet industry is thriving, but it also breeds problems such as spam bombing and phishing websites stealing information. In order to solve the most prominent problem of malicious bundled software download, this paper designs an interception and removal system for bundled software, and the system uses feature value matching, cloud killing, text analysis and other technologies to realize the protection functions of filtering before downloading, checking during installation, and supervising the use of new software. The system size is approximately 48.4% of the average of the similar products. In the function of preventing malicious bundled software downloads, the page detection accuracy index and resource occupation index of the system are 6% and 13% higher than the average scores of similar software, respectively.

Keywords: computer software and theory; bundling software; eigenvalue matching; system security; text analytics

文章编号: 1007-1423（2023）16-0099-05　　DOI: 10.3969/j.issn.1007-1423.2023.16.018

基于LabVIEW的物联网智能泳池水质监测系统设计

汤雅婧，韩　涛*，肖　波，薛　博

（湖北师范大学电气工程与自动化学院，黄石 435002）

摘要：针对传统水质检测效率低，对智能泳池水质监测系统进行设计。该系统是以STM32单片机为主控芯片，利用DS18B20温度传感器、水位传感器、TSW-30浊度传感器和pH传感器实时采集泳池内环境数据，具有阈值设定和报警功能。使用LabVIEW完成上位机监测实现数值与波形显示功能。通过算法控制和阈值比较控制继电器实现参数调节。采集的数据还可以在OLED屏上显示，也可以通过手机APP实现远程监测。

关键词：水质监测；LabVIEW软件；远程监测；STM32

0　引言

当下泳池水质安全事故频发对人身造成健康危害，水质管理成为重中之重，现代泳池能检测水质但难以反馈，监测效率低下，传统的水质监测过程中，人力资源浪费严重，时时进行抽查监测，不能详尽地表现出水质的变化过程，反而极易造成资源的浪费[1]。本设计将提供一个更完备的泳池水质控制系统，集游泳池水质检测控制、LabVIEW后台管理于一体的物联网智能泳池水质监测系统，对水质管理进行整合升级。

1　系统总体方案设计

系统由数据采集单元、主控单元、通信单元、LabVIEW上位机监测以及各执行机构组成，以STM32f103为主控芯片，具有一定的负载能力。数据采集单元由温度传感器、水位传感器、TSW-30浊度传感器、pH传感器组成，通过LabVIEW完成上位机监测实现波形输出，通过手机连接蓝牙查看采集数据信息及时反馈超标数据。变压稳压电路实现传感器数据稳定输出，执行机构则由加热棒、蜂鸣器、水泵、OLED显示屏、步进电机构成。STM32芯片与LabVIEW上位机以及各检测传感器的结合使用，使得本系统可作为水质在线参数检测仪使用。系统设计框图如图1所示。

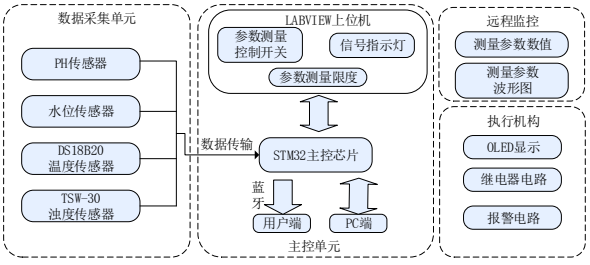


图 1 系统设计框图

2　硬件系统设计

智能泳池水质监测系统硬件部分由电源模块电路、数据采集单元、执行机构电路、OLED显示电路、A/D转换电路以及蓝牙通讯模块组

收稿日期：2023-04-25　**修稿日期：**2023-05-29

基金项目：国家自然科学基金面上项目（62071173）；2022年湖北师范大学国家级大学生创新创业训练计划项目（202210513010）

作者简介：汤雅婧（2002—），女，湖北孝感人，本科，主要研究方向为智能控制；*通信作者：韩涛（1987—），男，湖北黄石人，博士，校聘教授，研究方向为多智能体系统的协同控制，E-mail：taohan@hbnu.edu.cn；肖波（1987—），女，湖北黄石人，硕士，讲师，研究方向为网络控制与智能控制；薛博（2003—），男，湖北大冶人，本科，主要研究方向为智能控制

成。其中执行机构为本系统硬件核心。

2.1 数据采集电路

2.1.1 pH及温度传感器模块

pH传感器模块是通过检测被测物中氢离子浓度并转换成相应的可用输出信号，通过pH值反映水的酸碱度，使用BNC接头与pH复合电极进行连接，并且模块拓展有温度传感器DS18B20接口，选用不锈钢防水型温度传感器，通过与STM32单片机PA0引脚进行连接，pH传感器的I/O口则连接单片机的PB12引脚，通过测量热电阻的该变量来感知，将电阻的变化量转换成电压信号，再通过信号放大器放大，可直接输出0～5 V模拟电压信号[2]，做到将pH传感器电极信号放大三倍输出，通过单片机ADC采样后进行电压采集，进而将输出电压等比例地转化为3.3 V以下的电压，通过线性关系反推出真实电压值。pH值数据还需通过温度补偿来获取更为精确的参数数据，初次使用模块或更换pH电极传感器后，需用校正缓冲液对模块进行pH校准。温度、pH值采集电路如图2所示。

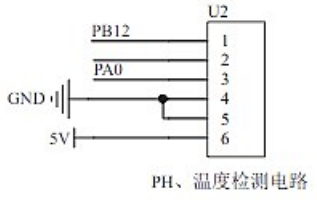


图2 温度、pH值采集电路

2.1.2 浊度检测电路

使用TSW-30浊度传感器，其内部含有光敏二极管和红外辐射二极管，通过溶液中的透光率和散射率来综合判断浊度情况。通过内部的红外线对管检测光线的透过量，将光强度转化为电流的大小。水体浑浊程度越大，透过的光线越少，被光接收端转换成的电流就小，反之电流越大。

浊度电流信号通过串联210 Ω电阻转换成0～5 V电压信号，采用ADC转换实现数据的输出，通过模块上10 K蓝色电位器的旋钮对数字量输出触发阈值进行调节，超过阈值时D1指示灯被点亮。且pH传感器内部处理方式与浊度传感器相似，浑浊度采集电路如图3所示。

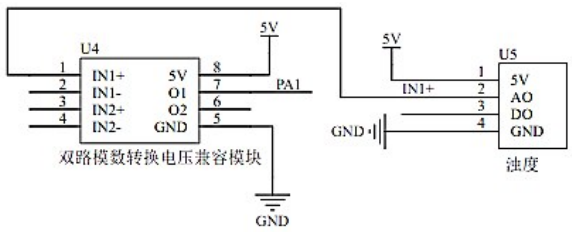


图3 浑浊度采集电路

通过最小二乘法拟合得到浊度与电压计算曲线满足式(1)，$R^2 = 0.9$。

$$Y = -865.68X + 3291.3 \quad (1)$$

2.1.3 水位检测模块

水位检测是通过一系列暴露的平行导线线迹测量水滴或水量大小来判断水位。当有水接触到平行导线时，模块会输出一个高电平信号，再通过模数转换成数字直观反映出测量数据，利用函数进行模拟电压值到水位的转换。当无水接触到平行导线时，会输出一个低电平信号，驱动芯片运行对信息进行采集，通过继电器自动闭合性控制负载电路，同时反馈数据。超过阈值时蜂鸣器触发，达到水位报警的功效。数值可直接读取并显示在OLED显示屏上。

2.2 通讯模块

使用JDY-31蓝牙进行单片机与手机之间的无线通讯，有效传输距离为30 m，可进行短距离信息传输，通过UART串口通讯，将单片机检测到的水温、水位、pH值、浑浊度信息传至手机中。通过USB转TTL电脑串口助手可收到蓝牙测量数据，接收界面显示连接成功即可读取测量值。

2.3 执行机构电路

执行机构包括OLED显示屏、继电器以及相应负载，通过继电器通断实现负载启停、步进电机、蜂鸣器、加热棒以及水泵等负载共同组成电路核心，温度通过加热棒调节，浊度和pH值通过步进电机和水泵调节，步进电机控制试剂阀门，水位通过水泵进行调节，当检测到的pH值、温度、浑浊度、水位超出阈值范围时相应继电器动作，蜂鸣器报警，提醒工作人员监督并采取相应措施。

3　软件系统设计

本系统中软件系统设计由STM32系统程序、LabVIEW上位机软件以及蓝牙通讯构成。软件系统设计流程如图4所示。

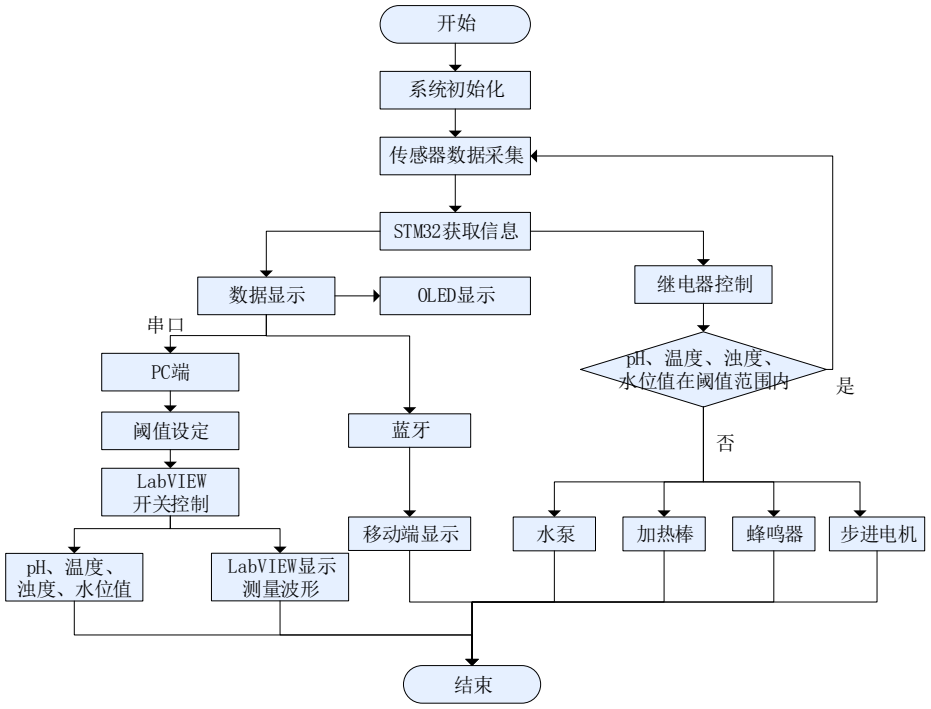


图4　软件系统设计流程

3.1　STM32单片机主控设计

采用STM32f103c8t6作为主控芯片，它是一款基于ARM Cortex-M内核STM32系列的32位微控制器，使用J-link仿真器直接用计算机供电进行程序的下载和辅助调试，通过蓝牙进行无线通讯技术连接，实现物联网操控。

