现代计算机

XIANDAI JISUANJI

第29卷第8期（总第776期）

半月刊（1984年创刊）

2023年4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

现代计算机 第29卷 第8期（总第776期） 2023年4月

2023年4月 第29卷

第8期 （总第776期）

ISSN 1007-1423
CN 44-1415/TP

现代计算机

MODERN COMPUTER

中山大学出版社 主办

现代计算机
XIANDAI JISUANJI
第29卷 第8期（总第776期）2023年4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*20*2023-04

Modern Computer
(Vol. 29, No. 8; Apr. 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）08-0001-011 DOI: 10.3969/j.issn.1007-1423.2023.08.001

基于改进YOLOv5算法和边缘设备的电动车违规载人检测

孙福临[1]，李振轩[1]，梁允泉[2]，董苗苗[2]，葛广英[3*]

（1. 山东省光通信科学与技术重点实验室，聊城 252059；
2. 聊城大学物理科学与信息工程学院，聊城 252059；3. 聊城大学计算机学院，聊城 252059）

摘要： 严禁电动车违规载人是新交通法规中的重要内容，针对目前缺乏有效的检测电动车违规载人算法的现状，设计了一种基于改进的YOLOv5目标检测算法与边缘设备相结合的电动车违规载人检测系统。首先构建电动车行驶数据集；其次以YOLOv5网络模型为基础，引入轻量化网络Mobilenetv3、ECA-Net注意力机制、Slim-Neck结构和SPPFCSPC空间金字塔池化结构，提升针对电动车违规载人的检测精度，并且与原算法做消融实验；最后将改进后的算法部署在边缘设备Jetson Nano上进行实时推理。通过分析实验数据，改进后算法的参数量下降为原YOLOv5n的18%，在Jetson Nano上其推理速度提升了62%，最快推理速度可以达到17 FPS。改进后的算法在Jetson Nano上可以在提升检测精度的同时大幅提高推理速度，满足在不同场景下进行边缘部署的需求。

关键词： YOLOv5；边缘设备；Jetson nano；电动车违规载人检测；轻量化网络

0 引言

这些年来，随着社会经济的不断发展，汽车的保有量不断攀升，燃油价格走高，机动车出行越发不便，越来越多的人选择了电动车这一出行方式。正是因为许多人在驾驶电动车出行时缺乏安全意识，有关部门于2022年出台了新的交通法规，对电动车驾驶人员的行为做出了严格限制，其中明确规定电动车严禁违规载人。目前来说，仅仅依靠交警很难完成对于电动车违规载人行为的查处，并且缺乏相关的能辅助交警检测的电子系统或设备。因此，结合计算机视觉以及深度学习技术来构建一套可以判别电动车违规载人的系统很有必要。

21世纪，伴随着深度学习技术突飞猛进的发展，立足于深度学习的目标检测算法也迎来了长足发展，主要包含了以YOLO、SSD[1]、Efficientnet[2]系列算法为代表的单阶段检测方法和以R-CNN[3]、Faster R-CNN[4]、Mask R-CNN[5]等算法为代表的两阶段检测方法这两大类目标检测算法。其中两阶段检测的网络依赖于算法提前产生的目标候选框，而单阶段检测方法则不需预先生成目标候选框。

自2015年YOLOv1[6]被提出以来，YOLO系列算法得到了广泛发展，YOLOv1算法利用回归的思想，使用一阶网络直接完成了分类与位置定位两个任务。YOLOv2[7]在此基础上加入了Batch Normalization，High Resolution Classifier，Anchor Boxes等方法，成功地提升了网络收敛速度和网络定位的精准度。2018年YOLOv1的作者提出了YOLOv3[8]，它是前作的改进，最大的改进特点包括使用了残差模型Darknet-53来提升主干网络的深度，以及为了实现多尺度检测采

收稿日期： 2022-12-07 **修稿日期：** 2023-01-26

基金项目： 中央引导地方科技发展专项基金（YDZX 2017370000283）

作者简介： 孙福临（1999—），男，山东泰安人，硕士生，研究方向为图像处理和模式识别和深度学习；李振轩（1998—），男，山东济宁人，硕士生，研究方向为图像处理和模式识别和深度学习；梁允泉（1997—），男，山东日照人，硕士生，研究方向为图像处理和模式识别和深度学习；董苗苗（1995—），女，山东聊城人，硕士生，研究方向为图像处理和模式识别和深度学习；*通信作者：葛广英（1964—），男，山东聊城人，博士，教授，硕士生导师，从事领域为计算机视觉、模式识别和人工智能等，E-mail：ggysd@126.com

用了FPN架构。YOLOv4[9]在原来的YOLO目标检测架构的基础上，采用了很多优化策略，在数据处理、主干网络、网络训练、激活函数、损失函数等方面都有不同程度的优化，有效增强了网络的学习能力，降低了成本。YOLOv5则是于2020年通过Ultralytics提出，该网络一经问世便广受好评。与YOLOv4相比，YOLOv4不仅目标检测的准确率不相上下，而且其强大的推理能力、快速的训练过程，以及轻量级的模型大小使其非常适合于前端部署和快速目标检测，这项优点使得YOLOv5广受青睐。

综上所述，正是由于YOLOv5在推理速度和准确率方面都十分优秀，而针对于电动车违规载人检测的轻量级算法部署在模型推理速度和准确率方面均有较高的需求，因此本文在YOLOv5的基础上，进一步针对其网络结构进行不断优化，通过更换轻量级主干网络Mobilenetv3、增加高效的注意力机制ECA-Net、更换轻量级的颈部网络Slim-Neck、空间金字塔池化结构SPPFCSPC等方法以达到尽可能减少参数量和计算成本的目的，进而实现提高FPS和检测精度的目标，最终通过筛选模型的准确率和推理速度等指标来确定和验证所提出模型的优越性，并将其部署在前端嵌入式设备Jetson nano上进行实时检测。

1 YOLOv5网络模型概述

自从2020年Ultralytics发布YOLOv5以来，该算法经过了多次更迭和改进，目前最新的YOLOv5-6.2共有五种不同参数量的版本(N,S,M,L,X)，他们的不同之处在于控制其网络的宽度和深度的参数不同，如果想要完成边缘计算的模型部署，首选模型最小且参数量最少的YOLOv5n部署实验，其结构如图1所示。

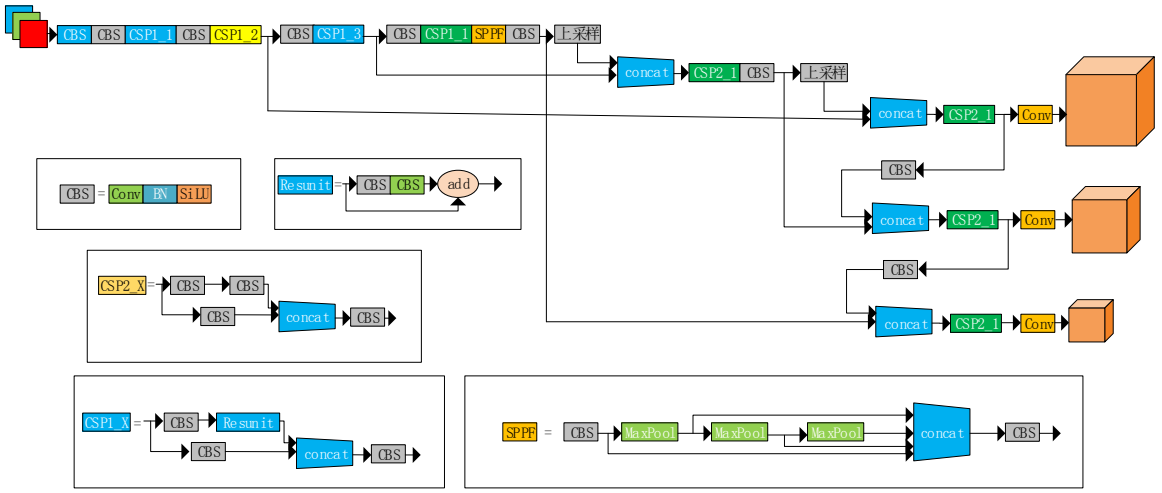


图1 YOLOv5结构

YOLOv5n的整个结构由四部分组成：①由各种卷积网络组成的主干网络Backnbone；②用于预测目标框和目标类别的检测头Head；③介于Backbone和Head之间的Neck结构；④用于输出检测结果的预测层。

YOLOv5n算法的工作原理：

(1)输入端采用了自适应图片缩放、自适应锚框计算、Mosaic数据增强对数据集进行操作以提高目标检测的精确度。

(2)主干网络删除了Focus结构来压缩网络提高速度，使用CSP1_X残差结构对图像特征进行提取，使用相较于SPP[10]结构而言速度更快的SPPF结构进行特征提取，提高感受野。

(3)Neck网络使用了FPN[11]+PAN[12]结构并结合CSP2_X结构来融合图像特征，得到预测的特征图。

(4)输出端使用了DIOU_Loss损失函数，在IOU_Loss和GIOU_Loss的基础上，改进了边界

框中心点距离的信息。

2　电动车违规载人模型改进

2.1　YOLOv5n主干网络改进

随着深度学习的不断发展，深度学习模型的网络深度持续增加，虽然很大程度上提升了精度，却也使得模型对计算资源的要求越来越高，运行速度变慢[13]。因此本文想要寻求一种以YOLOv5为基础的轻量化网络模型，作为适用于边缘设备的解决方案。

在YOLOv5n的主干网络中，主要由Conv模块和使用了CSP架构(cross stage partial network)的C3模块组成，它们的主要工作就是提取输入图片的有效特征信息。在实验中发现，YOLOv5的作者设计的通过调节网络宽度系数和深度系数的方式来减少网络参数实现模型轻量化的方法，会导致训练出的模型的目标检测准确度降低，同时误检率和漏检率升高。因此，本文寻求一种可以在减少网络参数量的同时提高特征提取能力的方法来实现轻量化，故此引入Mobilenetv3网络结构对YOLOv5n的主干网络所使用的Conv模块和C3模块进行替换，达到同时缩减模型参数量和提升检测准确率的目的。

2.1.1　Mobilenetv3工作原理

Mobilenetv1[14]是由Google团队于2017年提出的专注于移动端或者嵌入式设备的轻量级CNN网络。使用深度卷积(depthwise convolution，DW)在准确率小幅降低的前提下大大减少模型参数与运算量。紧接着，2018年Google团队又提出了Mobilenetv2[15]，基于前作增加了两个新的改进点，使用倒残差结构(inverted residuals)来增加网络深度，减少DW卷积所造成的特征信息损失；在结构的最后一层采用线性层(linear bottlenecks)取代了ReLU激活函数，降低对低维特征的信息损失。

2019年提出的Mobilenetv3[16]网络在结合前作的基础上重新设计了耗时层，加快了推理速度，引入了轻量级注意力机制SE-Net(squeeze-and-excitation networks)并且重新设计了激活函数(hard-sigmoid)。

深度可分离卷积(depthwise separable convolution，DSC)包含了深度卷积(depthwise convolution，DW)以及逐点卷积(pointwise convolution，PW)两个部分，图2展示了其卷积过程。相较于传统卷积而言，DSC卷积的参数量和计算量都大幅减少，对比如下：

$$\frac{W_1}{W_2}=\frac{D_K\times D_K\times M\times D_F\times D_F+M\times N\times D_F\times D_F}{D_K\times D_K\times M\times N\times D_F\times D_F}=\frac{1}{N}+\frac{1}{D_K^2} \tag{1}$$

其中：W_1代表了DSC卷积的计算量；W_2代表了传统卷积的计算量。Mobilenetv3网络中卷积核尺寸多数为5 × 5。所以DSC卷积的计算成本仅为传统卷积的1/25。

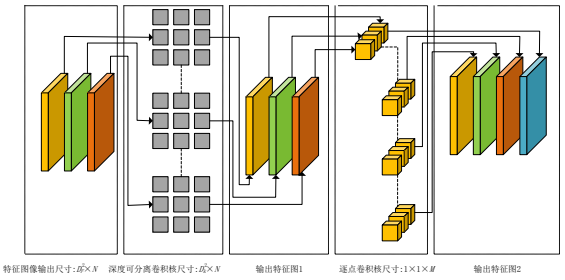
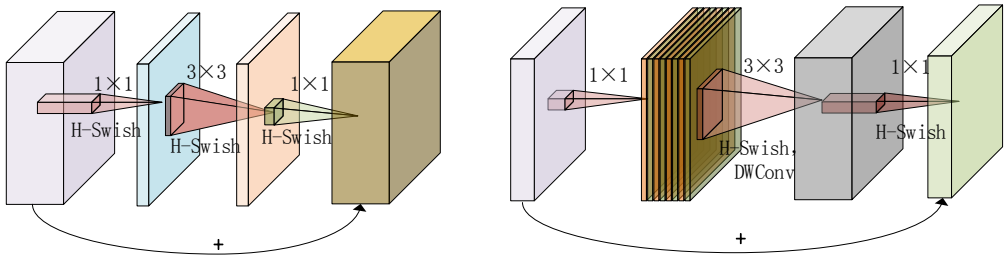


图2　深度卷积示意图

如图3所示，在Mobilenetv3的倒残差结构中先使用1*1卷积扩张通道数，在高维度中使用DW卷积，最后使用1*1卷积缩减通道数。这样做的目的在于通过第一次的升维操作将Tensor维度拉高，低纬度Tensor计算速度快，但是无法提取足够多的特征信息，因此将通道数目扩张，而针对高维Tensor使用传统卷积计算量太大，因此使用DW卷积既可以有效完成特征提取又可以减少运算量，随后再使用1*1卷积来压缩数据重新让网络变小，实现轻量化计算。正因如此，Mobilenetv3可以用于构建轻量化网络[17]，从而适配于边缘计算。

2.1.2　Mobilenetv3激活函数改进

之前在Mobilenetv2都是使用ReLU6激活函数。现在比较常用的是swish激活函数，即swish = $x \times \sigma(x)$。使用swish激活函数确实能提高网络的准确率，但是它存在一些问题。首先就是其计算和求导时间复杂，仅仅是进行$\sigma(x)$计算和求导就比较困难了。其次就是对量化过



（a）残差结构　　（b）倒残差结构

图 3　残差结构与倒残差结构示意图

程非常不友好，特别是对于移动端的设备，为了加速一般都会进行量化操作。为此，作者提出了一个叫做*h*-swish的激活函数。

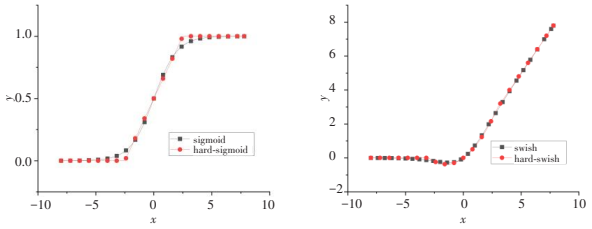
h-swish激活函数与*h*-sigmoid有些相似之处：

$$h\text{-sigmoid} = \frac{\mathrm{ReLU}(x+3)}{6} \tag{2}$$

$$h\text{-swish} = x \times h\text{-sigmoid} \tag{3}$$

$$h\text{-swish} = x\frac{\mathrm{ReLU}(x+3)}{6} \tag{4}$$

从图4可以看到swish激活函数的曲线和*h*-swish激活函数的曲线非常相似。Mobilenetv3的作者在原论文中提到，经过将swish激活函数替换为*h*-swish，sigmoid激活函数替换为*h*-sigmoid激活函数，对网络的推理速度是有帮助的，并且对量化过程也很友好。



（a）Sigmoid vs *h*-sigmoid　　（b）Swish vs *h*-swish

图 4　激活函数对比

2.2　针对于Mobilenetv3注意力机制的改进

常用的注意力机制一般分为通道注意力机制和空间注意力机制两种，在Mobilenetv3中添加了轻量级注意力机制SE-Net(squeeze-and-excitation networks)[18]，属于通道注意力机制。如图5所示，在输入SE注意力机制之前，特征图的每个通道的重要程度都是一样的，通过SE-Net之后，不同颜色代表不同的权重，使每个特征通道的重要性变得不一样，使神经网络重点关注某些权重值大的通道。

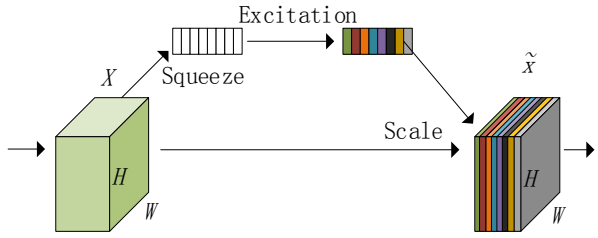


图 5　SE注意力机制

SE注意力机制的实现步骤如下：

（1）Squeeze：通过全局平均池化，将每个通道的二维特征(*H***W*)压缩为1个实数，将特征图从[*h*,*w*,*c*]转变为[1,1,*c*]。

（2）excitation：给每个特征通道生成一个权重值，论文中通过两个全连接层构建通道间的相关性，输出的权重值数目和输入特征图的通道数相同。特征图张量保持不变。

（3）Scale：将前面得到的归一化权重加权到每个通道的特征上。Mobilenetv3论文中使用的是乘法，每个通道分别乘以权重系数。

SE-Net的核心思想是通过全连接网络根据loss损失来自动学习特征权重，而不是直接根据特征通道的数值分配来判断，使有效的特征通道的权重增大。然而SE-Net注意力机制不可避免地增加了一些参数和计算量，SE-Net中的降维会给通道注意力机制带来副作用，并且捕获所有通道之间的依存关系效率不高，而且是不必要的。因此，本文使用SE-Net的改进版ECA-Net[19]注意力机制对其进行替换。

ECA注意力机制模块直接在全局平均池化层之后使用1*1卷积层，去除了全连接层。该模

块避免了维度缩减，并有效捕获了跨通道交互。并且 ECA-Net 只涉及少数参数就能达到很好的效果[20]。

ECA-Net 通过一维卷积 layers.Conv1D 来完成跨通道间的信息交互，卷积核的大小通过一个函数来自适应变化，使得通道数较大的层可以更多地进行跨通道交互。

自适应函数为

$$k = \left| \frac{\log_2(c)}{\gamma} + \frac{b}{\gamma} \right| \tag{5}$$

其中：$\gamma = 2, b = 1$。

ECA-Net 注意力机制如图 6 所示。

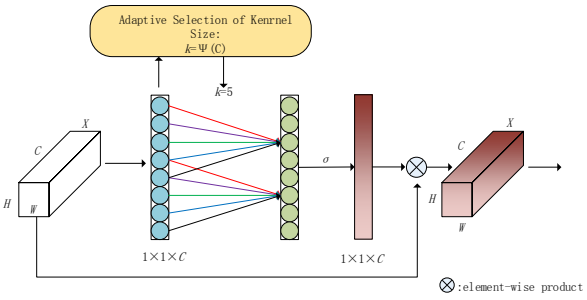


图 6　ECA 注意力机制

ECA 注意力机制的实现过程如下：

（1）将输入特征图经过全局平均池化，特征图从[h,w,c]的矩阵变成[1,1,c]的向量。

（2）根据特征图的通道数计算得到自适应的一维卷积核大小 kernel_size。

（3）将 kernel_size 用于一维卷积中，得到对于特征图的每个通道的权重。

（4）将归一化权重和原输入特征图逐通道相乘，生成加权后的特征图。

2.3　兼顾精准度与高效的 GSConv 和颈部网络改进

2022 年 6 月，重庆交通大学和不列颠哥伦比亚大学共同提出了新的卷积模块 GSConv[21]，阐述了目前的卷积机制的优劣，PW 卷积的精度高但是计算成本高，DSC 卷积可以有效缓解高计算成本[22]，但是其缺点也十分明显，其输入图像的通道信息在计算过程中是分离的，这会导致精度出现下降。在此基础上提出了 GSConv 结构，通过 GSConv 引入了 Slim-Neck 方法，在减少计算成本的同时维持精度，更好地维护了模型准确性与速度的平衡。其结构如图 7 所示。首先使用卷积核尺寸为 1*1 的 PW 卷积运算压缩通道数，得到特征图f_1，再经过 DW 卷积得到特征图f_2，将二者拼接得到f_3，而后通过 shuffle 操作将 PW 卷积生成的信息渗透到 DW 卷积生成的信息当中，得到最终的特征图f_4，这样做既能减少计算成本又可以最大限度降低 DW 卷积对模型精度造成的影响，使最终的卷积结果接近于 PW 卷积。

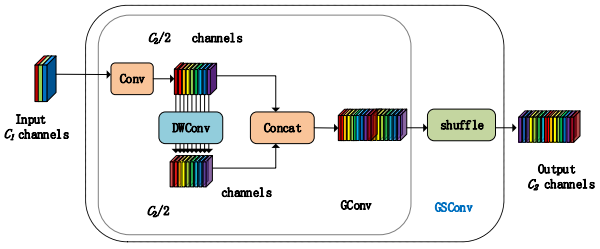


图 7　GSConv 结构

为了加快模型的运算速度，特征图会在 Backbone 中经历如下的过程：空间信息会逐步向通道传输，特征图的宽度和高度被不断压缩而通道数逐渐增多，这会导致语义信息的丢失。PW 卷积最大程度地留存了通道之间的隐藏链接，而 DSC 卷积则完全切断了这些隐形链接，由于 GSConv 结合了 PW 卷积和 DSC 卷积的特点，所以尽可能多地保留了这种隐形链接，但是如果在 Backbone 和 Neck 中都是用该结构的话，会使网络模型层不断加深，这会大大增加推理计算所耗费的时间，而特征图在传输到 Neck 时，已经变得较为细长，不再需要进行变换，冗余信息少且不需要压缩，因此在 Neck 部分使用 GSConv 效果最好。值得一提的是，原论文中认为在使用 Slim-Neck 的基础上使用注意力机制和空间金字塔池化结构的效果会更好，因此本文也针对这两处做了改进[23]。

使用轻量级卷积网络 GSConv 代替 PW 卷积构建 GSbottleneck，其结构如图 8(a)所示。相较于原先的 bottleneckCSP，其运算量只有 60%～70%，两者的学习能力旗鼓相当。在此基础上，我们使用一次性聚合方法来构建跨级部分网络（GSCSP）模块 VoV-GSCSP，结构见图 8(b)。VoV-GSCSP 在保持了精准度的基础上，降低了

网络结构的复杂性和计算成本。与 YOLOv5 原先的 C3 模块相比，其 FLOPS 降低了大约 15.7%。

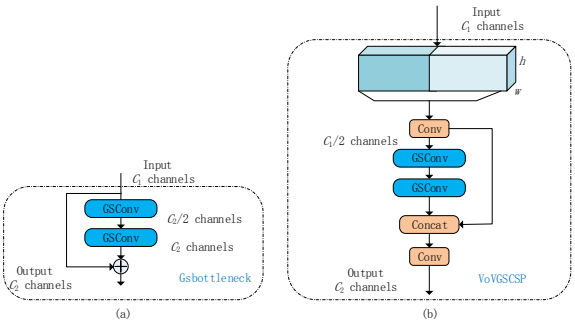


图 8　GSbottleneck 结构和 VoV-GSCSP 结构

本文在此基础上使用 GSConv 替换原 Conv 结构，使用 VoV-GSCSP 结构替换 C3 结构，构建 Slim-Neck 结构，如图 9 所示。

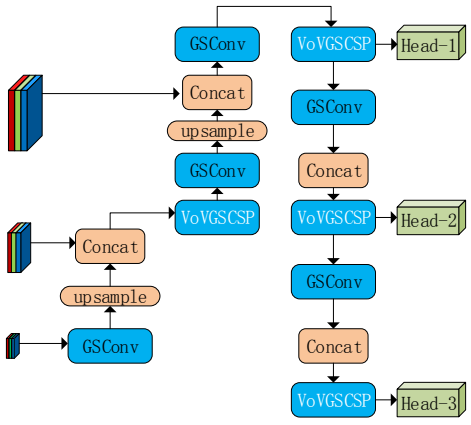


图 9　Slim-Neck 结构

2.4　空间金字塔池化结构(SPPFCSPC)

空间金字塔结构(SPP)最早由何凯明提出，旨在解决输入图像的尺寸问题，该结构包含三个显著优点，一是可以忽略输入尺寸并产生固定长度的输出；二是使用多尺度空间容器进行 Maxpool；三是可以在不同尺度上提取特征图的特征信息，有效提高了识别准确率。在 YOLOv5-6.0 中其作者提出了 SPPF 结构，相较于 SPP 模块而言，SPPF 模块使用了多个小尺寸的池化核，通过级联的方式来替代大尺寸的池化核，经过实验证明，该方法可以在保留原有的融合不同尺寸的特征图的功能下，加快推理速度，可以节省大约 51.6% 的推理时间。

YOLOv7 在 SPP 的基础上提出了一种名为 SPPCSPC 的金字塔池化结构，其结构如图 10(a) 所示。该结构与原 SPP 结构相比，输入端要先通过一个卷积核尺寸为 1*1 的卷积压缩通道数得到特征图 f_1，然后再通过两个卷积核尺寸为 3*3 和 1*1 的卷积得到特征图 f_2，然后通过 SPP 结构，得到特征图 f_3，再将 f_3 通过两个卷积核尺寸为 3*3 和 1*1 的卷积得到 f_4，与输入端经过一个 1*1 卷积后得到的 f_5 进行 Concat 操作，得到的结构再通过一个 1*1 卷积得出最终结果。

本文所使用的 SPPFCSPC[24] 结构由博主迪菲赫尔曼于他自己的博客提出，其结构如图 10(b) 所示。该结构立足于 SPPF 对 SPP 结构的改进方法，将 YOLOv7 所提出之 SPPCSPC 结构中最大值池化的部分修改为类似于 SPPF 的多个小尺寸

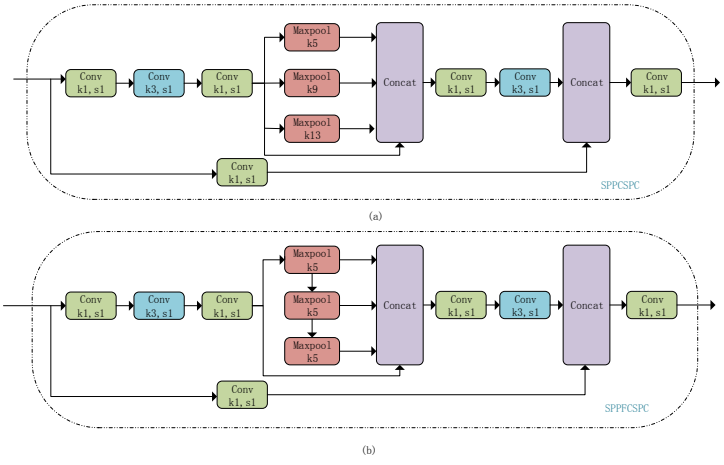


图 10　SPPCSPC 和 SPPFCSPC 结构

卷积核级联的方式，本文经过实验可以证实，在参数量没有改变的情况下推理时间缩减为SPPCSPC的72.7%。

2.5　改进的YOLOv5目标检测算法

综上所述，根据以上的改进策略完成网络模型的搭建。如图11所示，输入的电动车违规载人检测图像首先经过Backbone的Mobilenetv3-ECA模块，通过其中的DWconv卷积完成特征提取，然后经过ECA-Net注意力机制完成通道维度的信息融合，特征信息被送入空间金字塔池化结构SPPFCSPC，在不同尺度进行特征提取，而后送入颈部网络，通过轻量化的Slim-Neck结构，不同大小的特征层之间会再进行特征的进一步处理，最终将特征信息送入头部网络进行目标的检测。

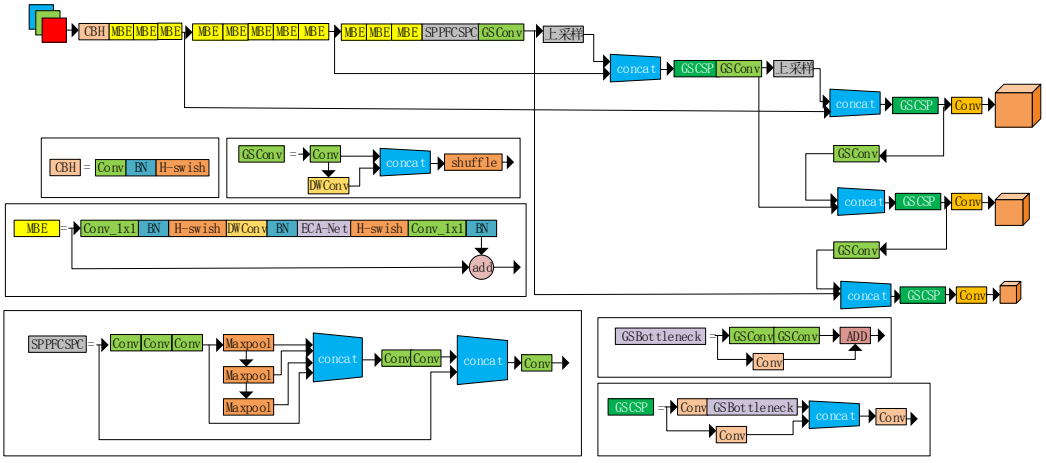


图11　YOLOv5-Mobilenetv3-ECA结构

3　实验方案

3.1　实验数据来源

因为目前尚未出现统一且公开的电动车违规载人的数据集，因此本文所使用的实验数据均通过网络爬虫和使用摄像头在不同道路对行驶的电动车进行拍摄获取，所获取的数据包括了不同的角度、不同的明暗色调、不同的道路场景和车辆稠密程度。总共有图片12951张，其中涵盖了电动车驾驶的各种尺寸的目标，使用LabelImg软件，将检测对象分成两类：overload、safe。根据分类对待检测对象进行标注。对于模糊的物体和远离摄像头的物体则不进行标注。

将数据集打乱后随机进行挑选和分类，最终选取了11291张图片作为训练集，验证集和测试集各830张图片，将数据集送入网络进行训练。

3.2　实验环境

本次实验的实验平台基于Windows10操作系统搭建，Intel(R)Xeon(R)Silver 4210R CPU@2.40GHz型号的CPU共有两个。GPU显存为8GB的Quadro RTX 4000系列，使用PyTorch深度学习框架，编程语言为Python3.8，网络模型使用YOLOv5n，将改进之后训练完成的网络模型结合摄像头和边缘设备组成电动车违规载人检测系统，系统界面如图12所示。

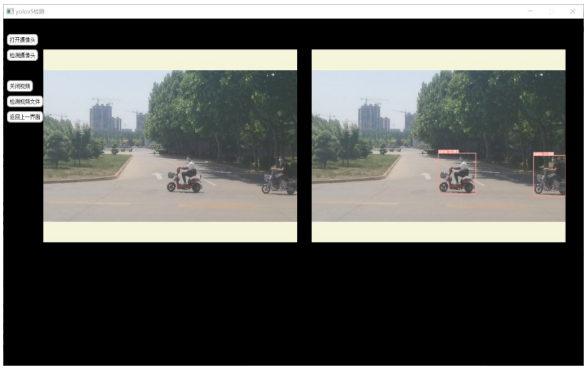


图12　系统展示界面

3.3 实验结果及分析

本文选用精确率(precision,P)、召回率(recall,R)、平均精确度(mean average precision,mAP)、参数(parameters,Par)、浮点运算数(giga floating-point operations per second,GFLOPs)、帧速(frames per second,FPS)这六个指标作为改进后网络模型的评价标准,其中帧速指标使用CPU对图片进行推理而测定,目的是检验在低算力的边缘设备上改进模型的推理速度和计算能力。

本文将YOLOv5n作为电动车违规载人检测的baseline,以此为基础增加各类改进策略进行消融实验,来验证不同改进策略在使用本文的数据集进行训练所生成的模型的检测性能的优化效果。为了验证本文针对YOLOv5n所使用的两种优化策略,在保持相同的硬件配置、数据集和超参数的情况下,加入Mobilenetv3模块、ECA注意力机制、Slim-Neck结构和SPPFCSPC模块,通过实验来验证改进策略的有效性,其结果见表1和表3。

表1 YOLOv5n消融实验

算法	Mobilenetv3	ECA	Slim-Neck	SPPFCSPC	mAP@0.5/%	FPS
YOLOv5s					94.0	4.0
YOLOv5n					94.4	5.9
Changed(1)	√				93.2	7.36
Changed(2)	√	√			94.3	9.8
Changed(3)	√	√	√		94.7	10.2
本文算法	√	√	√	√	**96.2**	**10.8**

通过结果分析可知,与YOLOv5n相比,引入Mobilenetv3后,参数量减少为原先的42%,FPS提升了24%。继续引入ECA-Net注意力机制后,其参数量下降为原始YOLOv5n的18%,检测速度相较于YOLOv5n也提升至原先的1.66倍,在mAP@0.5这个指标上与YOLOv5n基本持平。而通过对比Changed(1)和Changed(2)算法可以看出,Mobilenetv3-ECA结构的算法在mAP@0.5上提升了1.1%,在参数量上缩减了57%,并且FPS提升了0.33%。由此可以证明在低算力平台下,Mobilenetv3结构可以有效降低参数量,节约推理时间,而引入ECA-Net注意力机制来替换Mobilenetv3中原有的SE-Net注意力机制可以进一步降低网络参数量,并且提高了网络的特征提取能力,提升了网络中不同尺度之间特征融合的能力,弥补了单纯使用轻量化网络降低网络参数而导致的网络损失部分特征提取能力不足,以及检测精度下降的缺点。通过比较Changed(2)算法和Changed(3)算法可以看出,mAP@0.5和FPS均有所上升,这是由于引入了轻量化的颈部结构Slim-Neck,减少了颈部网络的参数量从而加快了推理速度。通过比对本文算法和Changed(3)算法可知,在mAP@0.5指标上提升了1.5%,这是因为在Slim-Neck结构前加入了空间金字塔池化结构SPPFCSPC,通过SPPFCSPC结构在不同尺度上完成特征提取的特征图在送入Slim-Neck后由于其冗余信息少的特点可以取得更好的结果。

图13展示了两车重合场景、树木遮挡场景、人物重合场景下,消融实验的检测结果对比,可以看出消融实验组检测结果的准确率要低于本文算法,并且会出现错检和漏检的现象,而本文所提出的算法在以下场景均有较好的检测效果。

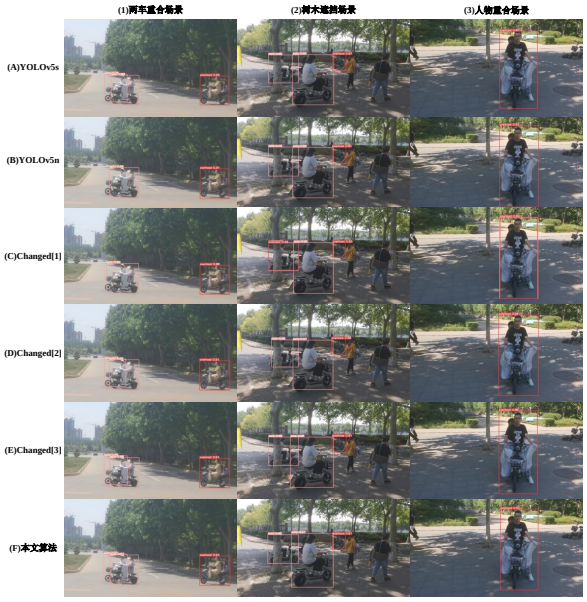


图13 消融实验对比示例

3.4 边缘设备部署实验与结果分析

为了验证本文所提到的改进模型在边缘设备上的推理能力,我们将模型部署在边缘设备上进行实验。

3.4.1　边缘设备简介

本文的边缘设备使用NVIDIA公司于2019年发布的Jetson Nano开发套件，Jetson Nano使用了ARM系列的4核64位CPU和NVIDIA公司的128核集成GPU，最高可提供472 GFLOPS的计算性能，拥有4 GB容量的LPDDR4存储器，具有高功效、低功耗的特点。

该产品使用了基于Ubuntu18.04的JetPack4.4.1镜像系统，拥有流畅的图形操作界面，支持NVIDIA CUDA Toolkit 10.0，以及cuDNN 7.3和TensorRT等库，也支持主流的深度学习框架如TensorFlow，PyTorch等，同时还安装了计算机视觉的框架，如OpenCV和ROS。完全兼容这些框架的Jetson Nano（见图14），非常适合作为承载AI推理测试和实时目标检测的轻量化边缘计算设备，设备的具体配置如表2所示。

Jetson Nano

图14　NVIDIA Jetson Nano

3.4.2　部署实验

表3展示了在Jetson Nano设备上部署模型进行推理得到的性能指标对比结果。

通过表3的实验结果可以发现，YOLOv5s在YOLO Parsing time、Inference time、FPS这些指标上都要远低于YOLOv5n，其中YOLOv5n的FPS提升了72.6%。这是因为不管是参数量还是网络深度，YOLOv5n都要远小于YOLOv5s，因此推理速度较快；通过对比YOLOV5n和Changed(1)可以看出，在YOLOv5n的基础上使用Mobilenetv3结构替换backbone中原有的卷积模块后可以发现，新的网络模型参数量大幅度减少伴随着mAP@0.5也有降低，这是由网络参数量下降使得网络的特征提取和融合能力下降导致的。在Changed(2)算法中，其FPS相较于YOLOv5n提升了72.9%，然而YOLO Parsing time却并未有明显提升，这是因为虽然使用了轻量化网络Mobilenetv3降低了网络参数量，却导致了网络层数大幅增加，使得推理过程中对于YOLO层的解析速度提升并不明显。针对Changed(1)的两个缺陷，在Changed(2)算法中使用ECA-Net注意力机制替换SE-Net注意力机制，一方面增强网络的特征提取能力；另一方面减小网络深度，减少YOLO层的解析时间。Changed(2)相较于YOLOv5n而言，其mAP@0.5基本持平且FPS也提升了72.9%，并且得益于网络深度的减少，YOLO Parsing time也降低了8.3%，取得了良好的推理速度。通过对比Changed(2)算法和Changed(3)算法的结果可以看出，增加了Slim-Neck结构后其参数量降低了30%，FPS增长了5.8%，inference time下降了13.9%，进一步加快了推理速度，验证了轻量化

表2　NVIDIA Jetson Nano配置

配置	系统	开发环境	CPU	GPU	RAM	CSI摄像头
细节	Ubuntu18.04	Python=3.7 C++, CMake OpenCV-4.0	64位四核ARM A57	128核NVIDIA Maxwell	4 GB 64位LPDDR4 25.6 GB/s	MIPI CSI-2 x2

表3　实验结果

模型	Parameters	mAP@0.5/ %	GFLOPs	YOLO Parsing time/s	FPS	Inference time/ms	Modl Layers
YOLOv5s	7129567	94.0	16.4	51.39	5.19	176.0	384
YOLOv5n	1872157	94.4	8.6	29.54	8.96	101.5	213
Changed(1)	793781	93.2	1.2	29.12	13.20	67.4	319
Changed(2)	337786	94.3	0.8	27.09	15.50	60.8	283
Changed(3)	236426	94.7	0.6	28.63	16.40	52.3	259
本文算法	**343626**	**96.2**	**0.7**	**26.42**	**15.60**	**55.9**	**282**

颈部网络Slim-Neck的有效性。通过观察本文算法的数据可以看出，在Changed(3)算法的基础上引入SPPFCSPC模块后，mAP@0.5增长了1.5%，验证了SPPFCSPC结构可以有效增强网络的特征提取能力，提高检测的精度。可以看出在Slim-Neck的基础上引入SPPFCSPC结构后，本文算法与Changed(2)算法相比，在参数量(Parameters)超过Changed(2)算法的情况下，GFLOPs却小于Changed(2)算法，在map@0.5取得1.9%的增长后FPS基本持平，可以验证Slim-Neck与SPPFCSPC结合后在网络推理速度和检测精度二者之间取得了良好的平衡性。本文算法在取得较高的检测精度的同时维持了较高的检测速度，其检测速度可以满足实时目标检测和前端部署的需求。

本文选取了100组数据进行推理实验，将每组数据的推理时间做比对，其结果如图15所示。相较于其他消融实验的算法，本文算法的推理速度较快且推理过程最为稳定，其推理时间浮动集中在59 ms到64 ms之间，通过对比可以验证本文所提出的算法具有良好的推理速度和不错的稳定性。

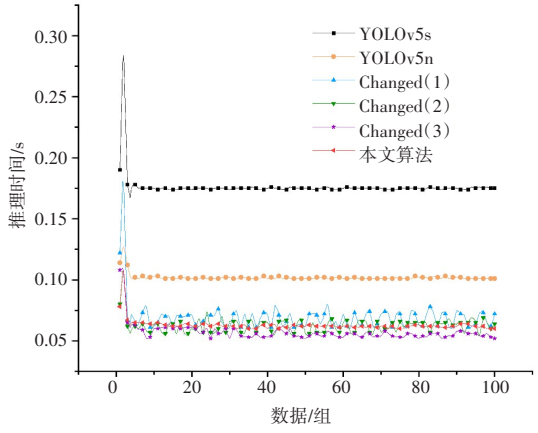


图15 100组数据推理时间折线图

4 结语

综上，本文基于YOLOv5目标检测算法，通过加入Moblienetv3轻量化网络、ECA-Net注意力机制、Slim-Neck结构和SPPFCSPC空间金字塔池化结构的方式对YOLOv5进行改进，来实现电动车违规载人的检测效果，将改进后的算法和NVIDIA的边缘设备Jetson Nano相结合，通过边缘计算的方式测试了改进算法的检测效果，并且比较了不同的网络模型在Jetson Nano上的推理速度和检测精度。根据实验结果分析，可以在边缘设备Jetson Nano上运行YOLOv5目标检测算法，且本文所提出的算法比YOLOv5n的检测速度更快，模型参数量较少。为交警提供了一种在复杂交通环境下查处电动车违规载人的解决方案，为了更好地帮助解决复杂交通环境下的各类实时检测问题，后续工作也将基于边缘设备Jetson Nano展开。

参考文献：

[1] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox etector[C]//Proceedings of the 2014 European Conference on Computer Vision, LNCS 9905. Cham:Springer,2016:21-37.

[2] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//Proceedings of the International Conference on Machine Learning. PMLR,2019:6105-6114.

[3] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(4):640-651.

[4] 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 & Machine Intelligence, 2017, 39(6): 1137-1149.

[5] HE K M, GEORGIA G, PIOTR D, et al. Mask R-CNN[EB/OL]. arXiv:1703.06870,2017.

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

[7] REDOMN J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017:17355115.

[8] REDOMN J, FARHADI A. YOLO v3: an incremental improvement [EB/OL]. arXiv: 1804.02767, 2018.

[9] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. arXiv:2004.10934,2020.

[10] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for

visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.
[11] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Piscata way: IEEE, 2017:2117-2125.
[12] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:8759--8768.
[13] 韩冲,王俊丽,吴雨茜,等. 基于神经进化的深度学习模型研究综述[J]. 电子学报, 2021, 49(2): 372-379.
[14] HOWARD A, ZHU M L, CHEN B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications [EB/OL]. arXiv: 1704.04861, 2017.
[15] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetv2: inverted residuals and linear bottlenecks [C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City: IEEE, 2018: 4510-4520.
[16] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 1314-1324.
[17] 钟志峰,夏一帆,周冬平,等. 基于改进YOLOv4的轻量化目标检测算法[J]. 计算机应用, 2022, 42(7): 2201-2209.
[18] HU J, LI S, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 2011-2023.
[19] WANC Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA: IEEE, 2020: 11531-11539.
[20] 郝巨鸣,杨景玉,韩淑梅,等. 引入Ghost模块和ECA的YOLOv4公路路面裂缝检测方法[J/OL]. 计算机应用: 1-7[2022-10-14]. http://kns.cnki.net/kcms/detail/51.1307.TP.20220904.1507.002.html.
[21] LI H L, LI J, WEI H B, et al. Slim-neck by GSConv: a better design paradigm of detector architectures for autonomous vehicles [EB/OL]. arXiv: 2206.02424, 2022.
[22] 梁允泉,董苗苗,齐振岭,等. 基于边缘设备和改进YOLOv5算法的车牌号码识别[J]. 现代计算机, 2022, 28(17): 16-22.
[23] 宋甜,李颖,王静. 改进YOLOv5s的车载红外图像目标检测[J]. 现代计算机, 2022, 28(2): 21-28.
[24] 迪菲赫尔曼. 空间金字塔池化改进SPP/SPPF/SimSPPF/ASPP/RFB/SPPCSPC/SPPFCSPC [EB/OL]. https://yolov5. blog. csdn. net/article/details/12635 -4660?spm=1001.2014.3001.5502.

Detection of illegal manning of electric vehicles based on improved YOLOv5 algorithm and edge devices

Sun Fulin[1], Li Zhenxuan[1], Liang Yunquan[2], Dong Miaomiao[2], Ge Guangying[3*]

(1. Shandong Key Laboratory of Optical Communication Science and Information Technology, Liaocheng 252059, China;
2. School of Physics Science and Information Technology, Liaocheng University, Liaocheng 252059, China;
3. School of Computer Science and Technology, Liaocheng University, Liaocheng 252059, China)

Abstract: It is an important content in the new traffic laws and regulations to prohibit the illegal carrying of people by electric vehicles. In view of the current lack of effective algorithms to detect the illegal carrying of people by electric vehicles, a detection system for illegal carrying of people by electric vehicles is designed based on the combination of improved YOLOv5 target detection algorithm and edge equipment. Firstly, the data set of electric vehicle driving is constructed; Secondly, on the basis of YOLOv5 network model, the lightweight network Mobilenetv3, ECA Net attention mechanism, Slim Neck structure and SPPFCSPC spatial pyramid pool structure are introduced to improve the detection accuracy for illegal passenger carrying of electric vehicles, and ablation experiments are conducted with the original algorithm. Finally, the improved algorithm is deployed on the edge device Jetson Nano for real-time reasoning. By analyzing the experimental data, the parameters of the improved algorithm are reduced to 18% of the original YOLOv5n, and the reasoning speed on the Jetson Nano is increased by 62%, with the fastest reasoning speed reaching 17FPS. The improved algorithm can greatly improve the reasoning speed while improving the detection accuracy on the Jetson Nano, and meet the needs of edge deployment in different scenarios.