完成系统初始化通过传感器采集数据，初始化包括STM32芯片、定时器清零、中断、ADC中断、时钟、串口等初始化，调用程序检测水温、水位、pH值和浑浊度，显示在OLED屏幕上，数据同步传输到PC端进行阈值大小设定[3]。并可通过蓝牙连接到手机用户界面，实现数据查询统计以及分析。串口通讯连接到PC端LabVIEW上位机，对参数进行实时监控且可手动调控上限阈值。采用继电器实现系统的自动通断，传感器产生输入电信号直到采集数据满足控制指标，反馈信息超标时继电器会自动闭合。对监测的目标设置动态变化范围，一旦超过预定范围，系统会自动启动相关设备的运转，用以调节泳池的多项参数。整体结构框图如图5所示。

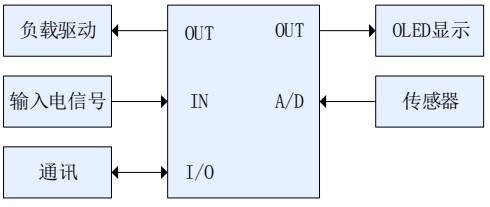


图5　整体结构框图

3.2　LabVIEW上位机设计

传感器数据通过连接STM32串口发送给电脑上位机，界面显示包括四部分数值、阈值、警报和波形[4]。通过visa属性的设定进行串口号和波特率的设置，建立可视化窗格更有利于编程调节，通讯完成后传入节点，节点信息设定

完成通过判断进行字符串到数组的整定。接收到16位的字符串通过字符串子集函数截取前4位数作为pH值，保留一位小数精度，后四位为浊度值，接着依次为温度和水位值，按照此方法最终通过截取数组得到相应的数据。

编程面板共分为四部分：串口通讯、字符串转换、数据比较以及显示输出[5]。其中串口通讯如图6所示，字符串转换如图7所示。

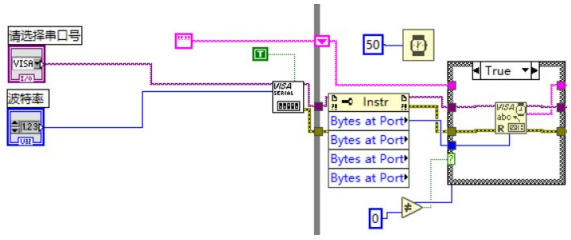


图6 串口通讯

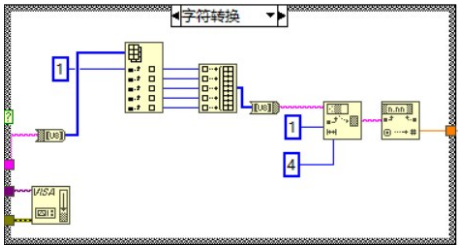


图7 字符串转换

串口连接后数值同步，可调节参数限度显示水质各参数波形。根据波形趋势判断数值变换，进行模拟分析和数值预测，显示水质的酸碱性、水温、水位以及浑浊度，超出阈值范围警报灯亮起，蜂鸣器报警相应执行机构动作。通过建立可视化View界面实现PC端上数据显示和远程监测功能，手机端与PC端实现数据同步。

3.3 温度采集

使用DS18B20不锈钢温度传感器，T1引脚为温度传感器信号输出口，通过软件进行温度补偿。模块与单片机GPIO PA0引脚相连，仅通过一个单线接口发送或接受信息，脉冲触发产生信号，复位脉冲跟着存在脉冲出现，表明准备好接发数据。检测到高电平时，延时后会通过60～240 us低电平信号的存在脉冲，单片机检测到电信号由此进行数据读写，以此来测量温度采集数据；若存在脉冲没有产生则做超时处理，在接收到低电平复位脉冲后进行复位处理。

3.4 浊度以及pH值采集

通过单片机时钟电路、复位电路以及模数转换的结合。数据采集时利用单片机内部的A/D转换器，通过将输入的模拟信号按规定的时间间隔采样，与一系列标准的数字信号相比较，直至两种信号相等为止。

系统首先进行ADC GPIO初始化，开启PA时钟和ADC时钟。选用PA1、PA2、PA3、PA4口为模拟输入，使用ADC四个转换通道，复位ADC1后进行分频因子设置，配置ADC时钟为8分频，即9 MHz，也就是转换时间为9 us；将使能ADC复位校准，ADC1转换的电压值则通过MDA方式传到SRAM，开启软件转换，转换完成后即可通过转换公式计算出真实pH值，通过多次转换求取平均值减小误差。而浊度传感器也是通过ADC转换进行数据采集，处理方式与pH传感器基本相同。

3.5 蓝牙串口通讯

蓝牙串口模块相当于手机端和单片机无线通信的媒介，蓝牙模块的RXD、TXD引脚分别对应连接单片机TXD、RXD引脚实现数据的收发。通过SPP串口透传协议建立无线通讯，配对成功后，使用AT指令串口调试助手进行蓝牙数据传输，确保实现蓝牙透传和串口通讯功能，使用AT指令修改蓝牙设置时，需要保证蓝牙没有处于通信状态，发送“AT+UART=9600,0,0”，“AT+PIN[1234]”等命令进行基本配置设定[6]，保证手机和单片机之间的稳定传输。

4 系统调试与验证

通过系统建立通讯后，对上位机进行多次数据的验证，对采集到的水进行水质分析，因为误差产生较小波动和振荡，最终显现出采集到的数据波形，PC端数据显示如图8所示。从页面上就可以看到温度、浊度、水位、pH值的读数，从条状图或仪表盘也可以很清楚数据所在位置，同时反应到OLED屏幕上，当数值超过阈值，警报灯亮起，通过pH灯判别酸碱度，从而提醒工作人员警惕。经过多次测量和验证保证误差在5%以内，系统测试表明，本系统能够

有效地对水质进行多参数监测。

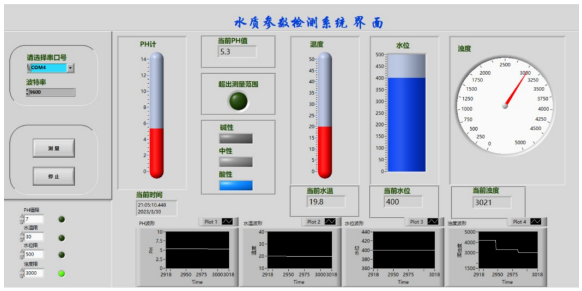


图 8　PC端数据显示

5　结语

本设计结合LabVIEW虚拟仪器技术实现水质参数监测，能有效地对包括pH值、浑浊度值、温度、水位四项参数进行远程实时监测，并在LabVIEW界面实时显示数值和波形。该系统实现了实时数据显示、自动调节和阈值报警功能。融合物联网技术通过手机APP实现远程控制，采用无线传输能够对水质参数进行监测并及时进行反馈，有较高的应用价值。

参考文献：

[1] 张开云. 基于单片机的实时水质监测系统设计[J]. 网络安全技术与应用,2020(6):75-66.

[2] 张亦飞,王靓. 水质参数在线监测系统的研究[J]. 科技资讯工程技术,2011(7):73.

[3] 闫翔,任立煌,钱颖. 无线智能水质监测系统的设计[J]. 电子信息科技风,2019(8):106.

[4] 刘时宇,肖波,韩涛,等. 一种基于虚拟仪器的电参数测量系统设计[J]. 自动化与仪表,2021,36(11):40-48.

[5] 谭华. 基于Labview监控的远程无线多参数水质监测系统设计[J]. 网络安全技术与应用,2022(8):36-37.

[6] 朱艳生,刘金亭. 基于蓝牙透传的串行通信电路设计[J]. 重庆科技学院学报(自然科学版),2020,22(4):97-99.

Design of an IoT smart swimming pool water quality monitoring system based on LabVIEW

Tang Yajing, Han Tao*, Xiao Bo, Xue Bo

(School of Electrical Engineering and Automation, Hubei Normal University, Huangshi 435002, China)

Abstract: In view of the low efficiency of traditional water quality detection, the intelligent swimming pool water quality monitoring system was designed. The system was based on STM32 single-chip microcomputer as the main control chip, using DS18B20 temperature sensor, water level sensor, TSW-30 turbidity sensor and pH sensor to collect real-time environmental data in the pool, with threshold setting and alarm functions. Use LabVIEW to complete the host computer monitoring and realize the numerical and waveform display function. Parameter adjustment was realized by algorithm control and threshold comparison control relays. The collected data can also be displayed on the OLED screen, and remote monitoring can also be realized through the mobile phone APP.

Keywords: water quality monitoring; LabVIEW software; remote monitoring; STM32

文章编号: 1007-1423(2023)16-0104-05 DOI: 10.3969/j.issn.1007-1423.2023.16.019

基于Unity3D的交通试验中心导览系统的设计与开发

王文心[1*]，邓欣睿[2]，秦永欣[3]

（1. 东南大学交通学院，南京 210000；2. 东南大学计算机科学与工程学院，南京 210000；
3. 东南大学机械工程学院，南京 210000）

摘要：随着交通强国战略的提出，交通人才培养过程中实验教学任务显得尤为关键。而交通相关实验所需场地广、仪器大，不利于学习了解。因此引入虚拟现实和增强现实技术，以东南大学交通学院试验中心为例，采用Unity3D引擎，结合3DsMax、Vuforia等技术，虚拟化实习场地和仪器，增强视听效果，构造出交通试验中心导览系统。将“以物为本”转化为“以人为本”的导览方式，富有交互、体验感，满足交通学院试验中心的导览、学习需求，是一次计算机辅助的交通运输专业多维教学方式的探索。

关键词：增强现实；虚拟实验室；Unity 3D；交通运输

0 引言

近年来，虚拟现实技术和增强现实技术发展迅速，在各个行业和领域的应用更加广泛。计算机收集各种辅助信息，使得用户与真实场景实现交互。随着交通强国战略的提出，交通基础设施建设也显得更为重要，而实验室正是基础设施建设的摇篮[1]。

虚拟实验室能提供直观的实验室场馆全貌和相应的实验仪器，方便学生获取相关实验信息，学习实验操作，实现与实验场馆设施的远程互动[2]。本文结合虚拟现实和增强现实的技术特点，将其应用于东南大学交通学院试验中心各实验室的虚拟展示和交互之中，设计与实现了一个基于Unity 3D的实验室虚拟漫游软件，打破了空间和时间的限制，实际运用在相关教学及宣传普及领域。

1 相关技术

1.1 虚拟现实技术(VR)

虚拟现实技术(virtual reality，VR)以计算机技术为基础，综合了电脑仿真、传感器、网格并行处理、人工智能等多种技术，通过给用户同时提供多种感官信息，使用户仿佛身临其境。依靠计算机系统，用户可以自定义生成一个三维空间。用户在该环境中，借助多种输入输出设备，感知和探索客观世界。借助于虚拟现实，用户可以突破时间空间的限制，优化感官感受，提高对客观世界的认识水平。虚拟现实的关键技术主要包括：人机交互技术、传感器技术、动态环境建模技术、系统集成技术和三维图像的实时刷新技术[3]。

1.2 增强现实技术(AR)

随着计算机水平的提高，增强现实技术(augment reality，AR)是发展来的一门新技术。它借助计算机技术，将构建的辅助虚拟信息如模型等叠加到真实世界，使虚拟的物体信息和真实的环境信息出现在同一个画面或空间。用户可以感知被呈递的信息，获得与真实世界漫游相似的体验感。增强现实系统通过分析大量输入数据，获取场景中各种位置信息，使生成

收稿日期：2023-04-24 **修稿日期：**2023-05-23

基金项目：江苏省大学生创业创新训练计划项目(202210286262Y)

作者简介：*通信作者：王文心(2001—)，女，福建三明人，在读本科，专业为道路桥梁与渡河工程，E-mail：213203719@seu.edu.cn；邓欣睿(2001—)，女，四川遂宁人，在读本科，专业为计算机科学与技术；秦永欣(2001—)，女，安徽马鞍山人，在读本科，专业为机械工程

的虚拟物体以合适的姿态精确地定位到真实场景中的特定位置。增强现实的关键技术主要包括：跟踪注册技术、显示技术、人机交互技术[4]。

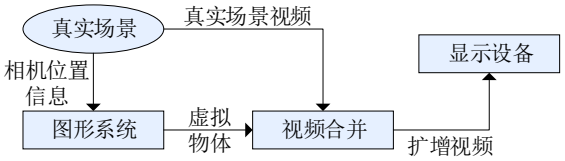


图1　增强现实技术

1.3　3DsMax

3DsMax是一款强大的三维动画渲染和制作软件，在模型制造领域有不可替代的作用。它功能种类丰富、灵活性高，可用来制作多种建筑模型、工业产品结构图及效果图等。该软件可模拟不同环境、不同风格的渲染效果，拥有强大的材质编辑功能。

1.4　Vuforia

Vuforia SDK封装了底层用来图像识别的计算机视觉模块，Vuforia为开发者提供了一系列参数设置。开发者只需配置参数，然后基于底层的识别算法便可开发出自己想要的AR程序。Vuforia功能丰富，如图像追踪、物体追踪、环境追踪等。

2　软件总体设计

整个设计软件效果：进入软件后，调用手机摄像头，对准已导入数据库的实验室环境、实验器材或展板时，系统能够在现实场景之上出现叠加音视频、文本、3D模型或VR体验等增强效果。在界面指示下，用户可以多维了解试验中心相关知识和实验器材的操作方法，也能跟随导览，走近感兴趣的下一目标点，继续游览学习。

2.1　虚拟环境构建

2.1.1　3D建模

建筑信息模型(BIM)现在是交通基础设施建设的新工具，3DsMax能构建出三维虚拟、内含完整信息集成的模型。借助内置基础的模型库进行打形，对简单的几何模型进行调整与组合，可以快速对三维建模搭建一些最基本的骨骼框架，真实比例还原实验仪器。将模型转换为可编辑多面体，然后使用点、线、面的选取修改功能来细化模型，最后修改贴图和材质，赋予模型细腻有质感的外观。

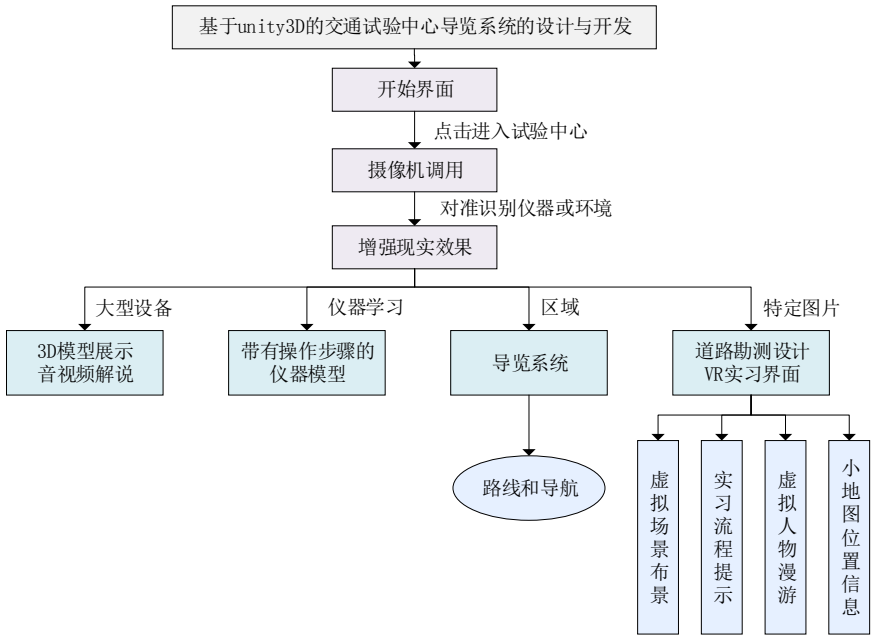


图2　软件整体设计架构

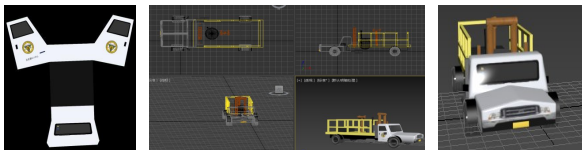


图 3 土体原位测试车建模

2.1.2 虚拟场景

为了达到全真模拟的效果，利用鱼眼摄像机拍摄，软件中进行图像识别特征点拼接，获取试验中心全景图。贴到Unity3D中的天空盒上，将整个视角放置于天空盒中，把二维布局区域虚拟成为现实的三维效果，保证立体感、真实性。

图 4 全景图

虚拟部分地形通过terrain用不同笔刷加之柔化效果，绘制出山峦与河流，细节上添加树木、房屋、车辆等模型，使地形自然逼真。

光照渲染采用实时烘培，这是一种在运行时动态计算光照的渲染方式，可以让游戏场景看起来更加真实、细节更加丰富。但是通过调整发现本软件所需的计算量较大，实时渲染可能会因设备问题跟不上导览进程。因此，部分大量计算且精密程度要求不高的场景，改为了前向渲染以提高效率。

通过Unity3D中的粒子系统，模拟实验器材操作时的环境效果，如室内操作时的粉尘、室外作业时的雨雪天气。如将单一粉尘形状作为基本粒子，设置预设脚本控制发射数目、形状、路线以及生命周期内的各种变化，最终实现仿真场景。

图 5 特效渲染

2.2 人机交互

2.2.1 自主漫游场景

VR灵境模块之中，使用者通过第三人称视角进入虚拟环境中体验交通学院教学实践体系中重要组成部分——道路勘测设计实习。

图 6 道路勘测设计实习

第三人称视角下，可以看到虚拟人物。人物拥有完整骨骼动画，有着丰富的动作库，在不同状态下有对应的动作效果。移动脚本挂载在人物身上，使用者通过摇杆操控方向行进，按键可跑动、跳跃。同时系统中有重力模拟器效果，从高处坠落或跳跃时，编入的g值会保证效果真实。

碰撞检测器能保证人物和物体有着真实存在感，避免出现“穿模”现象。在模型内部堆叠透明材质立方体，为其物理状态赋予刚体组件。刚体相遇即根据各自物理属性如摩擦系数、反弹性能得到不同效果。如：虚拟人物可以乘坐并操控试验车辆；在添加风组件之后，由于刚体组件及碰撞体的添加，树木等会出现随风摇曳的效果。

除人物控制外的其它物体移动可通过两种方式完成：一是通过Unity3D中的算法，可将实习场地中复杂的山川河流、道路桥梁等关系简化为带有一定信息的网格，在这些网格的基础上通过一系列的计算来实现AI自动寻路效果；二是通过Animation动画控制物体的大小、坐标、方向等物理性质随时间轴不断变化或循环，将组件赋予物体，以完成实时运动变化。