Keywords: YOLOv5; edge equipment; Jetson nano; illegal manned detection of electric vehicles; lightweight network

文章编号: 1007-1423(2023)08-0012-08 DOI: 10.3969/j.issn.1007-1423.2023.08.002

基于LDA-BiLSTM模型和知识图谱的电影影评文本挖掘研究

杨秀璋[1]，武 帅[1,2*]，廖文婧[1]，项美玉[3]，于小民[4]，周既松[1]，赵小明[1]

（1.贵州财经大学信息学院，贵阳 550025；2.南京农业大学信息管理学院，南京 211800；
3.贵州财经大学大数据应用与经济学院（贵阳大数据金融学院），贵阳 550025；
4.贵州财经大学贵州省经济系统仿真重点实验室，贵阳 550025）

摘要：针对传统方法仅从宏观层面对电影产业和影评进行计量统计和文字描述研究，无法有效挖掘高质量电影的主题和观众的评价，缺乏深层次的语义知识挖掘，提出一种基于LDA-BiLSTM模型和知识图谱的电影影评文本挖掘方法。首先，采集《你好，李焕英》电影的影评文本并进行预处理；其次，抽取影评的共现特征词，利用知识图谱挖掘电影质量及口碑相关特征词的关联关系，深入分析高质量电影的影响因素；最后，构建LDA-BiLSTM模型实现影评的情感分析，通过LDA模型提取影评的关键特征词，利用长短时记忆网络捕获长距离依赖关系，从而精准预测影评情感类别。实验结果表明，提出的方法能有效挖掘电影影评的情感特征词和关联关系，所提出LDA-BiLSTM模型的精确率、召回率、F_1值和准确率依次为0.9839、0.9805、0.9822和0.9805，其结果优于其它机器学习和深度学习模型，为我国高质量电影挖掘提供学术思路，具有一定的研究价值。

关键词：文本挖掘；LDA-BiLSTM模型；知识图谱；电影分析；深度学习

0 引言

随着社会水平不断发展，人们的生活质量不断提高。物质需求得到满足后，人们更加倾向于追求精神文化需求。电影作为文化传播的重要媒介，一定程度上成为大家满足精神文明追求的主要载体，日益成熟的中国电影产业和市场增添了许多乐趣。对电影的评价在一定程度上成为人们日常生活中的情感释放渠道和慰藉心灵的一种方式。随着自媒体技术的不断发展，电影网络评价已成为观影后的习惯行为，人们会将观影的整体感觉、电影质量、感情以及感想通过评论的方式予以呈现。因此，针对中国电影市场和影评的文本挖掘具有重要的研究意义，一定程度上有效反映电影的票房和质量，为观众提供更丰富和喜欢的电影，能够深入挖掘影响中国电影市场的本质和现状[1]。

目前，国内电影产业分析的方法主要以文字描述和计量统计为主，通过票房统计、黄金时段统计、影片描述以及评分统计从宏观层面对电影进行研究，缺乏一种从微观层面深度挖掘电影影评和影片质量的方法，传统方法无法有效剖析电影票房“高开低走”的真正原因，并且无法有效挖掘高质量电影主题和观众的评价，缺乏深层次的语义知识挖掘。情感分析

收稿日期：2022-12-23 **修稿日期：**2023-03-21

基金项目：贵州省科技计划项目（黔科合基础〔2019〕1041、黔科合基础〔2020〕1Y279）；贵州省教育厅青年科技人才成长项目（黔教合KY字〔2021〕135）；贵州财经大学2021年度校级项目（2021KYQN03）

作者简介：杨秀璋（1991—），男，贵州凯里人，助教，硕士，研究方向为Web数据挖掘、知识图谱、人工智能，E-mail：1455136241@qq.com；*通信作者：武帅（1994—），男，江苏淮安人，硕士，工程师，研究方向为文本挖掘、图书情报学；廖文婧（1985—），女，贵州贵阳人，硕士，副教授，研究方向为数据挖掘、机器学习；项美玉（1992—），女，贵州贵阳人，硕士，讲师，研究方向为金融数据分析、数据挖掘；于小民（1982—），男，贵州贵阳人，硕士，助教，研究方向为人工智能、数据分析；周既松（1985—），男，江苏淮安人，硕士，助教，研究方向为数据分析、软件工程；赵小明（1992—），女，贵州凯里人，硕士，助教，研究方向为Web数据挖掘、大数据分析

(sentiment analysis)和主题挖掘(topic mining)技术能有效挖掘文本的主观情感倾向和内容的关键主题，现已被广泛应用于文本挖掘、舆情分析、推荐系统等领域[2-4]。因此，本文提出一种基于LDA-BiLSTM模型和知识图谱的电影影评文本挖掘方法。本文以2021年度“春节档”票房冠军《你好，李焕英》为例，利用知识图谱挖掘电影质量及口碑相关特征词的关联关系，深入分析高质量电影的影响因素，并构建LDA-BiLSTM模型实现影评的情感分析，通过LDA模型提取影评的关键特征词，利用长短时记忆网络捕获长距离依赖关系，从而精准预测影评情感类别。通过文本挖掘，本文进一步分析高质量和高口碑原创影片受观众青睐的缘由，为我国的高质量电影和票房挖掘提供学术思路，这将促进中国电影产业为观众提供更好的精神粮食。

1　相关研究

1.1　中国电影产业文本挖掘研究

在电影分析中，已有部分学者开始尝试使用数据挖掘的方法进行电影产业宏观分析。王纯[5]针对院线市场票房进行数据分析，进而研究我国电影产业企业的主体分布情况。沈建军等[6]结合猫眼平台的数据进行短视频营销与中国电影票之间的实证分析。庞林源等[7]运用格兰杰分析模型构建票房和评分两者关系，进而探究当前国产电影发展势头。杨秀璋等[8-9]运用大数据及可视化技术深度剖析当前中国电影产业发展现状，研究表明多类型题材电影有利于中国电影产业的发展。王锦慧等[10]通过构建多元回归分析模型分析各因素对票房的影响，得出中国电影在北美市场应加强渠道建设、强化发行网络以及注重首周票房。

当前，不仅有学者对电影票房进行深度研究，同样也有对电影评论或电影文本进行深度分析的，进而探究中国电影产业的发展趋势。冯莎[11]结合情感词典对《乘风破浪》电影影评进行文本情感分析，得出积极情绪占51.36%、消极情绪18.85%、中性情绪29.8%。王智威等[12]对86部韩国高质量电影进行文本分析，明确韩国电影中存在多元化中国形象。胡一伟等[13]对电影“彩蛋”进行文本分析，发现“彩蛋”能勾连不同内容的影片或故事，同时也能延伸电影在其他市场的商业活力。李静[14]针对《纽约时报》的电影影评进行文本分析研究，发现中国喜剧电影在北美市场因为文化差异形成较大的文化隔阂，较难实现喜剧效果重现。

然而，上述研究主要是采用数据挖掘的方法对电影产业进行宏观分析，或是利用文本挖掘的方法进行数据探究，缺乏一种将文本挖掘和中国电影产业发展、电影质量剖析相结合的研究，也缺乏构建电影主题的知识图谱，详细挖掘电影影评对票房的影响，利用深度学习模型挖掘影评来分析中国电影市场的发展变化。

1.2　情感分析

情感分析(sentiment analysis)旨在挖掘社交媒体平台中用户对某个事件或观点的情感评价，通常包括情感倾向分析(积极或消极)、情感时序分析、意见推理等[15-17]。本文主要针对电影《你好，李焕英》影评的情感倾向进行分析。当前该领域划分为三种典型的方法，包括基于情感词典和规则的情感分析、基于机器学习的情感分析、基于深度学习的情感分析[18]。

传统方法通过情感词典和规则进行匹配、映射和识别实现情感倾向分析，王志涛等[19]提出一种基于词典和规则集的中文微博情感分析方法，结合情感词典对微博文本进行从词语到句子的多粒度情感计算。然而，该类方法智能化程度较低，需要大量的手工标注，缺乏深层次的语义挖掘[20]。随后，基于机器的情感分析方法被提出，它能在一定程度上解决缺乏完整标注的语料库带来的分析局限性。李高翡等[21]提出一种异质集成学习的文本情感分析方法，通过朴素贝叶斯、支持向量机和随机森林实现情感分析。李婷婷等[22]利用SVM和CRF实现多特征组合并完成微博情感分析。然而，机器学习方法较难分析海量的社交媒体文本，会受到信息噪声的感染，并且缺乏深层次的语义关联和长距离依赖分析。

随着神经网络和人工智能的迅速发展，各行各业开始利用深度学习技术开展研究。基于深度学习的情感分析能够解决语料标注问题，且准确率较高，通过深层次语义挖掘来提升情感分析的性能。杨秀璋等[23]提出一种融合情感

词典的BiLSTM-CNN和Attention情感分类算法，该方法通过融合情感词典的特征提取方法优化特征词的权重，并优于现有的其它方法。夏辉丽等[24]提出一种利用自注意力双向分层语义模型进行网络文档情感分析，提高了深度学习模型的求解速度和准确度。杨春霞等[25]利用深度BiLSTM和图卷积网络实现方面级情感分析。

然而，上述方法缺乏对电影影评的情感分析，无法精准捕获文本信息中带有情感色彩的特征词，忽略了情感特征对模型的提升，也缺乏利用文本挖掘技术深度剖析高质量电影的原因。基于此，本文设计并实现LDA-BiLSTM模型实现电影影评的情感分析，并构建影评主题知识图谱来深入分析高质量和高口碑电影的缘由，进而论证中国电影市场正逐步从“流量至上”向“质量至上”转变。

2 模型设计与实现

该部分主要介绍本文提出方法的模型设计与实现过程。

2.1 模型总体框架

为有效挖掘中国高质量电影的缘由，本文提出一种基于LDA-BiLSTM模型和知识图谱的电影影评文本挖掘方法，其研究框架如图1所示。主要包括两大部分，第一部分是利用知识图谱挖掘电影质量及口碑相关特征词的关联关系；第二部分是构建LDA-BiLSTM模型实现影评情感分析。步骤如下：

（1）利用Python和Xpath技术构建网络采集器，自动抽取电影影评信息，以《你好，李焕英》高质量电影为示例，采集猫眼2021年2月1日至5月3日期间共计335933条影评文本数据。接着，对采集的数据集实现预处理和特征提取。

（2）提取电影影评的关键特征词并作为实体，再利用共现关系来构建电影的知识图谱，最终实现主题挖掘和关联分析，有效分析高质量电影的关键主题及关联关系。

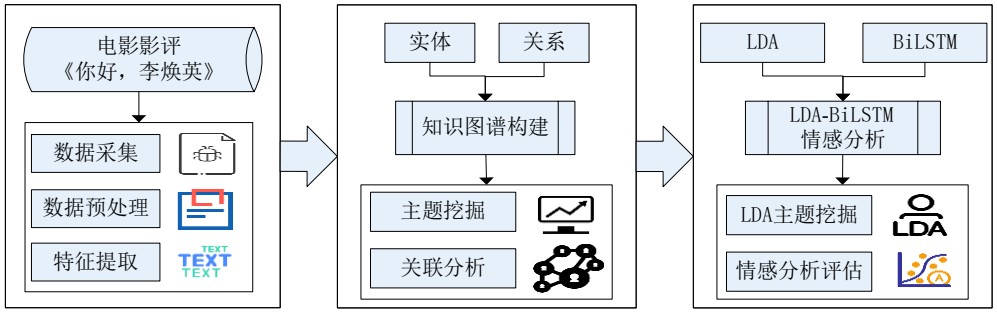
（3）设计融合LDA和BiLSTM的模型，利用LDA提取影评的关键情感特征词，再使用Word2Vec将特征映射成词向量，最终构建BiLSTM模型实现影评的情感分析，输出结果为Softmax分类器预测的积极情绪和消极情绪。

2.2 LDA-BiLSTM模型

本文提出一种融合LDA和BiLSTM的情感分析方法，模型网络结构如图2所示。具体实现过程如下：

首先，利用LDA模型提取不同评论文本的情感特征词。LDA（latent dirichlet allocation）是一种文档主题生成模型，由Blei等[26]在2003年首次提出，是一种三层贝叶斯结构，包括主题、文档和主题词三层结构，其中文档到主题、主题到词都服从多项分布。如图2所示，LDA模型从每篇文档D对应的多项分布θ中抽取每个单词对应的主题z，再从主题z中抽取一个单词w，其主题对应的多项分布为φ，最终重复N_d次，直至遍历文档中每个单词，自此成功抽取电影《你好，李焕英》影评的主题特征词，这些情感特征词能有效表征整个数据集。

其次，构建双向长短时记忆网络（Bi-directional long short-term memory, BiLSTM）模型实现情感分析。整个模型由前向LSTM和后向LSTM组成，能从两个方向捕获长距离依赖关系，其



（a）数据采集和预处理　（b）知识图谱文本挖掘　（c）深度学习情感分析

图1 基于LDA-BiLSTM模型和知识图谱的电影影评文本挖掘框架

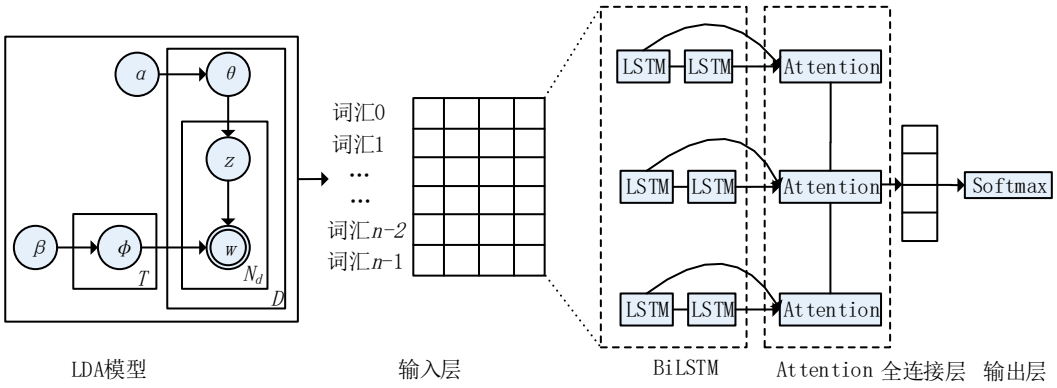


图 2　LDA-BiLSTM模型网络结构

计算公式如(1)至(3)所示。式中，$\boldsymbol{x}_t$表示输入词向量；w_1至w_6表示权重参数；f表示激活函数；$\overrightarrow{h_t}$表示t时刻前向LSTM层的状态，$\overleftarrow{h_t}$表示t时刻后向LSTM层的状态；y_t表示BiLSTM的最终输出结果。

$$\overrightarrow{h_t} = f(w_1 \times \boldsymbol{x}_t + w_2 \times \overrightarrow{h_{t-1}}) \tag{1}$$

$$\overleftarrow{h_t} = f(w_3 \times \boldsymbol{x}_t + w_5 \times \overrightarrow{h_{t+1}}) \tag{2}$$

$$y_t = g(w_4 \times \overrightarrow{h_t} + w_6 \times \overleftarrow{h_t}) \tag{3}$$

最后，经过LDA和BiLSTM模型计算的结果输入注意力层进行权重加成，最终输出至Softmax分类器完成电影影评的情感分类任务。整个结果将预测为积极情绪和消极情绪，从而实现对高质量电影《你好，李焕英》影评的情感挖掘。此外，本文与常见的机器学习和深度学习模型进行对比实验。

3　基于知识图谱的影评文本挖掘

由于文本数据具有体量大、类型复杂、来源多样等特征，因此使用知识图谱来挖掘文本具有重要意义。整个实验过程包括主题共词分析和知识图谱构建两部分。

3.1　主题共词分析

共词分析(co-word analysis)旨在利用文本数据中共同出现、表征文本数据主题的关键词来反映文献各主题间的关联，进而确定文献的热点主题。为更好地分析电影《你好，李焕英》的主题以及研究影片质量，本文采用共词分析法，构建共现矩阵，如公式(4)所示。当两个主题词同时出现在一段文本信息中，则认为存在共现关系，其边所对应的权重加1；反之，不存在共现关系，其权重为0。

$$y = \begin{cases} +1, & \text{主题词}A,B\text{存在共现} \\ 0, & \text{主题词}A,B\text{不存在共现} \end{cases} \tag{4}$$

结合共词分析结果，本文统计出电影《你好，李焕英》的前40组共现高频词，见表1。结果整体显示，影评观众对整部电影的评价较高，尤其是对电影结尾部分和“母爱”主题的评价称赞较多。从共现高频词组可看出，电影《你好，李焕英》的剧情反转具有一定的戏剧张力，有效将电影中母女之间的爱意、祝福层层递进到观众的感官和情绪的爆发点，让观众感同身受，产生情感的共鸣。

表 1　电影高频共现主题词

排序	主题词A	主题词B	频次
1	电影	结尾处	127
2	电影	强推	127
3	贾玲	真情流露	126
4	哭	发自内心	126
5	非常	精彩	125
6	好看	影评	125
7	感动	评分	125
8	搞笑	优秀	125
9	不错	人物	124
10	自己	亲身经历	124
11	故事	真情实感	123
12	感觉	精彩	123
13	成功	影片	122
14	作品	亲情	122
15	导演	用心	122
16	最后	悲剧	122
17	题材	母爱	122
18	孩子	煽情	121

续表1

排序	主题词A	主题词B	频次
19	孝顺父母	好好	121
20	感人	理解	121
21	泪点	精彩	120
22	眼泪	值得一看	120
23	反转	到位	120
24	作品	亲情	120
25	我们	没想到	119
26	打动	泪点	119
27	孩子	煽情	119
28	必须	感人	119
29	亲情	感情	119
30	部分	细节	118
31	期待	李焕英	118
32	非常	满意	117
33	喜欢	回忆	117
34	好看	成功	117
35	情节	值得一看	117
36	感动	说实话	116
37	好看	细腻	116
38	作品	细节	116
39	最后	很棒	115
40	经历	泪点	115

3.2 影评知识图谱分析主题关联关系

知识图谱(knowledge graph)通过实体(节点)和关系(边)分析实体之间的关联。本文通过共词分析挖掘影评主题特征词，计算各节点的度(degree)、网络密度(density)、聚类系数(clustering coefficient)后，再利用Gephi构建《你好，李焕英》主题社交网络关系知识图谱，最终生成如图3所示的影评知识图谱。

由图3可知，整个知识图谱包含782个核心节点，1383条主要关系边。模块化计算结果为0.528，平均加权度468.96，图中圆圈表示核心主题词，连线表示共现关系，连线越多表示其共现次数越高，反之越少。整部电影以喜剧为主，同时夹杂着“泪”与“笑”，最终形成以“电影”“妈妈”“贾玲”“我们”“李焕英”“笑”和“哭”为核心主题群的网络关系，突出整部电影笑中带泪、母爱感人的主题。该电影以怀旧为感情基调，将人性亲情、历史背景和现实主义相融合，为观众提供不同年龄段的心灵慰藉，表达了中华儿女对亲情的眷恋，对家庭的归属。

综上所述，本文提出基于知识图谱的影评文本挖掘方法能有效分析电影影评，挖掘出观众对亲情主题的关注，最终奠定影片的高口碑和高票房，也是国产电影向质量发展的重要转变。

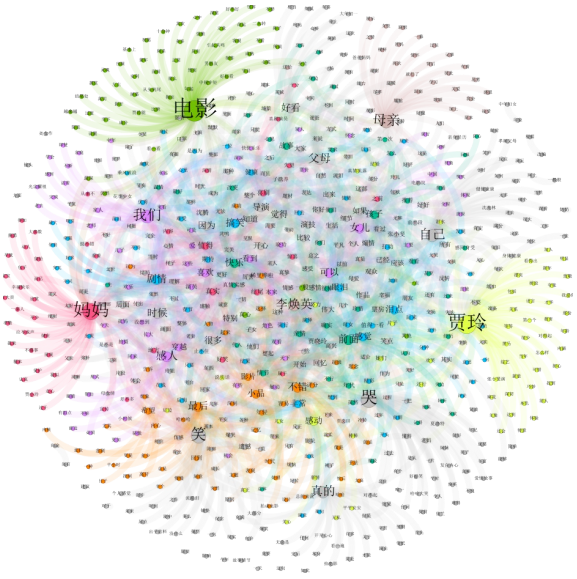


图3 基于知识图谱的电影影评主题分析

4 基于LDA-BiLSTM的影评情感分析

该部分主要介绍影评情感分析实验，包括评估指标、模型参数、实验结果和对比分析等。

4.1 评估指标和模型参数

本文共采集电影《你好，李焕英》猫眼2021年2月1日至5月3日期间共计335933条影评文本数据，将评分为0～2.5的影评划分为消极，2.5～5的划分积极，并按照一定比例随机划分为训练集、测试集和验证集。然后，通过文本分词、停用词过滤和特征提取后，构建LDA-BiLSTM模型，利用精确率(Precision)、召回率(Recall)、F_1值(F_1-score)和准确率(Accuracy)评估实验，其计算公式如(5)至(8)所示，再按各类样本数量进行权重加成。

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

本文模型超参数如表 2 所示，整个实验结果为十次实验结果的平均值。

表 2　模型超参数设置

参数	名称	值
L	文本序列长度	100
d	词向量维度	500
T_lstm	LSTM 隐藏层数量	128
T_dense	分类器类别数量	2
lr	学习率大小	0.001
p	Dropout 参数	0.3
Batchsize	批量梯度下降数	25
AF	激活函数	Adam

4.2　本文模型的实验评估

本文结合情感词典来构建 LDA 模型，并实现积极情绪和消极情绪影评的情感主题挖掘，图 4 和图 5 分别展示了两种情绪的主题特征词，积极情绪以“好看”“不错”“感动”“值得”“支持”等为主，消极情绪以“无聊”“一般”“烂片”“遗憾”“一般般”等为主。

图 4　影评积极情感主题特征词云图

图 5　影评消极情感主题特征词云图

接着构建融合注意力机制和 BiLSTM 的模型，通过该模型学习经过 LDA 提取的情感特征词，并利用 Word2Vec 映射成词向量，最终完成对电影影评的情感分析。

本文模型的实验结果如表 3 所示，影评积极情感的精确率、召回率和 F_1 值依次为 0.9953、0.9844 和 0.9898；消极情感的精确率、召回率和 F_1 值依次为 0.6659、0.8707 和 0.7546，最终精确率、召回率、F_1 值和准确率为 0.9839、0.9805、0.9822 和 0.9805。此外，由于消极影评相对较少，因此其对应的 F_1 值相对较低，但不影响本文 LDA-BiLSTM 模型的性能，LDA 模型充分抽取关键主题情感词从而提升性能，4.3 小节的对比实验将突出本文模型的优势。

表 3　基于 LDA-BiLSTM 的影评情感分析结果

类别	*Precision*	*Recall*	F_1-score	*Accuracy*
积极	0.9953	0.9844	0.9898	0.9844
消极	0.6659	0.8707	0.7546	0.8707
平均	0.9839	0.9805	0.9822	0.9805

4.3　机器学习和深度学习模型对比实验

该部分将详细对比本文模型与经典机器学习模型（包括逻辑回归、SVM、随机森林、朴素贝叶斯）和深度学习模型（BiLSTM、BiGRU、CNN、TextCNN），其结果如表 4 所示。

表 4　各模型影评情感分类的实验结果对比

类别	模型	*Precision*	*Recall*	F_1-score	检测时间/s
机器学习	逻辑回归	0.9386	0.9507	0.9404	9.31
	SVM	0.9386	0.9510	0.9365	12.59
	随机森林	0.9344	0.9477	0.9380	98.11
	朴素贝叶斯	0.9360	0.9500	0.9348	10.13
深度学习	BiLSTM	0.9643	0.9548	0.9587	76.88
	BiGRU	0.9591	0.9516	0.9548	65.08
	CNN	0.9597	0.9490	0.9535	110.79
	TextCNN	0.9714	0.9487	0.9579	174.57
	本文方法	**0.9839**	**0.9805**	**0.9822**	**156.63**

由表 4 可知，本文方法的精确率为 0.9839，召回率为 0.9805，F_1 值为 0.9822，检测时间为 156.63 s，整个实验结果均优于现有的机器学习

和深度学习模型。通过对比本文方法与其它方法的F_1值，发现本文方法分别比逻辑回归、SVM、随机森林和朴素贝叶斯模型提高4.18、4.57、4.42和4.74个百分点；分别比BiLSTM、BiGRU、CNN和TextCNN提高2.35、2.74、2.86和2.43个百分点。通过对比实验，进一步证明融合LDA和BiLSTM及注意力机制的模型能更好地挖掘《你好，李焕英》电影影评的情感色彩，并且检测时间未呈现指数级增长，突出本文模型的有效性。

最后，详细比较不同方法经过LDA模型提取情感色彩特征词前后的实验结果，如图6所示。该实验结果显示，经过LDA模型提取的情感特征词能更好地被机器学习或深度学习训练，其实验结果的F_1值均有一定程度的提升。其中，逻辑回归、SVM、随机森林和朴素贝叶斯模型分别提升0.80、1.27、0.55、1.06个百分点，BiLSTM、BiGRU、CNN和TextCNN模型提升0.52、0.72、0.77和0.71个百分点。综上所述，LDA模型对情感主题词的提取能在一定程度上提升情感分类模型的效果，更好地促进模型学习电影评论的情感色彩，预测情感倾向。

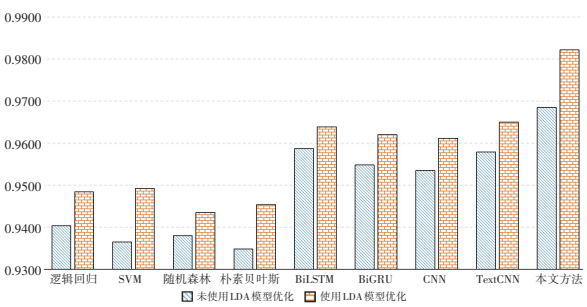


图6 使用LDA模型优化的对比实验结果

5 结语

本文针对传统方法仅从宏观层面对电影产业和影评进行计量统计和文字描述研究，无法有效挖掘高质量电影的主题和观众的评价，缺乏深层次的语义知识挖掘的问题，提出一种基于LDA-BiLSTM模型和知识图谱的电影影评文本挖掘方法。实验结果表明，本文方法能有效挖掘电影影评的情感特征词和关联关系，所提出LDA-BiLSTM模型的精确率、召回率、F_1值和准确率依次为0.9839、0.9805、0.9822和0.9805，其结果优于其它机器学习和深度学习模型，并且通过详细的对比实验证明LDA模型提取情感色彩的特征词能更好地被深度学习模型学习，最终为我国高质量电影挖掘提供学术思路，具有一定的研究价值。

在后续工作中，将进一步融合迁移学习、图神经网络来分析电影影评数据，并针对消极情绪语料分布不均匀的问题实现数据增强，为我国电影产业的高质量影片拍摄和剧情设计提供支持。

参考文献：

[1] 马蹦非，刘超一．“仪式传播”与“传播仪式”：春节档电影的文化反思：兼评电影《你好，李焕英》[J]．戏剧(中央戏剧学院学报)，2021(2)：36-43.

[2] 白静，李霏，姬东鸿．基于注意力的BiLSTM-CNN中文微博立场检测模型[J]．计算机应用与软件，2018，35(3)：266-274.

[3] 杨秀璋，武帅，夏换，等．基于主题挖掘和情感分析的“新冠肺炎疫情”舆情分析研究[J]．计算机时代，2020(8)：31-36.

[4] 王晰巍，张柳，文晴，等．基于贝叶斯模型的移动环境下网络舆情用户情感演化研究：以新浪微博“里约奥运会中国女排夺冠”话题为例[J]．情报学报，2018，37(12)：1241-1248.

[5] 王纯．我国电影生产企业的主体分布、格局与问题：基于2005-2019年度院线市场票房排名前30位影片企业数据分析[J]．当代电影，2021(3)：87-93.

[6] 沈建军，吴春集．短视频营销对电影票房的影响研究：基于猫眼平台数据的实证分析[J]．电影文学，2021(3)：18-24.

[7] 庞林源，王欢．国产电影票房与网络评分关系探讨：基于票房数据和豆瓣电影评分的分析[J]．北京电影学院学报，2020(12)：60-64.

[8] 杨秀璋，武帅，夏换，等．大数据视域下2019年中国电影产业分析研究[J]．电影文学，2020(23)：17-29.

[9] 杨秀璋，夏换，于小民，等．基于社交网络和决策树的中国电影产业研究[J]．电影文学，2019(5)：3-12.

[10] 王锦慧，杨勇．影响中国电影北美市场的票房因素研究：以在北美上映的242部中国电影为例[J]．当代电影，2020(11)：97-103.

[11] 冯莎．豆瓣电影评论文本的情感分析研究：基于2017年电影《乘风破浪》爬虫数据[J]．中国统计，2017(7)：30-33.

[12] 王智威,何镇飚.论近年来韩国高质量电影对中国形象的构建:文本分析法[J].电影文学,2020(20):26-32.
[13] 胡一伟,浦翰林.游走于文本框架内外的电影"彩蛋":一个伴随文本现象分析[J].电影文学,2020(12):23-27.
[14] 李静.中国类型电影的北美市场文化认同差异:《纽约时报》影评文本(1979—2017)实证分析[J].华侨大学学报(哲学社会科学版),2020(2):151-160.
[15] 杨秀璋,武帅,张苗,等.基于TextCNN和Attention的微博舆情事件情感分析[J].信息技术与信息化,2021(7):41-46.
[16] 李然,林政,林海伦,等.文本情绪分析综述[J].计算机研究与发展,2018,55(1):30-52.
[17] 余传明,原赛,王峰,等.大数据环境下文本情感分析算法的规模适配研究:以Twitter为数据源[J].图书情报工作,2019,63(4):101-111.
[18] 杨秀璋,刘建义,任天舒,等.基于改进LDA-CNN-BiLSTM模型的社交媒体情感分析研究[J].现代计算机,2022,28(3):29-36.
[19] 王志涛,於志文,郭斌,等.基于词典和规则集的中文微博情感分析[J].计算机工程与应用,2015,51(8):218-225.
[20] 陈晓东.基于情感词典的中文微博情感倾向分析研究[D].湖北:华中科技大学硕士论文,2012.
[21] 李高翡,张洋,杨新凯,等.基于集成学习的文本情感分析研究[J].计算机应用研究,2020,37(S1):50-51.
[22] 李婷婷,姬东鸿.基于SVM和CRF多特征组合的微博情感分析[J].计算机应用研究,2015,32(4):978-981.
[23] 杨秀璋,郭明镇,候红涛,等.融合情感词典的改进BiLSTM-CNN+Attention情感分类算法[J].科学技术与工程,2022,22(20):8761-8770.
[24] 夏辉丽,杨立身,薛峰.利用自注意力机制的大规模网络文档情感分析[J].计算机工程与设计,2021,42(9):2642-2648.
[25] 杨春霞,宋金剑,姚思诚.基于深度BiLSTM和图卷积网络的方面级情感分析[J].计算机工程与科学,2022,44(10):1893-1900.
[26] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003(3):993-1022.

Research on text mining of movie reviews based on LDA-BiLSTM model and knowledge graph

Yang Xiuzhang[1], Wu Shuai[1,2*], Liao Wenjing[1], Xiang Meiyu[3], Yu Xiaomin[4], Zhou Jisong[1], Zhao Xiaoming[1]

(1.School of Information, Guizhou University of Finance and Economics, Guiyang 550025, China;
2. School of Information Management, Nanjing Agricultural University, Nanjing 211800, China;
3. Guiyang School of Big Data and Finance, School of Big Data Application and Economics, Guizhou University of Finance and Economics, Guiyang 550025, China;
4. Key Laboratory of Economics System Simulation of Guizhou, Guizhou University of Finance and Economics, Guiyang 550027, China)

Abstract: Since the traditional method only conducts quantitative statistics and text description research on the film industry and reviews from the macro level, it cannot effectively mine the themes of high-quality films and audience evaluations and needs deep semantic knowledge mining. This paper proposes a model based on LDA-BiLSTM and the knowledge graph for text mining of movie reviews. Firstly, this paper collects the film review text of "Hi, Mom" and preprocesses it. Secondly, it extracts the co-occurrence feature words of film reviews, uses the knowledge graph to mine the relationship between film quality feature words, and profoundly analyzes high-quality movies. Finally, it builds the LDA-BiLSTM model to realize the emotional analysis of film reviews, extracts the critical feature words of film reviews through the LDA model, and uses the long-short-term memory network to capture long-distance dependencies to accurately predict the emotional category of film reviews. The experimental results show that the method proposed in this paper can effectively mine the emotional feature words and association relationships of movie reviews. The precision, recall, F_1-score, and accuracy of the proposed LDA-BiLSTM model are 0.9839, 0.9805, 0.9822, and 0.9805, which are superior to other machine learning and deep learning models. In short, this paper can provide academic ideas for mining our high-quality movies.

Keywords: text mining; LDA-BiLSTM model; knowledge graph; film analysis; deep learning

文章编号: 1007-1423(2023)08-0020-07 DOI: 10.3969/j.issn.1007-1423.2023.08.003

基于LSTM-TCN的地下水位数据修复及应用

袁志洪，陈 雨*

（四川大学电子信息学院，成都 610065）

摘要：准确预测地下水位是合理开发利用地下水资源的重要依据。由于数据采集过程中可能受到环境干扰或者采集装备发生故障等，导致数据缺失，进而影响模型的预测准确度。为了在含较多缺失值的地下水位数据上获得更准确的预测结果，首先，提出了基于长短时记忆网络(long short term memory，LSTM)-因果卷积网络(temporal convolutional network，TCN)的地下水位修复模型，通过学习原数据集中的时序特性和分布特点以改善数据集质量。然后，将多头注意力机制(multi-head attention mechanism，MA)与LSTM相结合进行地下水位预测，进一步提升LSTM模型的预测性能。最后，预测结果表明，使用LSTM-TCN方法修复后的数据集训练MA-LSTM模型，显著地提高了地下水位预测精度。

关键词：数据修复；生成对抗插补网络；因果卷积网络；地下水位预测；多头注意力机制

0 引言

水资源与人类的生存和发展息息相关，日益成为世界各地关注的热点之一。由于地表水缺乏，2008年德克萨斯州抽取地下水进行农业灌溉的量占总用水量的81.4%[1]。然而，随着灌溉农业的不断发展及人口数量的增加，过度开采地下水引发了水井枯竭，地表塌陷，农民农业经济收入减少[1]，生态系统稳定性被破坏[2]等问题。因此，准确预测地下水位对合理开发利用水资源，促进地区的农业经济发展等具有重要价值。

数据在采集过程中可能因为环境干扰或者采集装备发生故障等，使得观测数据集中存在数据缺失和异常值[3]，从而造成模型预测精度下降。为减少数据集因存在缺失值对预测精度的负面影响，有关学者采用了多种方法修复缺失的数据，这些方法包括：①删除法[4]，直接删掉数据集中的缺失数据，但当缺失率较高时，该方法往往会忽略缺失数据中所包含的关键信息，对模型的预测性能影响较大；②插补法，常采用三次样条插值[5]、均值插补、基于K近邻的插值对数据进行修复[6]，该方法关注于原始数据集的众数、中位数等统计学特征，却未考虑时间序列数据具有时序性这一关键特征，插补精度较低；③生成法[7-8]在数据修复中得到了较为广泛的使用，该方法通过学习原始数据集的分布特点生成新数据以填补缺失值，其中，生成对抗式插补网络(generative adversarial imputation networks，GAIN)中引入了提示矩阵以辅助鉴别器判断，促使生成器生成趋于真实数据分布的数据，以填补缺失值，可以较好地修复非序列数据或图像中的缺失值，但由于该方法未采取相关措施处理时间序列数据中的时序特征，使得GAIN模型在修复具有时序性的数据时难以发挥其效果。

因此，本文对GAIN模型的结构进行了改进，采用深度长短时记忆网络(long short term memory，LSTM)[9]作为生成器的核心部分，深度因果卷积网络(temporal convolutional network，TCN)[10]作为鉴别器的核心部分，提出了基于LSTM-TCN的地下水位数据修复模型。LSTM-TCN模型在插

收稿日期：2023-01-13 **修稿日期：**2023-02-21

作者简介：袁志洪(1998—)，女，贵州遵义人，硕士，研究方向为时间序列预测、遥感图像处理；*通信作者：陈雨(1976—)，男，四川成都人，硕士生导师，副教授，研究方向为GRACE重力卫星数据应用、遥感图像处理、深度学习，E-mail：ychen@scu.edu.cn

补缺失值时考虑了原数据集的时序特性和分布特点，所以经该方法生成的数据能更准确地反映真实地下水位数据的变化规律和局部特征，有效提高了数据集质量。由于LSTM网络[11-12]在地下水位预测上具有良好的表现，多头注意力机制(multi-head attention mechanism，MA)[13]可为模型的输入变量赋予不同的注意力权重，所以本文将两者相结合进行地下水位预测，以进一步提高LSTM网络的地下水位预测准确度。

1　基于LSTM-TCN的地下水位修复模型

LSTM-TCN是基于GAIN模型提出的地下水位数据修复模型，框架如图1所示。该插补方法的基本结构由生成器、鉴别器两个核心模块组成。其中，生成器网络通过学习真实地下水位数据在不同时间尺度上的特征信息和分布特点，生成更接近于真实数据的填充矩阵$\hat{\boldsymbol{X}}$，并将$\hat{\boldsymbol{X}}$传入到鉴别器网络中；鉴别器网络通过捕获$\hat{\boldsymbol{X}}$的时序特征和变化规律，分辨出该矩阵中哪些元素是真实值，哪些元素是生成值。通过迭代训练LSTM-TCN模型，直到鉴别器网络无法判别$\hat{\boldsymbol{X}}$中元素值的真伪为止，此时生成器的输出就是修复后的地下水位数据。

1.1　基于深度LSTM的生成器

本文构建的生成器网络模型如图2所示，该网络模型的输入包括含缺失信息的数据矩阵$\tilde{\boldsymbol{X}}=(\tilde{X}_1,\tilde{X}_2,\cdots,\tilde{X}_d)$、D维随机噪声向量$\boldsymbol{Z}$以及掩模矩阵$\boldsymbol{M}=(M_1,M_2,\cdots,M_d)$，输出$\hat{\boldsymbol{X}}$为$\tilde{\boldsymbol{X}}$与输入维度一致的填补矩阵。

$$\tilde{\boldsymbol{X}}_i=\begin{cases}*, & \boldsymbol{M}_i=0\\ X_i, & \boldsymbol{M}_i=1\end{cases} \tag{1}$$

$$\bar{\boldsymbol{X}}=G\left(\tilde{\boldsymbol{X}},\boldsymbol{M},(1-\boldsymbol{M})\odot\boldsymbol{Z}\right) \tag{2}$$

$$\hat{\boldsymbol{X}}=\boldsymbol{M}\odot\tilde{\boldsymbol{X}}+(1-\boldsymbol{M})\odot\bar{\boldsymbol{X}} \tag{3}$$

式中：*代指未观测到的数据；⊙表示同位元素进行相乘。$\boldsymbol{M}$中的元素值为$\{0,1\}$，表示该位置的对应元素分别是缺失值、真实值，因此，生成器生成的数据矩阵中既有真实值，也有生成

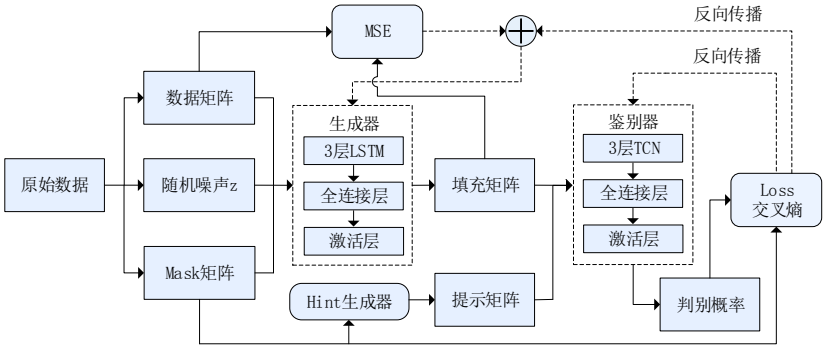


图1　基于LSTM-TCN的地下水位数据修复框架

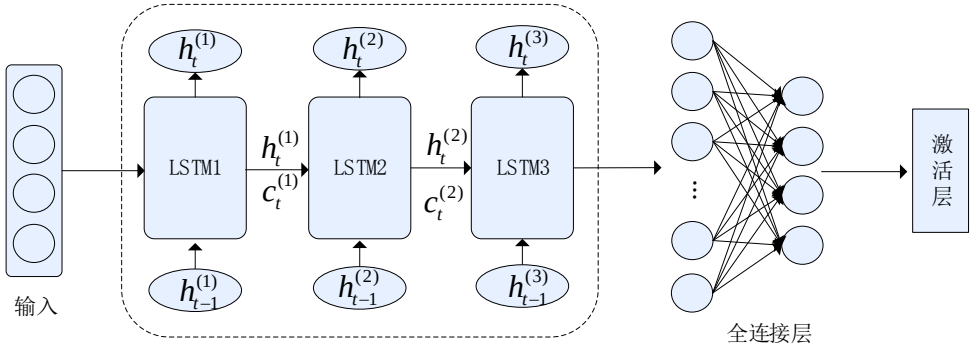


图2　生成器网络架构

值。生成器网络中各模块的作用如下：

首先，通过堆叠3个LSTM层以增加生成器网络中隐藏层到隐藏层之间的网络层数，将前一时刻的LSTM层输出的c_{t-1}和h_{t-1}作为后一时刻的LSTM层输入，以此控制向后层神经网络传递的信息，从而使得深度LSTM网络相较于浅层LSTM网络能更好地学习真实地下水位数据在不同时间尺度上的特征信息，促使生成器网络生成更符合原数据分布特点和时序性的假数据。然后，将深度LSTM网络的输出矩阵经过全连接层处理后，得到的矩阵维度等于$\tilde{\boldsymbol{X}}$的输入维度。最后，输出层中使用sigmoid激活函数，使得$\hat{\boldsymbol{X}}$的值域映射在[0, 1]之间，便于后续将$\hat{\boldsymbol{X}}$输入到鉴别器网络中进行训练。因此，所提出的生成器网络更适用于捕获原数据集的分布特性和时序特征，使生成的$\hat{\boldsymbol{X}}$更接近真实地下水位数据。

1.2 基于深度TCN的鉴别器

鉴别器网络模型的输入为生成器生成的填补矩阵$\hat{\boldsymbol{X}}$，输出的概率矩阵取值为[0, 1]且与$\hat{\boldsymbol{X}}$的维度一致。图3所示为鉴别器网络中的TCN网络结构，TCN网络用于提取地下水位数据中的时序特征和变化规律，其基本原理如下：

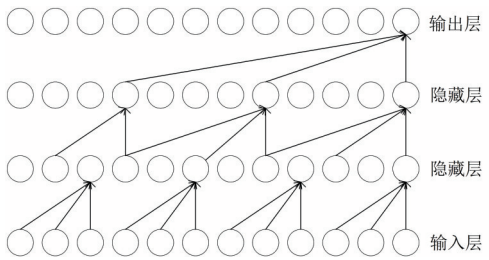


图3 TCN网络模型

TCN网络由因果卷积，膨胀卷积以及残差模块3部分组成。其中，因果卷积是TCN网络的关键特征，指的是当前t时刻的输出值y_t仅依赖于前一层t时刻和t时刻之前发生的元素卷积，具有因果性。膨胀卷积通过控制感受野大小来控制卷积核的学习范围，其表达式为

$$F(s)=(x*df)(s)=\sum_{i=0}^{k-1}f(i)X_{s-d\cdot i} \tag{4}$$

式中：$\boldsymbol{X}=\{x_1,x_2,\cdots,x_t\}$为输入值；$k$表示卷积核的大小；*为卷积运算符；$d$为膨胀因子。残差模块进一步优化了该网络模型，能较好地解决由于神经网络层数较深时出现的过拟合以及梯度消失等问题。因此，采用TCN网络作为鉴别器网络的一部分，有利于鉴别器网络判断$\hat{\boldsymbol{X}}$中的对应元素值的真伪。

1.3 提示机制

LSTM-TCN模型中的提示机制H，用于向鉴别器提示信息，促使生成器生成符合原数据集分布特点的数据，以插补缺失值。H定义为

$$H=B\odot M+0.5(1-B) \tag{5}$$

式中：D维随机变量B取值为{0, 1}；H取值为$\{0, 0.5, 1\}$。当H中的元素值为0或1时，用于提示鉴别器此值是生成的假数据(取值为0)还是真实数据(取值为1)；当元素值为0.5时，鉴别器自行判断此元素值的真伪。

1.4 评估函数

评估函数通过最大化正确预测$\boldsymbol{M}$的概率来训练鉴别器，最小化鉴别器能正确预测$\boldsymbol{M}$的概率来训练生成器，其公式定义如下：

$$L=\min_G\max_D E_{\hat{\boldsymbol{X},M,H}}\Big[(1-\boldsymbol{M})^{\mathrm{T}}\log\big(1-D(\hat{\boldsymbol{X}},H)\big)+\boldsymbol{M}^{\mathrm{T}}\log D(\hat{\boldsymbol{X}},H)\Big] \tag{6}$$

式中的G和D分别表示生成器、鉴别器。

2 多头注意力机制

Transformer模型[13]中的多头注意力机制是一种特殊的自注意力机制，它的基础单元是单头注意力机制，其第i个单头的注意值为

$$h_i=\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$三个矩阵的维度均为d_k，分别表示查询、键、值；缩放因子d_k为序列的输入维度，用于调节h_i的值，避免反向传播时产生梯度消失的现象。将多个单头注意力机制并联后经过线性层映射得到多头注意力机制，其计算公式如下：

$$\boldsymbol{M}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=(h_1\oplus h_2\oplus\cdots\oplus h_n)\boldsymbol{W}^0 \tag{8}$$