由于场景较大，加之实习场地林场森林覆盖率高、地形起伏大、视野不开阔，需要获取人物周边环境信息。这就需要以小地图形式用图标来代替真实的模型。

创建一个顶视角摄像机，计算出Unity中场景的长和宽，计算出模型在场景中的X与Y轴，然后计算X与Y的位置在场景大地图上的比例，按照比例给新创建Render Texture小地图赋值并进行边缘锯齿状问题的优化解决。当人物运动时，缩小的图标在地图上实时变化，即可清楚地从小地图中获取地理信息。

2.2.2　UI界面布设

为了便于用户在网络端参观实验室时实现不同场景的切换，我们利用Unity3D软件结合C#脚本实现这一功能。当选择不同的图片时，调用添加在该对象上Button组件的OnClick事件，使用SceneManager.LoadScene方法切换到对应场景。

在实验室场景漫游中，我们希望用户同时能够听到对应的语音讲解，实现首次加载进入场景时自动播放背景音乐，并且通过单击UI界面控制音乐的打开和关闭。使用AudioSource组件来控制音乐的播放，需要在Hierarchy窗口中创建一个Audio Source对象。然后在Inspector窗口中设置AudioSource组件的参数，例如是否在游戏开始时自动播放音乐(Play On Awake)，是否循环播放音乐(Loop)，音量大小(Volume)。

此外，视频可以用来在虚拟实验室中展示实验过程或者提供教学指导，用户可以通过观看视频来了解实验的步骤和注意事项。在Unity中，可以使用VideoPlayer组件播放视频。将VideoPlayer组件添加到一个游戏对象上，然后在运行时在游戏对象的纹理上播放视频。

当触发某个实验室物体时，系统调用弹出对话框对物体的功能进行介绍和说明。使用触发器(Trigger)来实现当物体进入触发器区域时调用事件。触发器是一个不可见的碰撞器，它可以检测其他物体是否进入了它的区域。在场景中创建一个空物体，并为其添加一个碰撞器组件(如Box Collider或Sphere Collider)。然后在碰撞器组件的检查器面板中勾选“Is Trigger”选项，将碰撞器转换为触发器。接下来，在物体上添加一个脚本，并在脚本中编写OnTrigger-Enter、OnTriggerStay和OnTriggerExit方法来处理触发器事件。当其他物体进入触发器区域时，将调用OnTriggerEnter方法；当其他物体停留在触发器区域时，将每帧调用一次OnTriggerStay方法；当其他物体离开触发器区域时，将调用OnTriggerExit方法。

图7　对话框效果

2.3　AR模块

2.3.1　AR识别

在Vuforia系统中创建所需要识别库，可以包括图像识别、多对象识别和实物识别。将识别对象导入识别库，在比率一栏会显示出星级。为提高识别效率，多次尝试优化识别对象的特征点，最终得到交通学院试验中心的识别库。生成unitypackage并导入Unity3D，打开Vuforia SDK，调用AR camera即可使用。

2.3.2　AR叠加融合

在AR camera识别出对象之后，软件将展示增强效果。对于基本实验器材或试验区域，利用UI设计部分中的Canvas画布半透明地展现实时场景上叠加的相关音视频、文字简介，并且配有相关对话框程序，使用者可根据需求选择软件服务。对于不方便当场使用的大型室外设备，叠加原比例三维模型并配有介绍，使用者可自行放大，翻转观看学习。对于不同的课程实习中需要学习掌握的实验仪器，叠加效果将识别出对应操作阶段，根据指示箭头指向下一步仪器需要操作的部位，指导使用者学习使用仪器。同时，在识别对应部分时还可以将系统跳转到对应VR灵境模块进行虚拟实习。

图8　AR识别效果

2.3.3 导览系统

导览系统依赖的是软件内部识别库和已写入的地理信息，不存在因为室内卫星定位信号不强而导致的问题。我们试图使用AR+GPS组件，此功能与地图框集成方向接口以创建实时增强现实路由，允许应用程序用户查看方向、导航说明、距离等。我们尝试使用此组件创建基于位置的导航AR体验，创建自定义路线Mapbox未映射的地方，创建自定义内容脚本来构建基于软件自身的独特体验。

3 结语

我们以研究虚拟现实技术和增强现实技术在室内导览系统中的应用为载体，基于东南大学交通学院试验中心，利用Unity3D引擎开发了一款视景仿真的实验室场馆漫游系统。设计控制方便、通用性强、有广阔的应用前景，其研究成果能宣传普及交通运输相关的一批重大科研装备的线下使用与线上虚拟体验，降低了实验教学经济和时间双重成本，有着重要的实际意义和应用价值。

进一步展望，本软件的整体构架和模式可以扩展到其它实验室、博物馆、体育场馆等室内体系[5]。开发场馆设施的场景交互漫游功能，实现用户交互、视点控制、场景空间置换等功能，让人们足不出户就能感受到AR\VR所构成的灵境世界。

参考文献：

[1] 蔡新民,刘金全,方毅.我国交通基础设施建设对经济增长的影响研究[J].经济纵横,2017(4):70-76.
[2] 谢婷玲.虚拟现实VR技术在教学资源可视化中的应用[J].电子技术,2022,51(7):310-312.
[3] 孙红杏.虚拟现实关键技术应用及研究[J].智库时代,2020(1):244-245.
[4] 王宇希,张凤军,刘越.增强现实技术研究现状及发展趋势[J].科技导报,2018,36(10):75-83.
[5] 赵吉宇,靳丽梅,王丽.虚拟现实技术在“互联网+教育”领域中的发展与展望[J].智慧农业导刊,2021,1(20):69-71.

Design and development of a navigation system for traffic test center based on Unity3D

Wang Wenxin[1*], Deng Xinrui[2], Qin Yongxin[3]

(1. School of Transportation, Southeast University, Nanjing 210000, China;
2. School of Computer Science and Engineering, Southeast University, Nanjing 210000, China;
3. School of Mechanical Engineering, Southeast University, Nanjing 210000, China)

Abstract: With the proposal of the strategy of building a strong transportation country, the experimental teaching task in the process of cultivating transportation talents becomes particularly crucial. However, transportation related experiments require a wide range of venues and large equipment, which is not conducive to learning and understanding. Therefore, virtual reality and augmented reality technologies are introduced. Taking the Experimental Center of the School of Transportation at Southeast University as an example, the Unity3D engine is used, combined with technologies such as 3Ds Max and Vuforia, to virtualize the internship site and instruments, enhance audio-visual effects, and construct a navigation system for the Traffic Experimental Center. Transforming the “object oriented” approach into a “people-oriented” approach to navigation is interactive and experiential, meeting the guidance and learning needs of the Experimental Center of the School of Transportation. It is an exploration of a computer-assisted multi-dimensional teaching method for transportation majors.

Keywords: augmented reality; virtual laboratory; Unity 3D; transportation

文章编号: 1007-1423（2023）16-0109-05 DOI: 10.3969/j.issn.1007-1423.2023.16.020

基于微信小程序的勤工助学管理系统研究与开发

谈伙荣，陈海宇*

（肇庆医学高等专科学校信息中心，肇庆 526020）

摘要：勤工助学活动是高校生活的重要组成部分，为学生提供经济资助和实践经验。为了提高勤工助学活动管理的效率和便捷性，设计一种基于微信小程序的勤工助学管理系统。该系统通过提供岗位发布、报名管理、签到、工时工资管理等功能，满足学生、发布者和管理员的需求，为其提供一个一站式管理平台。通过实现用户友好的界面、响应式设计和强大的后端服务，系统旨在满足各类用户的需求，进一步推动高校勤工助学活动的发展。

关键词：微信小程序；勤工助学管理；系统设计；系统实现；高等教育

0 引言

勤工助学活动具有较高的教育价值和实践意义，它不仅可以帮助学生缓解经济压力，降低家庭负担，还能培养学生的职业素养、组织协调能力及团队合作精神等综合素质[1]。然而，随着勤工助学活动的不断扩展，传统的管理方式已经难以满足高校对勤工助学活动的有效管理需求。微信作为国内具有广泛用户基础的社交软件，其小程序平台具有良好的普及性、便捷性和开发便利性，基于微信小程序开发勤工助学管理系统是一个具有可行性和广泛应用前景的解决方案。

本文旨在研究和开发一种基于微信小程序的勤工助学管理系统，以提高勤工助学活动的管理效率和服务质量[2]。首先，将对系统的需求进行详细分析，明确系统的功能、性能和技术选型等方面的要求。其次，本文将对系统的架构进行设计，并针对实现过程中可能遇到的难点提出相应的解决方案。最后，本文将对系统进行实现、测试和评估，以验证系统的实用性和有效性。

通过实现基于微信小程序的勤工助学管理系统，本文期望为高校提供一种高效、便捷、易用的勤工助学活动管理工具，推动高校勤工助学活动管理的现代化进程，为提高学生的综合素质和全面发展作出积极贡献。

1 需求分析

在开发基于微信小程序的勤工助学管理系统之前，首先需要对系统的需求进行分析。需求分析是软件开发过程中的关键环节，主要包括用户需求、功能需求和性能需求三个方面。

1.1 用户需求

用户是勤工助学管理系统的直接使用者，对于系统的需求分析起着至关重要的作用。本系统的主要用户包括三类：学生、发布者（教师或企业招聘者）和管理员。针对不同用户角色，系统需满足以下需求：

学生：可以快速查询并报名岗位，查看自己的报名、签到、工时和工资信息，及时获取

收稿日期：2023-04-17 **修稿日期：**2023-08-09

基金项目：广东省高等职业院校医药卫生专业教学指导委员会2021年教学改革课题（2021LX087）:《计算机应用基础》教学中课程思政实施办法的探索

作者简介：谈伙荣（1983—），男，广东肇庆人，硕士，讲师，研究方向为计算机教育、健康大数据教育、软件设计；*通信作者：陈海宇（1978—），女，广东罗定人，硕士，副教授，研究方向为计算机应用、大数据技术与应用，E-mail：Chenhaiyu@zqmc.edu.cn

相关通知，提出申诉或反馈意见。

发布者：可以便捷地发布和管理岗位信息，审核学生的报名申请，查看并管理学生的签到、工时和工资记录，及时处理学生的申诉和反馈。

管理员：可以对系统中的用户、岗位、报名、签到等数据进行监控和管理，维护系统的正常运行，处理纠纷和异常情况。

1.2 功能需求

根据用户需求，本系统需要实现以下功能模块：

用户模块：支持用户注册、登录、修改个人信息等操作，实现身份认证和授权。

岗位模块：发布者可以创建、发布、编辑和删除勤工助学岗位，学生可以查询和筛选岗位信息。

报名模块：学生可以报名参加岗位，发布者可以查看报名情况并进行审核，支持审核通过或拒绝操作。

签到模块：学生可以进行上下班签到，记录签到时间和地点，发布者可以查看签到记录。

工时工资模块：根据签到记录计算学生的工时和工资，发布者录入并审核工资信息，学生可以查看自己的工时和工资情况。

消息推送模块：系统向相关用户推送新岗位发布、审核结果等信息，保持用户的实时通知。

系统管理模块：管理员可以对用户和岗位进行管理，处理学生申诉和纠纷，以及系统数据统计等。

1.3 性能需求

性能需求是指系统在实际运行中所需满足的性能指标。本系统需要满足以下性能需求：

响应时间：系统应具备较快的响应速度，确保用户在使用过程中不会因为等待时间过长而影响体验。

可扩展性：系统应具备良好的可扩展性，以便在未来根据高校勤工助学活动的需求变化进行功能模块的添加和修改。

安全性：系统应保证用户数据和隐私安全，防止数据泄露和非法访问。

可用性：系统应在高峰期和其他压力环境下保持稳定运行，提供可靠的服务。

易用性：系统界面设计应简洁明了，操作流程合理，方便各类用户快速上手使用。

兼容性：系统应具备良好的兼容性，支持各种主流移动设备，以满足不同用户的需求。

通过对用户需求、功能需求和性能需求的详细分析，为开发基于微信小程序的勤工助学管理系统奠定了基础。在接下来的系统设计和实现过程中，需要充分考虑这些需求，并根据实际情况进行优化和调整，以确保系统能够满足高校勤工助学活动管理的需求。

2 系统架构

在明确了需求分析的基础上，接下来需要设计基于微信小程序的勤工助学管理系统的架构。系统架构是指系统的组成部分以及它们之间的关系，合理的系统架构设计有利于提高系统的开发效率和运行稳定性。本系统采用前后端分离的架构，如图1所示，包括微信小程序端(前端)和服务器端(后端)两部分。

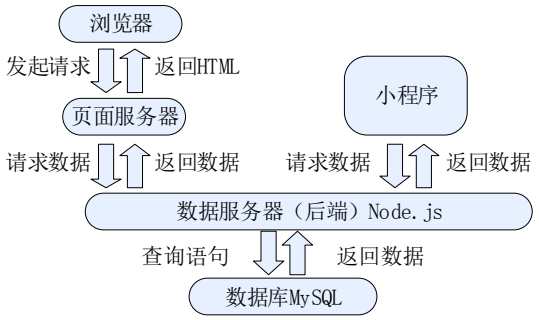


图1 前后端分离架构

2.1 技术选型

为实现系统的各项功能，本文在技术选型上做出如下选择：

前端技术：微信小程序采用微信官方提供的小程序框架进行开发，使用WXML、WXSS、JavaScript等语言和技术实现界面布局和交互功能[3]。

后端技术：后端采用Node.js作为开发语言，结合Express框架搭建服务器，提供API接口供前端调用[4]。数据库方面选择MySQL作为关系型数据库，用于存储和管理系统中的各种数据。

API设计：采用RESTful风格的API设计，规范API接口，方便前端调用和后端开发。

数据交互：使用JSON格式进行前后端数据交互，简洁易读，便于处理和传输。

2.2 系统模块划分

基于微信小程序的勤工助学管理系统可以划分为以下模块：

前端模块：包括用户模块、岗位模块、报名模块、签到模块、工时工资模块、消息推送模块和系统管理模块。这些模块负责实现系统的各项功能，以满足不同用户角色的需求。

后端模块：包括API模块和数据库模块。API模块负责处理前端的请求，实现业务逻辑和数据处理，同时与数据库模块进行交互。数据库模块负责对数据进行持久化存储和管理。

通过合理划分系统模块，可以更好地组织代码和功能，提高开发效率。在实际开发过程中，需要根据需求分析和系统架构进行相应的调整和优化，以确保系统的顺利实现和良好运行。

3 实现难点及解决方案

在开发基于微信小程序的勤工助学管理系统过程中，可能会遇到一些实现难点。以下列举了部分可能的难点以及相应的解决方案。

3.1 实现难点

实时消息推送：在勤工助学管理系统中，实时消息推送对于及时通知用户岗位信息、报名审核结果等信息具有重要意义，如何实现实时推送是一个难点。

签到定位：为确保学生的签到准确性，需要在签到模块中实现位置信息的获取与验证，如何获取准确的位置信息并进行验证是一个挑战。

数据安全与隐私保护：在管理系统中涉及到大量学生个人信息、工时和工资数据，如何保证数据的安全和用户隐私不被泄露是一个关键问题。

3.2 解决方案

实时消息推送：借助微信小程序的云开发功能，结合云数据库和云函数，可以实现实时推送消息。首先，将需要推送的消息内容存储在云数据库中；接着，通过云函数监听数据库变化，并调用微信小程序的消息推送接口实现实时推送。同时，可以设定推送权限，确保只有相关用户接收到通知。

签到定位：利用微信小程序提供的位置接口获取用户的经纬度信息，并将其与预设的岗位地点进行比较，若距离在可接受范围内，则视为有效签到。同时，可以设置一定的时间限制，防止学生提前或滞后签到。为提高定位准确性，可以结合多种定位方式，如GPS、Wi-Fi和蓝牙等。

数据安全与隐私保护：首先，在存储敏感信息(如学生姓名、手机号等)时，采用加密技术，如对称加密或非对称加密，防止数据泄露；其次，在开发API接口时，使用访问令牌(如JWT)对用户身份进行验证，防止未授权访问；另外，采用合理的数据库备份策略，确保数据安全可靠；最后，遵循相关法律法规，明确告知用户数据的使用范围和目的，以保护用户隐私。

通过以上解决方案，可以有效应对实现过程中可能遇到的难点，确保基于微信小程序的勤工助学管理系统的顺利开发和运行。

4 系统实现

在需求分析、系统架构及实现难点解决方案的基础上，进行基于微信小程序的勤工助学管理系统的实现。系统实现主要包括前端页面设计与实现、后端API设计与实现以及数据库设计三部分。

4.1 前端页面设计与实现

基于微信小程序框架，前端页面主要使用WXML、WXSS和JavaScript进行开发[1]。根据需求分析，前端页面设计包括以下几个部分：

登录与注册页面：实现用户的登录和注册功能，提供微信一键登录、账号密码登录等方式。

岗位列表与详情页面：展示勤工助学岗位的列表，支持筛选和搜索功能，点击可查看岗位详情及报名入口。

报名管理页面：学生可以查看报名记录并取消报名，发布者可以查看报名学生列表并进行审核操作。

签到页面：学生进行上下班签到，同时显示签到历史记录；发布者可以查看学生的签到记录。

工时工资页面：学生可以查看自己的工时和工资记录；发布者可以录入和审核学生的工时工资信息。

个人中心页面：展示用户基本信息，提供修改个人信息、查看消息通知、反馈与申诉等功能入口。

系统管理页面：管理员可以进行用户管理、岗位管理、数据统计等操作。

4.2 后端API设计与实现

后端采用Node.js和Express框架开发，负责处理前端请求、实现业务逻辑和与数据库交互。后端API设计遵循RESTful风格，包括以下几个部分：

用户相关API：包括用户注册、登录、获取个人信息、修改个人信息等接口。

岗位相关API：包括发布岗位、编辑岗位、删除岗位、获取岗位列表和详情等接口。

报名相关API：包括学生报名、取消报名、获取报名记录；发布者审核报名、查看报名列表等接口。

签到相关API：包括学生签到、查看签到记录；发布者查看签到记录等接口。

工时工资相关API：包括学生查看工时工资记录；发布者录入、审核工时工资等接口。

系统管理相关API：包括用户管理、岗位管理、数据统计等接口。

4.3 数据库设计

使用MySQL数据库进行数据存储和管理，数据库设计包括以下几个部分：

用户表：存储用户的基本信息，如用户ID、姓名、角色、联系方式等。

岗位表：存储岗位的基本信息，如岗位ID、岗位名称、发布者、岗位描述、工作地点、工作时间等。

报名表：存储学生报名的记录，包括报名ID、学生ID、岗位ID、报名时间、审核状态等。

签到表：存储学生签到记录，包括签到ID、学生ID、岗位ID、签到时间、签到类型(上班/下班)等。

工时工资表：存储学生的工时和工资记录，包括记录ID、学生ID、岗位ID、工时、工资、审核状态等。

消息通知表：存储系统消息通知，包括通知ID、接收者ID、通知类型、通知内容、发送时间等。

通过以上系统实现，基于微信小程序的勤工助学管理系统得以实现其功能，满足高校勤工助学活动管理的需求。在实际运行中，还需要不断地进行优化和维护，以保证系统的稳定性和可靠性。

5 结语

本文主要研究了基于微信小程序的勤工助学管理系统的设计与实现。首先，通过对勤工助学活动进行需求分析，明确了系统需求；接着，设计了系统架构，包括前后端分离的技术选型、系统模块划分等；然后，分析了实现过程中可能遇到的难点，并提出了相应的解决方案；最后，进行了系统的实现，包括前端页面设计、后端API设计和数据库设计。

基于微信小程序的勤工助学管理系统旨在提高高校勤工助学活动的管理效率和便捷性，为学生、发布者和管理员提供一个一站式的管理平台。通过实现岗位发布、报名管理、签到、工时工资管理等功能，系统有望满足各类用户的需求，进一步促进勤工助学活动的发展。

在实际运行中，需要不断地进行优化和维护，以保证系统的稳定性和可靠性。今后的研究可以在系统功能、性能优化、数据分析等方面进行拓展，提高系统的实用性和价值。

参考文献：

[1] 彭益全,吴彦宁,徐晓丽,等.资助育人视角下高校勤工助学精细化管理机制研究[J].高等农业教育,2019(6):44-49.