式中：$\oplus$表示向量之间进行拼接操作；$\boldsymbol{W}^0$为权重矩阵；n为总头数。

3　实验设计及结果分析

3.1　研究数据来源

观测井位于德克萨斯州奥加拉拉含水层，数据的时间范围为2001年6月7日到2022年4月8日，以周为单位进行提取，获得共1088条时序数据作为实验数据，经初步整理后发现，地下水位数据集中的缺失数据条数为62条，缺失率约为5.70%。由于缺失数据较多，因此预测前需要使用合理方法修复缺失数据，以提高后续地下水位的预测准确度。

地下水数据来源于德克萨斯州水资源开发委员会，引入的降水、温度和蒸发量数据均来源于PRISM卫星，分辨率为4 km。可将遥感图像中提取获得的气象数据和地下水位数据相结合，用于预测地下水位数据的动态变化[14]。因此，实验中使用修复后的地下水位数据、降水、温度及蒸发量数据整合后作为预测模型的输入值，以地下水位作为预测模型的输出值。

3.2　实验准备

进行地下水位预测前，将实验数据集的前80%作为训练数据集，后20%划分为测试数据集。并将输入数据和目标数据均归一化到[0, 1]之间，避免因数据尺度影响模型的预测效果。然后采用滑动窗口法对实验数据集进行处理，使用前10周的地下水位来预测未来1周的地下水位。

3.3　评价指标

为衡量本文所提模型的地下水位预测效果，使用均方根误差（root mean square，RMSE）和平均绝对误差（mean absolute error，MAE）作为评价指标，公式如下：

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| \tag{10}$$

*RMSE*定义为预测值与实际值之间的误差平方根的均值，反映了两者之间的偏差程度；*MAE*用于评估预测值与真实值之间的平均绝对误差。两者数值越小，模型预测性能越好。

3.4　地下水位数据修复

由于地下水位数据集中包含缺失值，导致不能直接评估不同插补算法的插补精度，因此，文中借助地下水位的预测结果来间接评估不同插补方法的插补精度。为验证所提方法的可靠性，文中的实验由两部分组成。

（1）从地下水位数据集中选择一部分完整连续的子序列作为样本数据，首先按照不同比例随机缺失样本数据以构造缺失子序列，然后使用各种插补方法填补缺失的子序列，将插补后的子序列与真实子序列进行对比，可直接验证不同插补方法在小数据集上的插补精度。

（2）对于含缺失值的整个地下水位数据集，首先使用不同插补方法对该数据集进行插补，然后将不同方法插补后的地下水位数据集与相关气象变量（降水、温度及蒸发量数据）整合后分别输入到不同模型进行训练及预测，最后通过地下水位的预测结果间接验证LSTM-TCN插补方法的有效性。

分别将LSTM、TCN网络进行组合后得到另外3种插补方法，即TCN-LSTM，LSTM-LSTM，TCN-TCN（生成器－鉴别器）。为验证所提LSTM-TCN方法的修复效果，将上述3种插补方法、GAIN网络以及常用插补方法作为插补实验的对照组。

首先，从原始地下水数据集中筛选出不含缺失值且连续的子序列作为样本数据，对应时间为2012年2月12日至2020年8月2日，长度为443条。然后，按照缺失率5%，10%，15%，20%，25%分别模拟随机缺失数据，图4展示了缺失率为5%的子序列未插补缺失值之前的效果。最后，使用多种插补方法对不同缺失程度的子序列进行插补，以评价指标RMSE来衡量各插补方法的修复效果，如表1所示。

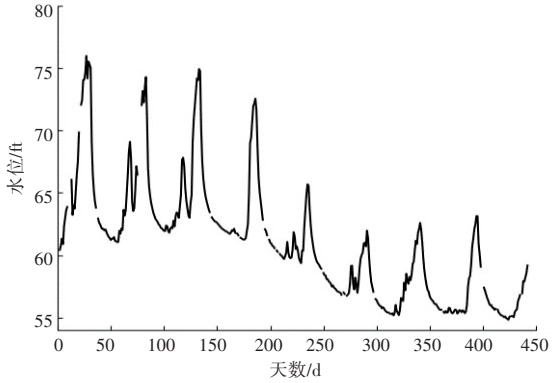


图4　缺失率为5%的子序列

表 1 不同缺失率下的子序列插补误差

插补方法	插补效果评价指标*RMSE*				
	5%	10%	15%	20%	25%
均值插值	0.9624	1.2821	1.5406	1.8823	2.2106
三次样条	0.2349	0.3754	0.4579	0.5107	0.5937
GAIN	0.2102	0.3569	0.4833	0.5999	0.7991
LSTM-TCN	**0.1045**	**0.2091**	**0.2484**	**0.3774**	**0.5336**
TCN-LSTM	0.3572	0.5399	0.5254	0.8335	1.0086
LSTM-LSTM	0.1054	0.2112	0.2312	0.3914	0.4541
TCN-TCN	0.2604	0.4475	0.5221	0.8901	0.9961

对比表 1 中*RMSE*的值可看出，对不同缺失率下的子序列进行插补时，本文提出的LSTM-TCN方法的插值精度均明显优于其余插补方法。在子序列的缺失率分别为5%、10%、15%、20%、25%时，相较于GAIN模型，采用该方法插补后其*RMSE*分别降低了50.29%、41.41%、48.60%、37.09%及33.22%。

3.5 基于MA-LSTM模型进行地下水位预测

影响地下水位的输入变量较多，且输入变量存在重要程度差异性。因此，本文将MA与LSTM相结合来预测地下水位，通过各单头注意力机制为输入变量赋予不同的注意力权重，使MA-LSTM模型可更好地捕获输入变量中的关键特征，从而进一步提高模型的预测准确度。为验证MA-LSTM模型的预测效果，实验中使用了几种经典预测模型：ARIMAX、LSTM、Bi-LSTM进行对照实验。

多元自回归移动平均模型(autoregressive moving average with extra input，ARIMAX)根据数据集的单位与检验和差分次数确定该模型的d为0，设置p和q值均为0～20之间，再结合贝叶斯准则[15]和网格搜索法得到ARIMAX模型d、p以及q值的最佳组合。以网格搜索法选择其余预测模型的超参数，如网络层数、训练批数、学习率、神经元个数等。

将LSTM-TCN方法修复后的地下水位数据集(缺失率为5%的子序列)及相关气象变量数据整合后输入到各预测模型中进行预见期为1周的地下水位预测，图5展示了各预测模型在修复后的子实验数据集上(2012年2月12日至2020年8月2日)进行地下水位预测时的误差。

由图5中*RMSE*和*MAE*的数值可知，使用修复后的子实验数据集分别进行训练及预测时，MA-LSTM模型对应的*RMSE*和*MAE*值均最小，该模型的地下水位预测效果最佳。MA-LSTM模型相较于Bi-LSTM模型、LSTM模型及ARIMAX模型，其*RMSE*分别降低了3.50%、14.25%、22.10%，其*MAE*分别降低了4.79%、15.56%、15.59%。

为进一步验证不同插补模型对地下水位预测结果的影响，分别使用各插补模型修复含缺失值的整个地下水位数据集，然后将修复后的整个地下水位数据集及相关气象变量数据整合后作为实验数据集，各自输入到AIMAX、LSTM、Bi-LSTM及MA-LSTM模型中进行地下水位预测，所得RMSE和MAE如表2所示。

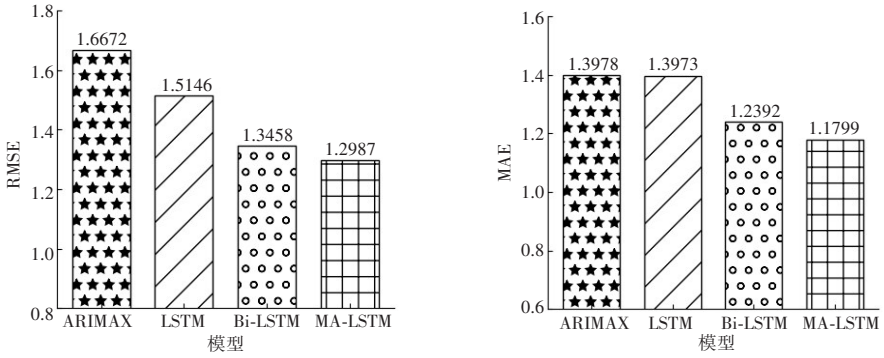


图 5 各预测模型的误差对比

表 2　不同插补方法修复整个序列后各预测模型的RMSE和MAE

插补方法	ARIMAX		LSTM		Bi-LSTM		MA-LSTM	
	RMSE	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*
均值插值	6.539	6.0236	1.5522	1.2346	1.5865	1.2564	1.3201	0.9635
三次样条	6.6382	6.1092	1.2517	1.0801	1.3559	1.1654	1.0477	0.9069
GAIN	6.9633	6.4925	1.2431	1.021	1.2488	1.1258	1.0477	1.0068
LSTM-TCN	**5.4065**	**4.8176**	**1.0788**	**0.8956**	**1.0287**	**0.8676**	**0.8249**	**0.6780**
TCN-LSTM	6.7772	6.323	1.2745	1.0731	1.3484	1.2239	1.205	0.9642
LSTM-LSTM	6.0628	5.5608	1.2105	1.0711	1.2257	1.0892	0.9128	0.7977
TCN-TCN	7.0549	6.6884	1.2346	1.0202	1.2764	1.1554	1.1415	0.8971

从表 2 可知，使用LSTM-TCN方法插补后的实验数据集分别训练ARIMAX、LSTM、Bi-LSTM、MA-LSTM模型，均获得了更高的预测精度。在各预测模型中，MA-LSTM模型的预测性能最佳，LSTM-TCN方法修复后的地下水位数据及MA-LSTM模型的预测值如图 6 所示。相较于GAIN网络，使用LSTM-TCN方法修复后的数据集训练MA-LSTM模型，能够使MA-LSTM模型的*RMSE*和*MAE*分别降低了21.27%、32.66%，获得了更准确的地下水位数据预测结果。因此，可认为使用LSTM-TCN模型修复含缺失值的地下水位数据集时，能更有效地提高数据集的质量，使得各预测模型能从修复后的数据集中学到更精确的特征信息，从而提高了模型的地下水位预测精度。

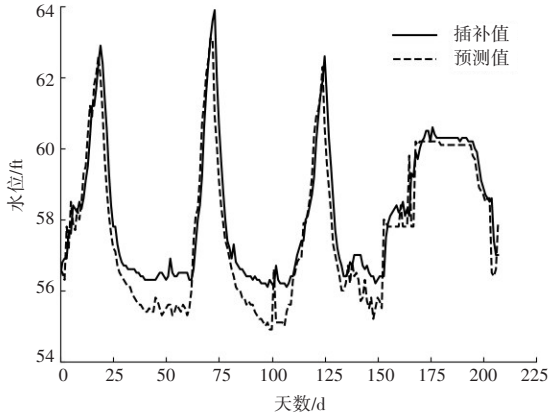


图 6　LSTM-TCN模型修复整个序列后的预测结果

4　结语

针对地下水位数据集中存在较多缺失值情况下的高精度地下水位预测需求，本文提出了基于LSTM-TCN的地下水位数据修复模型，改进了GAIN网络的结构。改进后，LSTM-TCN模型能够较好地学习原数据集的分布特点和时序特性，采用该修复模型对不同缺失率的地下水位数据集进行插补时，均能生成更接近真实值的假数据，显著提高了地下水位数据集的质量。然后，对于地下水位预测模型，本文引入了多头注意力机制为输入变量赋予不同的注意力权重，使得LSTM模型能更好地捕获实验数据集中的关键特征，进一步提高了地下水位的预测精度。

参考文献：

[1] 胡亚琼,刘静,廖丽莎. 美国德克萨斯州高地平原区地下水灌溉管理方法研究[J]. 灌溉排水学报,2019,38(1):96-100,115.

[2] MITCHELL MCCALLISTER D,MCCULLOUGH R,JOHNSON P,et al. An economic analysis on the transition to dryland production in deficit-irrigated cropping systems of the texas high plains[J]. Frontiers in Sustainable Food Systems,2021,5:531601.

[3] 晔沙. 数据缺失及其处理方法综述[J]. 电子测试,2017(18):65-67,60.

[4] 熊中敏,郭怀宇,吴月欣. 缺失数据处理方法研究综述[J]. 计算机工程与应用,2021,57(14):27-38.

[5] HONG S H, WANG L, TRUONG T K, et al. Novel approaches to the parametric cubic-spline interpolation[J]. IEEE Transactions on Image Processing,2012,22(3):1233-1241.

[6] 周庆平,谭长庚,王宏君,等. 基于聚类改进的KNN文本分类算法[J]. 计算机应用研究,2016,33(11):3374-3377,3382.

[7] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//NIPS´14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014: 2672-2680.

[8] YOON J,JORDON J,SCHAAR M. Gain: missing data imputation using generative adversarial nets[C]// Proceedings of the International Conference on Machine Learning, Stockholm, Sweden: PMLR, 2018: 5689-5698.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[10] BAI S, KOLTER J Z, KOLTUN V. An empirical

evaluation of generic convolutional and recurrent networks for sequence modeling [EB/OL]. arXiv: 1803.01271, 2018.

[11] 闫佰忠,孙剑,王昕洲,等. 基于多变量LSTM神经网络的地下水水位预测[J]. 吉林大学学报(地球科学版),2020,50(1):208-216.

[12] WANG Z, LOU Y. Hydrological time series forecast model based on wavelet de-noising and ARIMA-LSTM[C]//Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019:1697-1701.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates, Inc., 2017: 5998-6008.

[14] WUNSCH A, LIESCH T, BRODA S. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX)[J]. Hydrology and Earth System Sciences, 2021, 25(3):1671-1687.

[15] 郑丽,白宝玉. 基于时间序列的基坑地表沉降预测模型及其应用[J]. 长春师范大学学报,2015,34(10):75-80.

Restoration and application of groundwater level data based on LSTM-TCN

Yuan Zhihong, Chen Yu*

(College of Electronics and Information, Sichuan University, Chengdu 610065, China)

Abstract: Accurate prediction of groundwater level is an important basis for rational development and utilization of groundwater resources. Due to the possible interference of the environment or the failure of the acquisition equipment in the process of data acquisition, the data is missing, which affects the prediction accuracy of the model. In order to obtain more accurate prediction results on groundwater level data with more missing values, a groundwater level restoration model based on long short term memory (LSTM)-temporal convolutional network (TCN) is proposed to improve the quality of data sets by learning the time series and distribution characteristics of the original data sets. Then, the multi-head attention mechanism (MA) and LSTM are combined to predict the groundwater level to further improve the prediction performance of LSTM model. Finally, the prediction results show that using the LSTM-TCN method to train the MA-LSTM model can significantly improve the prediction accuracy of groundwater level.

Keywords: data imputation; generative adversarial imputation networks; temporal convolutional network; groundwater level prediction; multi-head attention mechanism

文章编号: 1007-1423（2023）08-0027-07 DOI: 10.3969/j.issn.1007-1423.2023.08.004

面向民航事故报告的异构图摘要模型研究

何元清*，郑 鑫

（中国民用航空飞行学院计算机学院，广汉 618307）

摘要：民航事故跟踪调查报告隐藏着大量民航安全信息，对其进行提炼与理解是后续安全监督与管理的基础。然而当前自动文本摘要模型主要集中在新闻等通用领域，移植到专业领域内会忽略很多民航领域关键信息。为此，提出了一种基于实体要素异构图的抽取式摘要模型EHGA，利用民航事故实体引导跨句关系图构建，提高模型对民航事故报告文本理解力，EHGA模型由三部分构成，分别为民航实体抽取模块、异构图注意力模块和文本筛选模块。首先，依据民航局颁发规范文件标注民航事故报告中实体，在实体抽取模块获得实体节点；随后根据实体与句子重叠构建实体句子边，将文本转换为由句子节点和实体节点构成民航事故异构图；最后采用异构图注意力机制学习句子特征表示，并使用Trigram blocking策略对句子进行排序，在充分保留文本语义信息时降低冗余。以861篇民航事故调查跟踪报告作为语料构建对比试验，结果表明，EHGA模型相比基于预训练模型在ROUGE评价中取得均值5.34%的性能提升。

关键词：民航事故调查跟踪报告；实体节点；异构图网络；图注意力机制；抽取式

0 引言

民用航空事故跟踪调查报告记录事故发生全过程，监管者会通篇阅读报告，提炼总结要旨，为本次事故提出原因分析和安全建议，同时指出监管疏漏之处，确定下一步工作重点。事故报告内容繁杂且专业性极强，目前主要依靠民航领域专家人工编写事故发生原因概要，但面对海量且迅速增长的航空事故跟踪调查报告，仅依靠专家不仅面临效率困境还容易出现分析疏漏。如何快速、深入、准确地整理事故原因提炼事件详情是制约事故报告利用率的关键问题。实现民航事故报告自动摘要可极大减轻专家阅读工作量，自主筛选出重点信息，对提升民航专家工作效率，推理事故影响因素，调整民航监管工作重点具有重大意义。

目前文本摘要技术已经广泛应用于新闻、微博用户发言、商业服务评价等领域。Cheng等[1]在CNN/Daily Mail中检索出大量新闻文章构建出新闻语料库，并为每个句子打上标准标签，以及创建了来自此新闻语料的词汇数据集。Zhou[2]等提出了端到端模型NEUSUM，首次将选择策略融入打分模型中，并在CNN/Daily Mail数据集中达到当时最好效果。Ming Zhong等[3]细分析数据集对神经网络摘要模型的影响因素，探讨通用领域模型移植到专业领域的可能性，展示了充分挖掘数据集以及增添外部知识对模型的重要性。随后，Ming Zhong等[4]提出Summary-level（篇章级）抽取式摘要的思想，即高质量摘要应当整体与原始文档在语义空间上尽可能相似，并在CNN/Daily Mail数据集上得到验证。研究者尽管在摘要领域取得不断进步，但大多集中在新闻领域。Zhong等[5]从美国退伍军人受伤后应激障碍诉讼案例中构建了法律领域摘要数据集，采用深度学习的抽取式方法获取

收稿日期：2023-03-07 **修稿日期：**2023-03-22

基金项目：四川省科技计划项目（2022YFG0027）：面向飞行技术分析与训练评估的民航飞行大数据智能处理关键技术研究与应用；中国民用航空飞行学院面上重点项目（ZJ2021-11）：多源数据驱动航空安全监管态势分析方法研究

作者简介：*通信作者：何元清（1967—），男，湖南常德人，教授，硕士生导师，研究方向为民航信息仿真、算法分析与设计，E-mail：hcaac@163.com；郑鑫（1995—），男，山东滨州人，硕士，研究方向为自然语言处理、深度学习

摘要，但出现语义缺失的问题；程坤等[6]针对中文新闻文本特点提出增加线索词、标题相似度等因素来改进MMR（maximal marginal relevance, MMR）[11]算法；施国梁等[7]提出专利文本领域摘要模型，指出专利文本结构复杂、内容繁多，而当前通用领域下的模型所生成的摘要内容单一重复且不够简洁流畅。以上研究工作主要集中在公共领域，部分研究者开展法律、专利等特殊领域文本摘要，但却难以直接移植模型，须对领域内数据集详细分析。

与上述领域研究相比，民航事故调查报告内容繁杂，包含事故详情，事故原因总结、专家意见等内容，文本中经常出现“飞行器名称、故障名称、零部件名称”等一系列专业词汇，这些内容直接影响摘要生成质量。这些因素导致现有的文本摘要模型在民航事故领域难以取得高质量摘要。因此，面向航空事故跟踪调查报告的自动文本摘要技术成为了切实且紧迫的需求。

针对上述问题，为深度挖掘航空事故跟踪调查报告文中语义关系，融入专业词汇指导摘要生成，提出基于实体要素异构图的抽取式文本摘要模型。基于图神经网络构建实体节点与句子节点多粒度异构图，结合注意力机制构建EHGA（entity heterogeneous graph abstract model）模型。针对EHGA模型有效性实验所采用的文本数据来自各个国家飞行事故调查局发布的民航事故调查跟踪报告，使用真实事故报告中事故详情部分作为输入。实验结果表明，EHGA模型通过引入实体节点内部信息补充，在进行抽取式摘要时能取得较不错的结果。于ROUGE[8]评分体系下显示出较好的摘要抽取能力，极大减轻摘要的冗余程度，扩大摘要信息覆盖范围，相比传统的序列到序列模型，在ROUGE-1、ROUGE-2和ROUGE-L上平均取得6.23%，4.67%和6.01%的性能提升。

1 EHGA模型介绍

针对民航事故调查跟踪报告文本设计基于实体要素异构图的文本摘要方法，总体架构如图1所示。模型分为3个主要部分，分别是实体抽取模块、融合实体要素的异构图注意力模块和句子抽取模块，本节将分别对以上部分进行详细介绍。

1.1 实体抽取模块

为使用实体要素来丰富句子之间的关联关系，采用BERT-BiGRU-CRF模型的方法获取民航事故调查跟踪报告中实体元素，其优势在于结合了BERT[9]模型和BiGRU（Bi-directional gate recurrent unit）模型的优点。BERT是已经在大型文本语料库中训练过的模型，其基于双向Transformer Encoder连接，内部采用多头注意力机制，可以高效获取文本中的语法结构和语义特征；之后利用BiGRU-CRF[10-11]模型标注实体，此模型充分考虑文本中上下文语句的连贯信息，从而使抽取的实体不是独立分类。

图2是实体抽取模块的总体架构图，以“泛非航空A332的黎波里复飞过程中坠毁”为例，

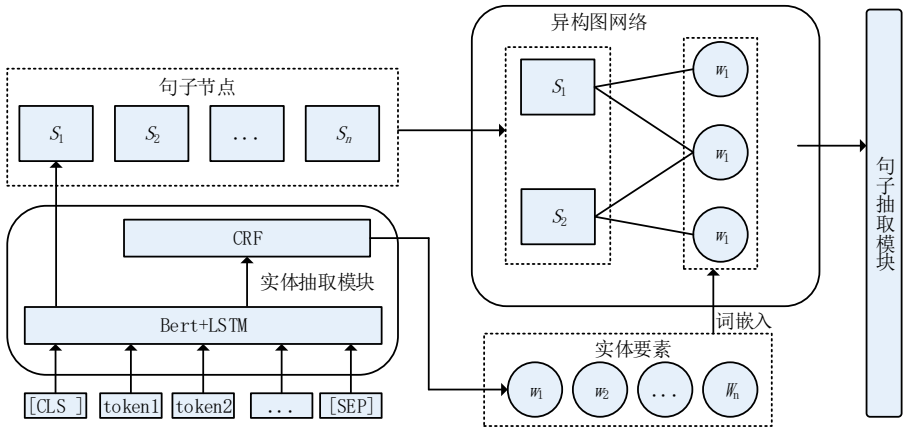


图1 实体要素异构图摘要模型

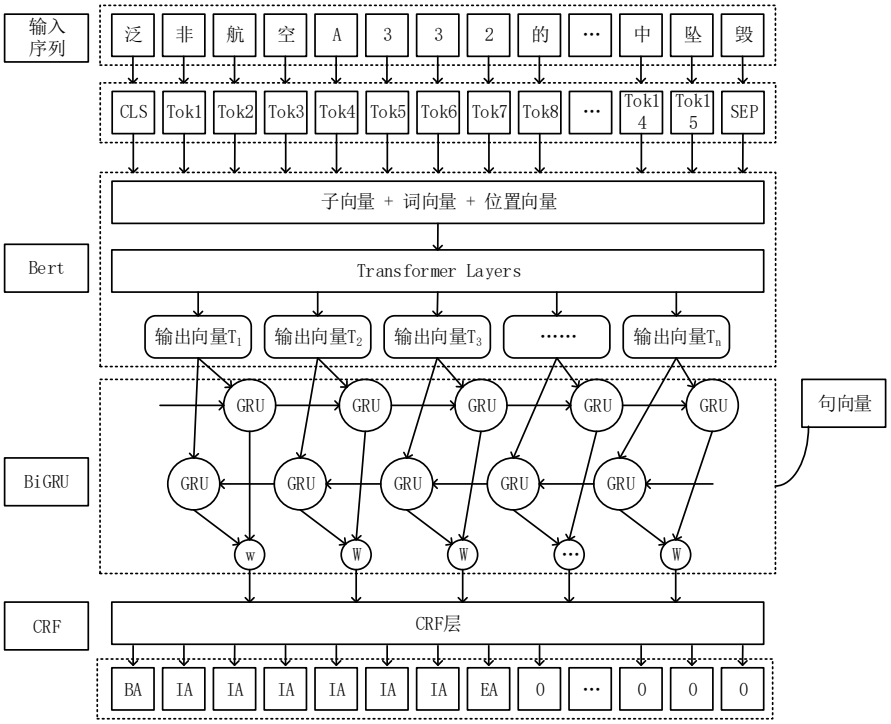


图 2　BERT-BiGRU-CRF模型结构

输入“泛非航空 A332 的黎波里复飞过程中坠毁”在BERT层映射为token值，然后在BERT层进行特征抽取获得输出向量，再到BiGRU层理解上下文语境，经过前后双向传播得到包含双层维度的输出向量，最后经过CRF层计算路径分数最大值，获得准确度最高的标注序列。

在获取实体要素的同时，所输入的文本数据被送入预训练模型BERT中生成对应向量，为更加充分地对上下文语义内容进行学习，将预训练获取到的向量送入基于BiGRU的特征提取层进一步凝合信息，之后便得到图模型所需的句向量。

1.2　异构图构建

传统图模型直接在句子间建立连接，而EHGA模型构建一个多粒度信息的异构图，以实体要素作为句子间的中介节点，由此形成的异构图拥有更丰富的语义信息。在此图中，有两种基本粒度类型节点：句子和实体，其中实体来自实体抽取模块。实体节点作为基本语义节点，代表词级信息；句子节点对应文档句子，代表全局信息。若实体出现在句子中，则将实体节点与句子节点连接，而句子结点间不直接相连，采用TF-IDF值作为边的初始值。

因此给定图 $\boldsymbol{G}=\{\boldsymbol{V},\boldsymbol{E}\}$，其中 $\boldsymbol{V}$ 表示节点集，$\boldsymbol{E}$ 表示节点之间的边，则异构图可以被定义为 $\boldsymbol{V}=\boldsymbol{V}_w\cup\boldsymbol{V}_s$ 和 $\boldsymbol{E}=\{e_{11},\cdots,e_{mn}\}$。其中，$\boldsymbol{V}_w=\{w_1,\cdots,w_m\}$ 表示文档中 m 个唯一实体，$\boldsymbol{V}_s=\{s_1,\cdots,s_n\}$ 对应文档中 n 个句子。$\boldsymbol{E}$ 为边的权重矩阵且 $e_{ij}\neq 0\left(i\in\{1,\cdots,m\},j\in\{i,\cdots,n\}\right)$，其含义为第 j 个句子包含第 i 个实体。

1.2.1　图节点表示

与其他模型相比，BERT通过位置编码和MLM(mask language model)得到符合上下文语境的词向量，使其更符合原文含义，因而采用BERT生成实体向量，句向量则需要充分考虑文本前后文采用BERT+BiGRU训练。令 $\boldsymbol{X}_w\in\mathbb{R}^{m\times d_w}$ 和 $\boldsymbol{X}_s\in\mathbb{R}^{n\times d_s}$ 表示实体向量和句子向量特征矩阵，$\boldsymbol{d}_w$ 表示实体向量维度，$\boldsymbol{d}_s$ 表示句向量维度。经过实体提取模块得到实体向量表示 $\boldsymbol{X}_z\in\mathbb{R}^{p\times d_z}$，可以得到由BERT所学习到的实体语义特征 l_w，和经过BiGRU获取句子级全局特征 g_s，最后经

过Average-polling层拼接，得到句向量的最终表示，具体如下：

$$\boldsymbol{X}_s = ave[l_w; g_s] \tag{1}$$

词节点出现的程度可以衡量文档的冗余程度，实体节点可以聚合更多句子的信息来丰富图结构信息。$\boldsymbol{X}_z \in \mathbb{R}^{q \times d_z}$表示实体节点语义特征矩阵，$q$是实体节点数量，$\boldsymbol{d}_z$是在民航事故调查跟踪报告中抽取实体的特征矩阵维数。

1.2.2 边表示

为进一步概括句子节点之间的关系，定义句子-实体边(如果一个句子包含一个实体)来模拟句子之间存在的丰富联系。句子节点可以通过实体节点建立彼此之间的联系，从全局层面观察全文句子隐含关联，实体与句子构成的边被称为 $\boldsymbol{wTs}$。

由此得到异构图 $\boldsymbol{G} = \{\boldsymbol{V}, \boldsymbol{E}\}$，$\boldsymbol{V} = \boldsymbol{X}_w \cup \boldsymbol{X}_s$，$\boldsymbol{E} = \boldsymbol{wTs}$。

1.3 异构图注意力机制模块

EHGA 模型通过引入图注意力网络(graph attention networks, GAT)[12]来更新语义节点表示，具体表现如图3所示。

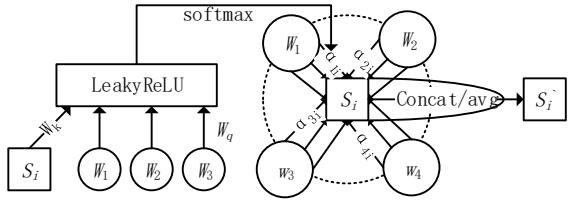


图 3 融入实体节点异构图注意力模型

民航事故调查跟踪报告正文作为输入，第i句的向量表示记为 $\boldsymbol{h}_i \in \mathbb{R}^{d_h}(i \in (1, \cdots, n))$，$\boldsymbol{h}_z \in \mathbb{R}^{d_h}(z \in (1, \cdots, m))$表示实体节点向量，$\boldsymbol{e}_{iz} \in \mathbb{R}^{n \times m}$表示实体节点与句子节点的边特征矩阵，则整个图注意力层设计如下：

$$\gamma_{iz} = LeakyReLU\left(W_a[W_q\boldsymbol{h}_i; W_k\boldsymbol{h}_z; e_{iz}]\right) \tag{2}$$

$$\alpha_{iz} = \frac{\exp(\gamma_{iz})}{\sum_{z \in N_i}\exp(\gamma_{iz})} \tag{3}$$

其中，W_a,W_q,W_k是可训练参数，γ_{iz}表示句子节点i与实体节点z之间的注意力权重计算，EHGA 模型对γ_{iz}进行归一化操作得到α_{iz}便于不同句子节点的重要性比较，如公式(3)。对于句子节点$\boldsymbol{h}_i$与其他相连的所有实体节点$\boldsymbol{h}_z$进行信息聚合，GAT层整体运算过程如以下表达式所示：

$$\mu_{iz} = \sigma\left(\sum_{c \in N_i}\alpha_{iz}W_v\boldsymbol{h}_c\right) \tag{4}$$

其中$\boldsymbol{\mu}_i$是句子节点$\boldsymbol{h}_i$在其所有邻接实体节点上学习到的向量表示，因此也具有特定的语义信息。为了在学习过程中提取更多特征，EHGA 模型采用多头注意力机制，如下所示：

$$\boldsymbol{\mu}_{iz} = \Big\|_{k=1}^{K}\sigma\left(\sum_{c \in N_i}\alpha_{iz}^k W^k\boldsymbol{h}_i\right) \tag{5}$$

考虑到图神经网络常见的过渡平滑以及梯度消失问题，EHGA 模型参考 transformers 中残差连接设计，避免因迭代次数过多而引起的梯度消失问题。因此在图注意力网络中句子节点$\boldsymbol{h}_i$的特征向量表示为

$$\boldsymbol{h}'_{iz} = \boldsymbol{\mu}_{iz} + \boldsymbol{h}_i \tag{6}$$

在每个图注意力层后，引入一个前馈网络(FFN)层对特征进行进一步压缩，获得最终的句子稠密向量表示，其计算过程如下。

$$\boldsymbol{H}_{iz} = FFN\left(\boldsymbol{E}_s + GAT(\boldsymbol{E}_s)\right) \tag{7}$$

1.4 句子选择模块

在真实句子选择时，往往会出现句子级分数较低但是整体摘要分数较高的情况，为了保证最终摘要结果的可读性和重要信息的覆盖度，EHGA 模型采用 Trigram blocking 策略。对所有候选句子依据概率排序，依次选择概率最高的句子，如果被选择的句子与当前摘要存在三元组重叠(trigram overlapping)[13]，则认为其冗余，反之则将其加入摘要，并从剩余候选句子中排除此句，反复进行以上操作直到满足摘要所设定的长度阈值。EHGA 模型采用交叉熵作为损失函数衡量真实摘要和预测结果之间的距离，损失函数公式为

$$L=-\sum_{s_i \in D} y_i \ln p(\boldsymbol{h}_i) + (1 - y_i)\ln\left(1 - p(\boldsymbol{h}_i)\right) \tag{8}$$

其中：y_i表示对应句子$\boldsymbol{h}_i$的真实标签，$\boldsymbol{y}_i = 1$表示第i个句子应该包含在摘要中。

2　实验及结果分析

2.1　数据集及实验环境介绍

2.1.1　数据集构建

本次实验使用来自各个国家飞行事故调查局所发布的民航事故调查跟踪报告数据集，包含由2010—2016年世界各地民航事故调查跟踪报告共861篇，并对文本进行清洗、标注，构建航空事故报告数据集，有效数据842对，数据集统计结果如表1所示。

表 1　数据集统计

数据集划分	报告文本数/篇	平均文本长度/字	平均摘要长度/字
训练集	642	783	210.4
验证集	100	812	180.6
测试集	100	764	201.2

实体数据集则通过民航局颁布的《民用航空器事故征候》《民用航空器征候等级划分办法》《事件样例》《民用航空器事故和飞行事故征候调查规定》《民用航空安全信息管理规定》等规范性文件确定实体名称，包括航空事件、航空事件原因、航空地面事件等类型规范实体名和报告中一些不规范实体名称，因此实体要素对于摘要的生成具有科学性与准确性。

2.1.2　实验环境介绍

本次实验CPU使用Intel Core I9-10900X，内存96 GB，GPU为Nvidia GeForce RTX 3090 24 GB一块。采用深度学习框架PyTorch，实验环境PyCharm，Python 3.8版本。EHGA模型使用预训练语言模型BERT初始化句子节点表示，其词向量的维度是768。对于实体的选择，每个文档选择前10个关键短语。在异构图注意力模块设置头数$K=8$。每个头中句子节点的隐藏向量维度为128，最终连接节点向量的维度为768。采用ROUGE（Recall - Oriented Un - derstudy for Gisting Evaluation）中的RG-1、RG-2和RG-L。

在训练过程中，实验设置训练的批量大小为32，训练轮次24，使用Adam优化器，设置学习率为5e-4。

2.2　实验结果分析

2.2.1　基准模型

为证明EHGA模型的有效性，将其与几个优秀的文本摘要模型进行比较。

（1）Lead-n：选取文中前n个句子作为文章摘要，常用于新闻领域。

（2）TextRank[14]：以句子间相似度构建图模型。

（3）Summer RuNNer[15]：是基于序列分类器的循环神经网络对句子分类训练模型，采用两层双向GRU（gate recurrent unit）和循环神经网络（recurrent neural network，RNN）来对句子进行编码。

（4）BERTSum[16]：采用预训练模型BERT获取文档中每个句子的句向量编码，通过贪心策略选择最优的top-n个句子。

2.2.2　模型检测

本实验以ROUGE评分体系作为文本自动摘要的评价标准，采用ROUGE-N（N为n-gram），ROUGE-L，ROUGE-S等数值作为对当前所得摘要的评价，其计算方式如下：

$$R_n=\frac{\sum_{s\in\mathbb{R}}\sum_{n\text{-gram}\in\mathbb{R}}Count_{\text{match}}(N\text{-gram})}{\sum_{s\in\mathbb{R}}\sum_{n\text{-gram}\in\mathbb{R}}Count(N\text{-gram})}\tag{9}$$

2.2.3　基线模型结果分析

本组实验是在民航事故调查跟踪报告数据集上进行EHGA模型与上述4个模型对比，结果如表2所示。可以看到，EHGA模型与其他模型相比ROUGE指标提升显著，证明EHGA具有更好的摘要效果。

表 2　基线模型对比试验

模型	ROUGE-1	ROUGE-2	ROUGE-L
Lead-n	18.54	14.21	14.76
TextRank	30.13	14.65	16.52
Summa RuNNer:	29.43	16.12	29.34
BERTSum	32.42	15.02	30.14
EHGA	**36.10**	**18.63**	**35.54**

EHGA模型通过采用异构图来融入内部信

息实体节点，可以有效丰富模型的语义信息，提高摘要性能，并且依照实体更贴近原文内容；同时图结构可以跨越简单上下文的关系而获得更远距离的语义信息，对抽取处更贴近原文的句子具有指导作用。与Lead-n模型相比，选取前n句作为摘要时更适合有总结句的文本，而航空事故报告是平铺叙，显然不适合。与TextRank相比，EHGA模型以实体要素作为句子关键程度的指标，重点关注的是句子，而TexkRank更加关注关键词，偏离原文主旨。与SummaRuNNer相比，EHGA模型引入实体要素辅助模型理解文本含义，而SummaRuNNer则只依靠神经网络学习文本特征，使得模型会过分关注某一方面而造成文本冗余。

2.2.4 消融实验

为验证EHGA模型中各个模块的效果而开展了消融实验，实验结果如表3所示。EHGA是在图神经网络(GNN)的基础上增加了实体要素节点，效果较GNN在ROUGE-1，ROUGE-2和ROUGE-L 3种评价指标上均有明显提高。说明增加实体要素可以使模型尽可能关注到与实体相关的句子，而达到专有名词指导文本摘要生成效果。其实验结果如表3所示。

表3 消融实验对比

模型	ROUGE-1	ROUGE-2	ROUGE-L
TextRank	30.13	14.65	18.52
GNN	33.42	16.02	30.14
EHGA	**37.10**	**19.63**	**35.54**

2.2.5 案例展示

为进一步展示EHGA模型的实验效果，以“美国航空公司MD-82飞机圣路易斯发动机起火事故调查报告”为例对输出摘要进行评价分析，具体如表4。

通过表4的样例展示，EHGA模型依据实体要素能够较为全面地概括报告内容，实体要素的加入能够在图注意力机制中筛选出与事故关联性强的句子，能够扩大信息的覆盖范围，关注不同层次信息，多维度概括文本内容，因此可见实体信息对于文本摘要的生成具有指导意义。

表4 抽取式摘要样例展示

实体要素	报告内容
事故调查报告正文	一架美国航空公司MD-82型客机(注册号N454AA)执行AA-1400航班准备从圣路易斯机场飞往芝加哥奥黑尔机场，机上共有138名乘客和5名机组人员。在准备起飞时，左侧发动机不能启动。维修人员被要求参与处理这一问题，他们试图手动启动发动机，但是没能成功。在第二次尝试手动启动发动机后，发动机达到了正常的运行状态，在滑行和起飞滑跑中各种参数也全部正常……
实体	美国航空公司、MD-82型客机、圣路易斯机场、发动机、火警、机组人员、飞行显示器、电力系统……
摘要	引起该事故的可能原因是美国航空公司的维修人员使用了不恰当的发动机手动启动程序，导致左侧发动机空气涡轮起动器控制阀(ATSV)非指令打开，以及随之而来的左侧发动机起火。由于机组人员在执行紧急检查单程序时被非关键任务打断，起火问题的解决受到了拖延。促成这一事故的因素是美国航空公司的持续监视和分析系统项目(CASS)有缺陷
EHGA摘要	一架美国航空公司MD-82型客机，在准备起飞时，左侧发动机不能启动。机组人员收到左侧发动机(JT8D)的火警报告。机组人员拉动发动机灭火手柄，关闭发动机，然而火情仍然继续。在复飞过程中，电力系统问题和前起落架无法放下的问题仍然没有解决。副驾驶执行了紧急起落架展开检查单，机长总结他们全部的液压系统失效

3 结语

EHGA模型针对民航事故调查跟踪报告，提出基于实体要素异构图注意力机制抽取式摘要模型。把词语、实体和句子构建为异构图，以注意力机制获得句子重要程度，联合评分机制获得最终摘要。实验证明，针对事故报告这一特定领域的摘要任务，融入实体要素能够提升摘要选择覆盖度和准确性，生成高质量摘要。同时也验证了，基于异构图网络进行文本数据分析，更加关注句子间隐含的深层关系。

同样，在研究过程中发现人工摘要存在大量总结式、概括式和推理式词语，这些无法在原文中找到对应，无疑给抽取式摘要带来极大的挑战。因此在下一步研究中，拟继续在异构图中添加更多外部知识，提升摘要性能。

参考文献：

[1] CHENG J P, MIRELLA L. Neural summarization by extracting sentences and words[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 484-494.

[2] ZHOU Q Y, YANG N, WEI F R, et al. Neural document summarization by jointly learning to score

and select sentences[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers),2018:654-663.
[3] ZHONG M,WANG D Q,LIU P F,et al. A closer look at data bias in neural extractive summarization models[C]//Proceedings of the 2nd Workshop on New Frontiers in Summarization,2019b:80-89.
[4] ZHONG M,LIU P,CHEN Y,et al. Extractive summarization as text matching[EB/OL]. arXiv:2004.08795,2020.
[5] ZHONG L,ZHONG Z,ZHAO Z,et al. Automatic summarization of legal decisions using iterative masking of predictive sentences[C]//Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law,2019:163-172.
[6] 程琨,李传艺,贾欣欣,等. 基于改进的MMR算法的新闻文本抽取式摘要方法[J]. 应用科学学报,2021,39(3):443-455.
[7] 施国良,周抒,王云峰,等. 基于改进多头注意力机制的专利文本摘要生成研究[J/OL]. 数据分析与知识发现：1-18[2023-03-25]. http://kns.cnki.net/kcms/detail/10.1478.G2.20221109.1511.006.html.
[8] LIN C Y. ROUGE:a package for automatic evaluation of summaries[C]//Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain,2004:74-81.
[9] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding[EB/OL]. arXiv: 1810.04805, 2018.
[10] ZHU Y,WANG G,KARLSSON B F. CAN-NER:convolutional attention network for Chinese named entity recognition[EB/OL]. arXiv: 1904.02141, 2019.
[11] SHUANG Z,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. arXiv:1508.01991,2015.
[12] PETAR V, CUCURULL G, CASANOVA A. Graph attention networks[EB/OL]. arXiv: 1710.10903,2017.
[13] LIU Y,LAPATA M. Text summarization with pretrained encoders[EB/OL]. arXiv:1908.08345,2019.
[14] WANG H Y,CHANG J W,HUANG J W. User intention-based document summarization on heterogeneous sentence networks[C]//Proceedings of the 24th International Conference on Database Systems for Advanced Applications, DASFAA 2019, Chiang Mai,Thailand,Part II 24. Springer International Publishing,2019:572-587.
[15] LIU Y,TANG M,DO Y,et al. Accurate ranking of influential spreaders in networks based on dynamically asymmetric link weights[J]. Physical Review E, 2017,96(2):022323.
[16] LIU Y. Fine-tune BERT for extractive summarization[EB/OL]. arXiv:1903.10318,2019.

Research on heterogeneous graph abstract model for civil aviation accident report

He Yuanqing*, Zheng Xin

(College of Computer Science, Civil Aviation Flight University of China, Guanghan 618307, China)

Abstract: The civil aviation accident tracking investigation report hides a large amount of civil aviation safety information, and refining and understanding it is the basis for subsequent safety supervision and management. However, the current automatic text summarization model is mainly concentrated in general fields such as news, and transplanting it to professional fields will ignore many key information in the civil aviation field. To this end, an abstract model EHGA(Entity Heterogeneous Graph Abstract Model) based on heterogeneous graph of entity elements is proposed, which utilizes the construction of cross sentence relationship diagrams guided by entities of civil aviation accidents to improve the model's understanding of civil aviation accident report text. The EHGA model is composed of three parts, namely, civil aviation entity extraction module, heterogeneous graph attention module, and text filtering module. Firstly, label the entities in the civil aviation accident report according to the specification documents issued by the Civil Aviation Administration, and obtain the entity nodes in the entity extraction module; Subsequently, an entity sentence edge is constructed based on the overlap of entities and sentences, and the text is converted into sentence nodes and entity nodes to form a heterogeneous graph of civil aviation accidents; Finally, a heterogeneous graph attention mechanism is used to learn sentence feature representation, and a Trigram blocking strategy is used to sort sentences, reducing redundancy while fully retaining text semantic information. A comparative experiment was conducted using 861 civil aviation accident investigation and tracking reports as corpus. The results showed that the EHGA model achieved an average performance improvement of 5.34% in ROUGE evaluation compared to the pre training model.

Keywords: civil aviation accident investigation and tracking report; entity node; heterogeneous graph network; graph attention mechanism; extraction

文章编号: 1007-1423（2023）08-0034-06 DOI: 10.3969/j.issn.1007-1423.2023.08.005

基于改进TF-IDF的电梯传媒广告推荐方法

陈彦彬[1, 2*]，杨泽华[2]，薛晓桂[3]，黄锦钿[4]

（1. 揭阳职业技术学院实训与信息中心，揭阳 522051；2. 广东博华科技有限公司工程技术研究中心，揭阳 522000；
3. 揭阳职业技术学院信息工程系，揭阳 522051；4. 韩山师范学院物理与电子工程学院，潮州 521000）

摘要：随着国家个人信息保护法的出台，电梯传媒终端精准广告投放面临多方面挑战，如何在不采集市民隐私信息的情况下，提高电梯传媒终端广告投放的精准度，为广告投放商提高经济效益是当前终端计算广告研究的重点，为此提出了基于改进TF-IDF的电梯传媒广告推荐方法。利用改进的TF-IDF对电梯点周边POI、居民等情况进行标签提取，构建了电梯传媒终端标签向量模型；然后利用商户对电梯的评分计算商户对标签兴趣度，最后构建商户兴趣模型对TOP-*N*部电梯终端广告位进行排序，推荐给商户。实验结果表明，该推荐方法的准确率、召回率等均优于采用传统TF-IDF算法的结果，而且不用采集市民个人隐私数据，具有较强推广应用效益。

关键词：词频–逆文本频率指数；传媒广告；标签评分；用户兴趣；个性化推荐

0 引言

大数据和人工智能技术的深入发展，极大地方便了人们的生产生活，但个人隐私信息也面临着泄露的困扰。网络分众传媒广告和互联网广告大多针对市民信息的收集、分析进行精准广告投放和推送。随着国家个人信息保护法的出台，传统网络广告投放、推送存在一定风险。但是电梯传媒终端广告领域由于信息传递的单向性，存在广告效率低下、广告资源浪费等问题[1]，使得电梯传媒终端广告技术发展缓慢，产业效益低下。

提高电梯传媒终端广告效率的方法中，构建个性化推荐系统最为常用[2]。推荐系统中最重要的是推荐算法[3]，林振荣等[4]提出基于TF-IDF（词频–逆文本频率指数）与用户聚类的推荐算法，在计算相似度中通过物品特征的TF-IDF值改进用户评分数据，从而生成推荐列表。付小飞[5]提出利用VSM算法构建用户画像实现对用户网络行为分析，并提出了混合推荐算法。岳志鹏[6]提出利用TF-IDF对微博用户进行兴趣关键词提取，进而挖掘用户兴趣。Ma等[7]提出对多源数据进行融合的方法，形成代表用户兴趣的关键词集合。

综上，目前在计算广告领域的推荐算法大多数集中于对用户行为进行挖掘，而对于电梯传媒广告来说，在用户不共享个人信息的情况下很难采集用户属性、行为等信息。另一方面，为提高TF-IDF值计算精度，需要对该算法进行改进。基于上述的思考，本文利用商家评分等行为信息，建立相关推荐算法。

本文提出了改进的TF-IDF，对电梯点周边POI、居民等情况进行标签提取，构建了电梯传媒终端标签向量模型；通过分析商户对电梯广告效果的评分行为，计算商户对标签喜好程度

收稿日期：2022-12-09 **修稿日期：**2022-12-24

基金项目：广东省科技厅驻镇帮镇扶村农村科技特派员重点派驻任务（KTP20210400）；广东省普通高校特色创新项目（2018KTSCX139）；揭阳市科技计划项目（2019070）

作者简介：*通信作者：陈彦彬（1987—），男，广东揭阳人，本科，学士，高级工程师，主要从事领域为电子信息系统开发、人工智能等嵌入式技术教学与科研，E-mail：chenyanbin01@126.com；杨泽华（1967—），男，广东揭阳人，本科，总经理，技术员，主要研究方向为计算机技术应用；薛晓桂（1987—），男，广东汕头人，硕士研究生，高级工程师，主要从事领域为大数据、人工智能等技术教学与科研；黄锦钿（1983—），男，广东揭阳人，博士，副教授，主要研究方向为工业工程等

和依赖程度，从而计算商户对标签的兴趣度；利用兴趣度与终端标签向量构建商户兴趣模型，在进行降序排序后将计算结果的TOP-N部电梯数据作为推荐列表。本文通过三组对比实验对算法模型进行验证分析。

1　基于改进TF-IDF标签体系提取算法

基于改进TF-IDF标签体系提取算法的思路是通过数据标注、商户基本信息以及电梯点周边POI等建立电梯传媒终端标签数据库，然后通过改进TF-IDF标签体系提取算法，提取出用于描述电梯传媒终端的标签体系关键词。

1.1　建立电梯传媒终端标签体系

“标签化”就是依据现有电梯传媒终端基本信息、电梯点周边POI以及商户评分数据等信息，通过算法构建电梯传媒终端标签体系，实现标签体系的“低交叉率”[5, 8-9]。建立电梯传媒终端的标签体系的常用技术包括提取关键词、识别去除停用词、词义扩充、词类扩充、重组等[5]，通过设置阈值过滤权重较小的标签，最终转化为表示电梯传媒终端的关键词集合。

电梯传媒终端的标签体系采用“层级法”，共设置三级，具体框架如图1所示。提取关键词的来源主要为各个电梯传媒终端的基本属性信息、周边POI以及商户发布的各个广告主题标题等。采集到的原始数据经过预处理、数据备份等数据清理环节后，形成唯一用户数据。

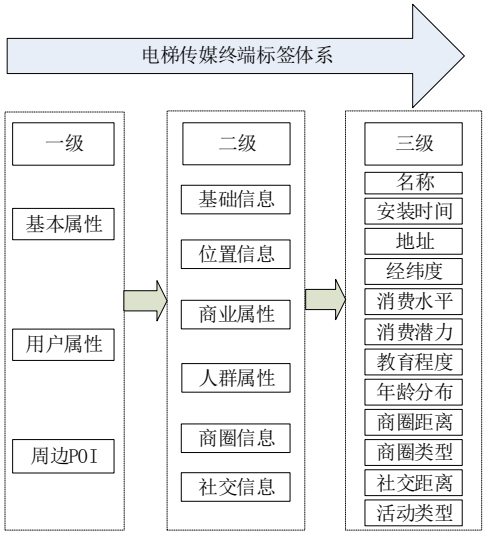


图1　电梯传媒终端标签体系框架

1.2　改进TF-IDF算法

TF-IDF是一种统计方法，常用于评估一个词语在文本数据或者文本数据集中的重要程度。TF-IDF的计算公式如下：

$$TF\text{-}IDF = TF_{i,j} * IDF_i = \frac{n_{i,j}}{\sum_k n_{k,j}} * \log\frac{N}{n} \quad (1)$$

式中$TF_{i,j}$表示候选关键词i在电梯j标签文本中出现的频率，IDF_i表示逆向文本频率指数，用于衡量候选关键词在整个数据集词条库中的区分能力。$n_{i,j}$表示候选关键词i在电梯j标签文本中出现的次数，$\sum_k n_{k,j}$表示电梯j标签文本词条的数量；N表示电梯标签文本集合的文本数，n表示电梯标签文本集合中包含电梯j标签文本中候选关键词i的电梯标签文本数。

由公式(1)可知，TF-IDF正比于候选关键词在电梯标签文本中出现的次数$n_{i,j}$，而反比于电梯标签文本中候选关键词的电梯标签文本数n。但在电梯传媒终端广告领域，存在有些候选关键词在文本集合中很少出现，n值很小，但计算的TF-IDF值很高的情况；甚至有时在所有标签文本中都没有出现。针对以上问题，本文对TF、IDF两部分的计算分别进行改进优化。

(1) 计算候选关键词i在电梯j标签文本集合中出现的总次数n_i，判断是否大于设定的标签过滤阈值n_0(本文中n_0=2)，若是，则按照公式(2)计算$TF_{i,j}$值；否则$TF_{i,j}$值为0，具体计算公式如下：

$$TF_{i,j} = \begin{cases} 0, n_i < n_0 \\ \dfrac{n_{i,j}}{\sum_k n_{k,j}}, n_i \geqslant n_0 \end{cases} \quad (2)$$

(2) 对IDF改进为电梯标签文本集合候选关键词总数与候选关键词在待分析标签文本中出现的次数比求对数，具体计算公式如下：

$$IDF_i = \log\frac{\sum_{i=1}^{m} n_i}{1 + n_i} \quad (3)$$

式中$\sum_{i=1}^{m} n_i$表示所有候选关键词在电梯标签文本集合中出现的次数总和；n_i表示电梯标签文本集合中候选关键词i出现的总次数；1是对分母为0的平滑处理。

综合TF、IDF两部分可以得到改进后的公式如下：

$$TF\text{-}IDF=\begin{cases}0, & n_i<n_0\\ \dfrac{n_{i,j}}{\sum\limits_k n_{k,j}}*\log\dfrac{\sum\limits_{i=1}^{m}n_i}{1+n_i}, & n_i\geqslant n_0\end{cases}\quad(4)$$

1.3 采用改进TF-IDF提取标签

在电梯传媒终端标签体系框架基础上，结合已有各类电梯传媒终端数据(基本属性、用户属性、周边POI以及各商户广告主题等)，用于本文提出的改进算法，可以提高计算精准度，防止总体权值过小的问题。标签提取算法的具体步骤：

输入：电梯传媒终端候选关键词数据集；

输出：电梯传媒终端标签关键词集合I。

(1) 设定标签过滤阈值n_0($n_0=2$)，判断候选关键词i在电梯标签文本集合中出现的总次数n_i是否小于n_0。若是，则$TF_{i,j}$=0；

(2) 计算候选关键词i的$TF_{i,j}$值；

(3) 计算所有候选关键词在电梯标签文本集合中出现的总次数和电梯标签文本集合中候选关键词i出现的总次数，并计算IDF_i值；

(4) 计算候选关键词i的TF-IDF值；

(5) 重复步骤(1)至步骤(4)，直到计算出所有候选关键词的TF-IDF值；

(6) 按照所有候选关键词的TF-IDF值降序排序，TOP-N个能够符合电梯传媒终端标签体系框架的关键词作为标签集合I。

2 电梯传媒广告推荐算法

张辉等[10]以及李尧[11]所提出的精准广告推荐方法均是从广告受众角度出发，而且需要采集受众位置、监控视频等个人信息。本文利用商户对电梯传媒终端的评分，构建商户兴趣模型，结合上文提取的电梯传媒终端标签向量，形成新的推荐方法。

2.1 相关数据集

为建立商户对标签的兴趣度，需要收集商户对电梯传媒终端的评价数据，从而计算出商户的标签兴趣度。

假设收集到s个商户对e个电梯传媒终端的评分数据，则可以构建商户-终端评价数据集。若该商户没有对终端进行评价，其值为0。否则为具体的评价分数。具体数据集见表1。

表1 商户-终端评价数据集

	终端1	终端2	终端3	终端4	…	终端e
商户1	3	0	5	2	…	2
商户2	4	4	0	5	…	1
商户3	5	2	1	4	…	0
…	…	…	…	…	…	…
商户s	0	4	3	2	…	5

2.2 计算商户对标签兴趣度

2.2.1 计算商户对标签的喜爱程度

根据上文提取到的终端标签集合的TF-IDF值，结合商户对终端的评分数据，可以计算出商户s对标签t的喜爱程度，具体公式如下：

$$rate(s,t)=\frac{\sum\limits_{e\in I_e}rate(s,e)*rel(e,t)+\bar{r}_s*k}{\sum\limits_{e\in I_e}rel(e,t)+k}\quad(5)$$

式中$rate(s,t)$表示商户s对标签t的喜爱程度，$rate(s,e)$表示商户s对终端e的评分，$rel(e,t)$表示终端与标签的相关程度(文中为标签的TF-IDF值)，$\bar{r}_s$为商户s的所有评分的平均值，k为平滑因子，目的是减少商户评分行为引起的预测误差，I_e电梯传媒终端集合。

2.2.2 计算商户对标签的依赖程度

商户对标签的喜爱程度是商户的主观表现，需要引入商户对标签的依赖程度对算法进行改进。商户对标签的依赖程度利用本文提到的改进TF-IDF算法进行计算。假设T为商户使用的标签集合，通过式(2)计算商户s使用标签t的TF值，记为$TF(s,t)$。式(2)中的分子表示商户s使用标签t的次数，记为$n(s,t)$；分母表示商户s使用所有标签的总次数，记为$\sum\limits_{t_i\in T}n(s,t_i)$。

为了惩罚热门标签越来越热，利用式(3)计算每个商户使用标签的IDF值，记为$IDF(s,t)$。假设S为商户集合，式(3)中的分子表示所有商户对所有标签的使用总次数，记为$\sum\limits_{s_i\in S}\sum\limits_{t_j\in T}n(s_i,t_j)$；分母表示所有商户对标签$t$的使

用总次数，记为$1+\sum_{s_i\in S}n(s_i,t)$。

综合$TF(s_i,\ t)$和$IDF(s_i,\ t)$，可得商户对标签的依赖程度计算公式为

$$D(s,t)=TF(s,t)*IDF(s,t) \tag{6}$$

可得，

$$D(s,t)=\frac{n(s,t)}{\sum_{t_i\in T}n(s,t_i)}*\log\frac{\sum_{s_i\in S}\sum_{t_j\in T}n(s_i,t_j)}{1+\sum_{s_i\in S}n(s_i,t)} \tag{7}$$

2.2.3　计算商户对标签的兴趣度

综合式(5)和式(7)，可以得到商户s对标签t的兴趣度计算公式为

$$P(s,t)=rate(s,t)*D(s,t) \tag{8}$$

2.3　构建广告推荐算法

本文基于电梯传媒终端标签向量与商户对标签的兴趣度，构建商户兴趣模型，形成电梯传媒广告推荐算法。

2.3.1　构建终端–标签矩阵

基于上文方法对各个终端提取标签的TF-IDF值，可以构建终端–标签矩阵，如公式(9)：

$$T_e=\begin{bmatrix}t_{11} & t_{12} & \cdots & t_{1m}\\ t_{21} & t_{22} & \cdots & t_{2m}\\ \vdots & \vdots & \vdots & \vdots\\ t_{n1} & t_{n2} & \cdots & t_{nm}\end{bmatrix} \tag{9}$$

式中n、m表示终端和标签的个数，元素t_{ij}($1\leqslant i\leqslant n$, $1\leqslant j\leqslant m$)代表终端i中标签j的TF-IDF值。如果终端没有相应标签，则其值为0。

2.3.2　构建商户–标签兴趣度矩阵

根据式(8)可以计算出商户对每个标签的兴趣度，构成了1行m列的商户–标签兴趣度矩阵，如公式(10)：

$$T_s=\begin{pmatrix}t_{11} & t_{12}\cdots & t_{1m}\end{pmatrix} \tag{10}$$

式中m表示标签的个数，元素t_{1j}($1\leqslant j\leqslant m$)代表商户对标签j的兴趣度。

2.3.3　商户兴趣建模

基于式(9)和式(10)，可以计算出商户对所有电梯传媒终端的兴趣程度，具体公式为

$$T(s,e)=T_s*T_e^T \tag{11}$$

对式(11)的计算结果进行降序排序，取TOP-N部电梯传媒终端给商户作为候选广告位，实现广告的精准投放。

综上，基于改进TF-IDF的电梯传媒广告推荐算法流程如图2所示。

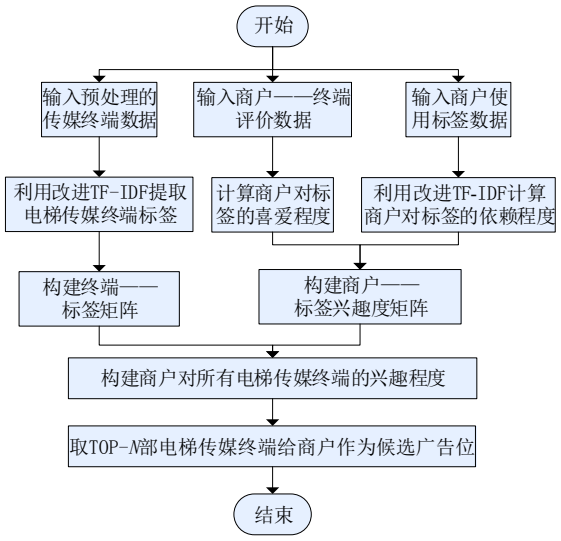


图2　电梯传媒广告推荐算法流程

3　实验分析与结果

3.1　实验设置

3.1.1　实验数据集

读取电梯传媒广告系统数据，经过预处理后作为实验数据集。数据集包括电梯传媒终端的基本属性信息、周边POI以及商户发布的各个广告主题标题等信息，总共99980条数据。数据集是商户、电梯传媒终端、评分、广告主题和用户使用标签数据集等各种类型信息数据集合，其概要信息见表2。实验数据集分为训练集70%和测试集30%，采用K折交叉验证进行训练，对计算结果采用TOP-N方法进行推荐。

表2　实验数据集概要信息

数据	数量/条
商户	14545
电梯传媒终端	2450
评分	37786
广告主题	44509
用户使用标签	690

3.1.2 评价指标

推荐系统的效果评估通常采用离线评估指标，主要有精确率(Precision)、召回率(Recall)和F测度值(F-measure)。计算公式分别如下：

$$Precision = \frac{|I \cap A|}{|I|} \tag{12}$$

$$Recall = \frac{|I \cap A|}{|A|} \tag{13}$$

$$F\text{-measure} = \frac{2*Precision*Recall}{Precision + Recall} \tag{14}$$

上述式(12)、式(13)和式(14)中，I表示通过算法推荐的TOP-N部电梯传媒终端集合，A表示商户感兴趣的电梯传媒终端集合，$|I|$表示通过算法推荐的TOP-N部电梯传媒终端的数量，$|A|$表示商户感兴趣的电梯传媒终端数量。

3.2 实验结果分析

为了验证基于改进TF-IDF的电梯传媒广告推荐算法的实效性，分别采用本文方法(记为方法1)、基于传统TF-IDF和本文提出的推荐算法(记为方法2)以及采用本文方法但没有考虑商户对标签依赖程度的推荐算法(记为方法3)进行实验对比分析，分别计算出三种算法的精确率、召回率和F测度值，从而比选最优性能的推荐算法。

实验结果如图3、图4和图5所示。通过对比分析可知，基于改进TF-IDF的电梯传媒广告推荐算法在精确率、召回率和F测度值三项指标上明显高于其他两种算法，因此，本文方法具有最优的推荐性能。但是随着N的增大，精确率会出现下降趋势，这是由于推荐数量的增加，导致排序靠后的终端也被推荐出来，但是该终端并不是商户喜爱的。实验表明，N的最优取值为15。

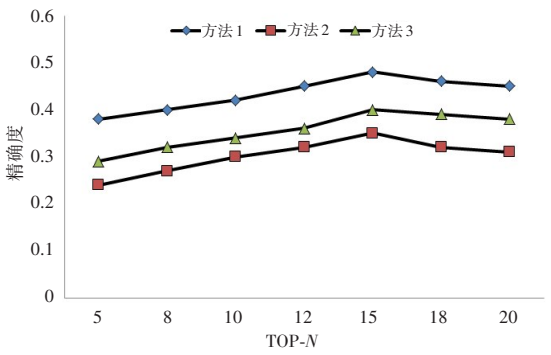


图3 精确率对比

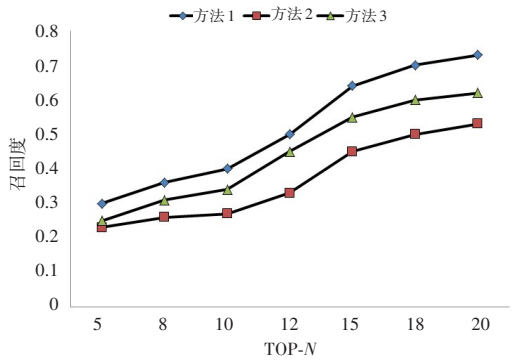


图4 召回率对比

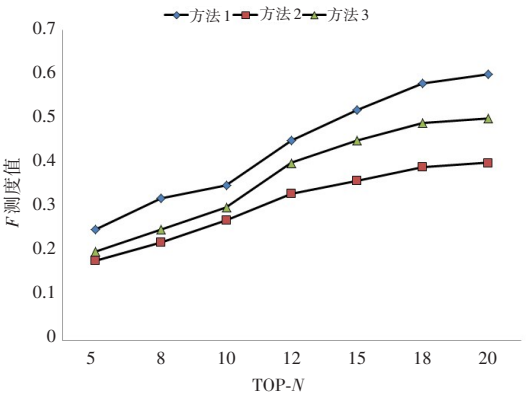


图5 F测度值对比

实验结果说明采用改进TF-IDF的电梯传媒广告推荐方法提高了终端广告推荐的精确率，且能够在不降低精确率、召回率和F测度值的情况下无需采集市民隐私信息就可以实现电梯传媒广告的精准推送，具有较好的实效性。

4 结语

本文提出了基于改进TF-IDF的电梯传媒广告推荐方法，利用改进的TF-IDF建立电梯传媒终端标签向量模型，并构建商户兴趣模型将TOP-N部电梯传媒终端作为候选广告位进行排序推荐。为验证本文提出算法的有效性，以现有广告系统数据集为基础开展对比实验。三个推荐算法的对比实验表明，基于改进TF-IDF的电梯传媒广告推荐方法能充分考虑商户兴趣，在精确率、召回率和F测度值等评价指标上明显优于其他两种算法，具有较好的有效性和适用性。

参考文献：

[1] 赵雅倩，李楠，刘慧敏，等. 广播电视开机广告精准投放系统[J]. 电视技术，2014，38(20)：44-47.

[2] PAL A，PARHI P，AGGARWAL M. An improved content based collaborative filtering algorithm for movie recommendations [C] //Proceedings of the 2017 Tenth International Conference on Contemporary Computing(IC3)，Noida，India. IEEE，2017：1-3.

[3] 秦鹏，贾洪杰，霍兴瀛，等. 融合大数据挖掘的用户个性化POI推荐方法[J]. 计算机仿真，2022，39(6)：355-358.

[4] 林振荣，黄虹霞，舒伟红，等. 基于TF-IDF与用户聚类的推荐算法[J]. 计算机仿真，2022，39(6)：341-345.

[5] 付小飞. 基于用户画像的移动广告推荐技术的研究与应用[D]. 成都：电子科技大学，2017.

[6] 岳志鹏. 用户兴趣点推荐方法和移动广告的投放研究[D]. 成都：西华大学，2016.

[7] MA Y，ZENG Y，REN X，et al. User interests modeling based on multi-source personal information fusion and semantic reasoning[C]//Proceedings of the 7th International Conference on Active Media Technology，Berlin，Heidelberg：Spring-Verlag，2011：195-205.

[8] 付关友，朱征宇. 个性化服务中基于行为分析的用户兴趣建模[J]. 计算机工程与科学，2005，27(12)：76-78.

[9] 毛晓星，薛安荣，鞠时光. 基于加权语义网和有效信息的个性化用户兴趣建模[J]. 计算机应用研究，2010，27(9)：3406-3408.

[10] 张辉，李华. 基于LBS的移动精准广告发布系统研究[J]. 计算机时代，2015(8)：31-33.

[11] 李尧. 基于智能视频分析的户外广告效果评估系统[D]. 西安：长安大学，2018.

Recommendation method of elevator media advertising based on improved TF-IDF algorithm

Chen Yanbin[1,2*], Yang Zehua[2], Xue Xiaogui[3], Huang Jintian[4]

(1. Training and Information Center, Jieyang Polytechnic, Jieyang 522051, China; 2. Engineering technology Research Center, Guangdong Bohua Technology Co., Ltd., Jieyang 522000, China; 3. Department of information Engineering, Jieyang Polytechnic, Jieyang 522051, China; 4. School of Physics and Electrical Engineering, Hanshan Normal University, Chaozhou 521000, China)

Abstract: With the introduction of the National Personal Information Protection Law, accurate advertising of elevator media terminal faces many challenges. How to improve the accuracy of elevator media terminal advertising without infringing on the privacy of citizens, and improve the economic benefits for advertising providers is the current focus of terminal computing advertising research. Therefore, a recommendation method of elevator media advertisement based on improved TF-IDF is proposed. The improved TF-IDF was used to extract the labels of POI and residents around the elevator points, and the label vector model of the elevator media terminal was constructed. Then use the merchant's score to calculate the user's interest in the label. Finally, a merchant interest model is constructed to sort the data of TOP-*N* elevators and recommend them to merchants. Experimental results show that this recommendation method is superior to the traditional TF-IDF algorithm in accuracy and recall rate, and does not need to collect citizens' private data, which has strong promotion and application benefits.

Keywords: TF-IDF; media advertising; label score; user interest; personalized recommendation

文章编号: 1007-1423（2023）08-0040-06 DOI: 10.3969/j.issn.1007-1423.2023.08.006

基于RBF函数的茶饮数据分析与预测

李锦朋[1]，黄贻望[1,2*]

（1. 铜仁学院大数据学院，铜仁 554300；2. 贵州省公共大数据重点实验室（贵州大学），贵阳 550025）

摘要：通过对企业销售数据的预测分析与处理可有效增强企事业的经济效益，采集某茶饮真实销售数据，分析数据的各个属性特征，通过降维方式，筛选影响销售额的变化因素作为数据分析的重要特征，将筛选后的特征作为输入特征向量，运用RBF支持向量机（SVM）方法对销售额进行动态预测，并用真实的交易额作对比分析，并对RBF函数SVM销售预测模型进行参数优化，通过对SVM模型预测结果误差和准确率进行仿真分析，验证了优化后模型的有效性，为企业数字化转型过程中的数据分析提供行之有效的方案。

关键词：SVM；核函数；损失函数；销售额预测

0 引言

随着数据技术的发展，企业转型数字化成为必然的趋势，如何去收集、挖掘、分析大数据加快企业的转型数字化发展是企业信息化的一个重要功能[1]。某餐饮品牌是贵州本土品牌，成立于2013年，结合本土各种好茶叶，酝酿出各种好口碑的奶茶，随着店面不断增加，销售数据也日益增加，现有简单的数据统计图应用无法支撑一个企业快速发展，根据某茶饮销售过程中产生的异构数据，利用支持向量机（support vector machine，SVM）小样本算法构建基于不同核函数的销售额预测分析模型，通过对不同核函数下 SVM 销售额预测模型的对比分析，得到参数调优后的 SVM 销售额预测值与实际销售额的值进行比较[2]。实验仿真表明，参数优化后的SVM可减少数据中噪声数据的影响，提高了销售预测模型的效率[3]。

实现某茶饮销售数据动态适时分析与预测具有重要的意义，能对企业未来的趋势进行风险预测，并能及时制定解决方案。通过公司的海量数据分析出产品与产品之间的关联模式、天气对企业销量的影响等诸多因素。通过图表观察数据的整体情况可探究历史企业整体运营情况、业务组成，以便了解企业每个业务的动态发展变化，所有店铺及单个店铺销售情况、消费者（口味、喜好）以及同行的经营状况等，从数据到实际生活等多个维度来定制数字化服务，从而实现企业的快速发展[4]。

主要贡献：①获取某茶饮历史销售数据集及时间段内地区气温温度；②在企业运营系统获取到的数据集进行预处理；③构建基于支持向量机的销售额预测模型；④将非线性SVM模型的预测销售额与真实销售额进行对比分析，有比较好的吻合度，说明模型具有较好的泛化性能。

1 某茶饮数据的支持向量机模型构建

1.1 模型定义

某茶饮销售额受到多种因素，如天气温度、消费者购买力、节假日、门店地域不同等影响，不同的门店位置、不同的人群购买力产生的销

收稿日期：2022-09-21 **修稿日期：**2022-12-10

基金项目：国家自然科学基金项目（62066040）

作者简介：李锦朋（1995—），男，贵州毕节人，本科生，研究方向为深度学习、数据分析等；*通信作者：黄贻望（1978—），男，湖南怀化人，教授，博士，硕士生导师，CCF高级会员，主要研究方向为人工智能、服务计算、形式化方法等，E-mail：458051500@qq.com

售额不同，选择消费者购买力、天气温度、节假日等特征值建立销售额关系的预测模型，利用支持向量机(SVM)方法可以实现销售额是否达到预期目标的预测，有效提升产品的销售布局和管理决策。

解决办法是根据已有的销售数据在模型中的多样性和学习能力之间寻求最好解决方案[11]，SVM解决海量数据中非线性问题的核心思想是原始的非线性可分数据X可找到一个非线性映射Φ，该映射Φ将非线性可分的原始特征空间投影到线性可分的高维特征空间F，从而在高维特征空间中实现样本的线性分类或回归[5]。由于SVM可以实现对特定训练样本的学习并分类识别，将SVM预测模型应用于销售额预测领域，通过对数据集的预处理，使用SVM可减少噪声数据对预测的影响并在分析过程中提高了SVM模型的准确性[6]。

设有M个数据样本的数据集$D=\{(x_i,y_i)\}_{i=1}^{m}$，其中$x_i\in R^d$是$d$维向量，表示每个数据样本的输入特征值向量，$y_i\in\{+1,-1\}$是每个数据样本的标签，表示样本属于的类别，则使用模型对该数据集中的样本进行预测的约束条件为

$$\begin{cases}\boldsymbol{\omega}^{\mathrm{T}}x_i+b\geqslant+1, & i=1,2,\cdots,m\\ \boldsymbol{\omega}^{\mathrm{T}}x_i+b\leqslant-1, & i=1,2,\cdots,m\end{cases}\tag{1}$$

将公式(1)合并为$y_i(\boldsymbol{\omega}^{\mathrm{T}}x_i+b)\geqslant+1,i=1,2,\cdots,m$，其中$\boldsymbol{\omega}=(\omega_1,\omega_2,\cdots,\omega_d)$为特征向量的权重向量，决定分类超平面的法向量；$b$为截距，表示超平面与原点之间的距离，记为$(\boldsymbol{\omega},b)$。

数据集中任意样本x到分类超平面$(\boldsymbol{\omega},b)$的距离公式写为

$$r=\frac{\left|\boldsymbol{\omega}^{\mathrm{T}}x+b\right|}{\|\boldsymbol{\omega}\|}\tag{2}$$

从而优化目标函数为

$$J(\boldsymbol{\omega},b)=\max_{(\boldsymbol{\omega},b)}\frac{\left|\boldsymbol{\omega}^{\mathrm{T}}+b\right|}{\|\boldsymbol{\omega}\|}\tag{3}$$

通过对$(\boldsymbol{\omega},b)$进行缩放使得$\left|\boldsymbol{\omega}^{\mathrm{T}}x+b\right|=1$，则将式(3)转化为式(4)：

$$J(\boldsymbol{\omega},b)=\min_{(\boldsymbol{\omega},b)}\frac{1}{2}\|\boldsymbol{\omega}\|^2\tag{4}$$

为降低基于SVM销售额预测模型的泛化误差[7]，引入松驰变量ξ_i，将优化目标转化为

$$\begin{cases}\min\limits_{(\boldsymbol{\omega},b)}\dfrac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum\limits_{i=1}^{m}\xi_i\\ \text{s.t. }y_i(\boldsymbol{\omega}_i+b)\geqslant1-\xi_i,\ \xi_i\geqslant0,i=1,2,\cdots,m\end{cases}\tag{5}$$

1.2　优化目标函数求解

通过引入拉氏(Lagrange)系数，构造拉氏函数，将式(5)化为无限制的优化问题，拉格朗日乘子$\alpha_i\geqslant0,i=1,2,\cdots,N$，拉氏函数如下：

$$L(\boldsymbol{\omega},b,\alpha,\mu,\xi)=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\mu_i\xi_i+\sum_{i=1}^{m}\alpha_i\left(1-\xi_i-y_i(\boldsymbol{\omega}_i+b)\right)\tag{6}$$

当满足对应的KKT条件时，

$$\begin{cases}1-\xi_i-y_i(\boldsymbol{\omega}^{\mathrm{T}}x_i+b)\leqslant b;\\ \alpha_i\geqslant0;\mu_i\geqslant0;\xi_i\geqslant0;\\ \alpha_i\left(1-\xi_i-y_i\left(\boldsymbol{\omega}^{T}x_i+b\right)\right)=0;\\ \mu_i\xi_i=0;\end{cases}\tag{7}$$

无约束优化问题式(6)转化为相应的强对偶问题：

$$\begin{cases}\max\limits_{(\alpha,\mu,\xi)}\min\limits_{(\boldsymbol{\omega},b)}L(\boldsymbol{\omega},b,\alpha,\mu,\xi)\\ \text{s.t. }\alpha_i\geqslant0,i=1,2,\cdots,m\\ \mu_i\geqslant0,i=1,2,\cdots,m\end{cases}\tag{8}$$

通过求解公式(8)得到原问题的优化解，见式(9)。

$$f(x)=\sum_{i=1}^{m}\alpha_iy_iK(x_i,x)+b\tag{9}$$

2　数据处理

2.1　数据集

数据集是采用贵州某餐饮品牌实时销售数据，某店面2015年1月1日至2021年7月31日每一天的销售额数据，数据集包含2826行10列的时间-销售金额数据。如表1所示。

2.2　特征选择

为防止多维属性的强关联对茶饮样本数据质量产生噪声，从而影响模型的可靠性，从一级品类、二级品类、商品名称、商品编码、单位、销售次数、销售数量、销售金额、退货数量、退货金额等10个特征中选择对预测销售额

表 1 茶饮销售数据源

日期	中类名称	商品名称	商品编码	单位	销售次数	销售数量	销售金额	平均气温	购买力
2015/1/1	柠檬系列	冰鲜柠檬水	SKU0100	杯	180	259	1036	14.0	1
2015/1/1	冰淇淋系列	蓝莓圣代	SKU0062	个	36	40	200	13.5	2
2015/1/1	柠檬系列	金桔柠檬茶	SKU0103	杯	29	39	273	11.0	3
				……					
2021/7/31	奶茶系列	燕麦奶茶(中杯)	SKU0058	杯	56	132	786	7.5	5
2021/7/31	甜品系列	无敌烧仙草	SKU0083	杯	23	69	420	7.5	2
2021/7/31	奶茶系列	原味奶茶(大杯)	SKU0118	杯	45	90	720	16.5	4
2021/7/31	热饮系列	鲜芋撞奶	SKU0168	杯	32	96	192	17.0	1

影响较大的特征，即样本空间属性的降维处理，也就是特征选择，从而降低预测过程的复杂性，同时由于是针对餐饮店销售额的预测，将加入影响销售的外界因素天气温度、购买力作为特征值，共计10个特征，通过降维到4个属性用于模型的训练。将气温、销售数量、购买力、节假日特征属性标签分别编号分为1～4，图1为各属性贡献值。

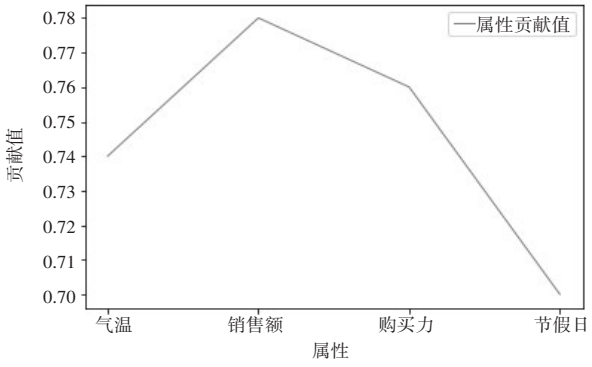


图 1 各属性贡献值

2.3 销售额的参数

影响销售额的参数有气温、购买力、节假日、销售数量，其中气温和购买力是长期影响销售额的因素。

气温数据从国家气象网上采集，客户群购买力从政府部门发布的统计数据可提供人群购买力的参考指标，比如人均收入、消费支出等。图2和图3是影响销售额的气温和购买力，销售数量与销售额呈正比，随着节假日到来，销售额也会随之增长，影响销售额的还有门店位置。

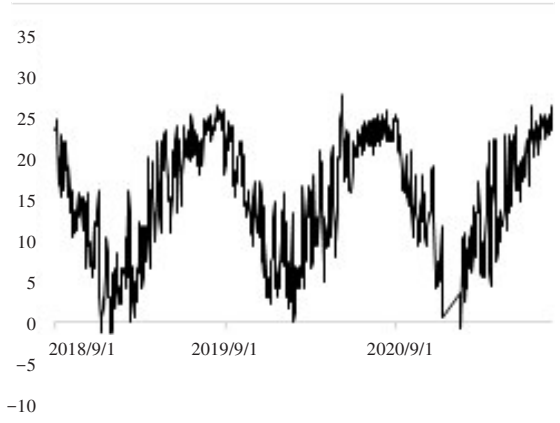


图 2 气温

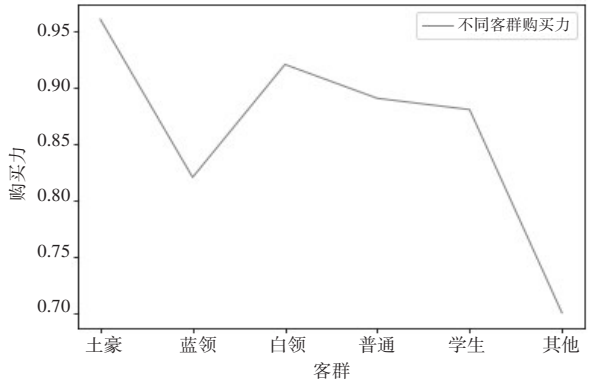


图 3 客户群购买力

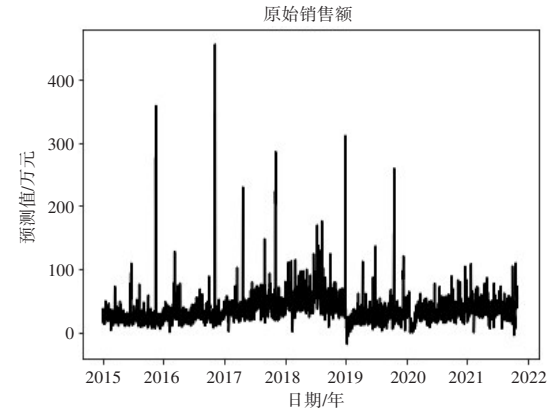
2.4 归一化

为解决因特征变化而导致的预测偏差，需要对数据集进行归一化处理，这里采用min-max

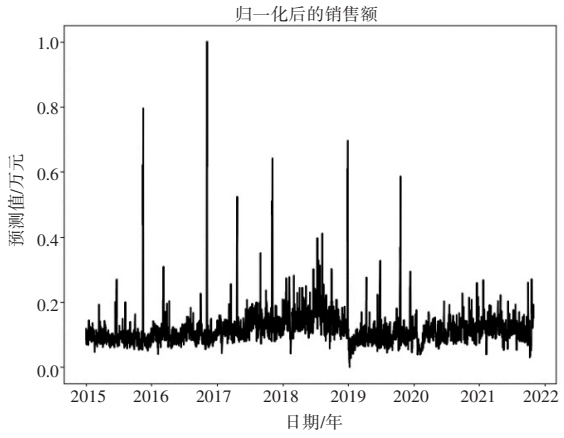
标准化[8]，如公式(10)所示。

$$x_{new} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{10}$$

其中：x_i为第i个样本数据属性值，x_{min}和x_{max}是属性的最小值和最大值。



(a) 实际销售额



(b) 归一化后销售额

图 4　标准化后的数据对比

归一化后的销售额数据可以提升模型精度和准确性，图4(b)是将实际销售额数据归一化后的结果。

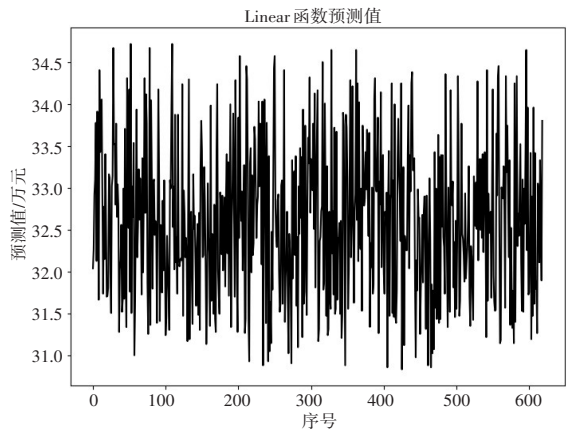
3　实验和分析

3.1　基于不同核函数的SVM销售额预测模型

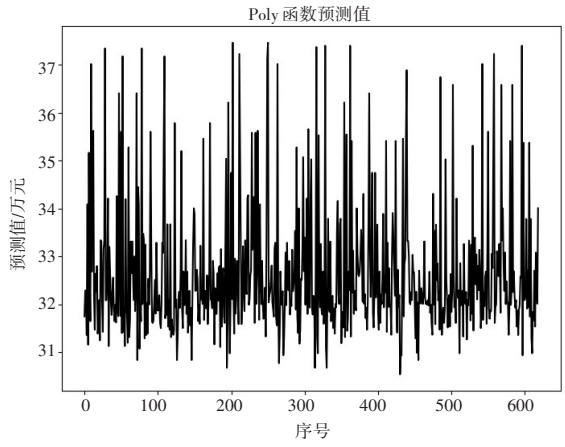
将2826条数据分为训练集和预测集，其中1978条数据作为训练集，848条数据作为测试集[9]。模型训练是基于线性核、多项式核和RBF核三种不同的核函数进行的，通过三种不同核函数构造SVM销售额数据的预测模型，其中RBF核为高斯核，对应的函数为高斯核函数(见表2)。

表 2　核函数的表达式

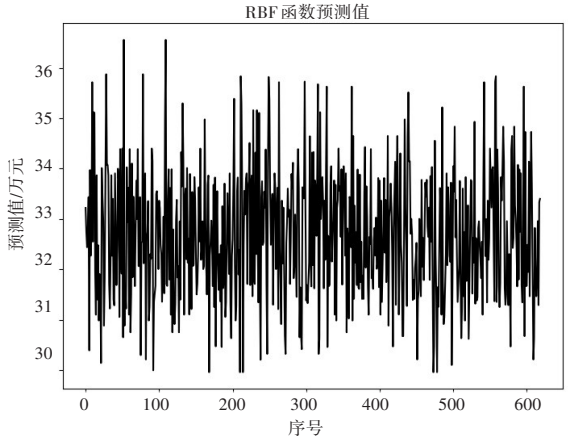
名称	表达式	参数
线性核	$K(x_i, y_i) = x_i^{\mathrm{T}}, y_i$	—
多项核	$K(x_i, y_i) = (x_i^{\mathrm{T}}, y_i)^d$	$d > 0$多项式次数
RBF核	$K(x_i, y_i) = \exp\left(-\delta \left\| x_i - y_i \right\|\right)^2$	$\delta > 0$高斯核的带宽



(a) 线性核函数预测的销售额



(b) 多项式核函数预测的销售额



(c) RBF函数预测的销售额

图 5　三种核函数销售额预测对比

选取均方误差(MSE)和平均绝对百分比误差(MAPE)两个评价指标分别从预测误差和预测精准度两个方面对不同核函数下的SVM销售额预测结果进行对比[10]，结果如表3所示。

表3 三种核函数销售额预测对比

核函数	*MSE*	*MAPE*	*Accuracy*/%
线性核函数	0.004202	0.1241	84.21
多项式核函数	0.004166	0.1231	86.62
RBF函数	0.004162	0.1184	89.12

据统计分析可知，均方误差(MSE)越小，表示预测值与真实值误差越小，即分类模型性能越好，也就是说模型的预测结果越接近真实值[11]，从表3可知基于高斯核函数(RBF)的支持向量机模型的预测销售额效果较其余两个函数的效果更佳[12]。

3.2 参数调优

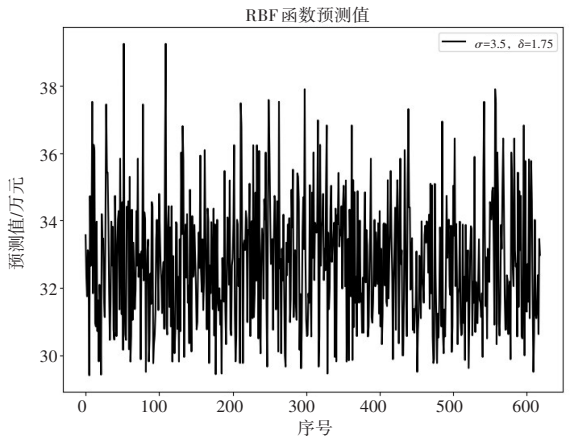
为降低预测销售额模型的预测误差，提高模型的泛化性能，现对RBF函数下的SVM销售额预测模型的参数进行优化。随机选取3组参数对(σ,δ)进行对比实验，其中σ为惩罚参数，δ为多项式函数的系数，对比结果如图6所示[13]。

对比表3和表4销售预测模型的*MSE*、*MAPE*和*Accuracy*，得到σ=3.00，δ=0.75时，RBF核函数*MSE*=0.004115，*MAPE*=0.0964，*Accuracy*=92.14%，表明SVM预测效果较好。

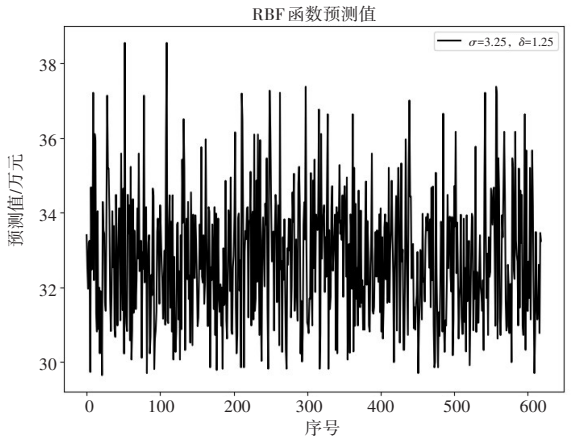
表4 基于不同核函数销售额预测对比

参数	*MSE*	*MAPE*	*Accuracy*/%
σ=3.50，δ=1.75	0.004122	0.1132	90.21
σ=3.25，δ=1.25	0.004142	0.1142	89.82
σ=3.00，δ=0.75	0.004115	0.0964	92.14

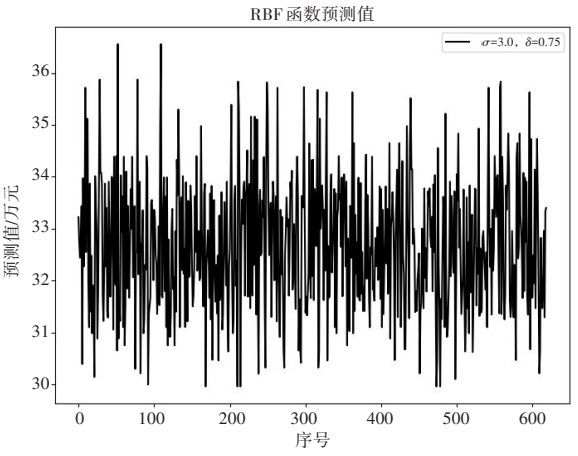
由表4可知，对参数调优前后*MSE*、*MAPE*的值进行对比，发现参数调优后模型预测效果更佳[14]。



(a) σ=3.5，δ=1.75



(b) σ=3.25，δ=1.25



(c) σ=3.00，δ=0.75

图6 基于不同参数的SVM的预测结果对比

4 结语

将销售数据作为训练集，训练不同核函数

下的SVM销售预测模型，并对RBF函数SVM销售预测模型进行参数优化，通过对SVM模型预测结果误差和准确率进行仿真分析，验证了优化后模型的有效性，有助于企业精准掌握客户喜好，针对不同的客户群采用不同的方案进行精准营销。

参考文献：

[1] MCKINNEY W. 利用python进行数据分析[M].2版. 北京:机械工业出版社,2018.