[2] 刘淑娟.高校档案馆(室)勤工助学学生管理问题与对策[J].上海理工大学学报(社会科学版),2022,44(1):103-108.

[3] 秦长春.微信小程序开发技术[M].北京:人民邮电出版社,2021.

[4] 刘江虹.Node.js实战:分布式系统中的后端服务开发[M].北京:机械工业出版社,2022.

(下转第117页)

文章编号: 1007-1423（2023）16-0113-05 DOI: 10.3969/j.issn.1007-1423.2023.16.021

基于自主加密芯片的智能家居控制系统设计

杨金龙*，马静怡

（郑州科技学院大数据与人工智能学院，郑州 450064）

摘要：提出了一种智能家居控制系统，该系统基于自主加密芯片，包括控制单元、加密芯片和连接单元。控制单元用于连接控制终端以接收来自控制终端的指令和数据，连接单元用于连接智能家居设备终端以输出控制指令。加密芯片用于将控制单元接收的来自控制终端的指令及数据加密后发送至连接单元。加密芯片由主控芯片和与之连接的算法协处理单元组成，算法协处理单元采用FPGA芯片，为主控芯片提供加密算法。通过在家居中设置不同类型的密钥对，并根据用户需要选择相应的密钥进行加密，从而保证了密钥的保密性和完整性。该智能家居控制系统采用基于自主加密芯片的技术，实现了更加多样化的加密算法，从而显著提升了系统的安全性。

关键词：自主加密；FPGA；控制系统

0 引言

智能家居系统是一种集成了计算机控制技术、传感器技术、通信技术、电子信息技术等并将其应用到传统家庭中的现代化系统[1]。智能家居系统具有集成度高、可靠性强、操作方便灵活、使用安全舒适、经济实惠、节能环保、安全可靠、人性化设计等优点。通过网络化综合智能控制和管理，智能家居系统实现了以人为中心的全新家居生活体验，为用户提供更加智能化、个性化的居住体验。

在智能家居的应用中，保障信息和数据的安全传输一直是一个需要解决的棘手问题，这是不可避免的挑战。由于目前市场上没有专门针对家庭中信息传递过程的安全产品，因此必须要对信息、数据传输进行严格的管理。其中的机密性是确保信息和数据传输有效的关键[2]，必须采用有效的身份验证措施。因此，对智能家居系统中常用的一些认证方式进行分析和研究，都是为了确保网络的安全性和可靠性。现如今在智能家居认证系统中，其主控芯片的算法参数可以进行调整，但是算法种类无法更换，正是由于该缺点的存在导致其在加密问题的广泛适配性上有一定缺失。

为了应对上述问题，人们一直在追求一种完美的技术方案。目前比较成熟的方法是通过使用加密芯片来对数据进行保护和处理。针对现有技术的局限性，我们提出了一种基于自主加密芯片的智能家居控制系统，以弥补现有技术的不足。该系统采用自主加密芯片对家庭内部进行加密处理，实现对家庭中重要数据和信息的保护，并且能够通过网络将加密后的文件上传至服务器上。该系统提供了更加多样化的加密算法，从而显著提升了其安全性水平。

1 总体设计

本文所设计的智能家居控制系统[3-4]由加密模块、控制模块以及相关的连接模块构成。控制器模块用来接通控制器端口以接收来自控制器端口的指令和数据；接口模块用来接通智能家居的端口以发出控制指令；加密芯片部分包括主控芯片以及算法协处理单元，其中算法协处理单元采用FPGA芯片，它是算法处理的核

收稿日期：2023-04-28 **修稿日期：**2023-05-24

作者简介：*通信作者：杨金龙（1990—），男，河南鹤壁人，硕士，助教，研究方向为精密仪器测量方面、网络通信的研究，E-mail：870015801@qq.com；马静怡（1995—），女，河南郑州人，硕士，助教，主要研究方向为电力自动化

心，为主控芯片提供算法加密。

1.1 系统控制模块

系统控制加密模块的硬件框图如图1所示。

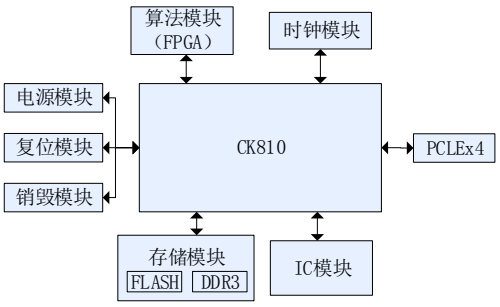


图1 系统控制加密模块的硬件框图

如图1所示，该控制模块以加密芯片CK810为控制处理核心，主控芯片通过PCIE接口将时钟模块、电源模块、复位模块、外部存储模块、IC模块和算法协处理单元连接在一起。多余的PCIE接口与控制单元、连接单元相连。

CK810在嵌入式系统开发中被广泛采用，其主要长处在于拥有卓越的主频表现、卓越的单位性能以及高效的功耗。该处理器具有很好的低功耗特性和良好的可移植性能，适合嵌入式操作系统移植。CK810芯片的移植开发采用了安卓系统，通过Java编程和底层程序开发，实现了在家居智能芯片的设计研究开发中更高效的应用。

本系统可以在单片机上运行多种算法，包括串行通信、并行通讯和并行处理等功能。FPGA芯片因其高度集成、可反复编程、高可靠性等优点，已成为当今硬件设计中最为重要的可选方案之一。算法模块采用FPGA芯片进行构建，为加密芯片提供了多样化的算法支持，解决了算法种类受限的问题。外部存储模块包括FLASH和DDR3，IC模块用于内部通过算法[5]生成的密钥，FLASH用来管理公钥和身份信息，而DDR3则存储有公钥与其使用者的身份匹配关系。

1.2 加密芯片的结构

加密芯片的结构框图如图2所示。

主控芯片由CPU核心、算法引擎接口、存储接口以及外面连接的数据端口构成。通过各种算法引擎连接到FPGA，实现多种算法处理，从而提高了系统的安全性。

算法引擎能够以硬件方式实现密码算法，主要通过控制电路和协议数据端口来实现；FLASH作为芯片启动引导程序、程序静态存储、密钥、证书和标志信息存储以及芯片参数等方面的重要工具，它能够确保程序的安全性和可靠性；作为芯片的主要服务接口，PCI-E接口提供了命令与密码服务的数据传输，而DDR3接口则连接了SDRAM芯片，同时还提供了对外PCI-E接口、USB2.0接口、DDR3接口、SPI接

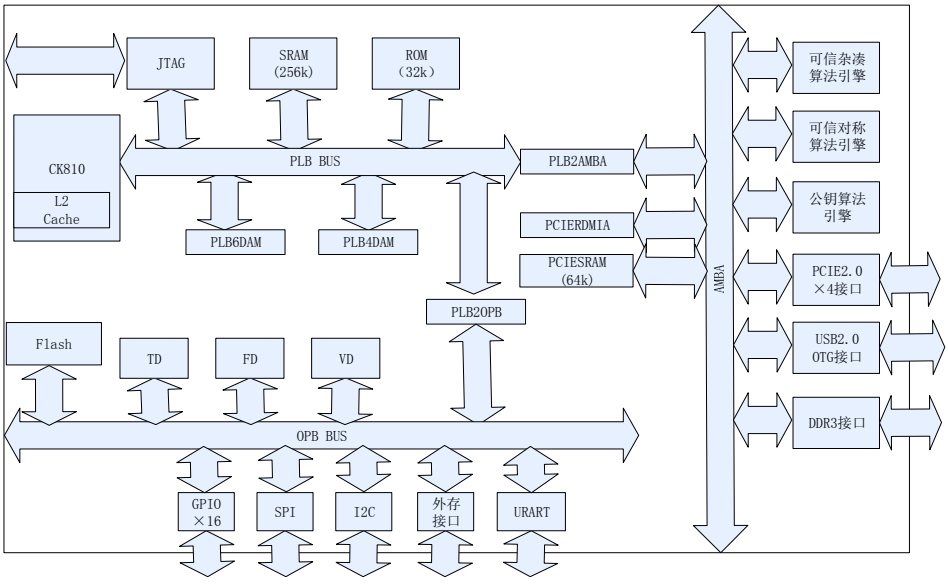


图2 加密芯片的结构框图

口、I2C接口、GPIO接口、UART接口和外部存储器接口；IC卡可通过I2C接口进行连接，而外部存储芯片则可通过外部存储器接口与之相连。

1.3　算法功能模块

控制单元[6]通过PCIE口与主控芯片算法模块进行传输，算法模块将信息反馈给控制单元，连接单元将用户加密信息传递给主控模块，并通过算法协处理单元进行加密信息处理，将结果反馈给连接单元，通过上述流程来实现控制终端与智能家居设备之间的绑定和身份认定过程。

其中加密处理过程是一种基于FPGA技术的硬件加密传输设备，它为公钥密码系统提供了在非安全网络上安全、准确地传输密钥的能力。网络加密卡具有良好的性能，可以很好地满足各种不同环境下对数据进行安全性和保密性要求。作为加密的重要组成部分，公钥密码算法采用TCP/IP协议作为一种全面的协议，针对网络加密卡的实际需求，对密钥交换协议进行了深入分析，并利用VHDL语言实现了AES算法，最终成功实现了公钥密码算法。同时针对不同类型的加密需要选择合适的加密算法和解密方法。类似的实现方式也包括非对称和杂凑等算法。