[2] 周志华,杨强. 机器学习及其应用[M]. 北京:清华大学出版社,2011.

[3] 杨永琪,徐欣蕾,陈曦. 基于SVM技术实现手写数字分类识别的研究[J]. 电脑知识与技术,2020,16(6):195-196.

[4] 江翠丽,曹滕飞,李长哲,等. 基于SVM的视频用户需求预测算法[J]. 2021,7(5):3-8.

[5] HAN J W,PEI J. 数据挖掘概念与技术[M].3版. 北京:机械工业出版社,2019.

[6] DING S,WANG Z Y,WU D S,et al. Utilizing customer satisfaction in ranking prediction for personalized cloud service selection[J]. Decision Support Systems,2017,93:1-10.

[7] 张益铭,徐晓钟,王智庆. 支持向量机与时间序列预测综述[J]. 计算机应用于软件,2010,27(12):127-130.

[8] 陈日进. 销售预测中指数平滑法与时间序列分解法的比较[J]. 统计与信息论坛,2019,35(12):256-311.

[9] JIANG C,CAO T,LI C,et al. Multimedia resources purchase and distribution optimization for 5G edge cloud network [J]. IEEE Access, 2020, 8:59340-59350.

[10] 艾瑞. 2020年中国企业数字化转型路径实践研究报告[R],2021.

[11] 张传强. 基于显著性的目标检测与识别算法[D]. 青岛:青岛理工大学,2016.

[12] 吴金花,孙德山. 加权支持向量回归的权值确定方法[J]. 计算机技术与发展,2009,19(6):135-137.

[13] 周游. 移动式重点车辆比对与预警算法研究[D]. 郑州:郑州大学,2017.

[14] 王欣欣. 混合自适应引力搜索优化的特征选择方法[J]. 计算机工程与应用,2017,53(12):166-171.

Analysis and forecast of a tea sales data based on RBF function

Li Jinpeng[1], Huang Yiwang[1,2*]

(1. School of Data Science, Tongren University, Tongren 554300, China;
2. Guizhou Provincial Key Laboratory of Public Big Data(Guizhou University), Guiyang 550025, China;)

Abstract: The economic benefits of enterprises can be effectively enhanced through the prediction, analysis and processing of enterprise sales data. collect the real sales data of a tea drink, analyze the attributes and characteristics of the data, and select the changing factors that affect sales as an important feature of data analysis through dimensional reduction. Taking the selected feature as the input feature vector, the RBF support vector machine(SVM) method is used to dynamically predict the sales volume, and the real transaction volume is compared and analyzed, and the parameters of the RBF function SVM sales forecasting model are optimized. Through the simulation analysis of the prediction error and accuracy of the SVM model, the effectiveness of the optimized model is verified. It provides an effective scheme for data analysis in the process of enterprise digital transformation.

Keywords: SVM; kernel function; loss function; sales forecast

文章编号: 1007-1423(2023)08-0046-06 DOI: 10.3969/j.issn.1007-1423.2023.08.007

基于协同多目标优化方法的流水车间组调度

肖秀梅*，王欣蕊

（云南师范大学数学学院，昆明 650500）

摘要： 针对带序列依赖的流水车间组调度问题(FSDGSP)，提出了一种协同多目标优化算法。首先，分析了FSDGSP的问题特点。其次，设计了适合问题求解的三维向量编码策略，以区分三种不同的优化子问题。再次，构建了一种协同搜索的多目标优化算法。最后，经过大量实验对比分析，并与其他算法相比较得出，该算法明显优于现有的经典多目标优化算法。

关键词： 流水车间组调度；多目标；能耗；协同算法

0 引言

近年来，随着生产力的快速发展，人们往往追求高效的方法来使企业获得更高的效益。其中带序列依赖的流水车间组调度问题(flow setup dependency group scheduling problem, FSDGSP)中提到的评估生产效率的指标是对其问题的突破，在工业上得到了广泛的应用，从而展开了激烈的研究。

FSDGSP作为单元制造系统中的一个重要调度问题，引起了学术界和实践者的极大关注。对于它的讨论也变得越来越多。如Costa等[1]研究具有阻塞约束的流水车间序列相关组调度问题的最小完工时间等。

目前解决FSDGSP的方法主要分为以下三类。

(1) 精确算法。提出了一种基于分支界定法的基于总流量时间准则的FSDGSP下界法。由于搜索效率相当低，对于小的问题，可以得到理论最优解。然而，对于中等大小的问题，在合理的时间内获得大规模问题的最优解是非常困难的。

(2) 构造性启发法。根据一定的调度规则，采用构造性启发式算法快速构造求解方案。一般来说，构造性启发式被用作初始化方法，为元启发式算法提供高质量的初始解。Reddy等[2]提出了在组内安排工件的启发式方法，以提高单元内机器的利用率。Neufeld等[3]认为每个组都是一份有时间延迟的工件。

(3) 元启发式算法。元启发式算法的通用性很强。各种搜索框架用于解决FSDGSP。Costa等[1]提出了一种自适应遗传算法，以最小化具有阻塞约束的FSDGSP的最大完工时间。Lin等[4]介绍了一种数学方法，用于求解具有无等待约束的FSDGSP。Li等[5]设计了一种混合和声搜索算法来解决FSDGSP问题，其目标是最小化总延误和平均总流量时间。Tavakkoli-Moghadam等[6]研究了一种基于分散搜索的元启发式算法，用于求解多准则的FSDGSP。

随着绿色经济的发展，能源消耗量逐渐成为评判生产效率的一个指标。然而在现有文献中，主要优化的是一个或两个生产目标。Shao等[7]提出PEDA来解决MDNWFSP-SDST问题，研究最大完工时间和等待时间之间的关系。Zhao等[8]提出了TS-CEA算法，研究加工时间和能耗之间的关系。何启巍等[9]提出了混合粒子群优化算法，来解决最大完工时间和总流经时间之间的关系。基于多目标优化方法的流水车

收稿日期： 2022-12-16 **修稿日期：** 2023-02-13

作者简介： *通信作者：肖秀梅(1975—)，女，山东聊城人，本科，讲师，研究方向智能优化方法，E-mail：1619568542@qq.com；王欣蕊(2000—)，女，山东聊城人，在读研究生，研究方向智能优化方法

间调度问题已经成为当前调度方向的主流趋势[10,11]，其中一个主要研究目标是最大完工时间。通过搜索能耗与多目标相关的论文，也可以观察到能耗是一个热点约束，而结合多目标与能耗的相关文献较少。多目标优化相关文献主题分布如图1所示。

1　问题描述

FSDGSP问题可以描述为：有n个工件需要在m台机床上依次加工，每个工件的加工工序一致，所有工件分配到指定组内。组内工件之间没有准备时间，组间工件之间需要准备时间。图2展示了5个工件在3台加工机床上的调度方案，其中工件1和3分到第一组，工件2/4/5分到第二组。由图2可见，工件1和3之间没有加工准备时间，第二组内的3个工件之间也没有加工准备时间。图2给出的调度方案的完工时间是425分钟。如果改变组内工件的排列顺序，会得到不同效果的调度方案，如图3所示。图3中工件1和3交换了一下位置，最终的makespan指标降低了15分钟。

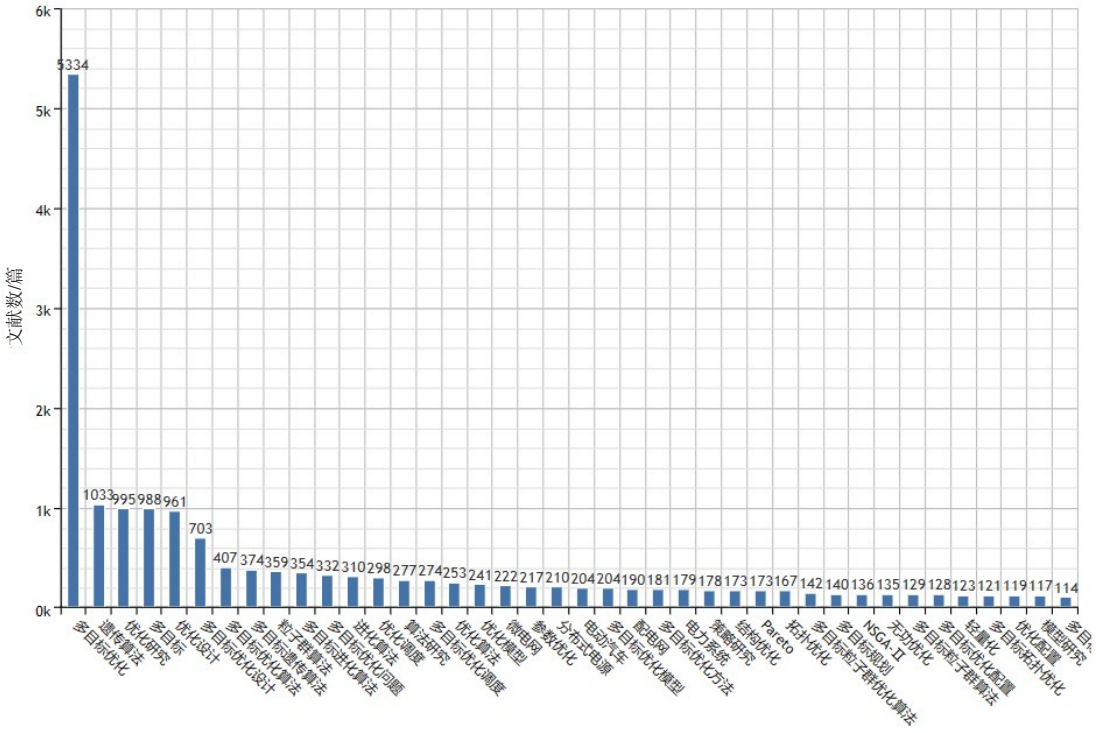


图1　多目标优化相关文献主题分布

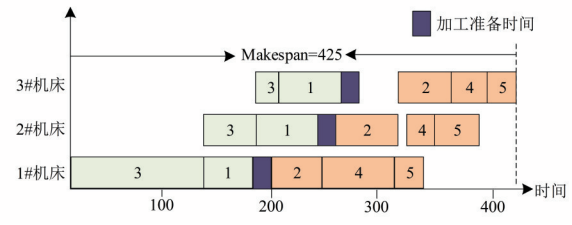


图2　调度方案1

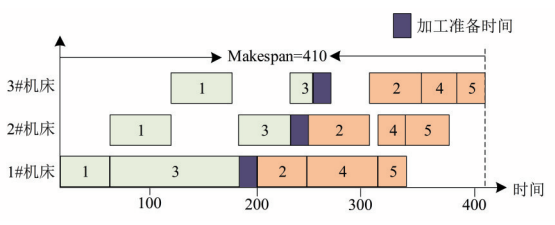


图3　调度方案2

2 算法设计

2.1 问题编码

首先进行种群编码，假设有三个子问题：所有组的分组序列，每一个组内的工件序列以及在所有机器上的速度序列。一个解可以表示为(μ,τ,v)。μ表示按顺序排好的一个组序列，τ表示在每一个组中的工件序列，v是一个速度等级矩阵，用于确定机器上处理每个作业的每个操作的速度。由于前期速度从未改变，所以一个解也可以表示为(μ,τ)。通常可以用机器甘特图的设计来实现。在图4所示的例子中，解的表示为组序列(1,4,2,3),工件的顺序为{(2,1),(7,8,6),(4,3),(5)}。

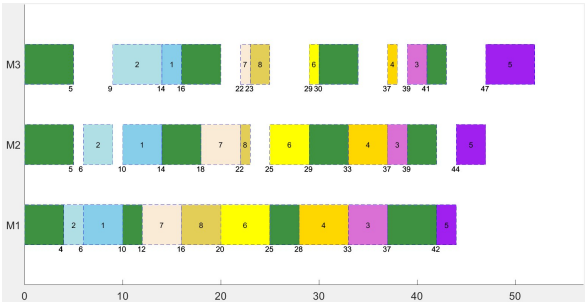


图4 甘特图编码

2.2 算法流程图

设计的多目标优化算法流程如图5所示。

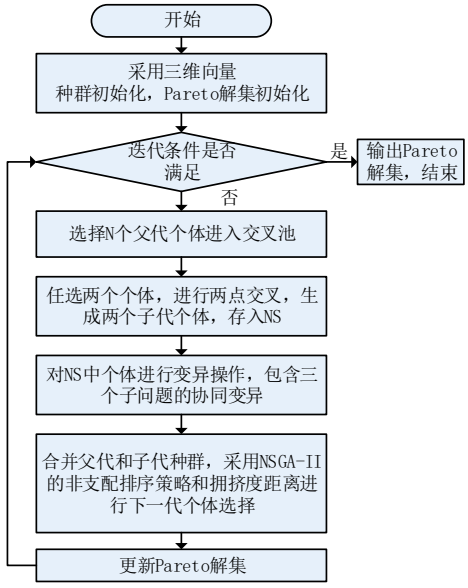


图5 CMOEA算法流程

3 实验分析

3.1 实验条件

该算法用MATLAB编程语言实现。机器配置参数如下：CPU型号为i7，内存为16 GB，操作系统为Win10。

3.2 实验算例

针对某纺织车间组调度流程开展算法测试分析。实例包含387个工件，6台机床，60个分组。所提算法CMOEA求解该类问题的Pareto解集如表1所示。

表1 CMOEA算法求解所得Pareto解集

solution number	Cmax	TFT	TEC
1	2507.434	494090.902	82820.928
2	2507.483	493932.763	82834.466
3	2507.483	493938.652	82823.640
4	2507.483	494093.267	82767.248
5	2508.076	493545.663	82826.039
6	2508.670	493355.265	82844.564
7	2508.670	493512.932	82830.713
8	2508.670	493577.140	82810.149
9	2513.062	493789.765	82800.722
10	2526.830	493303.375	82964.153
11	2528.310	492775.491	82947.889
12	2530.405	492777.586	82946.761
13	2531.149	492379.173	82968.483
14	2531.149	492531.481	82955.929
15	2531.149	492778.330	82925.531
16	2531.149	492874.246	82916.539
17	2536.915	491473.468	83033.899
18	2537.084	491474.144	83032.544
19	2538.081	491478.132	83022.129
20	2538.436	491318.373	83029.567
21	2538.436	491378.168	83028.810
22	2538.436	491457.397	83023.701
23	2538.436	491479.551	82986.773
24	2538.960	490835.414	83041.989
25	2540.316	490769.217	83044.973
26	2540.514	490770.206	83044.162
27	2540.730	490848.499	83038.842
28	2541.102	490757.564	83059.024
29	2541.102	490760.135	83044.560
30	2541.102	490786.421	83042.189
31	2541.102	490851.849	83014.709
32	2541.102	491272.797	83006.846

续表1

solution number	Cmax	TFT	TEC
33	2541.503	490322.665	83084.421
34	2541.503	490684.499	83047.214
35	2541.503	490763.308	83043.098
36	2541.503	490771.609	83040.981
37	2541.503	490775.151	83005.640
38	2593.874	486987.139	83464.111
39	2593.874	487012.366	83444.480
40	2593.874	487086.139	83429.632
41	2593.874	487681.396	83414.563
42	2604.734	485756.735	83416.416
43	2605.723	485815.087	83412.361
44	2605.847	485914.815	83412.070
45	2606.185	485842.318	83412.269
46	2606.364	485916.363	83378.193
47	2606.877	485679.761	83422.247
48	2606.877	485883.164	83372.740
49	2621.410	485286.755	83515.874
50	2622.806	485299.953	83509.171
51	2622.806	485340.680	83506.568
52	2622.806	485342.620	83488.225
53	2622.806	485344.394	83485.930
54	2624.101	485249.680	83519.931
55	2624.958	485300.252	83501.150
56	2624.958	485327.826	83497.309

3.3　算法对比

将CMOEA算法与其他三种最新算法进行了比较[12]，包括基于分解的多目标进化算法(MOEAD-SAS)[13]，基于知识的协同进化算法(KCA)[14]，基于支配关系的多目标遗传算法(NSGA-III)的改进版本[15]。

对这些算法的简要描述如下：

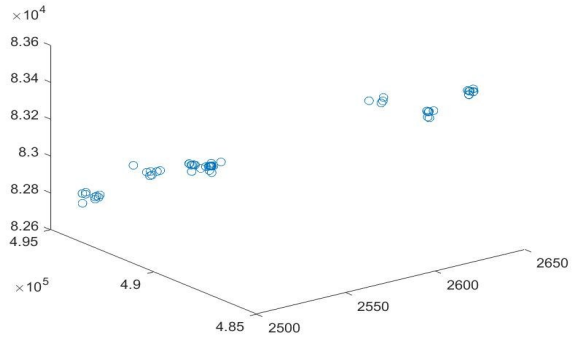
(1) 多目标优化问题(MOP)由基于分解的多目标进化算法(MOEA/D)分解为多个子问题。MOEA/D- SAS是MOEA/D的一种变体，采用基于角度的选择和基于分解的排序两种策略来实现多样性和收敛性的平衡。

(2) 针对高效节能的分布式流水车间调度问题，提出了KCA算法。其核心思想是对不同的子问题自适应地采用不同的搜索算子。

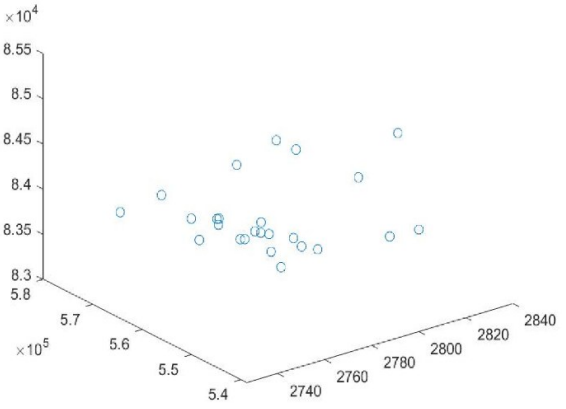
(3) NSGA-III是基于支配关系的多目标遗传算法(NSGA-II)的改进版，它提供了一组保持种群分布的参考点。NSGA-III用于解决多目标问题。

为了比较这些算法的差异性，本文使得所有竞争算法在同一计算环境中独立运行10次。最大运行时间是固定的，是K×n×m×δ毫秒。结果表明，在几乎所有的测试用例中，CMOEA在收敛性、分布性和超容量指标方面都比其他竞争算法得到的结果更好，这说明CMOEA明显优于其他竞争算法。

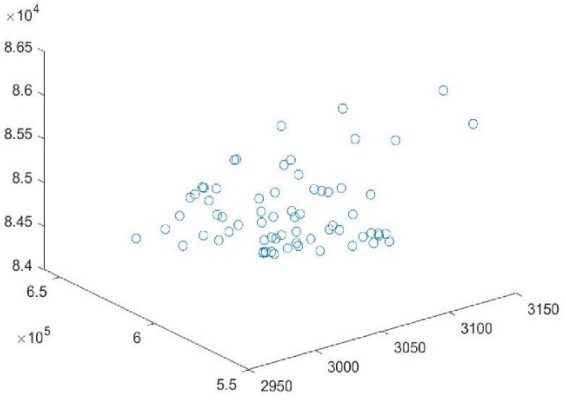
与其它三种算法的运行结果如图5(a)～(d)所示。



(a) CMOEA解集

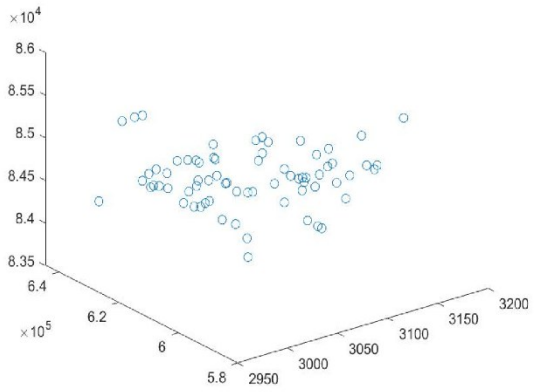


(b) KCA解集



(c) MOEADSAS解集

图5　算法求得的帕累托解集



(d) NSGA-III解集

图5 算法求得的帕累托解集(续)

通过与当前其它算法相比，CMOEA算法的最大完工时间、总流经时间、总能耗都是相对最小的，由此可以得出该算法的有效性。

4 结语

本文的主要工作：

(1) 在FSDGSP研究一两个目标的基础上，加入能量消耗作为优化目标，因此目标变成了三个。采用CMOEA算法提出求解多目标FSDGSP问题。

(2) 分析FSDGSP的问题特点，展示了问题特性。

(3) 对问题进行编码与解码，提出协同多目标优化CMOEA算法来寻找最优解集。与其他竞争算法相比较，得出本文提出的CMOEA算法具有明显优于现有竞争算法的搜索性能。

参考文献：

[1] COSTA A, CAPPADONNA F V, FICHERA S. Minimizing makespan in a flow shop sequence dependent group scheduling problem with blocking constraint[J]. Engineering Applications of Artificial Intelligence, 2020, 89: 103413.

[2] REDDY V, NARENDRAN T. Heuristics for scheduling sequence-dependent set-up jobs in flow line cells [J]. International Journal of Production Research, 2003, 41: 193-206.

[3] NEUFELD J S, GUPTA J N, BUSCHER U. Minimising makespan in flowshop group scheduling with sequence-dependent family set-up times using inserted idle times[J]. International Journal of Production Research, 2015, 53: 1791-1806.

[4] LIN S W, YING K C. Makespan optimization in a no-wait flowline manufacturing cell with sequence-dependent family setup times[J]. Computers and Industrial Engineering, 2019, 128: 1-7.

[5] LI Y, LI X, GUPTA J. Solving the multi-objective flowline manufacturing cell scheduling problem by hybrid harmony search[J]. Expert Systems with Applications, 2015, 42: 1409-1417.

[6] TAVAKKOLI-MOGHADDAM R, JAVADIAN N, KHORRAMI A, et al. Design of a scatter search method for a novel multi-criteria group scheduling problem in a cellular manufacturing system [J]. Expert Systems with Applications, 2010, 37: 2661-2669.

[7] SHAO W S, PI D C, SHAO Z S. A pareto-based estimation of distribution algorithm for solving multiobjective distributed no-wait flow-shop scheduling problem with sequence-dependent setup time [J]. IEEE Transactions on Automation Science and Engineering, 2019, 16: 1344-1360.

[8] ZHAO F Q, HE X, WANG L. A two-stage cooperative evolutionary algorithm with problem-specific knowledge for energy-efficient scheduling of no-wait flow-shop problem[J]. IEEE Transactions on Cybernetics, 2021, 51: 5291-5303.

[9] 何启巍,张国军,朱海平,等. 一种多目标置换流水车间调度问题的优化算法[J]. 计算机系统应用, 2013, 22: 111-118, 110.

[10] ZHAO F Q, HE X, ZHANG Y, et al. A jigsaw puzzle inspired algorithm for solving large-scale no-wait flow shop scheduling problems [J]. Applied Intelligence, 2020, 50: 87-100.

[11] LIU J, REEVES C R. Constructive and composite heuristic solutions to the P//$\sum C_i$ scheduling problem [J]. European Journal of Operational Research, 2001, 132(2): 439-452.

[12] HE X, PAN Q K, GAO L, et al. A greedy cooperative co-evolution algorithm with problem-specific knowledge for multi-objective flowshop group scheduling problems[J]. IEEE Transactions on Evolutionary Computation, 2021.

[13] CAI X, YANG Z, FAN Z, et al. Decomposition-based-sorting and angle-based-selection for evolutionary multiobjective and many-objective optimization[J]. IEEE Transactions on Cybernetics, 2017, 47: 2824-2837.

[14] WANG J J, WANG L. A knowledge-based cooperative algorithm for energy-efficient scheduling of distributed flow-shop [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 50: 1-15.

[15] DEB K, JAIN H. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, Part I: solving problems with box constraints [J]. IEEE Transactions on Evolutionary Computation, 2014, 18: 577-601.

(下转第56页)

文章编号: 1007-1423(2023)08-0051-06 DOI: 10.3969/j.issn.1007-1423.2023.08.008

Web时态对象模型研究分析

李淑霞，杨俊成*

（河南工业职业技术学院电子信息工程学院，南阳 473000）

摘要：研究Web内容要素中加入时态要素的Web时态对象模型，解决层次树结构对站点、栏目、子栏目和内容统一建模问题，根据数据与时序密切相关的特性，利用时态特征词对各时态要素进行自动提取及评估，提出利用LSTM网络学习时态节点之间的时态约束关系，将Attention机制引入预测模型；实验表明该体系能自动发现Web过时信息，在同类网站质量排序、时间感知的搜索排序等方面有着重要的应用前景，能极大地节约人力，提高Web信息质量。

关键词：深度学习；时态要素；时态对象模型；时态特征词；层次树

0 引言

随着Web日益成为人们发布和获取信息的主要渠道，Web数据的质量变得尤为重要。目前，网络中包含过时信息、存在时间不一致等问题是一个普遍现象，是导致Web数据质量低下的主要因素之一。这一现象较普遍存在于各级政府网站、企事业单位网站、门户网站中。例如：①链接标题显示“今日人民币汇率突破6.3”，而实际上这可能已经是昨天的价格；②栏目标题是本周新闻，但由于没有及时更新，栏目中包含有本周以前的信息；③招聘网站的信息栏目中包含有“最新”“急招”，但当用户打开链接后发现有些已经过期失效；④政府网站的“最新公告栏”中包含早已过时的公告信息。以上问题极大地损害了政府及企事业单位形象，影响了用户的体验，浪费了用户的浏览时间，更有甚者，可能会误导决策行为，造成生产、生活中不必要的经济损失。目前解决Web过时信息的问题主要采用人工逐个排检的方式。然而，面对不断增长的互联网海量数据，人工方式显然难以胜任。因此，迫切需要自动发现Web过时信息的理论系统、方法论及实用工具。

从以上问题可发现的共性科学问题是“保持Web时态一致性”——即在当前情境下，Web的各个时间要素不存在矛盾和歧义性，保持一致。目前国内外学者对Web时态一致性还没有开展系统的研究，更没有形成这方面的理论和方法。因此，本文研究在Web内容要素中加入时态要素的Web时态对象模型，对站点、栏目、子栏目和页面的内容及时态统一建模，以及各时态要素的自动提取和评估方法；研究Web时态对象模型中时态向量一致性约束关系、推理机制和代数运算系统，从而得到Web时态不一致的自动发现、分类和度量方法；建立Web时态一致性理论体系，提出自动发现Web过时信息的方法和工具，填补国内外该方面的研究空白，具有重要的理论价值。同时，该理论的建立具有广泛的应用前景。

近年来，时态Web日渐成为学者们关注的焦点。Web学术权威的国际会议WWW在2011年专门设立了“时态Web”Workshop——TWAW。Omar等[1]分析了文档中时间信息的类型，时间的表述方式及形式化，时间的标注等内容，指

收稿日期：2023-02-23 **修稿日期：**2023-03-13

基金项目：河南省科技厅科技攻关项目(222102210203)；河南省教育厅高等学校重点科研项目(22B520009)；全国高等院校计算机基础教育研究会纵向课题(2022-AFCEC-288、2022-AFCEC-295)

作者简介：李淑霞(1982—)，女，河南南阳人，硕士，讲师，研究方向为人工智能、嵌入式系统；*通信作者：杨俊成(1982-)，男，河南南阳人，硕士，副教授，从事领域为人工智能、大数据等专业教学工作，E-mail：juncheng@126.com

出了时态网络的研究方向，包括时空信息挖掘、时态检索、时间相似度与实时搜索等。Meng等[2]对Web时序特征进行分析，基于LSTM这个深度学习模型，提取了网页评论中的问答时间信息等。Rohoman等[3]通过挖掘和利用隐含的时间特征构建了一个检索系统，并在检索系统中对相关的查询进行简单的分类。沈思等[4]分析了如何加入时间因素的检索模型构建，同时研究时间如何提升检索结果。张梦妮[5]引入网页抽样方法，通过选取少量具有代表性的网页进行无障碍检测，压缩检测内容，降低人力开销，加快检测过程。张鹏程等[6]定义了时间属性序列图的形式语法，给出基于时间Buchi自动机的形式操作语义，并用实时规约模式度量了时间属性序列图的表达力。时态Web的相关成果为本论文的研究提供了理论基础。

在网站质量的评价方面，刘凯鹏等[7]研究了利用网页质量评价的新维度——社会性标注——以改进网页检索性能。王伟等[8]提出了一种网络资源敏感的性能诊断方法。陈传夫等[9]在采用层次分析法确定各级指标权重的过程中，构造了时效性指标的判断矩阵。Na等[10]利用网页新鲜度来评估网页质量，并从页面本身及其链入页面两方面来度量网页的新鲜度。Yang等[11]将内容新鲜度的概念形式化，提出了用最少的网络流量保持并优化内容新鲜度的方法。事实上，以上的测评指标均针对的是网站内容的整体质量和一般意义上的信息时效性，对于网页的时间一致性并未进行建模和度量。

在时间感知的Web网页信息检索方面，以PageRank为代表的基于链接分析打分方法并未考虑网页的时效性，故在时间感知搜索中，其排序存在一定的偏差[12]。因此，对已有的检索模型的时间维度的扩展与深化成为必然。近年来，不断出现基于时间信息的检索系统的研究成果，王晶晶等[13]结合DBpedia特殊语义网提出支持隐式时间查询的文档排名方法。本论文将在现有工作基础上，利用网页时间不一致度量，建立时间感知的Web网页信息检索模型。

在Web信息抽取方面，已有大量的研究工作。为了从网页中快速获得隐含的有用信息，白钰洁[14]提出一种基于开始定界符的Web信息抽取方法。为了更好地了解学术期刊或学术人物，刘子玉[15]提出基于Web的异构学术信息抽取与聚合方法的自动化算法框架，从而帮助研究人员从互联网大量的异构网页中迅速挖掘所需信息。为了能够帮助患者早日发现病情，为医生确诊提供数据支持，李利敏[16]设计了一个新的基于DOM树的Web信息抽取模型。寇月等[17]分析了DeepWeb结果页面的特点，提出了基于DOM树的自动实体抽取策略。本文主要采用基于时态DOM模型的Web信息提取方法，有关时间的正则文法匹配，以及基于模式代数的方法和时间概念本体方法，抽取网页多个时间维度。

1 Web信息时态敏感性分析与度量

Web信息对时间的敏感性各不一样，有的对时间十分敏感，如新闻与金融信息等；有的基本不敏感，如生活常识等。因此，需研究分析Web中不同栏目网页的时态敏感性特征及敏感程度的量化度量方法，对网页进行过滤筛选。经分析，新闻、金融等对时间十分敏感的信息，具备如下特征：在信息发表后的短时间内，网页访问量大，访问频率高；该类信息栏目更新频率高；信息中包含有大量时间信息。所以，本文依据以上特征，采用用户的访问模式、栏目更新模式和文本时间信息特征对Web信息进行敏感性度量，从动态和静态两方面反映网页的时态敏感性。

时态敏感度函数定义如以下公式所示：

$$DS_t = \lambda F(P_v) + \nu F(P_u) + \mu F(tw) \tag{1}$$

$$F(P_v) = \frac{\sum_{i=1}^{n} \delta(t_i) p_v(t_i)}{\max\limits_{1 \le i \le n} \delta(t_i) p_v(t_i)} \tag{2}$$

$$F(P_u) = \frac{\sum_{i=1}^{n} \theta(t_i) p_u(t_i)}{\max\limits_{1 \le i \le n} \theta(t_i) p_u(t_i)} \tag{3}$$

$$F(tw) = \frac{\sqrt{\sum\limits_{0 \le i < j \le f_{tw}} \left\| s_{tw}(j) - s_{tw}(i) \right\|^2}}{f_{tw} \times L} \tag{4}$$

其中：λ, ν, μ是权重函数，$\delta(t_i)$是随时间t_i变化的权重函数，$P_v = \left(p_v(t_1), \cdots, p_v(t_n) \right)$是用户的访问模式，是以网站栏目为单位，利用$t_i$时刻

栏目的整体访问频率$vf_s(t_i)$对t_i时刻访问频率$vf_w(t_i)$进行平滑，得到的平滑后访问频率$P_v(t_i)=\alpha\times vf_s(t_i)+\beta\times vf_w(t_i)$的时间序列。$P_u=\left(p_u(t_1),\cdots,p_u(t_n)\right)$是栏目的更新模式，是$t_i$时刻的更新频率$p_u(t_i)$随时间变化的时间序列。$\theta(t_i)$是随时间$t_i$变化的权重函数。$tw$是文本时间信息特征函数，$f_{tw}$为时间词词频，$s_{tw}(i)$为时间词$i$的位移，$L$为网页文本长度。

2　Web时态对象模型建立

采用层次树的形式，对Web信息内容和时态信息两者统一建模，树的根结点、中间节点和叶子节点分别代表网站、栏目或子栏目、网页，每一个结点由二元组(V_C,V_T)表示，V_C为文本向量，V_T为时态向量。V_C包括标题、网页链接、网页主题、网页文本等维度，V_T包含事件发生时间、发表时间、转载时间和过期时间等维度，以及描述事件的将来时、过去时等时态信息。

2.1　Web时态对象模型层次树

Web信息除包括文本内容信息外，还包含时态信息。现有方法对Web建模一般只关注其内容及其内容挖掘，而忽略了时态信息。本文拓展现有模型，在Web信息的内容要素中加入了时态要素，对网站、栏目、子栏目和网页页面进行抽象，在模型中将网站描述成一棵五层非空树，网站主页是根节点，栏目及其各级子栏目是中间结点，网页页面是叶子结点。根据SEO的优化原则，每个网页最多离网站首页四次点击就能到达，所以将网站描述成一棵五层非空树，而且叶子结点的深度最大值为5。

在网站树中，每一个结点用N_j^i表示，N_j^i由一个二元组$(\boldsymbol{V}_C,\boldsymbol{V}_T)$表示，$\boldsymbol{V}_C$为内容向量，$\boldsymbol{V}_T$为时态向量。其中，内容向量$\boldsymbol{V}_C=(C_{\text{title}},\boldsymbol{V}_{\text{url}},\boldsymbol{V}_{\text{topic}},\boldsymbol{V}_{\text{text}})$是一个关于网页$w$与其描述的事件$e$的4维向量，包括网页标题$C_{\text{title}}$、网页链接$C_{\text{url}}$、网页主题$C_{\text{topic}}$、网页文本$C_{\text{text}}$。对于根节点，其内容向量可表示为$\boldsymbol{V}_C=(C_{\text{title}},C_{\text{url}},0,0)$，$C_{\text{title}}$为网站名，$C_{\text{url}}$为网站主页链接；对于中间节点，$C_{\text{title}}$为栏目名称，$C_{\text{url}}$为栏目主页链接。时态向量$\boldsymbol{V}_T=(T_{\text{occur}},T_{\text{publish}},T_{\text{forward}},T_{\text{expire}})$是一个关于网页$w$与其描述的事件$E$的4维向量，包括事件发生时间$T_{\text{occur}}$、发表时间$T_{\text{publish}}$、转载时间$T_{\text{forward}}$和过期时间$T_{\text{expire}}$。

2.2　时间知识的时态层次模型

时间知识的时态层次模型用于时态向量的抽取和推理。时态层次模型用来描述时间实体的时间类型、时间表示、时态、描述事件等概念的层次关系，时间类型包括时间点、时间区间、时间频率(比如两周一次)；时间表示包括显式时间、隐含时间和相对时间；时态包括现在时、过去时、将来时等；描述的事件包括区间事件(如开会)、瞬时事件(如车祸)、周期事件(如：每周做一次报告)。

3　Web时态信息抽取及性能评估

3.1　时态对象模型中时态向量抽取

本文利用时态信息特征词自动提取各时态要素，内容向量包括网页链接、标题、主题和文本，在内容向量的抽取方面采用较为成熟的基于正则表达式、模式代数、DOM树、关键字匹配等方法进行抽取。文档向量中的主题维度采用基于LDA的主题模型或者采用基于概率的统计模型进行抽取。

时态向量根据时态层次模型进行抽取，首先将时间信息按照出现的形式分为显式时间表示、隐含时间表示和相对时间表示，如2012年1月1日为显式时间表示，“上周一”为相对时间，而隐含时间一般为节日和重大事件的名称，如“元旦节”；其次基于时间本体，通过参照时间和时间的时态信息，对时间信息进行统一的标准化；最后根据时间信息在文档中出现的移位、标准化值、类型(如时间区间、时间点、时间频率)推理出时间信息的语义，从而得到时态向量的每一个维度。

3.2　Web时态信息抽取性能评估

Web时态信息抽取复杂度主要取决于涉及的领域和抽取的场景，一般的门户网站涉及的领域较广，有新闻、财经、娱乐、体育、科技、教育、女性、时尚等多个栏目，抽取场景包含公司并购、恐怖活动新闻、股票大盘走势预测等。其评估主要是评估从Web实体(网站、栏

目、子栏目、网页)中抽取内容V_C、时态向量V_T的时态分量值的抽取任务复杂度和抽取性能，性能评估指标主要采用精度和召回率，对于性能评估可以采用单指标和综合指标评估方法，单指标即单独采用精度和召回率进行性能评估，综合指标评估即使用精度和召回率的综合值。同时采用CNN网络学习数据的基本特征，根据数据与时序密切相关的特性，引入LSTM网络学习时态节点之间的时态约束关系，并根据已知节点的时间推理未知节点，将Attention机制引入预测模型，对其赋予不同的关注度，形成具有特征偏好的模型，获得更优的预测效果。

4 Web时态一致性约束与推理机制

4.1 Web时态一致性约束机制

4.1.1 网页内时态一致性约束机制

网页的信息有些是没有时效性的，比如一些文史类信息，对这类信息根据它的发布时间规定一个过期时间，因为随着时间的增长它们会慢慢变得不新鲜，变成时态不一致。而对于有时效性的信息，通过结合时间的过去、现在、将来等属性进行语义分析，将信息大致分为预测信息、实时报道信息、回顾报道信息三类，这些信息往往会在一段时间后失去价值，过期时间与网页信息的时间敏感度息息相关，时间敏感度越高，过期时间越短。在相同的时间敏感度下，预测类信息的过期时间相对于信息发布时间最长，回顾报道信息的过期时间相对于信息发布时间最短。有些时效性很强的信息可能会因为发布延迟而在发布之时就过期了。

4.1.2 网页与栏目之间时态一致性约束机制

对于栏目内的子网页，如果栏目中不含时态约束信息，那么网页与栏目的时态一致性约束就等同于网页内部的时态约束；如果栏目中有时态约束信息，那么就首先根据栏目的时态信息提取出基本过期时间，这个过期时间是对网页的基本时间限制，只有在当前读取时间同时在基本过期时间和网页自身过期时间之前时，才认为网页与栏目是时态一致的。例如“本周新闻”栏目里的网页除了需自身时间有效外，还要满足本周这一时间限制。

4.1.3 同类网站相同栏目之间时态一致性约束机制

对于同类网站相同栏目的时间信息，它们经常会出现描述同一个事件的不同时间信息，可以对这些描述信息做时态一致性比较，另外也可以同时读取同类网站相同栏目间的所有网页，对栏目的时态一致性做比较，这些都可以为网站评价提供参考。

4.2 Web时态一致性推理

根据Web时态一致性约束机制构建逻辑推理算子，以推断时态信息和约束关系。

4.2.1 由Web已知时态分量值到未知维度时间值的推理

网页的各时间维度之间具有一定的逻辑关系，据此可由已知时间维度信息推出未知时间维度信息。比如知道了事件已发生的时间和信息的发布时间，就可以推出事件的写稿时间，因为写稿时间在发生时间和发布时间之间。同理，阅读时间在发布时间和转发时间之间，知道发布时间和转发时间后也可以推出大致的阅读时间。另外，若知道参照事件的发生时间和关联时间，就可以推出信息描述事件的发生时间。

4.2.2 相同主题网页信息的时间一致性推理

相同主题的时态信息一般是近似的，当它们具备时态一致性时，时态关联很明显。通过研究相同主题网页时态信息，可以推理未明确包含的时间信息。从一部分网页推出的时态信息，往往可以当作相关网页的相应未知时态信息。比如“今日新闻”栏目里的网页发布时间往往是在同一天。另外网页与它的父节点网页和子节点网页常常描述同一个主题，上层网页的发布时间和过期时间均较下层网页的发布时间更晚。相邻两层网页或者同一层的相邻网页之间具有相似的时间维度。

4.2.3 不同网站的相同栏目的时态统计推理

对于一些大型的时态一致性非常好的网站，它们相同栏目的时态信息也有很大的相关性。通过不同网站的相同栏目的统计分析与比较，可以从已知的时态信息，推理得到未包含时间信息的网页时态信息。比如同一事件的发生时

间在不同网站相同栏目里可以当作是一致的，从而是可以互推的。另外通过比较它们的时态一致性，还可以得出不同网站的时效性排名。

4.2.4 Web时态一致性公理系统和代数运算系统

针对Web时态对象模型中各节点(网站、栏目、子栏目、网页)二元组(V_C, V_T)中时态向量V_T的各时态分量值，及其各节点之间的约束与推理关系，包括节点自身各分量之间、父节点与子节点之间、兄弟节点之间、不同树的节点之间的约束与推理关系，可分别定义若干运算算子；基于这些运算算子，可建立公理系统描述显然成立的约束与推理知识，并讨论该公理系统的完备性和有效性。同时，基于这些时态运算算子，可建立Web时态代数运算系统，得到该代数系统的运算规则。基于这些运算规则，可以优化约束与推理运算。

4.3 性能测试与评估

本算法在Inter酷睿i7 12700F CPU，内存32 GB，操作系统为ubuntu 12.04，GPU为GeForce RTX 3090显卡上测试，以NTDO(时态数据对象的数目)、RTDO(时态数据对象占全部数据对象的百分比)，AET(单个数据操作的平均执行时间)作为测评指标，以TCHP-2PL对比在实时方面的差异，测试不同的操作对性能的影响。在实时网页上的性能比较如图1所示。

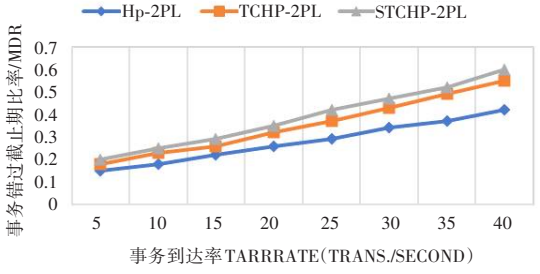


图1　三种实时并发控制协议在MDR上的比较

5 结语

本文对栏目与网页的时间敏感性特征进行分析，获得栏目与网页对时间敏感程度的度量方法，为Web时态一致性建模提供基础，提出加入时态信息要素的Web时态对象模型，对Web站点、栏目、子栏目和页面层次树抽象，提出时态向量的自动抽取方法，并研究Web时态一致性约束机制和推理机制，在基于时间感知的搜索排序、网站质量排序等方面有着重要的应用前景。

参考文献：

[1] ALONSO O, STRONGEN J, BAEZA-YATES R, et al. Temporal information retrieval: challenges and opportunities [C] //Proceedings of the 1st international temporal web analytics workshop (TWAW 2011), 2011: 1-8.

[2] MENG Y, RUMSHISKY A, ROMANOV A. Temporal information extraction for question answering using syntactic dependencies in an lSTM-based architecture[EB/OL]. arXiv: 1703.05851, 2017.

[3] RAHOMAN M M, ICHISE R. A proposal of a temporal semantics aware Linked Data information retrieval framework [J]. Journal of Intelligent Information Systems, 2017(11): 1-23.

[4] 沈思，李成名，吴鹏. 基于时态语义的Web信息检索实践进展与研究综述[J]. 中国图书馆学报，2018, 44(4): 109-129.

[5] 张梦妮. 面向网站无障碍评估的网页抽样方法研究[D]. 杭州：浙江大学，2018.

[6] 张鹏程，李必信，李雯睿. 时间属性序列图：语法和语义[J]. 软件学报，2010, 21(11): 2752-2767.

[7] 刘凯鹏，方滨兴. 一种基于社会性标注的网页排序算法[J]. 计算机学报，2010, 33(6): 1014-1023.

[8] 王伟，张文博，魏峻，等. 一种资源敏感的Web应用性能诊断方法[J]. 软件学报，2010, 21(2): 194-208.

[9] 陈传夫，唐琼，于媛，等. 网络上科学信息的时效性测量[J]. 情报学报，2009, 28(4): 610-617.

[10] DAI N, DAVISON B D. Capturing Page Freshness for Web Search [C] //SIGIR´10: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010: 871-872.

[11] YANG M H, WANG H X, LIM L, et al. Optimizing content freshness of relations extracted from the web using keyword search [C] //SIGMOD´10: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010: 819-830.

[12] CHO J, ROY S, ADAMS R E. Page quality: in search of an unbiased web ranking [C] // SIGMOD´05: Proceedings of the 2005 ACM

SIGMOD International Conference on Management of data,2005:551-562.
[13] 王晶晶,吴胜利.信息检索中支持隐式时间查询的文档排名方法[J].计算机工程与设计,2018,39(11):3380-3383.
[14] 白钰洁.基于开始定界符的自动Web信息抽取[J].微型电脑应用,2019,35(11):141-142,146.
[15] 刘子玉.基于Web的异构学术信息抽取与聚合方法研究[D].长春:吉林大学,2019.
[16] 李利敏.Web信息抽取与症状识别算法研究[D].郑州:河南大学,2018.
[17] 寇月,李冬,申德荣,等.D-EEM:一种基于DOM树的Deep Web实体抽取机制[J].计算机发展与研究,2010,47(5):858-865.

Research and analysis of Web temporal object model

Li Shuxia, Yang Juncheng*

(1. School of Electronic Information Engineering, Henan Polytechnic Institute, Nanyang 473000, China)

Abstract: This paper studies the Web temporal object model and adds temporal elements to the Web content elements. So as to solve the problem of unified modeling of site, column, sub column and content by hierarchical tree structure. According to the characteristics of data closely related to time series, this paper uses temporal feature words to automatically extract and evaluate various temporal elements, and proposes to use LSTM network to learn the temporal constraints between temporal nodes. At the same time, the Attention mechanism is introduced into the prediction model. Experimental results show that the system can automatically discover outdated Web information, and it has important application prospects in the quality ranking of similar websites and time aware search ranking, etc. It can greatly save manpower and improve the quality of Web information.

Keywords: deep learning; temporal element; temporal object model; temporal characteristic words; Hierarchical tree

(上接第50页)

Solving flow shop group scheduling by using collaborative multi-objective optimization algorithm

Xiao Xiumei*, Wang Xinrui

(School of Mathematics, Yunnan Normal University, Kuming 650500, China)

Abstract: This paper develops a collaborative multi-objective optimization algorithm(CMOEA) to solve the flow shop group scheduling problem with setup time constraints. First, the features of the considered problem are analyzed and described. Then, a three-dimensional vector is designed to represent each solution to list the three sub-problems. Next, a collaborative searching multi-objective algorithm is designed to solve the problem. Finally, after extensive experimental results and comparison with other algorithms, this algorithm significantly outperforms existing classical multi-objective optimization algorithms.

Keywords: flow shop group scheduling; multi-objective; energy consumption; collaborative algorithm

文章编号: 1007-1423（2023）08-0057-05 DOI: 10.3969/j.issn.1007-1423.2023.08.009

基于时释性和多维虚置换的代理重加密方案

周婉祎*，尹毅峰，王兆博

（郑州轻工业大学计算机与通信工程学院，郑州 450002）

摘要：信息技术的发展推动了数字化信息的应用，数据信息安全的问题也接踵而至，传统的代理重加密方案虽然可以解决数据安全共享的问题，但是不能解决数据共享的效率低和没有时效性的问题。因此提出了一种多用户数据信息共享的代理重加密方案，利用时释性、多维虚置换和代理重加密的结合实现用户之间的信息共享，数据请求者无法在指定时间之前获得关于数据的信息。通过安全性分析和性能评估，表明具有安全性能高、计算开销少的优点。

关键词：时释性；代理重加密；多维虚置换；对称加密

0 引言

随着信息科技和网络的飞速发展，数字化信息不断地涌入人们的生产和生活中。虽然数字化信息的便捷高效功能极大地促进了社会的发展，但是数据信息在传输或者保存的过程中，一旦被恶意篡改，就会被不法分子所利用，这不仅会影响到个人的安全和利益，而且还会影响企业乃至国家的安全和利益[1]。为此，代理重加密被提出，它为上述问题提供了一种新的解决方法，代理重加密是一种密文转换功能的加密密码体制，通过半可信的代理服务器可以将数据拥有者的密文转换成数据请求者的密文[2-5]。在重加密的过程中，不能通过密文和转换密钥获得任何关于明文的信息[6-7]，但是在有些场景对时效性有要求的情况下[8]，此技术还不够完善，这时就需要代理重加密所生成的密文具有时释性的特点。

Zheng等[9]利用一种基于类型的代理重加密实现细粒度共享，提出了一种新的数据共享方案，可以实现一般云计算环境下的差异访问控制，并且保护隐私泄露。Liang等[10]提出了一种支持定时释放的条件代理广播重加密方案，支持对重加密的条件委托，能够抵抗代理和接收者勾结。但是由于非对称加密算法效率没有对称加密算法高，不适合大型数据文件的加密。

针对以上的需求，提出了一种基于时释性的多维虚置换代理重加密方案，将多维虚置换机制、时释性和代理重加密相结合，不仅可以解决对时效性有要求的问题，而且又可以提高效率。用户通过多维虚置换生成安全的对称密钥，使用对称加密算法加密数据文件得到数据密文，然后使用代理重加密算法生成重加密密文。保证了方案的效率和安全性，提高了性能。

1 模型设计

为了保护数据信息的安全，更好地解决时效性问题，为其设计出多用户信息共享模型，如图1所示。

收稿日期：2022-12-20 **修稿日期：**2022-12-30

基金项目：国家自然科学基金项目（U1804263）；河南省自然科学基金面上项目（202300410508）；河南省自然科学基金项目（222300420371）；河南省高等学校重点科研项目（22A520047）；河南省科技攻关项目（142102210081）

作者简介：*通信作者：周婉祎（1997—），女，河南南阳人，硕士研究生，研究方向为信息安全、对称密码学，E-mail：zhouwanyiy@163.com；尹毅峰（1971—），男，河北饶阳人，博士，教授，研究方向为信息安全、对称密码学；王兆博（1997—），男，河南郑州人，硕士研究生，研究方向为信息安全、对称密码学

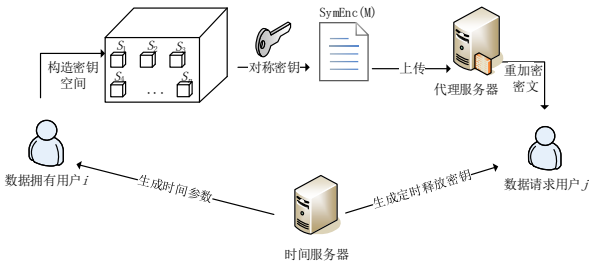


图 1 多用户信息共享模型

数据拥有者通过构造多维密钥空间，可以生成对称密钥，执行对称加密算法可以将数据明文加密成数据密文，并且上传到代理服务器中，由第三方可信的时间服务器生成时间参数，发送给数据请求者，可以使密文有时效性。

2 方案设计

本文用到的符号和其对应的含义如表1所示。

表 1 文中的符号和含义

符号	含义
sk_{ts}/pk_{ts}	时间服务器的私钥/公钥
sk_i/pk_i	用户i的私钥/公钥
sk_j/pk_j	用户j的私钥/公钥
T_G	定时释放的密钥
k_i^0	用户i的初始密钥
k_i'	用户i的初始密钥阵列
$k_i[n]$	用户i的密钥控制阵列
$IVPF(i)$	缺少第i个参数的虚置换函数
$MVPF$	完整的虚置换函数
$VITE^{(m)}$	迭代m次的虚迭代函数
C_T	一级密文
C_P	对称密钥密文
C_{TPRE}	重加密密文
$RK_{i\to j}$	用户i到j的代理重加密密钥

针对传统的代理重加密方案存在的缺陷，本文引入了时间服务器生成时间参数，并且结合了多维虚置换机制，提高传统方案的机密性和效率，根据多用户数据共享模型可以实现的方案具体阐述如下。

（1）系统参数初始化。选取素数q，阶数为q的有限循环群G和G_T，一个双线性映射$e:G\times G_T\to G_T$，g是G的生成元。定义哈希函数$H_0:\{0,1\}^*\to Z_q^*$，$H_1:\{0,1\}^*\to G$，$H_2:\{0,1\}\to G$，$H_3:G_T\to\{0,1\}^L$，从Z_q^*中随机选取一个元素r作为时间服务器的私钥$sk_{ts}=r$，并且计算时间服务器的公钥$pk_{ts}=g^r$。

（2）初始密钥生成。从Z_q^*中随机选取元素X_i，X_j分别作为数据拥有用户i和数据请求用户j的私钥，令$sk_i=X_i$，$sk_j=X_j$，则所对应的公钥分别是$pk_i=g^{X_i}$，$pk_j=g^{X_j}$。

（3）定时释放密钥生成。时间服务器可以生成时间参数$T\in\{0,1\}^*$，以此来表示发布的时间。通过时间服务器的私钥sk_{ts}和时间参数T，可以得到定时释放密钥T_G，如公式(1)。

$$T_G=H_1(T)^r \tag{1}$$

（4）对称密钥生成。此阶段需要构造出多维密钥空间，使用基于流密码的轻量级密码算法，构造多维虚置换函数，生成对称密钥，具体的步骤描述如下所示。

① 密钥空间构造。将数据拥有用户i的私钥k_i^0和私有密钥阵列k_i'通过哈希映射生成密钥控制阵列$k_i[n]$，将其发送给其他的用户，那么每个用户就可以拥有除了自身的私钥之外的不完整虚迭代函数$IVPF(i)$，如公式(2)。

$$k_i[n]=\prod_{a=1}^{n}\left(k_i^0\oplus H\left(k_i'[a]\right)\right) \tag{2}$$

数据拥有用户i把$k_i[n]$代入到不完整虚迭代函数中，就可以得到一个拥有完整密钥参数的多维虚迭代函数$MVPF$，即n维密钥空间，如公式(3)。其中每一个小空间都映射着一个安全子系统。

$$S=MVPF\left(k_1[n],k_2[n],\cdots,k_i[n],\cdots,k_n[n]\right) \tag{3}$$

② 迭代和置换。当每个用户随机选择密钥控制阵列中的m个密钥元素，m个安全子系统分别记作$\gamma_1,\gamma_2,\cdots,\gamma_m$，按照公式(4)进行迭代的操作，就可以获得安全的对称密钥K。

$$K=\gamma_m\left(VITE^{(m-1)}\right) \tag{4}$$

（5）一级密文生成。用户可以根据上一个阶段产生的对称密钥，用公开的对称加密算法实现对数据明文M的安全加密，可以得到所需要

的数据密文C_1，如公式(5)。选择随机数$\alpha \in Z_q^*$，$\omega \in G_T$，令$\varphi = H_0(M,\alpha)$，经过公式(6)计算可以得到密文C_2，然后使用时间服务器的公钥pk_{ts}、时间参数T，由公式(7)得到密文C_3。最后将所得到的这些密文经过公式(8)计算，得到一级密文C_T，同时上传到代理服务器中。

$$C_1 = SymEnc_K(M) \tag{5}$$

$$C_2 = g^{H_3(g^\alpha,\omega)} \tag{6}$$

$$C_3 = \omega \cdot e\left(H_1(T), pk_{ts}\right)^{\varphi} \tag{7}$$

$$C_T = H_2\left(C_1 \| C_2 \| C_3\right) \tag{8}$$

(6) 加密对称密钥。这个步骤是加密对称密钥生成密文，用于之后代理重加密生成重加密密文的过程，将对称密钥和用户的公钥通过公式(9)计算，就可以得到此对称密钥的密文C_P，并且上传到代理服务器中。

$$C_P = pk_i^{\alpha} \cdot K \tag{9}$$

(7)重加密密钥生成。将数据拥有用户i的私钥sk_i和数据请求用户j的公钥pk_j，通过哈希映射生成转换密钥RK_1，如公式(10)。然后把时间服务器生成的时间参数T和公钥pk_{ts}经过公式(11)的计算得到RK_2，最后由公式(12)得到重加密密钥$RK_{i\to j}$，并且上传到代理服务器中。

$$RK_1 = sk_i\left(H_3(g^\alpha, pk_j)\right) \tag{10}$$

$$RK_2 = \omega \oplus e\left(H_1(T), pk_{ts}\right)^{\alpha} \tag{11}$$

$$RK_{i\to j} = (RK_1, RK_2) \tag{12}$$

(8) 信息共享。输入对称密钥密文C_P和重加密密钥$RK_{i\to j}$，以及时间服务器的时间参数T，通过公式(13)可以计算出重加密密文C_{TPRE}。这个过程是由一个半信任的代理商进行的，不会泄露任何关于密文的消息，其安全性得到了保证。

$$C_{\text{TPRE}} = C_P \| e\left(H_2(T), RK_{i\to j}\right)^{\alpha} \tag{13}$$

数据请求者对重加密密文进行解密，可以得到明文M，首先将重加密密文C_{TPRE}和数据请求者的私钥sk_j进行公式(14)解密后得到对称密钥K，再使用K解密数据密文C_1，获得明文，如公式(15)。

$$K = Dec_{sk_j}(C_{\text{TPRE}}) \tag{14}$$

$$M = SymDec_K(C_1) \tag{15}$$

3　安全性分析和性能评估

3.1　安全性分析

3.1.1　重放攻击保护

在本方案中，只有在设定的时间有效期内数据请求方才可以解密获得明文，当数据请求者收到密文并将其解密后可以获得对应的反向密钥，因为该反向密钥存在一定的时效性，所以在这个时间到达之前，可以保存此反向密钥。然而当攻击者进行重放攻击时，解密信息获得相同的反向密钥后将丢弃该信息，不予传递此信息，并且不能获得服务器的认证，以此来确保时间陷门始终是限时的。因此本方案可以抵抗重放攻击。

3.1.2　中间人攻击保护

当攻击者介入用户之间，即使攻击者获得所有在公开网络信道传输的所有参数，但是其密钥信息即便是公钥也不会对外公开，因此中间人无法得到以密文形式传输的任何秘密信息。在生成对称密钥的过程中，并没有直接地传递其对称密钥，其中虚置换函数和多维密钥空间的构造都需要安全子系统，然而安全子系统并不能在公开信道中传输，所以攻击者就无法构建完整的多维密钥空间，也就不能进行后续的一系列操作。攻击者无法通过通信中产生的不完整迭代函数破译出用户的密钥阵列，同时由于多维虚置换函数具有自编译性，如果攻击者强行进行攻击破译的话，那么将会面临多维函数变异类，因此攻击者无法实施攻击。

3.1.3　选择密文安全

在任意的概率多项式时间内，解决DBDH问题的优势是可以忽略不计的，因此本文提出的方案是选择密文攻击安全。假设存在攻击者A和挑战者C。

A向C发出系统参数初始化、时间陷门生成、初始密钥的生成、对称密钥生成、一级密文生成、加密对称密钥、重加密密钥生成、代理重加密密文生成的相关询问，C向A返回查询的结果。C通过系统参数初始化和初始密钥生成算法，生成相应的公钥和私钥，以及时间服务

器的私钥，运行重加密密钥算法获得重加密密钥，并发送给A。A自己选择两条等长的明文M_0和M_1，发送给C，发起挑战。当A询问一级密文生成算法时，对称密钥是保密的。C选取随机比特$d \in \{0,1\}$，计算用于挑战询问的密文$C^* = Encryption(T_C^*, pk_{ts}^*, K^*, M_d, T^*)$返回给$A$。最终，$A$返回给$C$一个猜测值$d_1 \in \{0,1\}$，如果$d_1 = d$，则说明挑战成功。$A$在挑战中获胜的概率优势定义为$\varepsilon$，且$\varepsilon = \left| Pr[d_1 = d] - 1/2 \right|$可忽略不计，即$A$在概率多项式时间内猜对的概率可以忽略不计，那么$A$挑战失败，也可以推出本文的方案是选择密文安全。

3.2 性能评估

将本文方案的计算开销和文献[9]、文献[10]进行对比，设定T_p，T_m，T_e，T_h，$T_{E/D}$分别表示为双线性对、点乘运算、幂运算、哈希运算、一次对称加密/解密运算所需时间，根据表2中的比较结果可知，本文的方案在计算开销上小于文献[9]和文献[10]，因此我们所提出的方案具有较高的效率。

表2 各方案的时间复杂度对比

文献	密钥生成	代理重加密	总计
文献[9]	$2nT_e + nT_p$	$nT_m + nT_p$	$2nT_e + 2nT_p + nT_m$
文献[10]	$2nT_e + T_p + (2n+5)T_h$	$(n+1)T_e + 8T_p$	$(3n+1)T_e + 9T_p + (2n+5)T_h$
本文	$T_e + T_p + (n+2)T_h$	$nT_e + T_p + (n+1)T_h$	$(n+1)T_e + 2T_p + (2n+3)T_h$

4 结语

针对当前传统的代理重加密方案难以满足加密大量数据文件的效率和应对的时效性要求，设计出一种基于时效性的多维虚置换代理重加密方案，提供了高效灵活并且安全的数据共享服务。利用对称加密算法将数据明文加密，保护数据信息，在保证了数据安全和机密性的前提下，降低计算开销，提高其效率。相较于传统的代理重加密方案，这种方案适用于较大文件的数据共享，可以在短时间内完成，具有很大的优势和使用价值。

参考文献：

[1] MANZOOR A, LIYANAGE M, BRAEKE A, et al. Blockchain based proxy re-encryption scheme for secure IoT data sharing[C]//2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE, 2019: 99-103.

[2] AGYEKUM K, XIA Q, SIFAH E B, et al. A proxy re-encryption approach to secure data sharing in the Internet of things based on blockchain[J]. IEEE Systems Journal, 2021, 16(1): 1685-1696.

[3] DENG H, QIN Z, WU Q, et al. Flexible attribute-based proxy re-encryption for efficient data sharing[J]. Information Sciences, 2020, 511: 94-113.

[4] MAITI S, MISRA S. P2B: privacy preserving identity-based broadcast proxy re-encryption [J]. IEEE Transactions on Vehicular Technology, 2020, 69(5): 5610-5617.

[5] GE C, SUSILO W, BAEK J, et al. A verifiable and fair attribute-based proxy re-encryption scheme for data sharing in clouds[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 19(5): 2907-2919.

[6] ZHENG X, ZHOU Y, YE Y, et al. A cloud data deduplication scheme based on certificateless proxy re-encryption [J]. Journal of Systems Architecture, 2020, 102: 101666.

[7] GE C, SUSILO W, FANG L, et al. A CCA-secure key-policy attribute-based proxy re-encryption in the adaptive corruption model for dropbox data sharing system[J]. Designs, Codes and Cryptography, 2018, 86(11): 2587-2603.

[8] KAN G, JIN C, ZHU H, et al. An identity-based proxy re-encryption for data deduplication in cloud [J]. Journal of Systems Architecture, 2021, 121: 102332.

[9] ZHENG T, LUO Y, ZHOU T, et al. Towards differential access control and privacy-preserving for secure media data sharing in the cloud[J]. Computers & Security, 2022, 113: 102553.

[10] LIANG K, HUANG Q, SCHLEGEL R, et al. A conditional proxy broadcast re-encryption scheme supporting timed-release[C]//International Conference on Information Security Practice and Experience, Springer, Berlin, Heidelberg, 2013: 132-146.

（下转第66页）

文章编号: 1007-1423（2023）08-0061-06 DOI: 10.3969/j.issn.1007-1423.2023.08.010

基于区域位置特征数据搜索模型的电子签名笔迹检验研究

黄娟娟[1*]，薛宇航[2]，王　凡[1]

（1.湖南警察学院刑事科学技术系，长沙 410138；2.宁远县公安局桐山派出所，永州 425699）

摘要：探索动态特征与区域位置特征相结合的电子签名检测方法，提升对高水平练习摹仿电子签名的识别精度。研究采用静态分区法对电子签名所附的动态数据进行分析，并进行定量测试，通过数据统计分析，比较了区域位置特征划分前后同一人电子签名和不同人电子签名的动态数据的重合度。该方法能有效提升练习摹仿电子签名笔迹的识别精度，数据直观、准确、稳定，对电子签名笔迹识别的理论和实践具有一定的意义。

关键词：文件检验；动态特征；区域位置；电子签名笔迹；练习摹仿笔迹

0　引言

2004年出台，并于2015年修正的《中华人民共和国电子签名法》第2条规定：电子签名是指数据电文中以电子形式所含、所附用于识别签名人身份并表明签名人认可其中内容的数据[1]；第3条阐明了电子签名的合法性以及适用范围。近年来，电子签名在银行、保险、电子政务、商业合同、在线支付和在线结算等许多领域越来越普遍。

根据《电子签名法》第4条，要求电子签名“能够有形地表现所载内容，并可以随时调取查用的数据电文”，也就意味着电子签名笔迹不仅有有形的图像，还有记录书写过程的动态数据[2]。有形图像反映了笔迹的静态特征；动态数据主要是各类数位板记录的书写过程的详细数据，主要包括书写时间、行程位置、速度、加速度和压力，反映出笔迹的动态特征[3]。同时，电子签名的书写过程与普通签名笔迹不同。电子签名通常是通过在书写板和其他标记对象上移动特殊笔、手指或鼠标而形成的有形图像。由于记录下来的有形图像只是运动轨迹，能反映出来的书写信息远远比不上普通的书写原件，仅相当于复制件，且有的软件为减少后台数据储存量，对签名图像进行高程度压缩，其清晰度非常低，导致特征反映质量低，甚至远远不及复制件。因此，对于电子签名笔迹的鉴定如果只利用静态特征，按照传统笔迹鉴定程序开展鉴定，仅相当于对复制件开展鉴定，很难确保其科学性和准确性，因此必须与动态特征相结合，才能全面地反映书写过程，有效提升鉴定的准确性[4]。

国内外对于利用电子签名动态特征开展书写人同一性鉴别的研究主要是基于一些数据的算法，如函数算法改进以及与相关生物识别技术等方法的结合。国内大多数研究基本以书写时间和速度作为认证签名的分区标准[5-6]。研究表明，笔迹的动态特征也具有与静态特征相同的属性[7]。而将动态特征与汉字区域位置特征结合在一起分析电子签名笔迹书写人同一性的研究至今几乎没有。

本研究主要采用静态分区的方法对其所附的动态数据进行统计分析，研究了在起笔、收

收稿日期：2022-12-27　**修稿日期：**2023-02-13

基金项目：2021年湖南省大学生创新创业训练计划项目（湘教通〔2021〕197号）：电子签名动静态特征综合检验研究

作者简介：*通信作者：黄娟娟（1975—），女，湖南邵阳人，教授，硕士，研究方向为文件检验，E-mail：273207702@qq.com；薛宇航（1997—），女，湖南张家界人，本科，研究方向为刑事技术；王凡（2001—），男，湖南怀化人，在读本科，研究方向为刑事技术

笔、单字整体三个区域中书写速度、压力的变化规律，得出真实电子签名与摹仿电子签名在不同分区的区别，这种方法可有效提升高水平练习摹仿笔迹的识别准确度。研究发现，该方法结果直观，能通过图形直接观察到研究对象的内在联系，操作简单可行，具有创新性。

1 实验部分

1.1 实验器材

WACOM INTUOS数位板(INTUOS A6 USB)；SVC2004数据库。

从SVC2004数据库中选用20人的真实签名和20人的高水平练习摹仿签名数据。每个签名数据提供：有形图像和x坐标、y坐标、时间戳、按钮状态、方位角、高度和压力的详细数据。点位的压力值范围为0至2048级，1级等于9.8×10^{-4}N。

1.2 样本的预处理

选取真实签名数据(见图1和图2)和对应高水平摹仿签名数据(见图3和图4)，用Excel分别计算真实签名与高水平摹仿签名的速度和压力，作为写入时间的函数。

数据库中的所有数据按时间顺序排序，分为x轴、y轴位置数据和相应的压力数据。选择签名轨道上的任意点j，该点的速度可分为V_x、V_y，相邻两点的时间间隔为$\triangle t$，则可通过如下公式计算:

$$V_x(t_j)=\frac{x(t_j)-x(t_j-1)}{\triangle t} \tag{1}$$

$$V_y(t_j)=\frac{y(t_j)-y(t_j-1)}{\triangle t} \tag{2}$$

$$V_j=\sqrt{V_x^2+V_y^2} \tag{3}$$

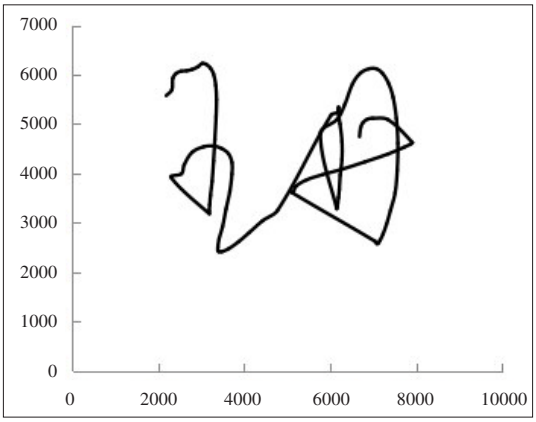


图1 数据库中真实签名1

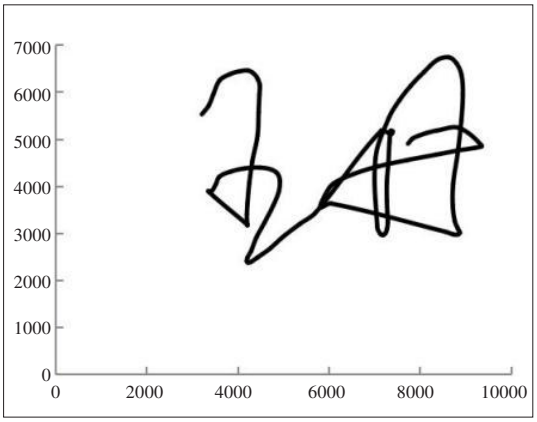


图2 数据库中真实签名2

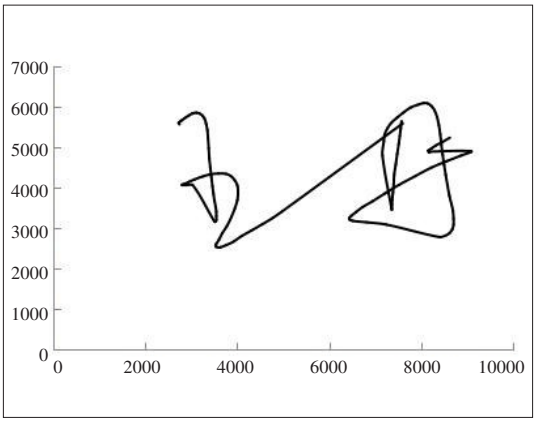


图3 数据库中摹仿签名1

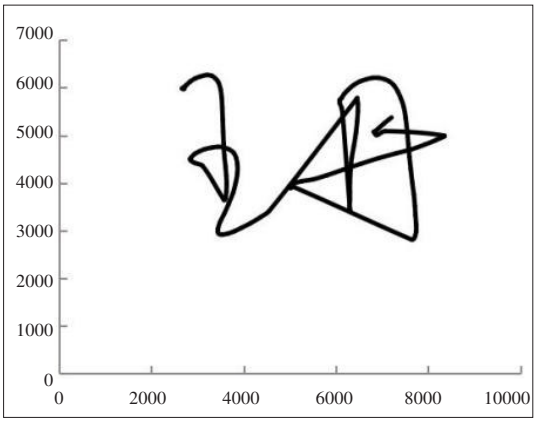


图4 数据库中摹仿签名2

2　结果与讨论

2.1　分区的原理及方法

2.1.1　分区的原理

练习摹仿签名笔迹是书写人先对被摹仿人的笔迹摹本进行比较分析，经过适当的摹仿练习后，脱离摹本凭记忆仿写形成的非正常笔迹。经过一段时间的摹仿练习，书写者能够感知和摹仿笔迹的大小、长度和结构关系，以及明显的书写顺序、写法特征，甚至其它一些细节特征，但不能准确摹仿出运笔的轻重缓急和复杂的连笔动作，大量研究表明，即使是高水平摹仿，在起收笔、字间连笔及单字压力、速度等方面存在明显差异。因此，从起收笔、字间连笔、单字这几部分进行分区研究，将静态特征与动态特征结合在一起分析是可行的。

2.1.2　分区的方法

第一区间为签名起笔至单字间连笔的起笔；第二区间为单字连笔起笔至连笔结束(无连笔情况分析起收笔)；第三区间为连笔收笔至单字书写完成。按区间分析动态数据。

2.2　实验数据的分析

计算得出签名笔迹书写过程中的线速度的大小和变化，再依据压力值绘制出曲线图。为了较为直观地表示签名笔迹动态特征，做出同一人多次书写签名笔迹的速度、压力分别随时间变化的曲线图(见图5～图8)，并按照动态特征值的大小，给笔迹中不同区间的笔迹部分着色，按特征值从大到小依次用红、橙、黄、绿、蓝、紫等颜色来表示，数值越大，所对应的数据坐标表示的颜色越深，直观表示特征值的具体变化情况(见图9和图10)。

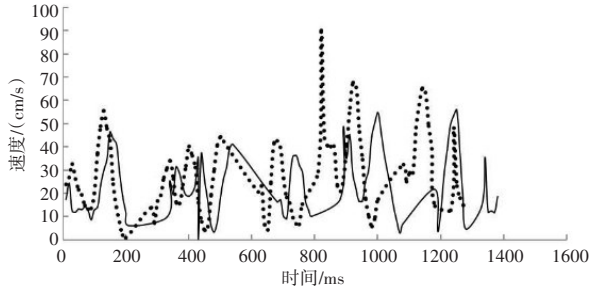


图5　真实签名1(虚线)、真实签名2(实线)的速度与时间变化曲线

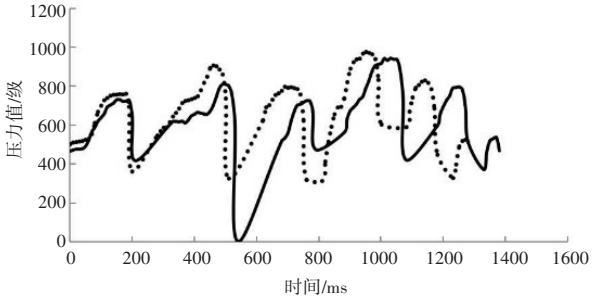


图6　真实签名1(虚线)、真实签名2(实线)的压力与时间变化曲线

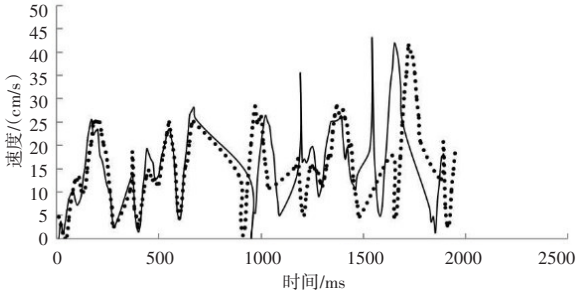


图7　摹仿签名1(虚线)、摹仿签名2(实现)的速度与时间变化曲线

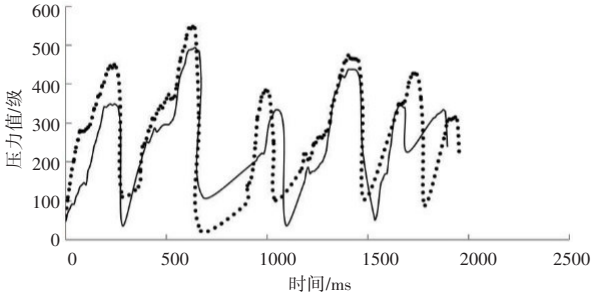


图8　摹仿签名1(虚线)、摹仿签名2(实线)的压力与时间变化曲线

图 9 真实签名1的压力变化

图 10 摹仿签名1的压力变化

使用Excel对真实签名、摹仿签名的原始数据分别进行计算，得出不同签名的重合度：同一人的动态特征图不仅具有相同的变化趋势，而且具有相似的变化幅度。动态特征图的形状高度相似，曲线之间的距离很小。签名笔迹的动态特征具有整体的特殊性，即当相同的笔画、部首和单个字符由不同的人书写时，其笔迹的动态特性是相似的，但其总和是根本不同的。也就是说，不同的书写人拥有属于自己系统的动态特征，从整体上来说，每个人都是不同的，真实签名与摹仿签名均有属于书写人自身特殊的动力定型；动态特征具有反映性和稳定性，反映了书写者的签名习惯(见表1)。

表 1　分区前重合度统计表

重合度	摹仿签名1/真实签名1	摹仿签名2/真实签名2	真实签名1/真实签名2	摹仿签名1/摹仿签名2
压力	40.60%	47.20%	90.70%	81.70%
速度	74.70%	51.00%	83.40%	77.10%

2.3　不同书写人的动态特征与区域位置特征的关系

2.3.1　不同书写人的书写压力与区域位置特征的关系

研究发现，相比于摹仿签名，真实签名的压力变化更丰富，书写时间更短，波峰、波谷更为密集，说明真实签名压力的动态特征变化更大，书写更流畅(见图11)。

同一人多次摹仿签名的压力、速度曲线图较相似，差异主要是由于摹仿时运笔受阻、心理因素或其他客观原因导致，高水平摹仿签名与真实签名均表现出各自动态特征的稳定性。可观察到，同一人签名笔迹曲线的起伏状态、波峰、波谷数量与比例关系基本一致(见图8)。

真实签名在第一区间初始起笔部分的压力值高于摹仿签名的压力值，第二、三区间逐渐增大，呈规律性起伏。真实签名单字部分数值基本同初始压力持平，较大的波峰波谷存在于单字笔画连笔处，曲线波动规律(见图9和图11)；摹仿签名单字压力由第一区间初始增加，同样第二、三区间随笔画规律增加和减少，但连笔部分压力较小，均接近于零，该特点体现出书写人有顿笔；收笔部分的真实签名的数值仍然大于摹仿签名的数值，与初始压力相持平(见图10和图11)。

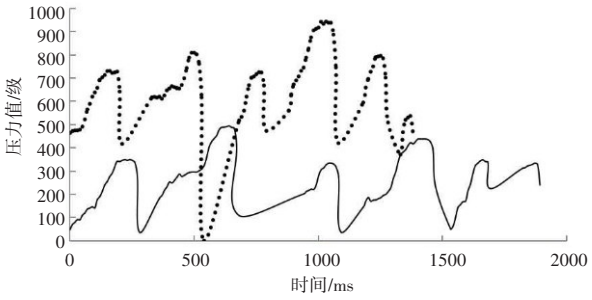


图 11 真实签名2(实线)、摹仿签名2(虚线)压力与时间变化对比

2.3.2　不同书写人的书写速度与区域位置特征的关系

第一区间，书写者在起笔部分书写本人姓名时，运笔流畅，一气呵成。即起笔时基本带有初速度(见图5)，练习摹仿的书写者因心理机制原因，会受自己本身动力定型、习惯动作的约束，起笔时通常初速度为零或接近于零(见图7)。

第二、三区间，真实签名的书写者和摹仿签名的摹仿者书写相同的笔画、偏旁或单字时，他们一般都遵守文字符号的书写规范，因

此在动态特征上，不同人之间有一定的相似之处[7]。而摹仿签名的单字整体呈现出速度缓的特点，最高峰值和速度平均值比例维持在2∶1左右。收笔部分真实签名的速度趋于初速度，摹仿签名速度则大于初速度，趋于整体平均速度(见图12)。

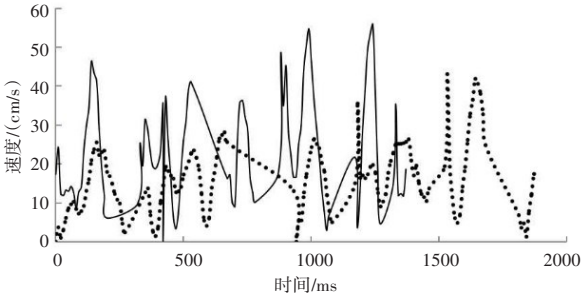


图12　真实签名2(实线)、摹仿签名2(虚线)速度与时间变化对比图

2.4　不同书写人动态特征中速度与压力的相互关系

同一人不同真实签名的压力与速度曲线图基本重合，差异体现在每次签名书写人因运笔受阻、心理因素或其他客观原因导致非本质差异。笔迹的动态特征存在相对稳定性，即同一人在书写同一笔画、同一偏旁、同一单字或同一签名时，笔迹的动态特征基本不变。然而，书写不是机械的重复，不同笔迹会产生不可避免的平行偏移[7](见图5和图6)。

练习摹仿签名具有较高的书写熟练程度，并且特征清晰。大部分的笔画都很流利，但有些笔画在书写中存在顿笔，笔尖压力平缓，有些连笔较僵硬，细小的连续笔画表现出抖动和不连贯等现象。真实签名与练习摹仿签名的书写者速度与压力在各区间(除字与字间的连笔部分外)均表现出压力大的速度也较大，压力小速度也小的特征；但字与字间的连笔部分相反(见图7和图8)。

2.5　动态特征结合区域位置特征分析对识别准确度的影响

从实验结果可知，真实签名和摹仿签名在三个不同分区中最主要特征为起笔时的初速度、单字整体的平均速度与压力的波峰与波谷的压力最值，及收笔处的压力值。据此可对不同分区的重合度进行计算，再利用Excel中的算法对真实签名、摹仿签名的原始数据分别进行计算，求各分区的均值，最终得出不同类型签名的重合度(见表2)。

表2　分区前、后重合度对比统计表

	分区前压力重合度	分区前速度重合度	分区后压力重合度	分区后速度重合度
摹仿签名1/真实签名1	40.60%	74.70%	35.32%	43.21%
摹仿签名2/真实签名2	47.20%	51.00%	40.11%	39.98%
真实签名1/真实签名2	90.70%	83.40%	93.88%	88.67%
摹仿签名1/摹仿签名2	81.70%	77.10%	85.31%	82.30%

由表2可见，结合区域位置特征进行动态特征的统计分析，同一人的真实签名或摹仿签名的重合度均有提升，说明分区分析后，能有效提升同一人笔迹的认定概率。此外，摹仿签名与真实签名的重合度有5%~7%的下降，说明结合区域位置分区分析，摹仿签名的识别得到了改进。

3　结语

高水平练习摹仿签名笔迹的鉴别一直是笔迹鉴定实践中的难点，电子签名的有形图像因其难以清晰、全面反映书写人的书写习惯，仅凭有形图像更是难以准确鉴别高水平练习摹仿笔迹。结合动态特征，虽能提升其识别的准确度，但差异不够明显。通过研究，发现按起笔、收笔及单字三个区间进行分区分析动态特征，将动态特征与区域位置特征相结合，能有效提升高水平摹仿笔迹的识别率。该方法数据直观、准确、稳定，对电子签名笔迹识别的理论与实践都具有一定意义。

参考文献：

[1]　史丕功. 关于电子商务合同相关法律问题的探讨[J]. 中国商贸,2012(1):159-160.

[2]　涂舜. 电子签名笔迹的动态特征:种类、价值及其运用[J]. 证据科学,2019,27(2):242-255.

[3]　向明. 电子签名笔迹在法庭科学检验中面临的问题

与现状[J]. 法制博览,2022(4):51-53.
[4] 彭兆君,刁文韬. 套摹电子式电子签名笔迹鉴定问题分析[J]. 法制博览,2022(22):92-94.
[5] 黄飞腾,郝红光,陈维娜,等. 电子签名笔迹的量化检验研究[J]. 辽宁警察学院学报,2020,22(2):70-78.
[6] 曾苗苗,王相臣. 电子签名笔迹时长特征和压力特征的实验研究[J]. 广东公安科技,2018,26(3):23-26.
[7] 陈晓红,贾玉文. 签名笔迹动态特征的理论研究[J]. 中国司法鉴定,2007(1):40-43.

Research on electronic signature handwriting inspection based on data search model of regional location features and dynamic features

Huang Juanjuan[1*], Xue Yuhang[2], Wang Fan[1]

(1. Department of Criminal Science and Technology, Hunan Police Academy, Changsha 410138, China;
2. Tongshan Police Station of Ningyuan County Public Security Bureau, Yongzhou 425699, China)

Abstract: To explore the method of combining dynamic characteristics with regional location characteristics to check electronic signature, so as to improve the identification accuracy of simulated electronic signature. Methods: The static partition method was used to analyze the dynamic data attached to the electronic signature, and the quantitative test was carried out. Results: Through statistical analysis of data, the variation rules of velocity and pressure in the dynamic characteristics of real signature and imitation signature were analyzed in different regions, and the coincidence degree of the dynamic data of the same person's electronic signature and different person's electronic signature was compared before and after the partitioning of regional location characteristics. Conclusions: This method can effectively improve the recognition accuracy of high level imitation handwriting, and the data is intuitive, accurate and stable, which is of certain significance to the theory and practice of electronic signature handwriting identification.

Keywords: document inspection; dynamic features; regional location; electronic signature handwriting; practice imitating handwriting

(上接第60页)

Proxy re-encryption scheme based on timed-release and multi-dimensional virtual permutation

Zhou Wanyi*, Yin Yifeng, Wang Zhaobo

(School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China)

Abstract: The development of information technology has promoted the application of digital information, and the problem of data information security has followed. Although the traditional proxy re-encryption scheme can solve the problem of secure data sharing, it cannot solve the problem of low efficiency and timeliness of data sharing. Therefore, a proxy re-encryption scheme for multi-user data information sharing is proposed, using a combination of timed-release, multi-dimensional virtual permutation and proxy re-encryption to achieve information sharing between users, where the data requestor cannot obtain information about the data before a specified time. The security analysis and performance evaluation show that it has the advantages of high security and low computing cost.