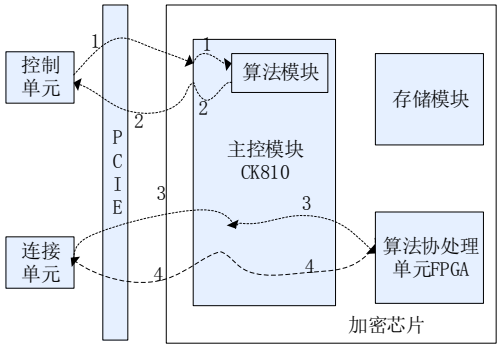


图3　控制模块的数据传输示意图

通过引入上述算法协处理单元，不仅丰富了加密芯片的算法种类，而且在面对各种不同的加密需求时，能够灵活地加载相适配的算法，从而显著提升其广泛适用性。

2　功能测试

2.1　硬件设计

本文采用的测试平台由上位机与测试平合硬件系统组成：上位机选用具有Windows操作系统的PC机，实现对测试过程的发起与控制；测试平合硬件系统由FPGA测试基板组成，为多芯片同步测试提供必要的工作环境。PC机与FPGA[7]测试基板之间通过USB线连接。

考虑到测试的效率与稳定性，本文选用FPGA-DE3开发板作为测试平合的FPGA测试基板[8]，它通过内部FPGA芯片的Nios II CPU按收与下发上位机命令，并在测试结束后将测试结果上传，完成测试平台系统上位机与下位机间的数据交互。测试中FPGA测试基板主要用到的DE3开发板自带器件包括LED品示、USB通信接口、DDR IT SDRAM插槽、HSTC扩展接口、电源、晶振、按键等。

2.2　软件设计

测试平台软件质量会直接影响批量测试板及FPGA测试基板性能的实现，本文设计的软件系统应具有良好的层次性、可复用性、可扩展性，同时还应该满足系统的标准化与开放化的要求。

本文采用层次化设计思想将测试平台的软件系统分为三层，分别为提供上位机端的应用层、FPGA测试基板端的控制层以及测试实现层。通过Nios II CPU将上位机与下位机、用户与被测芯片紧密联系起来，共同保证了测试平合运行的稳定性。系统软件层次划分如图4所示。

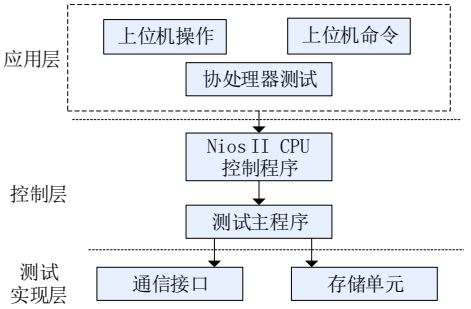


图4　软件层次划分

（1）应用层。上位机端的应用层是测试平合系统软件的最顶层，也是直接与用户接触的层次。用户通过相应测试命令实现对测试流程的发起；操作界面完成测试项的选择，重要信息及测试结果的显示；用户通过上位机直接控

制芯片协处理器的测试流程。

（2）控制层。FPGA 测试基板端的控制层是测试平台系统软件的中间层。它以 USB 接口读写驱动为基础，通过内部 Nios II CPU 完成对应用层命令的接收，实现应用层与测试实现层的数据信息交换，并对 芯片内部各模块的测试顺序进行控制。

（3）测试实现层。批量测试板端的测试实现层直接与芯片硬件接触，负责内部模块功能测的实现。它以芯片通信接口和存储单元读写驱动为基础，完成对芯片功能测试程序的执行，并编写模块测试程序。

2.3 协处理器功能测试

为了验证协处理器功能，本文以处理 AES[9] 算法功能为例，具体的实现流程如图 5 所示。

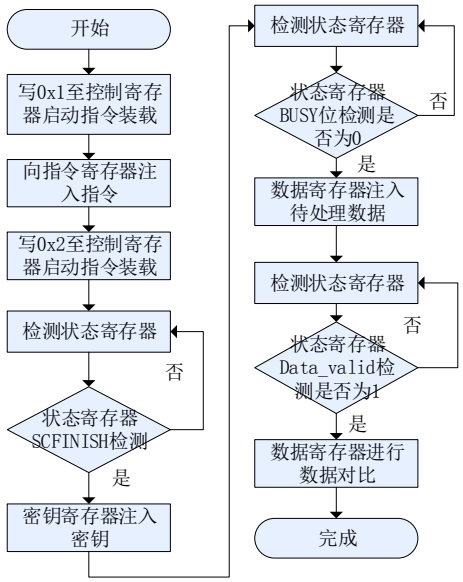


图 5 算法协处理单元流程

（1）算法注入。协处理器内部需要写入相应算法[10-12]方可进行对算法处理能力的测试，上位机在测试前的首要工作就是从 Coprocessor Test.date 中读出需要进行功能测试的算法并注入协处理器。

（2）指令注入。在算法写人协处理器后，上位机向控制寄存器注入控制命令并完成启动指令的装载准备；向指令寄存器注入指令，完成协处理器的配置、密钥选取等数据信息加密前必须进行的准备工作；向控制寄存器注入控制命令完成指令的注入操作。

（3）参数注入。在确定上位机已经将算法、指令注入完成后，根据算法类型的不同，依次向协处理器注入密钥、IV 等参数并通过检测相应的状态寄存器来判断是否注入成功。当检测状态位空闲时注入待测试的明文，这样就完成了上位机向协处理器的参数注入。

（4）读出结果并比对。上位机在接收到状态寄存器发出处理完成的信号后，从数据寄存器中将协处理器运算后得到的密文读出，与相应算法的标准密文进行比对，并将结果显示在上位机界面上。

3 结语

本文在查阅相关的国内外智能家居系统的基础上，针对我国传统主要控制器件的算法种类不可替代和无法针对不具有广适性的加密情况，设计了具有多样化加密算法、高稳定性的采用了自主保密芯片的智能家居控制系统，并通过搭建的测试平台进行了软硬件的功能测试。结果表明，本文采用的自主加密芯片的智能家居系统设计能够实现最理想的效果。

参考文献：

[1] 刘玉磊. 智能家居产业的标准化发展[J]. 电子元器件与信息技术,2022,6(10):113-116.

[2] 刘亚,苏皖,杨刚,等. 基于加密芯片的健康管理高安全系统的构建与应用[J]. 中国医疗设备,2022,37(1):82-85.

[3] GUOLI S,XINYI G,WENBO W,et al. A machine learning-based underwater noise classification method [J]. Applied Acoustics. 2021,184:108333.

[4] 李勤勤,刘乐钰. 基于STM32的智能家居控制系统设计研究[C]//中国城市科学研究会数字城市专业委员会轨道交通学组.智慧城市与轨道交通2022.北京:中国城市出版社,2022:123-127.

[5] 何安平,郭慧波,冯志华,等. 基于异步电路设计的RSA算法加密芯片[J]. 计算机工程与设计,2019,40(4):906-913.

[6] 黄晓咪. 基于ARM架构的安全加密专用SOC的研究与实现[D]. 桂林:桂林电子科技大学,2022.

[7] 王明. 面向加密芯片高安全性测试电路设计与实现[D]. 南京:南京邮电大学,2021.

[8] BO X,BABAR H,YIRU W,et al. Smart home control system using vlc and bluetooth enabled AC light bulb for 3D indoor localization with centimeter-level precision[J]. Sensors,2022,22(21):8181.
[9] 张馨方,周江华.基于轻量型AES加密算法的浮空器平台数据传输方案[J].计算机测量与控制,2023,31(6):183-190.
[10] 刘宇峰.AES算法的优化设计及FPGA实现[D].石家庄:河北科技大学,2019.
[11] 陈晓宇.高级加密标准AES算法的分析与优化改进[D].南充:西华师范大学,2020.
[12] 许韫韬,吕志刚,黄义国,等.基于STM32单片机的AES算法优化与实现[J].自动化仪表,2020,41(7):56-60.

Design of intelligent home control system based on autonomous encryption chip

Yang Jinlong*, Ma Jingyi

(School of Big Data and Artificial Intelligence, Zhengzhou University of Science and Technology, Zhengzhou 450064, China)

Abstract: A smart home control system based on autonomous encryption chips is proposed, which includes a control unit, an encryption chip, and a connection unit. The control unit is used to connect the control terminal to receive instructions and data from the control terminal, the connection unit is used to connect the smart home device terminal to output control instructions, and the encryption chip is used to encrypt the instructions and data received by the control unit from the control terminal and send them to the connection unit, The encryption chip includes a main control chip and an algorithm co processing unit connected to it. The algorithm co processing unit adopts an FPGA chip, and the algorithm co processing unit is used to provide encryption algorithms for the main control chip. The intelligent home control system based on autonomous encryption chips can provide more diverse encryption algorithms, improving security.

Keywords: autonomous encryption; FPGA; control module

(上接第112页)

Research and development of work-study management system based on WeChat mini-program

Tan Huorong, Chen Haiyu*

(Information Center of Zhaoqing Medical College, Zhaoqing 526020, China)

Abstract: Work-study programs are an important part of university life, providing students with financial support and practical experience. In order to improve the efficiency and convenience of work-study program management, a work-study management system based on WeChat mini-program is designed. This system provides functions such as job posting, application management, attendance tracking, and wage management, which meet the needs of students, employers, and administrators, and provide a one-stop management platform for them. By implementing a user-friendly interface, responsive design, and powerful backend services, the system aims to meet the needs of various users and further promote the development of work-study programs in higher education.

Keywords: WeChat mini-program; work-study program management; system design; system implementation; higher education

文章编号: 1007-1423(2023)16-0118-03 DOI: 10.3969/j.issn.1007-1423.2023.16.022

云机房系统数据安全管理机制研究与实现

李宇锋*

（东华理工大学软件学院，南昌 330013）

摘要：提出了一种云机房管理方法及相应的验证方法。首先阐述了现有普通机房管理方法弊端，随后给出了云机房管理机制及实现，主要包括网络设备系统安全、教师管理机系统和数据安全、服务器系统数据安全以及建立健全云机房日常教学管理制度。结果表明此方法能够确保系统数据安全性，提高云机房管理水平。

关键词：桌面云；系统数据安全

0 引言

桌面云，是基于融合架构的新型桌面模式，通过深度整合服务器虚拟化、桌面虚拟化及存储虚拟化，只需桌面云一体机和云终端两种设备，即可实现云平台的快速交付，为用户提供操作体验及软硬件兼容性媲美PC，更安全、更高效的云桌面。桌面云将原来绑定在PC上的桌面、应用和数据全部集中到数据中心，然后通过SRAP桌面交付协议，将操作系统界面以图像的方式传送给前端的接入设备，包括云终端、笔记本、普通PC、智能终端等，只要网络是可达的，用户就可以通过各种类型的终端去访问位于服务器上的个人桌面，让数据保护更安全，桌面管理更高效，用户访问更灵活[1]。

使用普通物理电脑，核心重要的数据都存储在电脑的硬盘里面，也就是说在本地。随着现在网络的快速发展，这也带来了一些问题，人们的数据很容易通过网络传播出去，也受到网络黑客的攻击，一些学校还要避免内部资料的泄密，使用传统PC这些数据的安全性和可靠性达不到保障，在突发情况下如果数据没有备份，就存在数据丢失的风险。所以，如何解决这些问题是网络管理员和广大学校用户所关心的。每个学校都有一些部门的业务性质超级重要，比如研究院对数据安全性的要求性很高，一些重要的数据关系到一个学校的发展，一些科研成果不能被不法分子窃取，并且为了机密，有些只能在一个终端上面进行操作，这样对学习实验的环境就有了限制，达不到学习实验的灵活性，这都和桌面的安全性相关，所以，桌面虚拟化可以实现这些桌面的安全性保障，可以实现任意地点、任意时间在保障数据安全的情况下的学习实验。

1 云机房系统数据安全管理机制

云系统数据安全研究技术，包括登录用户、权限的控制，密码管理、重要数据备份，网络访问策略设计与编写，教学软件在虚拟系统中的正确配置。

2 云机房系统数据安全管理

2.1 云机房网络设备的安全措施:本机防攻击[3]

(1)配置畸形报文攻击防范

```
<HUAWEI> system-view
[HUAWEI] sysname DeviceA
```

收稿日期：2023-04-12 **修稿日期：**2023-04-24

基金项目：东华理工大学实验技术开发项目(DHSY-202256)

作者简介：*通信作者:李宇锋(1977—),男,江西吉安人,本科,助理实验师,研究方向为计算机应用,E-mail:14296466@qq.com

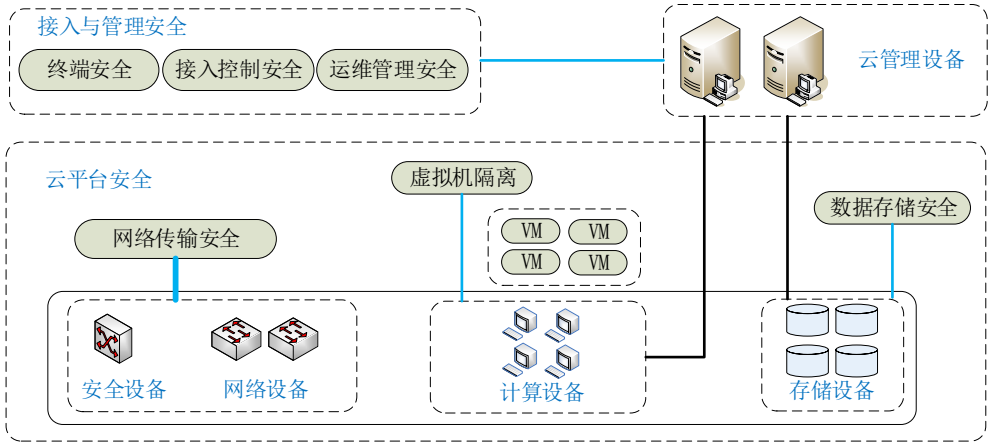


图 1　云机房系统构架[2]

[DeviceA] anti-attack abnormal enable

（2）配置分片报文攻击防范，分片报文的限制速率为15000 bit/s

[DeviceA] anti-attack fragment enable

[DeviceA] anti-attack fragment car cir 15000

（3）配置泛洪攻击防范

配置TCP SYN泛洪攻击防范，TCP SYN报文的限制速率为15000 bit/s

[DeviceA] anti-attack tcp-syn enable

[DeviceA] anti-attack tcp-syn car cir 15000

配置UDP泛洪攻击防范，对特定端口发送的UDP报文直接丢弃

[DeviceA] anti-attack udp-flood enable

配置ICMP泛洪攻击防范，ICMP泛洪报文的限制速率为15000 bit/s

[DeviceA] anti-attack icmp-flood enable

[DeviceA] anti-attack icmp-flood car cir 15000

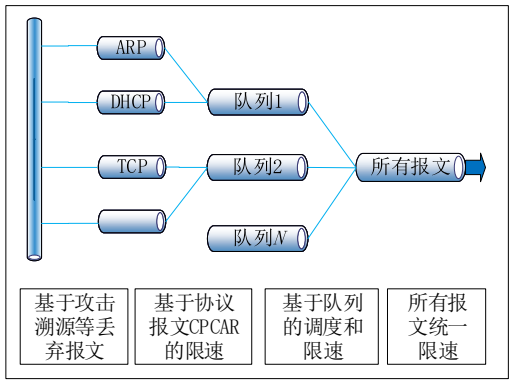


图 2　三层交换机防攻击示意图

2.2　教师管理机系统和数据安全

教师机控制权限高于学生终端机，利用管理软件和操作系统进行策略设计、权限分配，使用极域多媒体教学软件[4]进行数据的安全传输，指导教师正确使用教师管理机和提高安全防范意识等措施来保障教师机的系统安全。

（1）独立系统设计：教师机拥有自己独立的操作系统，与云机房系统互不相关，确保两者的数据互不干扰。

（2）在教师机上配置了还原卡，利用还原卡的特点，教师机的系统能够根据教学上的不同需求设置成不同的还原模式：立即还原、每天还原、每月还原，开放模式[5]，提高了系统的安全性。

（3）对系统用户进行了策略配置，根据不同老师的需求进行了相应的权限设置。

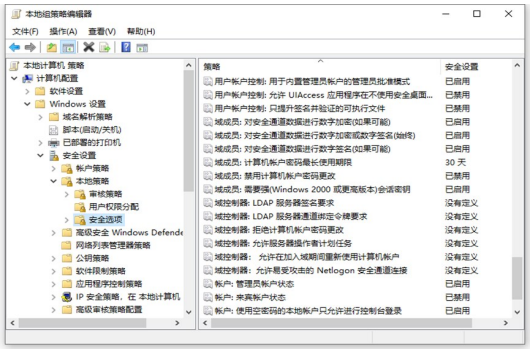


图 3　管理机系统本地用户与域用户配置界面

管理机(教师端)，管理机系统独立于终端虚拟机系统，教学管理软件与云系统和虚拟系统分离，教师机系统的安全策略设计，教学数据独立存储、管理。

2.3 服务器系统数据安全

通过技术手段切断网络访问系统底层数据，在关键网络设备内进行安全策略设置、访问限制，服务器底层系统访问设置，利用虚拟化技术阻隔云系统数据平台与教学数据平台的互访。

配置域服务器对终端登录用户进行权限设置和访问控制云系统数据安全研究技术，登录用户、权限的控制，密码管理、重要数据备份，网络访问策略设计与编写，教学软件在虚拟系统中的正确配置(系统数据安全机制)。

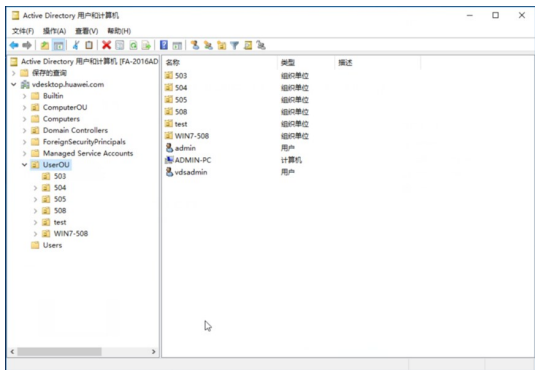


图 4 域服务器用户管理配置

2.4 建立健全云机房日常教学管理制度

包括制定实验员日常管理制度，教师、学生实验室规章制度，云机房应急预案制度建设。

3 结语

在日常实验教学中，学生对于云系统的访问权限将受到限制，只能访问其必要数据，但在教学过程所需要的数据不受影响，网络访问也将进行严格管控，防止非法数据的侵入，能够提高云系统数据的安全级别；教师在教学过程中管理学生终端不受影响，正常的教学数据可安全有效地进行双向传输，但对云系统的访问权限将受到限制，在满足教学需要的前提下确保云系统数据的安全；限制云系统接入互联网，将不会受到来自互联网的攻击，但教师机可以连接互联网，保证教学需求；使用虚拟机技术手段解决多系统，多个实验环境场所的自由切换，实现不同实验教学任务。通过建立管理制度来约束管理人员、教师和学生在日常教学中的行为规范，最大程度减少人为因素所导致的安全隐患。

参考文献：

[1] 桌面云|虚拟桌面基础架构(VDI)私有云部署[EB/OL]. https://www.northsoar.com/show-49.html.
[2] FusionAccess 桌面云解决方案[Z].
[3] 技术支持[EB/OL]. 华为企业业务网站. https://support.huawei.com/enterprise/zh/index.html.
[4] 极域电子教室软件快速上手指南[M/OL]. https://wenku. baidu. com/view/0faf7e785f0e7cd185253628.html.
[5] 噢易 OSSV8 系统自定义安装指导手册[M/OL]. https://www. os-easy. com/uploads/soft/230402/1-230402231Q7.pdf.

Design and implementation of community sales platform based on Web

Li Yufeng*

(Software College of East China University of Technology, Nanchang 330013, China)

Abstract: The project proposes a cloud room management method and corresponding verification method. Firstly, the disadvantages of the existing common computer room management methods are described, and then the management mechanism and implementation of the cloud room are given, including network equipment system security, teacher management computer system and data security, server system data security and the establishment and improvement of the cloud room daily teaching management system. The results show that this method can ensure the security of system data and improve the management level of cloud room.

Keywords: desktop; system data security