Keywords: timed-release; proxy re-encryption; multi-dimensional virtual permutation; symmetric encryption

图形图像

文章编号: 1007-1423（2023）08-0067-08 DOI: 10.3969/j.issn.1007-1423.2023.08.011

单图像超分辨率多尺度特征融合网络

赵光辉，杨晓敏*

（四川大学电子信息学院，成都 610065）

摘要：提出一种用于单图像超分辨率任务的多尺度特征融合网络，该网络包含特征下采样、多尺度特征融合和增加额外约束方案。首先，构建了包含Pixel-Unshuffle操作的特征下采样模块获得浅层多尺度特征，进一步学习深层多尺度特征融合，并通过多尺度特征融合提高超分辨率性能。此外，使用包含Pixel-Unshuffle操作的特征下采样模块从SR图像中重建LR图像，并计算其与LR图像之间的损失增加额外约束。大量的实验证明，提出的方法与现有的方法相比表现出具有竞争力的性能。

关键词：图像超分辨率；深度学习；Pixel-Unshuffle；特征融合

0 引言

单幅图像超分辨率（SISR）是计算机视觉领域中一项重要任务[1]，其目标是将低分辨率（LR）图像重构为高分辨率（HR）图像。在军事、医学、公共安全、计算机视觉等领域具有重要的应用前景。近年来，随着深度学习的兴起，基于卷积神经网络的超分辨率模型得到了积极的探索，并在各种基准数据集上取得了先进的性能表现。Dong等[2]首先提出了基于卷积神经网络（CNN）的方法，并取得了令人印象深刻的效果。随后，出现了更多基于CNN的SR方法[3-7]，并表现出了良好的性能。

尽管基于CNN的模型已经得到充分探索，但现有的方法仍存在一定的局限性。残差学习缓解了随着网络深度增大带来的梯度消失现象[8]，单纯的增加网络深度已经难以获得突破性的提升，而且网络深度的增加往往伴随着模型训练的困难和昂贵的计算资源[9]。一些经典的深度网络模型，如EDSR[4]、DBPN[9]和RCAN[3]，已经很难通过加深模型层来提高模型性能。因此，如何建立更高效的SR模型仍然是一个重要的问题。此外，图像的超分辨率是低分辨率图像的上采样任务，很少有研究者研究如何利用下采样特征来提高SR效果。文献[9]提出了采用迭代上投影单元和下投影单元的纠错反馈机制网络。文献[5]利用迭代下采样产生反馈，实现超分辨率增强。尽管他们必须采用迭代方案来抵抗信息丢失带来的性能下降，但证明了通过特征下采样操作提高SR性能的可行性。此外，最有效的学习策略尚不明了，很难确定学习到的端到端的非线性映射就是最优解[10]，因此，如何找到LR/HR/SR图像之间潜在的联系，寻求更精确的损失函数是一项有意义的研究工作。

通过特征下采样获取多尺度特征融合是一种有效改善视觉识别性能的方案，而现有的超分辨率模型几乎没有探索特征下采样方法在超分辨率任务中的应用。超分辨率是将低分辨率输入映射到高分辨率输出的过程，所以特征下采样对于图像超分辨率是一种逆直觉行为，另外下采样操作往往伴随着信息损失，因此它很少应用到超分辨率方法中。我们对特征下采样

收稿日期：2022-12-05 **修稿日期：**2023-01-03

基金项目：四川省科技厅科学基金（2021YFH0119）

作者简介：赵光辉（1998—），男，河南驻马店人，硕士研究生，研究方向图像处理；*通信作者：杨晓敏（1980—），女，四川广安人，博士，教授，研究方向为计算机视觉、图像处理，E-mail：arielyang@scu.edu.cn

方法在超分辨率中的应用进行了探索，利用Pixel-Unshuffle操作降低特征的尺度，同时避免信息损失，并通过多尺度特征融合增强特征表示。探索了一个多尺度特征融合额外约束方案的高效模型。具体而言，我们构建了一个由Pixel-Unshuffle、最大池化层和3*3卷积层组成的高效下采样模块(pixel-unshuffle downsampling block，PDB)来获取多尺度特征；提取浅层多尺度特征之后，利用残差注意力和Pixel-Shuffle[11]构建的RPUB模块（RCAB and pixel-shuffle up-sampling block)学习深层特征，注意每个RPUB模块都包含一次特征上采样操作，进而获得深层多尺度特征，并通过多尺度特征融合(multi-resolution feature fusion，MRF)模块将获取到的特征与对应尺度特征融合。最后，参考文献[12]中提出的方法，利用PDB模块将SR图像重构为低分辨率图像，该方法可以提供额外的约束条件，减少可能的函数空间。

本文的工作总结如下：

(1) 设计了一种高效的下采样模块(PDB)来生成浅层多尺度特征，该模块由Pixel-Unshuffle操作和非线性层两部分组成。与步长卷积相比，Pixel-Unshuffle运算可以在避免信息损失的同时降低特征分辨率，强大的非线性运算可以提取更复杂的局部特征；

(2) 利用残差注意力模块和Pixel-Shuffle构建了RPUB模块学习深层特征，生成深层多尺度特征，有效增加了网络深度；

(3) 提出了一种高效的多尺度特征融合模块(MRF)来融合多尺度特征。浅层多尺度特征表征关注纹理结构信息，深层多尺度特征更关注抽象本质。这种融合方法使本文网络绕过共有的信息，更多地关注特殊的特征，从而实现更具有判别性的特征学习。

1 相关工作

1.1 基于CNN的超分辨率网络

Dong等[2]首先提出了基于CNN的SR算法(SRCNN)，并取得了优于传统插值算法的性能。在这项开创性的工作之后，许多更深、更高效的超分辨率模型出现了。Haris等[9]提出了DBPN，利用传统算法迭代反投影提出了包含多个迭代的上采样层和下采样层形成误差反馈机制，通过不断迭代产生更优的解。Zhang等[3]研究了一种针对高精度图像SR的通道注意机制，提出了残差通道注意网络(RCAN)，使非常深的网络训练难度降低，但这些方法都有各自的局限性，很难学习到更精确的解。

1.2 Pixel-Unshuffle

Pixel-Unshuffle是Pixel-Shuffle的逆变换，Pixel-Shuffle是一种上采样方法，可以替代插值或反卷积的方法对特征图进行放大。主要功能是将低分辨的特征图，通过卷积和多通道间的重组得到高分辨率的特征图。Pixel-Shuffle是Shi等[11]提出的基于特征抽取和亚像素卷积将特征映射从LR空间转换到HR空间的有效方案。相应的，Pixel-Unshuffle是一种降低特征图尺度的下采样方法，它将形状张量(*，$C, H \times r, W \times r$)中的元素重新排列为形状张量(*，$C \times r^2, H, W$)，其中$r$是一个降尺度因子。图1展示了Pixel-Unshuffle的过程。通过该操作能得到包含原始特征全部信息的子特征，降低了特征分辨率[13]。

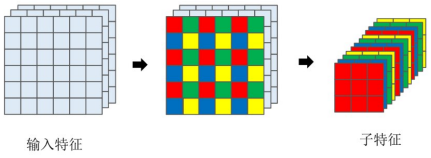


图1 Pixel-Unshuffle操作的可视化

1.3 对偶学习

不同于现有的学习范式，对偶学习方法同时训练两个相互的对偶模型，形成闭环。对偶学习的核心思想是建立对偶模型与原始任务模型之间的双向反馈通道，可以有效地通过信息反馈来完善特征信息，进而不断提升模型的性能。许多人工智能应用场景都涉及两个互为对偶的任务，例如机器翻译中从中文到英文翻译和从英文到中文的翻译互为对偶[14-15]、语音处理中语音识别和语音合成互为对偶[16]、图像理解中基于图像生成文本和基于文本生成图像互为对偶、问答系统中回答问题和生成问题互为对偶，以及在搜索引擎中给检索词查找相关的网页和给网页生成关键词互为对偶。这些互为对

偶的人工智能任务可以形成一个闭环，使对没有标注的数据进行学习成为可能[17]。我们借鉴了双重学习方法，构建了一个双重任务，为我们的SR模型提供额外的约束。

2　模型框架

本文提出了一个增加约束的多尺度特征融合网络（additional constraint in multi-resolution feature fusion network，ACMF）来处理超分辨率任务，整个网络结构如图2所示，其中虚线矩形框内为对偶任务。提取的浅层特征输入到特征下采样模块，通过不断向下进行获取多尺度特征，多尺度特征融合部分通过多分辨率特征融合进行特征学习，图像重建部分将学习到的特征恢复到图像中，对偶模型将从SR图像重构LR图像，提供额外的损失函数约束。值得注意的是，在输入网络提取浅层特征之前，我们先使用双三次插值算法将LR图像放大到SR的尺度。

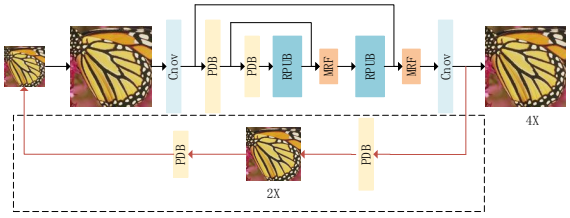


图 2　ACMF网络的结构图（x4）

2.1　浅层特征提取和重建部分

在将LR图像输入网络之前，我们先使用双三次插值算法将其放大到SR尺度。如图2所示，我们利用一个3*3卷积作为浅层特征提取部分，然后将每个PDB模块得到的浅层特征输送到MRF模块进行多尺度特征融合。需要注意的是，最后一个PDB模块的输出直接输入到RPUB模块学习深层多尺度特征，由于网络最终输出的深层特征已经被放大至SR尺度（详见2.2节），我们的图像重建部分只有一个3*3卷积，没有上采样层。重构层将学习到的特征恢复为图像。

2.2　多尺度特征融合

多尺度特征融合部分是网络的主干部分，包括PDB、RPUB和MRF三部分。多尺度特征融合过程的特征映射可视化如图3所示。下面分别介绍这些模块的详细细节。

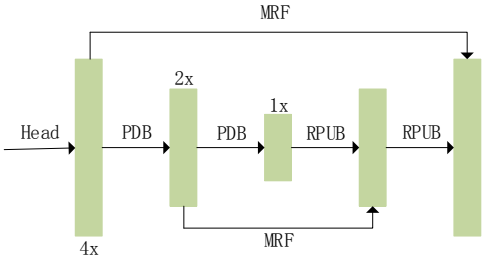
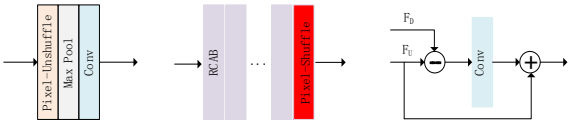


图 3　多尺度特征融合过程



（a）PDB模块　（b）RPUB模块　（c）MRF模块

图 4　模块结构图

2.2.1　PDB模块

PDB模块的细节如图4（a）所示。该模块由三部分组成：Pixel-Unshuffle、最大池化层和3*3标准卷积层。先前的研究[18-19]表明，低分辨率特征可以增强高分辨率表示。PDB模块作用于浅层特征获得浅层多尺度特征，通过特征融合增强特征表示。

Pixel-Unshuffle可以简单理解为Pixel-Shuffle操作的逆变换。它可以将原始特征变换为多个更低尺度的子特征，与原始特征相比，子特征包含了原始特征的全部信息，同时降低了特征的尺度。与步长卷积或插值相比，Pixel-Unshuffle在降低特征分辨率的同时避免了信息的丢失。在Pixel-Unshuffle操作之后，我们需要对低尺度子特征进行非线性操作以提取更好的局部特征。参考Sun[13]的工作，我们选择最大池化来处理子特征。这个过程描述为

$$F_{\text{out}} = M\left(P^{u}\left(F_{in}\right)\right) \tag{1}$$

其中：F_{out}代表输出特征，P^{u}表示Pixel-Unshuffle操作，M代表最大池化层（步长为1），F_{in}表示输出特征。这里F_{out}是通过非线性运算后得到的低尺度特征，与原始特征的子特征相比，它能表达更突出的局部特征。然后利用标准的1*1卷积

层来减少特征通道的数量。

2.2.2 RPUB模块

我们选择残差通道注意力模块(RCAB)[3]作为网络的基本块，使用B个RCAB块堆叠来增加网络深度。通过RCABs得到深层学习特征之后，利用Pixel-Shuffle来放大特征，如图4(b)所示，这个过程可以表述为

$$F_{\text{out}} = P\left(R_{b-1}(F_{b-1})\right) = P\left(\cdots R_0(F_0)\cdots\right) \quad (2)$$

其中：P代表Pixel-Shuffle操作，R_{b-1}代表第$(b-1)$个RCAB块，F_{b-1}则代表该块相应的输入。通过以上流程之后，就得到了放大后的深层特征，需要注意的是，每一次的深层特征放大都对应前面的一次PDB。

2.2.3 多尺度特征融合模块(MRF)

我们利用PDB模块和RPUB模块分别获得浅层多尺度特征和深层多尺度特征。接下来需要一种高效的融合方法来融合多尺度特征，所以我们构建了多尺度特征融合模块，该模块的细节如图4(c)所示。用F_D表示由PDB模块获取到的浅层特征，用F_U表示由RPUB模块得到的深层特征，这个融合过程可描述为

$$R_{UD} = F_U - F_D \quad (3)$$

$$F_{\text{out}} = C(R_{UD}) + F_U \quad (4)$$

其中：C表示标准的3*3卷积层，R_{UD}表示一个源中存在而另一个源中不存在的信息，F_{out}表示经过融合后得到深度残差特征。该融合模块能使网络绕开不同源特征中的相同部分，更加关注不同特征源中的特殊部分。该模块能提高网络的判别能力，并在融合不同来源特征的同时进行残差学习，与单纯的加法操作和拼接操作相比，有利于提高特征学习能力[20]。

2.3 增加约束的对偶学习

对偶学习方法同时学习两个相反的映射，通过对偶模型间的相互反馈提高初始任务模型的性能。在我们的网络中，初始任务模型的任务是训练SR模型。相应地，对偶任务是学习从SR到LR的下采样映射。参考文献[12]中的方法，我们引入了一个对偶学习任务，从SR图像重建LR图像。为了保证与SR网络的一致性，对偶学习任务通过PDB模块实现下采样映射(见图2中虚线框部分)。

如果重建的SR图像无限接近真实的HR图像，则相应的从SR图像下采样得到的图像应该是近似LR图像。意识到这一点，就可以通过增加额外约束来减小函数空间，优化训练时的损失函数。训练过程中，除了计算SR图像与HR图像之间的损失(L_P)之外，还引进了LR图像与对偶模型重建出的图像之间的损失(L_D)，最终损失函数(L)由L_P和L_D共同确定。其表达式为：$L = L_P + \lambda L_D$。训练中设置$\lambda = 0.1$。

3 实验结果

3.1 实验环境

3.1.1 数据集和指标

参考前人的研究工作[21-23]，我们选择数据集DIV2K[22]和Flickr2K[4]作为训练数据来训练我们的网络模型。选择以下标准基准数据集进行测试:Set5[24]，Set14[25]，B100[26]，Urban100[27]和Manga109[28]。选择了图像评价指标PSNR和SSIM[29]来衡量重建图片质量。将图像转换到YCbCr空间，并在Y通道上评估图像质量。对于BI退化模型，我们通过实验分别重建了比例因子为x4，x8的图像。

3.1.2 训练参数设置

我们将成对的LR和HR图像裁剪成大小为48×48的块作为训练数据集，并使用[3-4]中的数据增广方法。在将LR图像输入网络之前，我们先利用双三次插值对LR图像进行上采样，上采样因子为SR因子。设$r = 2$为Pixel-Unshuffle中的比例因子。对于4xSR，我们在RPUB模块中设置$B = 30$个RCAB块，并设置$\log_r(4)$个PDB模块和RPUB模块；对于8×SR，我们在RPUB模块中设置$B = 30$个RCAB块，并设置$\log_r(8)$个PDB模块和RPUB模块。采用ADAM优化器更新网络参数权重，指数衰减速率设置为$\beta_1 = 0.9$，$\beta_2 = 0.999$，初始学习率设置为10^{-4}，经过106次迭代的余弦退火学习率衰减到10^{-7}。

3.2 结果对比

为了证明所提方法的有效性，我们将训练模型重建图像的评估结果与以下经典方法进行比较：ESPCN[11]，SRGAN[30]，LapSRN[6]，Dense-

Net[31]，EDSR[4]，DBPN[9]，RCAN[3]，SAN[32]，RRDB[33]和DRN[12]。表1列出了所有与BI退化模型的对比结果。从结果来看，我们的ACMF网络与以上方法相比，可以取得更好或相当的性能。对于4xSR，ACMF在Set14和Manga109上表现最好，SSIM指标在Set5上表现出最好结果。特别是对于8xSR，ACMF在大多数基准数据集上的性能都远远优于其他方法，这也证明了我们的网络具有更强的泛化能力。

对于图像的视觉质量比较，我们将重建的SR图像与其他方法重建的图像进行视觉效果对比。

表1　BI退化模型下各基准数据集下的结果比较

Method	Scale	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSDS100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	4	28.42/0.810	26.10/0.702	25.96/0.667	23.15/0.657	24.92/0.789
ESPCN		29.21/0.851	26.40/0.744	25.50/0.696	24.02/0.726	23.55/0.795
SRGAN		29.46/0.838	26.60/0.718	25.74/0.666	24.50/0.736	27.79/0.856
LapSRN		31.54/0.885	28.09/0.770	27.31/0.727	25.21/0.756	29.09/0.890
DenseNet		32.02/0.893	28.50/0.778	27.53/0.733	26.05/0.781	29.49/0.899
EDSR		32.48/0.898	28.81/0.787	27.72/0.742	26.64/0.803	31.03/0.915
DBPN		32.42/0.897	28.75/0.786	27.67/0.739	26.38/0.794	30.90/0.913
RCAN		32.63/0.900	28.85/0.788	27.74/0.743	26.74/0.806	31.19/0.917
SAN		32.64/0.900	28.92/0.788	27.79/0.743	26.79/0.806	31.18/0.916
RRDB		**32.73/0.901**	28.97/0.790	**27.83/0.745**	**27.02/0.815**	31.64/0.919
DRN-S		32.68/0.901	28.93/0.790	27.78/0.744	26.84/0.807	31.51/0.919
ACMF		32.71/**0.901**	**28.97/0.790**	27.80/0.744	26.99/0.811	**31.64/0.920**
Bicubic	8	24.39/0.657	23.19/0.568	23.67/0.547	20.74/0.515	21.47/0.649
ESPCN		25.02/0.697	23.45/0.598	23.92/0.574	21.20/0.554	22.04/0.683
SRGAN		23.04/0.626	21.57/0.495	21.78/0.442	19.64/0.468	20.42/0.625
LapSRN		26.14/0.737	24.35/0.620	24.54/0.585	21.81/0.580	23.39/0.734
DenseNet		25.99/0.704	24.23/0.581	24.45/0.530	21.67/0.562	23.09/0.712
EDSR		27.03/0.774	25.05/0.641	24.80/0.595	22.55/0.618	24.54/0.775
DBPN		27.25/0.786	25.14/0.649	24.90/0.602	22.72/0.631	25.14/0.798
RCAN		27.31/0.787	25.23/0.651	24.96/0.605	22.97/0.643	25.23/0.802
SAN		27.22/0.782	25.14/0.647	24.88/0.601	22.70/0.631	24.85/0.790
DRN-S		27.41/0.790	25.25/0.652	24.98/0.605	22.96/0.641	25.30/0.805
ACMF		**27.41/0.791**	**25.38/0.653**	**24.99/0.605**	**23.05/0.644**	**25.37/0.806**

注：加粗表示指标最高，红色字体表示次高

图5和图6展示了对比结果，对于图像“img002(x4)”，与其他方法产生的图像整体模糊相比，我们的图像具有更高的对比度。对于图像“img015(x4)”，大多数方法都容易产生伪影或生成的轮廓不够清晰。相比之下，本文方法恢复了更准确的细节和更尖锐的轮廓。对于图像“img024(x4)”，本文方法实现了其他比较方法无法恢复的线细节。对于图像“img070(x4)”，可以清楚地看到其他方法生成的图片中城堡线轮廓的失真，而我们的方法恢复了接近真实的线轮廓。此外，在图6中，本文模型重建的SR图像对于8×SR具有更清晰的边缘和形状。这些对比结果进一步证明了本文方法可以提取更复杂的特征，并且重建出更接近真实的图像。

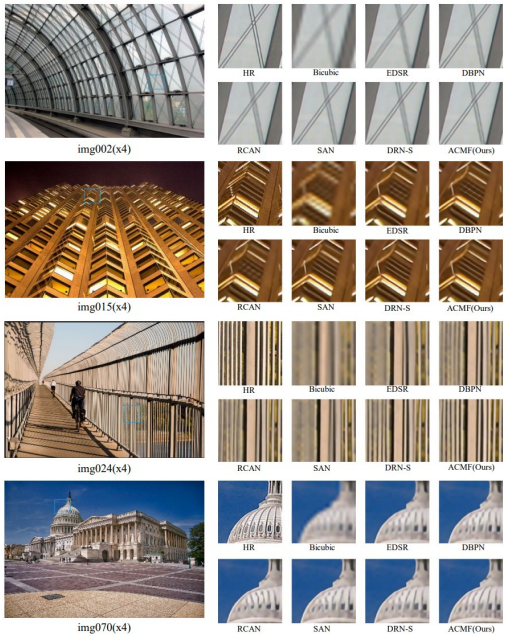


图 5　4xSR视觉效果对比

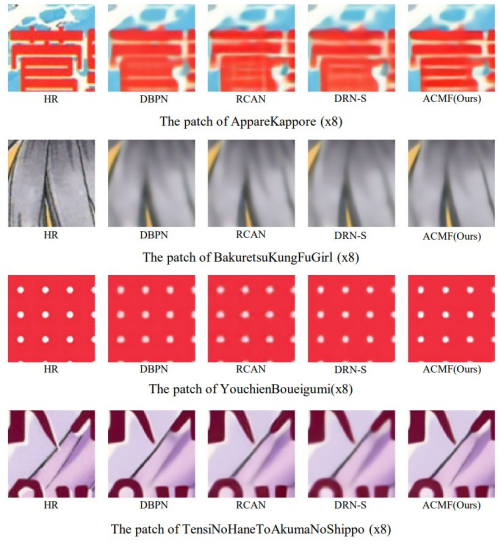


图 6　8xSR视觉效果对比

3.3　消融实验

通过进一步的实验证明了所提方法的有效性。为了便于比较，以步长为2的卷积作为PDB模块的基准；以加法特征融合作为MRF模块的基准，分别研究了PDB模块和MRF模块对实验结果的影响。4×SR不同组合消融结果见表2。

3.3.1　PDB模块的影响

前节讨论了PDB相比于步长卷积，既能保证信息的完整性，又能与Pixel-Shuffle操作对称互补。从表2的结果来看，仅使用PDB模块的模型比基准模型（使用步长卷积和加法特征融合的模型）的PSNR结果高出0.13 dB；使用PDB模块和MRF模块的模型比仅使用MRF的模型PSNR结果高出0.03 dB。这主要是因为PDB中的Pixel-Unshuffle操作可以在降低特征分辨率的同时避免信息丢失，下采样过程保留了相对完整的特征信息。得益于互为逆运算的上下特征采样操作，下采样特征和上采样特征形成信息对称互补。这些结果表明，我们的PDB模块可以学习到更完整的信息，提高SR的性能。

表 2　不同模块对网络模型的影响

模块	结果1	结果2	结果3	结果4
PDB 步长卷积	√	√	√	√
MRF 加法融合	√	√	√	√
PSNR（Set5）	32.53	32.68	32.66	32.71

3.3.2　MRF模块的影响

前面讨论了MRF有利于提高特征学习能力，并在融合不同特征源时进行残差学习。从表2的结果来看，仅使用MRF模块的模型比基准模型（使用步长卷积和加法特征融合的模型）的PSNR结果高出0.15 dB；使用PDB模块和MRF模块的模型比仅使用PDB的模型PSNR结果高出0.05 dB。这是因为MRF更关注不同特征源之间的特殊部分，同时进行残差学习，能够提高模型辨别能力，学习到更好的特征。这些结果表明MRF模块提高了特征融合效果。

4　结语

提出了一种附加约束的多尺度特征融合网络（ACMF）。ACMF框架有效地挖掘了浅层下采样特征信息，并进行了高效的多尺度特征融合，提高了SR任务的性能。我们利用Pixel-Unshuffle操作来降低特征分辨率，同时避免信息丢失，并通过多尺度特征融合增强特征表示。此外，还学习了获取重构LR图像的对偶任务，以增加额外的约束来优化损失函数。在基准测试上的大量实验表明，我们的模型表现出相当有竞争力的性能。

参考文献：

[1] BEYERER J, LEÓN F P, FRESE C. Machine vision: automated visual inspection: theory, practice and applications[M]. Springer, 2015.

[2] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(2): 295-307.

[3] ZHANG Y, LI K, LI K, et al. Image super-resolution using very deep residual channel attention networks [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 286-301.

[4] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 136-144.

[5] LI Z, YANG J, LIU Z, et al. Feedback network for image super-resolution [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3867-3876.

[6] LAI W-S, HUANG J-B, AHUJA N, et al. Deep laplacian pyramid networks for fast and accurate super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 624-632.

[7] ZHANG Y, TIAN Y, KONG Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2472-2481.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[9] HARIS M, SHAKHNAROVICH G, UKITA N. Deep back-projection networks for super-resolution [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1664-1673.

[10] ULYANOV D, VEDALDI A, LEMPITSKY V. Deep image prior [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9446-9454.

[11] SHI W, CABALLERO J, HUSZÁR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1874-1883.

[12] GUO Y, CHEN J, WANG J, et al. Closed-loop matters: dual regression networks for single image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2020: 5407-5416.

[13] SUN B, ZHANG Y, JIANG S, et al. Hybrid pixel-unshuffled network for lightweight image super-resolution[EB/OL]. arXiv: 2203.08921, 2022.

[14] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4320-4328.

[15] XIA Y, TAN X, TIAN F, et al. Model-level dual learning[C]//Proceedings of the International Conference on Machine Learning. Stockholm: ACM, 2018: 5383-5392.

[16] HE D, XIA Y, QIN T, et al. Dual learning for machine translation[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016: 820-828.

[17] XIA Y, QIN T, CHEN W, et al. Dual supervised learning[C]//Proceedings of the International Conference on Machine Learning. Sydney: ACM, 2017: 3789-3798.

[18] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5693-5703.

[19] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3349-3364.

[20] MEI Y, FAN Y, ZHOU Y, et al. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2020: 5690-5699.

[21] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network [C]//Proceedings of the 28th International Confer-

ence on Neural Information Processing Systems, 2015:1135-1143.
[22] TIMOFTE R, AGUSTSSON E, VAN GOOL L, et al. Ntire 2017 challenge on single image super-resolution: methods and results[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017:114-125.
[23] ZHANG K, ZUO W, ZHANG L. Learning a single convolutional super-resolution network for multiple degradations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE, 2018:3262-3271.
[24] BEVILACQUA M, ROUMY A, GUILLEMOT C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding[C]// Electronic Proceedings of the British Machine Vision Conference 2012, 2012:135.1-135.10.
[25] HUANG G, LIU S, VAN DER MAATEN L, et al. CondenseNet: an efficient denseNet using learned group convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE, 2018:2752-2761.
[26] MARTIN D, FOWLKES C, TAL D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]//Proceedings Eighth IEEE International Conference on Computer Vision. Vancouver:IEEE, 2001:416-423.
[27] HUANG J-B, SINGH A, AHUJA N. Single image super-resolution from transformed self-exemplars[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston:IEEE, 2015:5197-5206.
[28] MATSUI Y, ITO K, ARAMAKI Y, et al. Sketch-based manga retrieval using manga109 dataset[J]. Multimedia Tools and Applications, 2017, 76(20):21811-21838.
[29] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4):600-612.
[30] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu:IEEE, 2017:4681-4690.
[31] TONG T, LI G, LIU X, et al. Image super-resolution using dense skip connections[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice:IEEE, 2017:4799-4807.
[32] DAI T, CAI J, ZHANG Y, et al. Second-order attention network for single image super-resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019:11065-11074.
[33] WANG X, YU K, WU S, et al. Esrgan: enhanced super-resolution generative adversarial networks[C]// Proceedings of the European Conference on Computer Vision(ECCV). Berlin:Springer, 2018:63-79.

Image super-resolution using additional constraint in multi-resolution feature fusion network

Zhao Guanghui, Yang Xiaomin*

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: A multi-resolution feature fusion network with additional constraint for image SR is provided in this paper. The network contains feature downsampling and multi-resolution feature fusion as well as an augmentation constraint scheme. Specifically, we obtain multi-resolution features through pixel-unshuffle downsampling block (PDB), then fuse these multi-resolution features with the deep features. We utilize the multi-resolution feature fusion (MRF) module to improve image SR performance. In addition, we use the pixel-unshuffle downsampling block (PDB) module to estimate LR images from SR images, which provide an extra constraint on SR model. Extensive experiments demonstrate that our method achieves competitive performance against existing methods.

Keywords: super resolution; deep learning; pixel-unshuffle; feature fusion

文章编号: 1007-1423（2023）08-0075-08　　DOI: 10.3969/j.issn.1007-1423.2023.08.012

高精度三维视觉技术在智能制造上的应用研究

张广宇*，徐　罕，魏亚峰，赵雨琨，李守委，吉　勇

（无锡中微高科电子有限公司，无锡 214035）

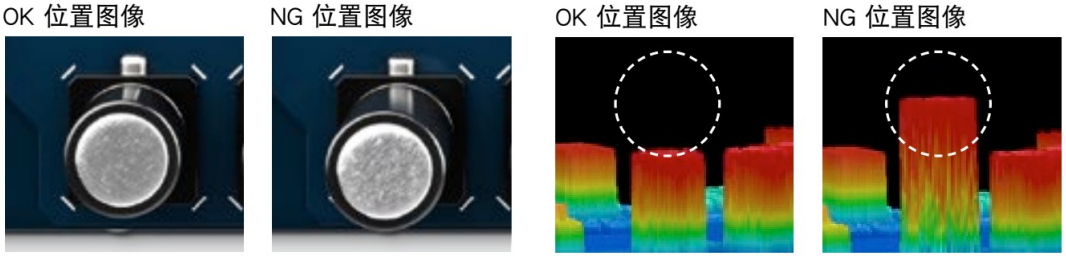
摘要：随着科技的发展，智能制造装备的高效自动化检测功能对机器视觉三维重建技术及其重建精度与效率提出了更高要求。分析了高精度三维视觉重建算法测量原理与关键技术，对线激光三角测距法、相位轮廓测量术、光度立体视觉法三种方法在智能制造装备应用中可能存在的问题，并对智能制造装备三维视觉技术的发展方向进行了展望。

关键词：三维成像；高精度测量；智能制造；立体视觉

0 引言

机器视觉在现代智能制造领域中的重要性越来越高[1-2]，传统的2D视觉技术将三维空间的待测物品通过投影变换映射到二维图像上[3-4]，丢失了成像空间坐标系的Z轴信息，导致与深度信息相关的视觉检测无法进行。如图1所示，不同高度的电子元件在2D成像模式下差异不明显，但通过三维重建算法采集待测品的深度信息，并通过对点云图像信息的处理，即可实现三维视觉的自动化检测。

在智能制造装备中，三维重建技术被广泛应用于测量产品高度、体积、表面角度、平面度等信息，并可对透明或颜色相近的物体进行特征区分[5-6]。三维重建算法种类很多，按照拍照视角数量可分为多目法与单目法；按照是否借助信号调制光源分为主动成像法与被动成像法；按照成像原理是否借助光学原理分析可分为光学成像法与非光学成像法。智能制造领域对三维重建精度与效率均有较高要求，对检测范围要求较小[7-8]。通过分析三维重建算法的特点，本文着重对适用于智能制造装备的三角测距法、相位轮廓测量术、光度立体法这三种高精度高效的三维重建算法进行研究，分析算法的技术原理以及在智能制造不同领域的应用[6]。



（a）2D成像效果　　（b）3D成像效果

图 1　2D成像效果与3D成像效果对比

收稿日期：2022-12-20　　**修稿日期：**2023-02-22

作者简介：*通信作者：张广宇（1990—），男，江苏沛县人，硕士，工程师，主要从事自动光学检测算法研发工作，E-mail：351894071@qq.com；徐罕（1989—），男，江苏淮安人，本科，工程师，主要从事晶圆级先进封装研究及封装自动化设计相关工作；魏亚峰（1996—），男，山西怀仁人，硕士，助理工程师，主要从事功率器件封装研究工作；赵雨琨（1996—），男，江苏南京人，硕士，助理工程师，主要从事自动化设计相关工作；李守委（1988—），男，山东临沂人，硕士，高级工程师，主要从事系统级封装研发相关工作；吉勇（1985—），男，江苏泰州人，本科，高级工程师，主要从事高可靠封装设计及管理相关工作

1 三维视觉重建算法原理

通过对三维视觉重建算法原理进行分析，从而深入了解算法的物理条件与实施方式，帮助快速定位该算法在智能制造的应用领域。

1.1 线激光三角测距法

三角测距法是一种应用较广的三维成像方法[9]，三角测距系统通过激光器发射线激光在空间中与相机内部构建两个相似三角形，计算标定参数获取两个三角形对应信息。当产品在Z轴方向上有深度变化，则空间三角形发生变化，通过计算相机内部三角形产生的相应变化即可获取产品位移量。目前工业产品表面3D信息提取主要基于三角测距法的3D相机采用线激光方式，国内外均有不同品牌厂商可以生产，技术方案比较成熟。3D相机机械结构大多如图2(a)所示，使用时激光通过柱面物镜调制，方向垂直于待测品。高精度CMOS相机倾斜固定角度采集激光图像，为防止环境光源干扰，需在镜头前添加激光波长滤波器。相机采用移轴镜头，防止虚焦，增加可检测深度范围。使用方式与线阵相机一样，待测品与相机间保持相对匀速运动。

具体工作原理如图2(b)所示，ΔABO是空间三角形，ΔFEB是相机内三角形，且$\Delta ABO \backsim \Delta FEB$，易知$AO = \frac{AB \cdot BF}{FE}$，其中AB为相机限定的已知长度，$BF$为相机镜头的焦距$f$，$FE$可由像元尺寸与像素坐标相乘得出，从而完成深度信息的采集。当待测品表面发生高度变化时，变化量y的公式为

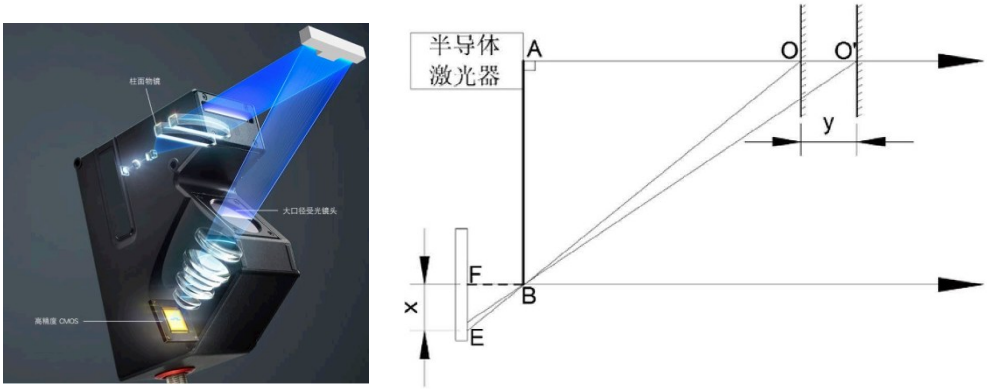
$$y = AO' - AO = \frac{f \cdot l_{AB}}{Pixel_{size}}\left(\frac{1}{Pos'} - \frac{1}{Pos}\right) \quad (1)$$

其中：l_{AB}为AB的长度，$Pixel_{size}$为像元尺寸，Pos'为高度变化后激光在图像中的成像坐标，Pos为高度变化前激光在图像中的成像坐标。由式(1)可知，基于三角测距法的3D相机检测精度与像元尺寸与检测范围相关，像元尺寸越小，检测精度越高；检测范围越小，检测精度越高。实际项目中线激光相机的选型需根据检测需求与检测能力权衡确定。

1.2 相位测量轮廓术

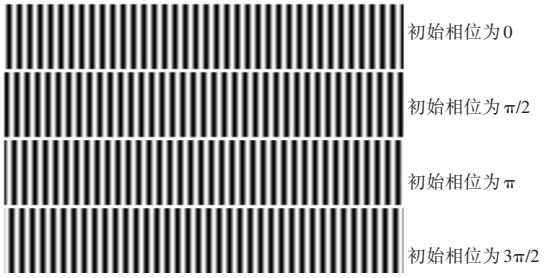
在使用正弦编码的光栅条纹检测三维信息的方法中，相位测量轮廓术(phase measuring profilometry，PMP)是面结构光三维重建较为重要的方法[10-11]，目前在工业产品三维重建中有广泛的应用，国外已有多家品牌商量产了基于该技术的3D相机，在保证Z轴计算精度高的同时，提高了可测量高低落差的范围，并且可一次性重建整个待测面的三维信息，与基于三角测距法的线激光3D相机相比，提高了待测品的检测效率。

PMP使用投影仪分n次投射相位间隔$\frac{2\pi}{n}$的正弦光栅条纹，例如使用四步相移法提取光栅相位时，$n = 4$，相邻投射的正弦光栅相位差为$\frac{2\pi}{4}$，如图3(a)所示。

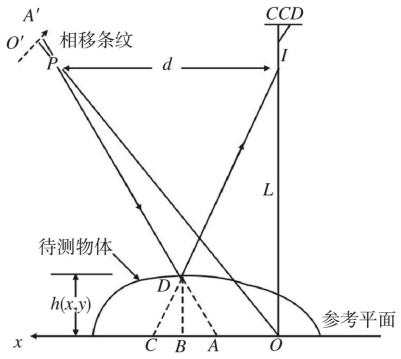


(a) 线激光结构图　　(b) 线激光测量原理图

图2 线激光测量仪结构原理



（a）四步相移投射光栅序列



（b）PMP光路原理示意图

图3　PMP工作原理

当正弦光栅条纹投射到待测品表面时，正弦光栅的相位随表面高度信息而发生相应偏移，其中光路原理如图3(b)所示，正弦光栅由P点发出照射到待测品上，相机由I点采集到光栅图像，取坐标为(x,y)的D点，以光路构建三角形可得$\Delta PDI\backsim\Delta CDA$，由此易知$\dfrac{L-h(x,y)}{h(x,y)}=\dfrac{d}{CA}$，推导出：

$$h(x,y)=\frac{CA\cdot L}{CA+d} \tag{2}$$

同时相位差$\Delta\varphi(x,y)=\varphi_C-\varphi_A=2\pi f\cdot CA$，则$CA=\dfrac{\Delta\varphi(x,y)}{2\pi f}$,带入上式可得：

$$h(x,y)=\frac{L\cdot\Delta\varphi(x,y)}{2\pi fd+\Delta\varphi(x,y)} \tag{3}$$

其中：$h(x,y)$为D点的Z轴高度值，$\Delta\varphi(x,y)$为D点的相位差值，L为相机到参考平面的距离，d为P点到I点的距离，f为参考平面上光栅频率，可由相邻光栅间距得出。使用相移法求取，相机采集到光栅图像的光强分布函数为

$$I_i(x,y)=A(x,y)+B(x,y)\cdot\cos\left[\varphi(x,y)+\delta\right] \tag{4}$$

$A(x,y)$为待测面背景光强，$B(x,y)$为待测面调制度，$\varphi(x,y)$为待求相位，δ为光栅相移量，i表示第i步相移。四步相移就是对(x,y)点投射四次相位差为$\dfrac{\pi}{2}$的光栅，求取四次光强值，分别是：

$$\begin{cases}I_1(x,y)=A(x,y)+B(x,y)\cdot\cos\left[\varphi(x,y)\right]\\ I_2(x,y)=A(x,y)-B(x,y)\cdot\sin\left[\varphi(x,y)\right]\\ I_3(x,y)=A(x,y)-B(x,y)\cdot\cos\left[\varphi(x,y)\right]\\ I_4(x,y)=A(x,y)+B(x,y)\cdot\sin\left[\varphi(x,y)\right]\end{cases} \tag{5}$$

由方程组求解后：

$$\begin{cases}\varphi(x,y)=\arctan\left(\dfrac{I_4-I_2}{I_1-I_3}\right)\\ A(x,y)=\dfrac{I_1+I_2+I_3+I_4}{4}\\ B(x,y)=\dfrac{\sqrt{(I_1-I_3)^2+(I_4-I_2)^2}}{2}\end{cases} \tag{6}$$

这样就可将光栅图像上每个像素的相位值$\varphi(x,y)$求出，但被截断在arctan的值域范围内。路径跟踪算法中基于网络规划的最小费用流法能较好保持与原始相位的一致性，并且受噪声影响较小，是比较好的相位展开算法。所以使用路径跟踪算法将包裹相位展开成真实相位，将解出的$\Delta\varphi(x,y)$带入式(3)，即可求出D点的Z轴高度信息。

1.3　光度立体视觉法

光度立体视觉基于明暗度的光度立体法由阴影恢复形状法发展而来，可精确计算出非刚体待测品表面的法向及Z轴信息，同时可避免表面轮廓、反射率及纹理等信息对测量结果的干扰[12]。该三维重建方法测量深度信息效率高，且照明方式部署较为简单，硬件成本较低。同时保留了较多的高频信息，是智能制造领域高效三维重建的重要方法。

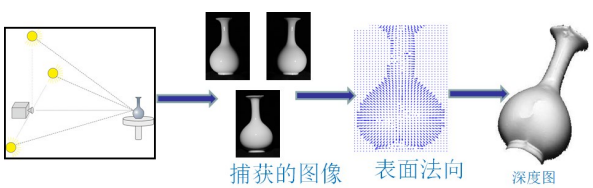


图4　光度立体法原理

光度立体法通过求取每点表面法向量间接计算出深度信息，最初是基于标准朗伯(Lambertian)面提出的。其工作原理如图4所示，即相机在固定位置拍摄一组n个照明方向下的待测品朗伯面图像，每个方向照明的光源均在无穷远且光强为I，记录n次照明下待测像素点的灰度$\boldsymbol{E}(x,y)=\left(E_1(x,y),\cdots E_n(x,y)\right)^{\mathrm{T}}$，若该像素对应点的单位法向量为$\boldsymbol{N}(x,y)=(n_x,n_y,n_z)^{\mathrm{T}}$，$n$次光源方向单位向量为$\boldsymbol{L}=\begin{bmatrix} l_{1x} & l_{1y} & l_{1z} \\ \vdots & \vdots & \vdots \\ l_{nx} & l_{ny} & l_{nz} \end{bmatrix}$，设该点反射率为$\rho(x,y)$，根据标准朗伯体表面反射模型有：

$$\boldsymbol{E}(x,y)=\rho(x,y)\cdot I\cdot\boldsymbol{L}\cdot\boldsymbol{N}(x,y) \tag{7}$$

方程中$\boldsymbol{E}(x,y)$、$\boldsymbol{L}$为已知量，$\rho(x,y)$与I为未知常数，后面可通过向量归一化过滤掉。待求法向量的$\boldsymbol{N}(x,y)$有三个未知量，需建立至少三个方程的方程组进行求解，即$n\geqslant 3$方可满足求解需求。同时由线性代数相关知识，$\boldsymbol{L}$矩阵中至少有三个行向量需满足不共面条件，求得的$\boldsymbol{N}(x,y)$才有唯一解。

对式(7)进行整合，得：

$$\boldsymbol{E}(x,y)=\boldsymbol{L}S(x,y) \tag{8}$$

$$S(x,y)=\rho(x,y)I\cdot\boldsymbol{N}(x,y) \tag{9}$$

由于$\boldsymbol{N}(x,y)$为单位法向量，故：

$$\boldsymbol{N}(x,y)=\frac{\rho(x,y)I\cdot(n_x,n_y,n_z)}{\rho(x,y)I\cdot\sqrt{(n_x{}^2+n_y{}^2+n_z{}^2)}}=\frac{S(x,y)}{\left\|S(x,y)\right\|} \tag{10}$$

同时，由式$\boldsymbol{E}(x,y)=\boldsymbol{L}S(x,y)$可得：

$$S(x,y)=\left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\cdot\boldsymbol{L}^{\mathrm{T}}\cdot\boldsymbol{E}(x,y) \tag{11}$$

由此可计算出点(x,y)的法向量$\boldsymbol{N}(x,y)$，根据法向量与梯度的关系，求出点(x,y)的梯度：

$$\begin{cases} p(x,y)=-\dfrac{n_x}{n_z} \\ q(x,y)=-\dfrac{n_y}{n_z} \end{cases} \tag{12}$$

由梯度信息计算待测品表面Z轴信息的算法主要有多路径积分法、傅里叶基函数法、最小二乘法，其中最小二乘法是光度立体法求解中较为常用的方法。在任意连续的曲面中$Z=f(x,y)$，(Z_x,Z_y)为曲面在x轴与y轴上的偏导数，当曲面$Z=f(x,y)$无限接近实际待测品表面函数时，有：

$$\min_Z\iint\left((Z_x-p)^2+(Z_y-q)^2\right)\mathrm{d}x\mathrm{d}y \tag{13}$$

这是一个泛函中求极限函数的问题，令$F=\dfrac{(Z_x-p)^2+(Z_y-q)^2}{2}$，由欧拉–拉格朗日(Euler-Lagrange)方程可得：

$$\frac{\partial F}{\partial Z}-\frac{d}{dx}\frac{\partial F}{\partial Z_x}-\frac{d}{dy}\frac{\partial F}{\partial Z_y}=0 \tag{14}$$

其中：$\dfrac{\partial F}{\partial Z}=0$，$\dfrac{\partial F}{\partial Z_x}=Z_x-p$，$\dfrac{\partial F}{\partial Z_y}=Z_y-q$，带入上式后得到由深度和表面梯度构成的泊松方程：

$$\nabla^2 Z=div(p,q) \tag{15}$$

通过对泊松方程求解即可获取待测品表面的深度信息。

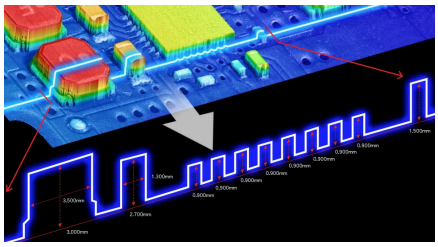
三个方向的照明可能会因凸起始终存在阴影区导致阴影区的法向量无法获取，所以工业产品借助光度立体法实现三维重建时一般采用四个方向以上的照明方式求解法向量。

2 三维重建技术应用研究

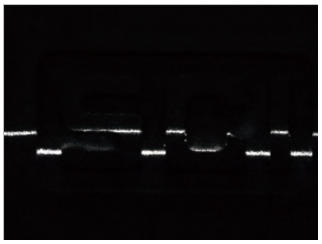
2.1 线激光三角测距法噪点问题研究

线激光三角测距法依赖于产品反射线激光的成像效果。激光随待测品表面高度变化而在成像时产生位移，借助三角测距法即可获取激光位于产品单列的高度信息，随后将每一列的高度信息进行拼接，即可生成三维图像。理想状态成像效果如图5(a)所示。

在相同的曝光时间与光圈条件下，产品表面激光线成像亮度仅与产品表面对激光的反射率相关。实际采集到的激光图像如图5(b)所示，若某些区域反射率过大，导致激光成像过亮；或某些端点区域的激光反射角出现跳点，通过三角测距法计算该点高度出现较大偏差。如图6(a)所示，测量得到的原始点云数据进行三维渲染后可见三维视图表面有很多噪点。

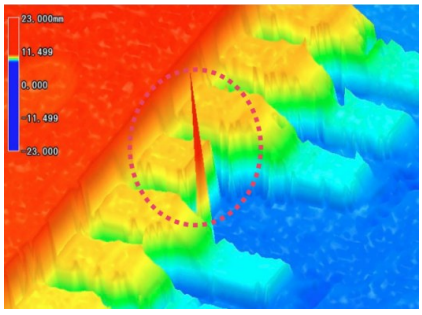


（a）线激光成像示意图

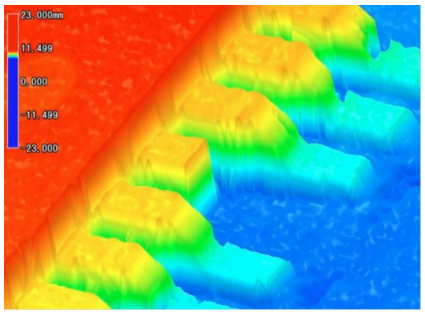


（b）线激光成像图

图 5　线激光成像



（a）三角测距噪声干扰



（b）噪声干扰过滤后效果

图 6　三角测距法噪声过滤

通过先验知识可知，正常先进制造产品无尖峰装凸起或凹陷，可通过对原始三维数据进行滤波处理，实现噪点的去除。通过分析点云数据特征可知，边缘梯度信息与深度信息非常重要，所以使用自适应双边滤波算法对点云数据进行处理，经三维渲染后的效果如图 6(b)所示，噪点干扰已获得较好的滤波效果。

2.2　线激光三角测距法凸台结构应用研究

三角测距法通过斜视相机采集线激光在产品表面反射光计算每点的三维信息，在产品表面有凸台结构的条件下，由相机斜视导致的问题主要有两个方面。如图 7 所示，一方面是当激光反射光到镜头光路接触凸台弧角时，激光反射光在凸台弧角区域发生二次镜面反射，使相机采集到的激光位置失真，导致计算出该区域的位置信息产生较大偏移；另一方面当激光反射光被凸台完全遮挡时，相机无法获得激光图像，导致从此遮挡区域到凸台边缘的深度信息完全缺失，不能满足产品整面三维重建的需求。

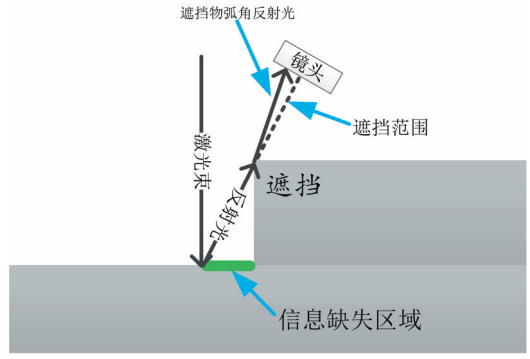


图 7　凸台结构对三角测距法影响

为提升三角测距法深度信息采集能力，减少凸台结构对三维重建的影响。首先根据先验知识，对凸台边缘区域的失真阶跃信息进行删除，防止对深度信息的计算产生影响。对于激光被完全遮挡的区域，使用不同采集方向的三角测距相机进行二次测量，并将两次测量的深度信息进行拼接，补足缺失区域。如图 8 所示，为满足智能制造领域对三维重建效率的需求，使用两颗检测方向相对的感测头同时测量，将结果拼接后渲染输出，解决凸台结构对三维重建的遮挡。

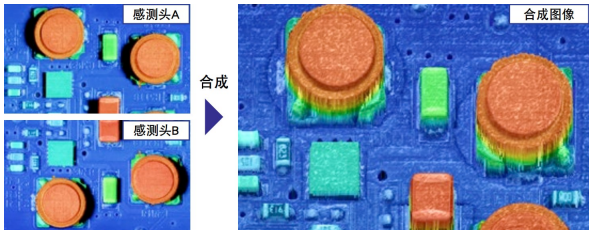
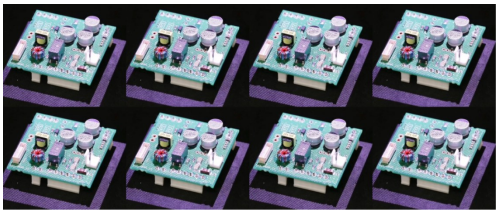


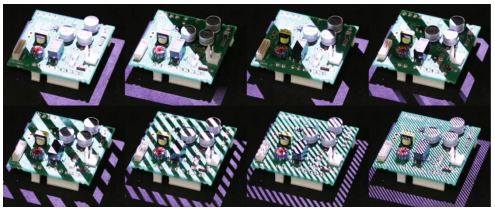
图 8　双检测方向三角测距法三维信息合成

2.3 相位轮廓测量术在非朗伯面应用研究

PMP应用于待测品表面三维重建将待测物表面假设为朗伯面，即将入射光在所有方向均匀反射，但实际检测中待测品表面通常会存在光滑的玻璃、金属等材质，反射光为镜面反射使局部光强超过测量范围导致重建失真，另外光栅图像中还会存在噪声干扰、环境光干扰等问题[13]。为降低这些问题的影响，实际应用中基于PMP的3D相机通常使用四个光栅光源从四个方向分别依次发射四组光栅，如图9为一个光栅光源发出的一组光栅，每个正弦光栅相移八次，并切换八个频率，应用多频外差相移算法对同一坐标求出四个Z轴信息，并对四个Z轴值进行极值抑制拟合。其中高频光栅用来降低噪声影响，低频光栅可最小化相位展开的运算复杂度，四个方向的光栅避免了单一方向光栅发生镜面反射使测量结果失真的问题。这样就避免了因为待测品表面结构、反射率的差异对三维信息的影响，并平滑了噪声干扰，从而得出精度较高的Z轴信息。



(a)八步相移光栅成像



(b)八频光栅成像

图9 多频外差相移光栅成像

2.4 光度立体法在彩色表面应用研究

光度立体法理论基础建立在产品为标准朗伯面的基础上，以产品表面对各方向光线的反射强度统一为前提，且对产品成像灰度值有高度依赖性。当光度立体法应用于彩色物体的三维重建，因彩色产品表面的颜色会影响成像灰度值大小，计算出彩色区域的法向量失真，会得到错误的深度信息。在此情况下需使用彩色光度立体法进行解决。彩色光度立体法使用三原色平行光从三个特定角度同时对产品照明，使用彩色相机获取产品图像后，通过参照平板与相同材质的采样球即可进行彩色表面产品三维重建。

彩色相机与灰度相机在相同分辨率与像元尺寸的条件下，彩色相机的三原色像元的单色像元尺寸更小，对噪声抑制能力不如灰度相机；另外现有的光源控制器与相机成像速度可满足光度立体法飞拍条件，所以彩色光度立体法实际应用在智能制造领域时，一般使用高分辨率的灰度相机进行飞拍，照明方式如图10所示，三原色平行光依次从四个角度对产品进行照明，相机直接获取单通道成像亮度，并完成产品三维重建。

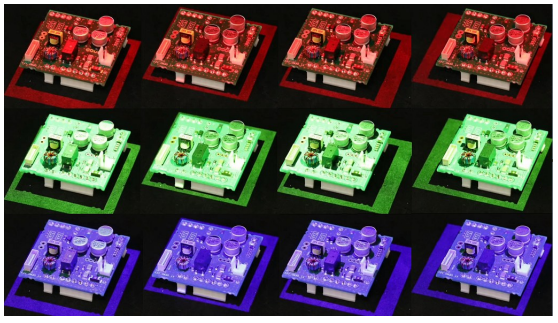
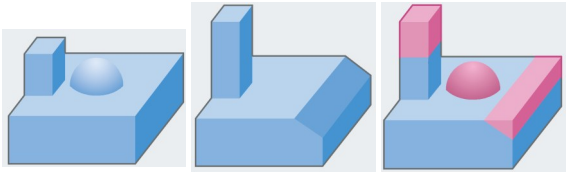


图10 彩色光度立体法照明方式

2.5 应用过程中三维数据处理方法研究

使用三维重建算法提取出产品表面深度信息后，通过构建基于深度值的二维矩阵，即可将3D图像转化为2D图像进行处理[14]。用于2D图像的算法均可以用来检测待测品表面深度缺陷信息。



(a)3D模板图 (b)3D检测图 (c)差分运算后缺陷结果图

图11 差分法在3D图像缺陷检测的应用

以常用的差分法为例，如图11所示，首先构建标准品的3D图像，对图像使用算法处理后注册为标准模板。然后对后面获取的3D检测图像进行相同算法处理，并进行特征定位后与标准模板图像做差分运算。差分法原理如图12所示，若待测品存在高度类别的缺陷，差分结果在缺陷位置会存在较大的数值，若数值超过设定的最大阈值，则在该处标记为缺陷位置。

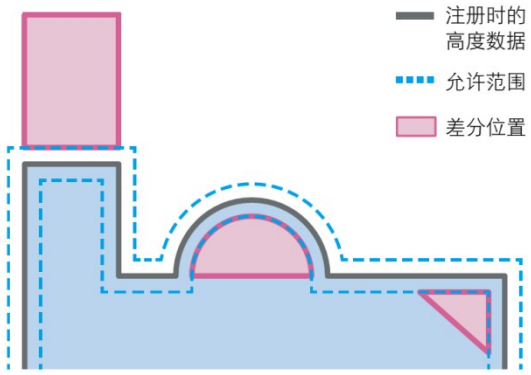


图 12　差分法原理图

3　结语

三维重建算法应用于智能制造装备领域中需要具有较高的测量精度与测量效率。本文分析了高精度高效的三维重建算法原理与实施方式，对线激光三角测距法、相位轮廓测量术以及光度立体法应用在智能制造装备中出现的问题进行了统计。并基于算法原理对应用问题出现的原因进行研究，给出问题解决方案，最后研究了三维数据的处理方法。在实际的工程应用中，为提高三维视觉测量的精度与鲁棒性，三维视觉相机通常具备两种以上的三维测量功能，对不同三维重建算法测得的深度信息进行比照与补偿，降低产品表面结构与噪声对测量结果的影响，是智能制造装备三维视觉未来发展的方向。

参考文献：

[1] 董宜平，谢达，钮震，等．基于FPGA的双源无轨电车的改进型YOLO-V3模型[J]．电子与封装，2022，22(8)：83-89.

[2] 吕璐，程虎，朱鸿泰，等．基于深度学习的目标检测研究与应用综述[J]．电子与封装，2022，22(1)：72-80.

[3] MIN Y，XIAO B，DANG J，et al. Real time detection system for rail surface defects based on machine vision [J]. EURASIP Journal on Image and Video Processing，2018.

[4] 唐震，戴莹，邹巧云，等．基于机器视觉技术的军用集成电路测试序列号读取装置[J]．电子与封装，2022，22(5)：77-81.

[5] MIN Y，XIAO B，DA N J，et al. Real time detection system for rail surface defects based on machine vision [J]. EURASIP Journal on Image and Video Processing，2018.

[6] 陈鑫，高博，龚敏．基于FPGA的虚拟听觉系统设计[J]．电子与封装，2022，22(6)：52-56.

[7] ZHANG B，YANG H，YIN Z. A region-based normalized cross correlation algorithm for the vision-based positioning of elongated IC chips [J]. IEEE Transactions on Semiconductor Manufacturing，2015，28(3)：345-352.

[8] 陈强．基于双目立体视觉的三维重建[J]．现代计算机(专业版)，2015(2)：66-69.

[9] CHANG F，DUAN Y，LIU M，et al. Coarse-to-fine adaptive illumination hard-adjustment for vision inspection system under uncertain imaging conditions [C] //2019 IEEE SENSORS. Montreal，QC，Canada，2020：19261523.

[10] 顾晓雪，郝国锋．基于LabVIEW的射频捷变收发器测试系统[J]．电子与封装，2022，22(5)：87-90.

[11] SRINIVASAN V，LIU H C，HALIOUA M. Automated phase-measuring profilometry of 3-D diffuse objects[J]. Applied Optics，1984，23(18)：3105.

[12] SAKAMOTO T，KAWANISHI T，IZUTSU M. Multi-Frequency Heterodyne System for All-Optical-Technology-Free Ultrafast Optical Waveform Measurement[C]//Proceedings of the 33rd European Conference and Exhibition of Optical Communication，Berlin，Germany，2007.

[13] LI M，ZHOU Z，WU Z，et al. Multi-view photometric stereo：a robust solution and benchmark dataset for spatially varying isotropic materials [J]. IEEE Transactions on Image Processing，2020，29：4159-4173.

[14] 罗卫，杨志成．基于三维点云的重建算法[J]．现代计算机(专业版)，2017(2)：54-57.

(下转第103页)

实践与经验

文章编号: 1007-1423(2023)08-0082-05 DOI: 10.3969/j.issn.1007-1423.2023.08.013

基于数字孪生的空军机场消防保障需求分析

胡嘉旭[1,2]，王海涛[1*]，于长明[2]，刘尧锋[2]，徐显海[2]

（1. 陆军工程大学野战工程学院，南京 210007；2. 95979部队，泰安 271200）

摘要：为提高空军机场消防保障能力，以机场消防保障任务为牵引，构建基于数字孪生的空军机场消防保障管理体系。对空军机场消防保障的现状进行分析，结合保障任务分析实际保障需求，对通过数字孪生实现机场火灾预警、资源可视化管理、指挥调度、态势显示、效能评估、数据共享等功能需求进行系统分析，为构建基于数字孪生的空军机场消防保障管理体系提供依据。

关键词：空军机场消防保障；数字孪生；需求分析

0 引言

军用飞机作为国防力量中的一种重要武器装备，具有不可替代的作用。随着备战打仗的要求不断提高、新装备不断列装部队，军用飞机执行任务、日常训练的强度大幅度增加，发生故障起火的概率极大地增加，若在战时，飞机还将面临遭袭导致起火。近几年，国内外飞机失事起火事故时有发生，消防救援的效果并不理想，往往造成飞机损毁和人员伤亡。起火后的救援对于及时止损起着关键作用，军用飞机火灾与民用飞机火灾相比有很大的区别，主要体现在军用飞机起火后具有极高的危险性，极大地增加了飞机火灾扑救的难度，分析存在问题和保障需求，对高效完成飞机火灾扑救任务起着重要作用。

数字孪生技术将物理实体和系统的属性、结构、状态、性能、功能和行为映射到虚拟世界[1]，为实现虚实之间动态交互、双向映射、实时连接提供了方法途径，为观察、控制和改造物理世界提供有效手段[2]。针对空军机场消防保障存在的问题和保障需求，基于数字孪生技术的可视化管理、虚实双向反馈和实时决策支持为突破口，本文对构建机场消防保障的数字孪生需求进行分析。

1 空军机场消防保障现状分析

1.1 机场消防保障问题分析

目前来看，传统的粗放式、被动式机场消防保障方法，在未来信息化条件下的高强度训练与作战，将表现得越来越力不从心。空军机场飞行任务重、飞机起降频率高，在战时和训练中发生事故导致飞机起火的概率大。机场消防保障力量在执行消防保障任务时，主要存在4点问题：

（1）机场消防保障对指挥员的应急处置反应速度有很高的要求，在缺少科学高效的空军机场消防保障系统的情况下，信息的迟滞、缺失、冗余都将导致指挥员决策效率低、指挥能力弱，对局势把控不全面。

（2）飞机火灾突发性强、发展速度快、不确定性高，在实际保障中，很难精准保障，加之消防保障资源配置分散，保障资源难以做到预

收稿日期：2022-11-23 **修稿日期：**2023-01-11

作者简介：胡嘉旭(1989—)，男，辽宁沈阳人，硕士，讲师，主要研究方向为装备虚拟仿真；*通信作者：王海涛(1978—)，男，安徽宿松人，硕士，副教授，硕士研究生导师，主要研究方向为装备虚拟仿真与虚拟维修，E-mail：648530594@qq.com；于长明(1978—)，男，辽宁沈阳人，本科，讲师，主要研究方向为机场消防保障；刘尧锋(1987—)，男，河南周口人，硕士，讲师，主要研究方向为机场消防保障；徐显海(1972—)，男，黑龙江东宁人，硕士，高级讲师，主要研究方向为装备保障

置预储、快速调度和准确供应[3]。

（3）飞机火灾受机型、装载武器、飞行速度、位置、气象环境等影响很大，其发展速度、规模不确定性大，对飞机火灾的评估很难做到科学准确和事先充分准备，只能以被动式的救援保障为主，缺少智能化的火灾预警，对火灾隐患的排除及火灾救援难度大。

（4）传统的依靠经验和预定方案的消防保障在灭火作战行动中缺少适用性，因为飞机火灾的发生具有唯一性，在一定程度上只有相似性，而没有相同的火灾事故，因此指挥员的决策，仅依靠原有的经验和预定方案，还不够准确。

综上所述，空军机场消防保障能力的转型和提升显得尤为迫切，已成为空军建立智慧场务保障体系中的重点部分。数字孪生技术的发展与应用，为空军机场消防保障的转型建设，带来新的契机和发展途径，符合建设智慧场务保障体系的实际需求。

1.2　机场消防保障资源分析

机场消防保障资源根据内容要素可分为保障力量、保障对象和保障环境，如图1所示。消防保障力量是主体，保障对象是受体，保障力量对保障对象实施保障的过程是在保障环境下进行的。

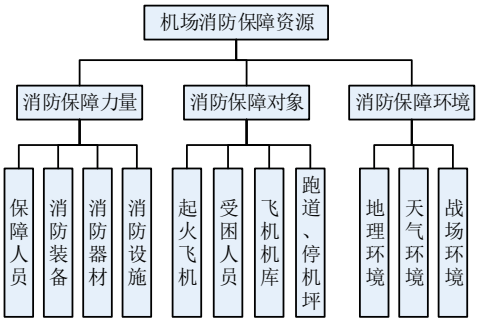


图 1　机场消防保障资源

（1）机场消防保障力量。机场消防保障力量是指执行机场消防保障任务的保障要素集合，包括保障人员、消防装备、消防器材、消防设施。保障人员包括消防员、装备驾驶员、指挥员及其他相关保障人员等；机场消防保障装备以机场主力泡沫消防车为主，担负机场消防保障主要救援任务；消防器材包括常规防护装备、消防梯、破拆工具、供水器材、灭火器、灭火剂等；消防设施主要包括机场固定的器材库、水源、取水设施等。

（2）保障对象。机场消防保障对象是指在机场范围内，发生火灾或存在消防安全隐患的人员、装备等，是消防保障任务的受体。其中，飞机和机载人员是机场消防保障的重点对象，除此之外，还包括机库、机场跑道、停机坪、加油坪等机场相关设施。

（3）保障环境。机场消防保障环境是指执行保障任务所在的环境，包括机场地理环境、周围天气环境以及战场环境。地理环境是指执行保障任务开展灭火行动的地点及相关地理因素；天气环境指影响火灾扑救和火势发展的天气因素，如风向、雨雪等天气；战场环境是指在作战背景下的消防保障，如飞机跑道损毁、存在未爆弹、敌军空袭等。

1.3　机场消防保障资源需求

提出空军机场消防保障资源需求是为了提升机场消防保障能力，提高指挥的速度和效率，使指挥员能够做出精准决策，使信息获取更及时，智能化水平更高。如图2所示。

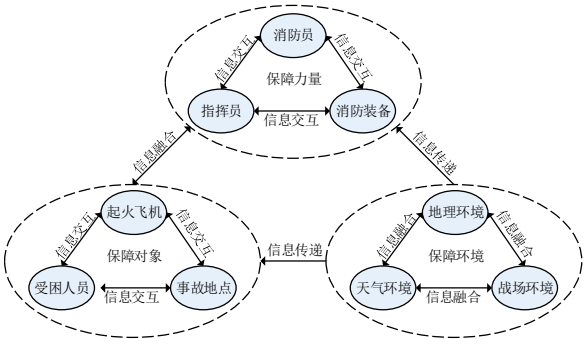


图 2　机场消防保障资源需求

（1）保障力量智能化。建立体系化的机场消防保障力量，在人员、装备等资源上布设信息通信模块，通过信息网络将消防保障力量各资源关联起来，实现保障力量的智能化管理调度，使保障力量的运用更加高效有序。

（2）保障对象同步化。建立保障对象与保障力量同步的信息交互体系，保障对象的状态信息可以实时地传达到保障力量终端，进行分析决策，为保障力量开展保障行动提供依据。

（3）保障环境透明化。通过布设传感器、摄

像头、北斗定位系统等设施，构建透明化的保障环境，使保障人员可以通过智能设备通视整个机场保障环境，同时，可以融合战场环境信息，对保障部队提供预警。

2 空军机场消防保障任务需求分析

针对机场消防保障任务，为解决存在的不足，提高保障水平，对消防保障任务需求进行分析。图3为空军机场消防保障任务需求结构图。

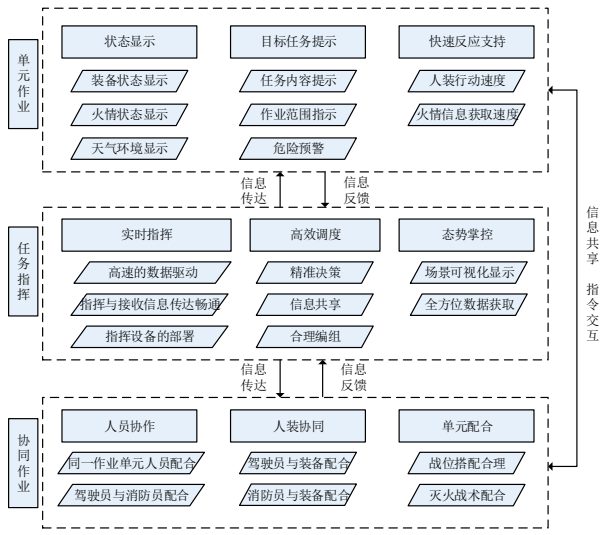


图 3 机场消防保障任务需求结构

2.1 单元作业需求

(1)状态显示。以消防员和消防车所构成的消防保障单元，在实施灭火作战行动时，以目前的方法，装备状态依靠仪表显示，火情和天气环境状态依靠消防员的经验来判断，存在一定的主观因素，尤其是对火情判断，没有精确的状态显示，仅靠经验判断，很难准确把握火势的发展，对灭火行动造成影响。准确的状态显示可以给消防员的行动提供依据，需要结合显示终端，通过数据传输，将装备、火情、环境等重要信息传递给作业末端，由末端作业单元进行具体操作，执行灭火作战行动。

(2) 目标任务提示。火场烟雾弥漫，视野受阻，在进行灭火行动时，火势、风向等因素的变化，影响消防员对火情的判断，忽视潜在的隐藏危险。明确的目标任务提示，可指引消防员有针对性地开展灭火和救援行动，有效地规避火场危险，控制安全距离，划定作业区域，安全、准确、高效地开展灭火行动。

(3) 快速反应支持。反应速度取决于消防员的出警速度和火灾信息的获取速度，在应对飞机火情时，消防员的出警速度一般情况下相对固定，主要受获取信息的速度影响，从火情的发现开始，到接警出动抵达火场展开行动的反应过程决定了能否及时控制火势。目前，灭火作战行动的开始，主要依靠自主发现火情，由驾驶员根据情况选择行进路线，抵达火场需开展火情侦察后才能展开救援，为了减少反应时间，需要从火情的发现、路线选择、火情侦察等环节上进行优化，提高对机场动态的监控，实时辅助引导灭火作战单元快速行动。

2.2 协同作业需求

(1) 人员协作。主要体现在消防员在执行灭火作战行动时，所担负的任务不同，相互之间需要密切配合。例如，用消防水枪灭火时，铺设水带与供水的时机把握，需要消防员之间密切配合操作；在进行破拆救援时，同样需要消防员进行掩护和压制火情。消防员之间的配合既关系到灭火作战行动，也影响到自身安全，因此十分重要，除了需要日常扎实的训练外，还需要智能化的信息提示，实现信息同步。

(2) 人装协同。消防装备在遂行灭火作战行动时，主要依靠喷洒灭火剂实现灭火，消防员对消防装备的运用，决定了能否在灭火剂耗尽之前完成灭火，人与装备的配合尤为重要。例如，车辆行驶至距离起火飞机多远时打开消防水炮进行灭火，打开过早距离不够，灭火剂利用率不高，打开过晚，延误救援时机；用消防水炮和消防水枪进行射水的时机把控，也决定了能否高效灭火。因此，人装的协同配合除了消防员的主观操作和判断外，还需要结合装备，即指挥系统的提示和引导。

(3) 单元配合。在进行灭火作战行动时，各作战单元进入火场展开行动的位置、时机和方法都需要进行统一部署，包括战位和战术的配合。在实际灭火作战行动中，受浓烟和起火对象的影响，往往视线受阻，各单元之间主要依靠对讲机实现相互沟通和受领指示，缺少自主的配合提示，同时，也需要各单元之间通过共享信息自主进行战斗配合。

2.3　任务指挥需求

（1）实时指挥。目前在开展灭火作战行动时，主要依靠现场指挥员的指挥，在浓烟和警笛等噪声的干扰下，很难准确判断情况，火灾扑救对时效性要求极高，指挥上的滞后会严重影响行动结果。因此，需要对指挥体系进行改进优化，主要应在现场信息的获取和指令的传递下达上进行改进。从传统的现场观察指挥，向现场指挥与后方数据采集分析后协同指挥发展；从传统的依赖对讲机进行指挥，向部署先进的指挥和接收终端发展。通过高速的数据驱动，实现任务的实时指挥，能够及时应对突发多变的火场情况。

（2）高效调度。在执行机场消防保障任务时，不同的作业单元散布在火场的不同区域，如何对各单元进行高效的指挥调度是火场指挥的难点。需要指挥员对作业单元进行合理的指挥编组，根据现场情况及时调整保障力量，充分利用消防保障资源；依托智能化的辅助决策系统，实现指挥员的精准决策，并通过信息共享，实现保障人员对火情信息的同步掌握。

（3）态势掌控。无论是单元作业、协同作业还是实现准确高效的指挥，都是在对火场态势全面掌控的前提下才能够有效开展实现的，目前，开展机场消防保障灭火作战行动时，还无法对火场态势实现全方位掌控，主要受制于现场视频监控和数据采集困难，火场消防保障力量和进展情况无法通过可视化实时展现在指挥员面前。实现对机场消防保障的全方位态势掌控，需要建立体系化的监控系统，确保在机场的任意位置发生火情都能被监控到；建立全方位的数据采集系统，从消防保障力量、起火对象、现场环境三个方面进行数据实时获取，通过数据处理、融合分析，将火场态势以三维可视化方式展现。

3　机场消防保障数字孪生建设需求分析

基于数字孪生的机场消防保障，通过数字驱动，建立虚拟模型，构建物理实体与虚拟模型的映射，1∶1还原物理实体，将物理实体的变化及运行规则和逻辑关系，实时地通过虚拟模型体现出来，同时，通过对虚拟模型的控制，关联物理实体，实现对现实世界的指挥控制和调度干预，实现虚拟与现实的数字孪生[4]。

3.1　机场火灾预警

通过布设高清摄像头、传感器等数据采集设备对机场运行状态进行监控，对飞机起火、机场设施起火等火灾事故，以及可能引发火灾的故障和问题能够及时发现，并向值班人员发出预警提示，便于消防救援人员第一时间做好消防救援准备，为火灾扑救争取宝贵时间，或能在火灾发生前消除安全隐患，避免事故发生。

3.2　机场消防保障资源可视化管理

为机场消防救援提供可视化的场景监控，通过对机场消防资源进行实时监控，使值班人员能够及时准确地掌握消防资源的运行和保障情况，主要包括对机场的全景监控、环境监测、装备设备监控、火情监控等消防资源的监控[5]，同时，通过运用北斗定位设备搭建一对多的指挥端和移动终端，实现人员装备的精准定位，通过无线通信网络，将视频监控画面、声音信号等实时数据传输到系统终端，便于指挥员根据现场情况进行指挥调度。如图4所示，为机场消防资源监控内容。

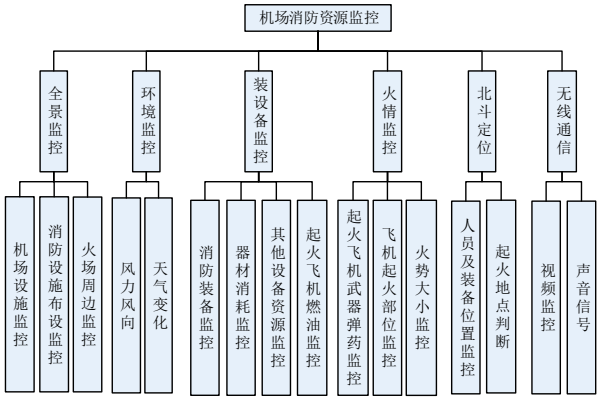


图4　机场消防资源监控内容

3.3　机场消防保障指挥调度

基于消防资源的监控和可视化显示，实现对机场消防保障过程的精准管控，对飞机起降、车辆调度、分组配合等进行系统分析，为指挥员提供最优解决方案，提高指挥调度的准确、高速、高效。通过自动分析，对人为指挥进行预警和纠正，对存在的隐患及时地预警提示，减少人为操作和指挥上的失误，对人为的调度实施系统干

预，例如，发生飞机火灾后，起火位置将对其它飞机的起降造成干扰，系统通过智能预警直接干预指挥，第一时间将信息传达至相关飞机驾驶员，并根据现场情况和各待起降飞机油量等相关状态，分析出最佳的应急起降方案。

3.4 机场消防保障态势显示

运用三维建模技术，构建火场地理环境、消防装备和救援人员模型，根据消防保障实时采集回传的数据信息，以数据融合技术为基础，通过数据驱动模型，将保障过程通过虚拟模型在显示器上1∶1地实时展现，实时地反映火场的情况，实现机场消防保障态势的动态显示。同时，通过现场监控设备，可实现虚拟场景和实际场景间的切换，给指挥员带来更直观清晰的态势掌控。

3.5 机场消防保障效能评估

基于数字孪生的机场消防保障管理的效能评估是指在执行机场消防保障任务时，通过定性和定量的方法来实现系统对预期目标的分析、计算和评价。系统通过基于效能评估的评估模型和评估指标，对机场消防资源部署、指挥层级设置、机场消防运行机制等情况进行评估分析，查找出机场消防保障管理中的隐患[6]。同时，通过对信息数据的综合分析，对机场火灾救援任务执行情况进行评估，从指挥调度、资源部署、救援效果等全方面总结查找出任务保障中存在的不足，提出改进意见，提升保障能力。

3.6 机场消防保障数据共享

机场消防保障涉及任务多，保障点位多，除灭火外，还担负飞行员抢救、破拆、疏散等相关任务，在执行任务中消防救援人员往往会与医疗救护人员及其他场务、机务人员配合，在紧急情况下，各类人员之间各自为战，专业技能和掌握的数据互不兼容，相互之间的协同和沟通将耗费一定时间。机场消防保障数字孪生的构建，可为整个救援保障提供数据共享，可与医疗救援、场务、机务人员共享数据，使保障人员可以更好地协同配合，对未来做好预判，将极大程度地提高救援人员的保障效率，并充分保证人身安全。

4 结语

基于数字孪生技术对机场消防保障管理体系进行升级改造，可以极大提升机场消防保障水平，对于保证飞行安全具有重要意义。根据机场消防保障数字孪生建设需求，对构建新的机场消防保障体系提供了依据和抓手，为保障体系的进一步改进打下基础。

参考文献：

[1] 陶飞，张贺，戚庆林，等. 数字孪生十问：分析与思考[J]. 计算机集成制造系统，2020，26(1)：117.
[2] 陶飞，刘蔚然，刘检华，等. 数字孪生及其应用探索[J]. 计算机集成制造系统，2018，24(1)：1-18.
[3] 杜红兵，吴风婷，胡辉，等. 民机飞行中火灾事故影响因素研究[J]. 消防科学与技术，2021，40(5)：759-762.
[4] 王海涛，胡嘉旭，于长明. 空军机场消防保障数字孪生系统初探[J]. 空军军事学术，2022(3)：92.
[5] 胡宗健. 机场视频监控技术与应用[J]. 智能城市，2021，7(10)：67-68.
[6] 张丽敏，裴建国. 消防指挥自动化系统效能评估指标体系研究[J]. 消防科学与技术，2010，29(4)：311-313.

Analysis of air force airport fire protection demand based on digital twins

Hu Jiaxu[1,2], Wang Haitao[1*], Yu Changming[2], Liu Yaofeng[2], Xu Xianhai[2]

(1. Field Engineering College, Army Engineering University, Nanjing 210007, China;
2. 95979 Troops, Tai'an 271200, China)

Abstract: In order to improve the capability of air force airport fire protection, the air force airport fire protection management system based on digital twins is constructed under the guidance of airport fire protection tasks. This paper analyzes the current situation of the air force airport fire protection, analyzes the actual support requirements in combination with the support tasks, and systematically analyzes the functional requirements of realizing airport fire early warning, resource visualization management, command and dispatch, situation display, effectiveness evaluation, data sharing, etc. through digital twins, so as to provide a basis for building the air force airport fire protection management system based on digital twins.

Keywords: air force airport fire protection; digital twins; requirement analysis

文章编号: 1007-1423（2023）08-0087-04 DOI: 10.3969/j.issn.1007-1423.2023.08.014

基于区块链与星际文件系统的电子证据存证研究

陈 潮*

（浙江警察学院计算机与信息安全系，杭州 310053）

摘要：电子证据作为八大类证据之一，在司法活动中占据越来越重要的地位，而电子证据的易复制、易损坏和易篡改等特点，导致其公信力相对不高。在研究区块链和星际文件系统的基础上，提出一种基于区块链与星际文件系统的电子证据存证模型，加密技术实现对电子证据的加密，星际文件系统实现电子证据的分布式存储，区块链存储电子证据的哈希值、关键词和在星际文件系统中的分片位置索引，同时在区块链上实现电子证据的搜索与验证。该存证模型确保电子证据的机密性与完整性，并实现对电子证据的授权访问，因此对于提高电子证据的公信力有一定的现实意义。

关键词：电子证据；区块链；星际文件系统；存证

0 引言

随着信息技术的不断发展及在各个社会领域的深入应用，人类社会产生的电子数据呈现爆炸式增长，公安机关在办理各类案件中提取海量的电子数据，部分电子数据成为案件侦查、诉讼和审判等司法活动中必须的电子证据。《中华人民共和国民事诉讼法》和《中华人民共和国刑事诉讼法》中明确规定电子数据可以作为证据，电子证据已被列为八大类证据之一，成为司法活动中的重要证据之一[1]。电子证据具有的易复制、易篡改、易损坏和保管监管难等特点，导致其在安全性和可靠性方面面临巨大的挑战。2018 年，最高人民法院发布《关于互联网法院审理案件若干问题的规定》，人民法院应当对当事人通过电子签名、区块链等证据收集、固定和防篡改的技术手段或通过电子取证存证平台认证收集的证据予以认定[2]。

区块链是一种分布式存储的共享数据库，采用区块和哈希链的数据结构，具有去中心化、防篡改和可溯源等特点，特别是其中的联盟区块链，适合公安机关跨部门跨警种使用，可以实现对电子证据的分布式存储，采用共识算法实现各节点数据的一致性，从而确保上链后数据的无法篡改。当前电子证据的种类和形式表现多样性，存储容量上既有以字节为单位的电子证据，也有以太字节为单位的电子证据，呈现数量级上的巨大差别，而区块链在存储大文件时存在冗余大、效力低和处理速度慢等问题，为解决区块链不适合存储容量较大的电子证据的问题，提出一种基于区块链和星际文件系统的电子证据存储模型，将电子证据及其哈希值分别存储在星际文件系统和区块链上，星际文件系统确保电子证据的去中心化和分布性存储，区块链实现对电子证据的查询和验证，从而有力地保障电子证据的司法效力。

1 区块链

区块链是数字加密货币比特币的一种数据结构，区块链技术是比特币的底层支撑技术，是计算机技术、点对点网络、分布式存储技术、共识算法、智能合约、密码技术和博弈论等众多理论和技术的集成创新，是近几年来最新最热门的信息技术之一。区块链由区块和哈希链

收稿日期：2022-12-14 **修稿日期：**2023-01-26

作者简介：*通信作者：陈潮（1980—），男，浙江诸暨人，硕士，副教授，研究方向为计算机网络、区块链，E-mail：chenchaocn@163.com

两个部分构成，数据结构是区块，比特币区块结构如图1所示，区块由区块头和区块体两个部分构成，区块头包含前一区块哈希值、区块版本号、时间戳、目标哈希（难度）、随机数和默克尔树根等，区块体由该区块所有交易的哈希构成的一个默克尔树根构成[3]。区块链网络由众多节点构成，每个节点存储全部或部分区块的信息。区块链分为若干类型，根据节点加入区块链网络是否需要许可，将区块链分为许可链和非许可链，非许可链即公有链，任何节点无需许可即可加入公有链，是一种完全去中心化的区块链，典型的公有链有比特币、以太坊等；许可链又分为私有链和联盟链，私有链一般用于一个单位或组织内部，是一种不具有去中心化特点的区块链；而联盟链一般用于跨单位、跨组织和跨部门之间，是一种具有部分去中心化的区块链。公安机关跨部门跨警种电子证据的存证，适合使用许可链，因此采用联盟区块链来搭建电子证据存证模型。

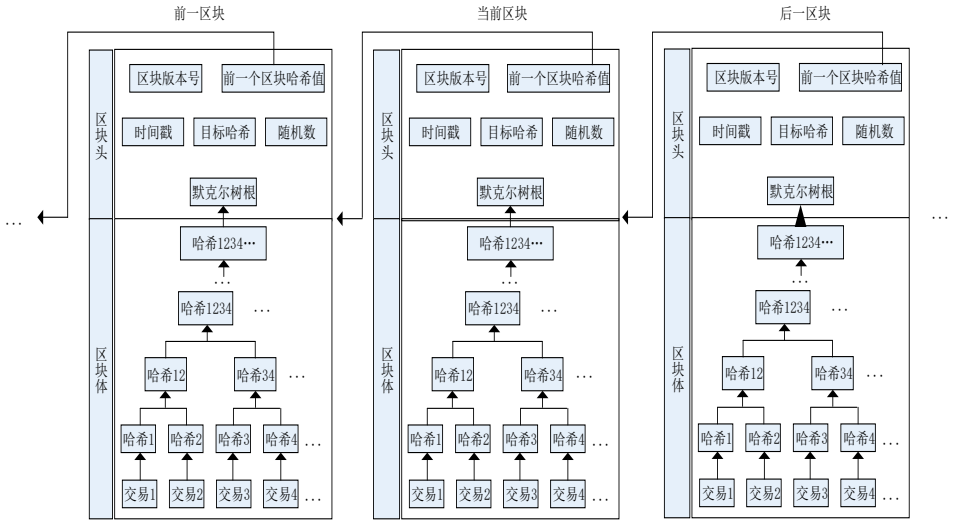


图1 区块链数据结构

2 星际文件系统

星际文件系统（interplanetary file transfer network，IPFS）是一种分布式数据存储协议，用来替代互联网中心化的HTTP协议，也是一种点对点的分布式文件存储系统，可以实现对各种类型文件的永久存储，星际文件系统具备的重复文件删除机制，能够避免文件的重复存储，从而节约存储空间资源，同时存储在星际文件系统中的文件具有唯一的哈希值，通过文件哈希值实现对文件的快速搜索和定位[4]。互联网采用传统的基于地址的寻址技术，星际文件系统采用基于内容的寻址技术，当一份文件提交到星际文件系统时，星际文件系统将文件分解为每个大小为256 K字节的数据块，数据块互相链接，并构成默克尔有向无环图。文件存储到星际文件系统时，返回一个内容标识符（content-addressed identifiers for distributed systems，CID），内容标识符是星际文件系统中标准的文件寻址格式，集合内容寻址、加密散列算法和自我描述格式等内容，根据内容标识符可以实现在星际文件系统中的文件搜索[5]。若一个文件重复存储到星际文件系统时，新产生的文件哈希值与之前存储的文件哈希值相同，星际文件系统将不再重复存储相同的文件。若一份文件进行修改后提交到星际文件系统，则会生成新的内容标识符。星际文件系统既有效地弥补区块链存储空间有限的不足，又利用其自身的分布式存储技术解决传统中心化存储存在的问题，从而确保电子证据的安全存储。

3　电子证据存证系统模型

3.1　电子证据存证系统模型架构

电子证据存证模型以区块链网络为基础网络架构，以星际文件系统为存储介质，区块链以公安局各派出所、各警种大队和各业务科室为节点，证书中心为执法办案民警颁发公私钥对，并实现对民警身份的认证，具体架构如图2所示，各节点之间通过点对点专用网络互联，同时每个节点也连接星际文件系统，星际文件系统实现对加密电子证据的分片存储与合并提取，实现电子证据的安全可靠存储，区块节点将电子证据哈希值、关键词和分片位置索引等信息打包成交易存储在区块链中，区块链实现对电子证据的查询和防篡改验证。

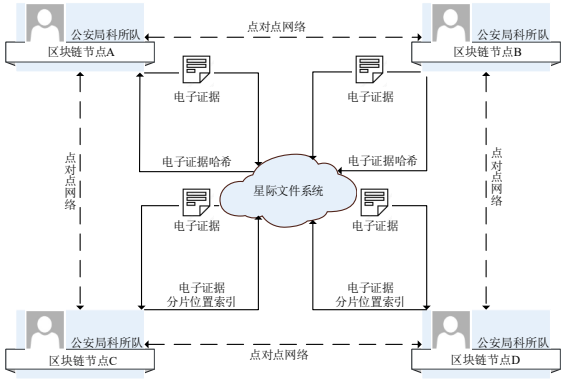


图2　电子证据存证模型架构

3.2　电子证据存证系统模型设计

电子证据存证系统由民警身份的注册和认证、电子证据存储、电子证据查询、电子证据下载和验证等部分组成。

3.2.1　电子证据存储

电子证据存证系统采用加密技术和星际文件系统实现对电子证据的安全可靠存储，支持日志、文档、图片、音视频和各类取证镜像文件的加密存储[6]。电子证据存储流程如图3所示，民警在客户端提交认证信息进行身份验证，验证通过后登入到电子证据存证系统，并接入到所在单位的区块链节点。在客户端计算电子证据哈希值和实现对电子证据的加密，加密后的电子证据上传到星际文件系统，星际文件系统对加密电子证据进行分片存储，同时将电子证据文件分片位置索引等信息返回给客户端，客户端将电子证据关键词、在星际文件系统中的分片位置索引等信息打包成一条交易提到所在单位的区块链节点，并通过点对点网络广播到区块链所有节点，最终该交易纳入到新的区块中，从而实现对电子证据的加密分布式存储[2]，确保电子证据的机密性和完整性。

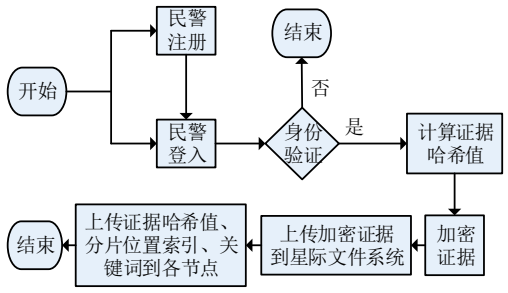


图3　电子证据存储流程

3.2.2　电子证据查询

电子证据查询由区块链各节点对区块信息进行关键词搜索，依赖于电子证据存储过程中保存在区块链上的电子证据在星际文件系统中的分片位置索引和电子证据相关关键词。电子证据查询流程如图4所示，民警在客户端经过身份认证后，生成需要查询的关键词，提交到民警所在单位的区块节点，节点在区块链上进行关键词搜索，若搜索成功，即可获取电子证据的哈希值和在星际文件系统的分片位置索引，下一步可以进行电子证据的下载和验证。

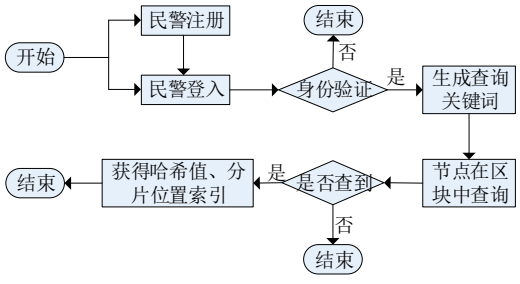


图4　电子证据查询流程

3.2.3　电子证据下载和验证

电子证据存证系统经过电子证据查询获得

电子证据在星际文件系统的分片位置索引，经过授权后，即可从星际文件系统下载电子证据，哈希校验来防止电子证据的篡改。电子证据下载与验证流程如图5所示，民警获得下载电子证据的授权后，向其所在单位的区块节点发送下载电子证据的请求，区块节点一方面获得原始电子证据的哈希值，另一方面根据电子证据在星际文件系统中的分片位置索引，从星际文件系统下载电子证据的所有分片，并生成完整的电子证据，对下载的电子证据计算出哈希值，与原始电子证据的哈希值进行比对，若一致，则说明电子证据未被篡改，从区块节点下载电子证据；若不一致，则说明电子证据已经被修改失效。

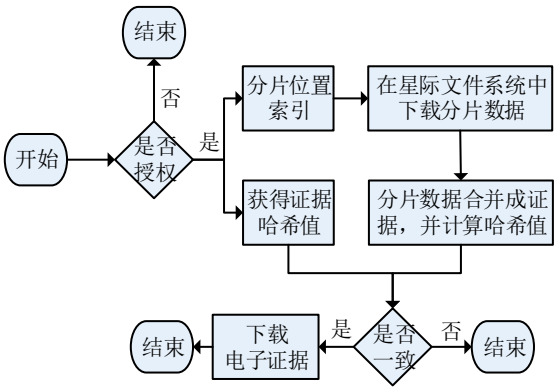


图 5 电子证据下载与验证流程

4 结语

针对传统中心化存储电子证据面临的易复制、易篡改和易损坏等问题，在研究区块链和星际文件系统工作原理的基础上，提出一种基于区块链和星际文件系统的电子证据存证模型，采用点对点网络、哈希计算、加密技术和星际文件系统等技术，实现对电子证据的安全存储、可靠下载、防篡改和自动验证功能，对于提高电子证据的公信力有一定的借鉴意义。

参考文献：

[1] 刘品新. 论区块链存证的制度价值[J]. 档案学通讯,2020(1):21-30.

[2] 杨坤桥. 基于区块链的电子数据存证平台研究[J]. 现代计算机,2021,27(9):36-40.

[3] NAKAMOTO S. Bitcoin: a peer-to-peer electronic cash system[Z/OL]. https://bitcoin.org/bitcoin.pdf, 2008.

[4] SUN J, YAO X, WANG S, et al. Blockchain-based secure storage and access scheme for electronic medical records in IPFS [J]. IEEE Access, 2020, 8: 59389-59401.

[5] KUMAR S, BHARTI A K, AMIN R. Decentralized secure storage of medical records using Blockchain and IPFS: a comparative analysis with future directions [J]. Security and Privacy, 2021, 4(5): e162.

[6] 郑清安,程仲汉. 基于区块链的电子数据存证系统[J]. 莆田学院学报,2022,29(5):61-65.

Research on electronic evidence storage and certification based on blockchain and interplanetary file system

Chen Chao*

(Department of Computer and Information Security, Zhejiang Police College, Hangzhou 310053, China)

Abstract: As one of the eight types of evidence, electronic evidence is playing an increasingly important role in judicial activities, but the characteristics of electronic evidence such as easy copying, damage and tampering lead to relatively low credibility of electronic evidence. Based on the research of blockchain and interplanetary file system(IPFS), an electronic evidence storage and certification model based on blockchain and IPFS is proposed. The encryption technology realizes the encryption of electronic evidence, and IPFS realizes the distributed storage of electronic evidence, the blockchain stores the hash value of the electronic evidence, keywords and the index of the segment location in IPFS, and at the same time realizes the search and verification of the electronic evidence on the blockchain. The evidence Storage and certification model ensures the confidentiality and integrity of electronic evidence, and realizes authorized access to electronic evidence, so it has certain practical significance for improving the credibility of electronic evidence.

Keywords: electronic evidence; blockchain; interplanetary file system; storage and certification

开发案例

文章编号: 1007-1423（2023）08-0091-07 DOI: 10.3969/j.issn.1007-1423.2023.08.015

基于HLS的MobileNet加速器实现

韦苏伦，陶青川*

（四川大学电子信息学院，成都 610065）

摘要：随着卷积神经网络的发展，越来越多的现代化智能应用出现在人们的生活中，其中，嵌入式场景的应用更加考虑低功耗与资源控制方面的设计。提出一种基于HLS高层次综合的卷积神经网络的加速器，使用流水加速、pingpong缓存、分组分块卷积等加速策略，充分发挥FPGA的计算优势。实验结果表明，在速度上相较i7 12th有了2倍左右的提升，在精度上相较RTX3060基本不变的情况下，功耗下降了10倍左右。

关键词：FPGA加速；卷积神经网络；高层次综合；低功耗

0 引言

作为一种人工神经网络，卷积神经网络被广泛应用于图像、语音识别[1]等各种智能识别系统。随着卷积神经网络的发展，各种各样的网络层出不穷，并且被应用到越来越复杂的场景当中。但随着网络复杂性的增加以及随之而来的庞大计算量，运行卷积神经网络的计算设备也需要更好的性能。传统的CPU并不适用于矩阵运算占主导的模型训练和推理，GPU虽然满足这一特性，但对于一些嵌入式设备来说还需要更低的功耗[2]。FPGA作为可编程逻辑器件，具有功耗低、性能高、灵活性好的特点，因此更加适用于卷积神经网络硬件加速的开发研究[3]，Verilog开发门槛比较高，开发周期相对较长，这极大影响了卷积神经网络在FPGA中部署的普及。

软件工程师们应该更多考虑的是大的架构，而非某个单独部件或逐周期运行，HLS工具[4]的出现也是源自于此，它是一种代码综合技术，具体是指采用C、C++等高级编程语言进行程序编写，而不是传统的Verilog语言，这大大提高了FPGA的开发速度。本文应用HLS高层次综合工具，基于轻量化的原则选择了MobileNetv2网络，在赛灵思的FPGA开发板Kria KV260上实现了一个卷积神经网络加速器，通过数据的串并转换，充分利用AXI总线带宽，利用pingpong缓存技术实现数据的读写与计算的并行操作，同时在卷积计算中使用分组分块计算进一步提高推理的速度。

1 开发平台与IP核的加速策略

1.1 开发平台与卷积神经网络

本文使用赛灵思的Kria KV260 FPGA开发板作为实验板卡，Kria KV260是赛灵思专为AI视觉设计的入门级FPGA开发板。它的设计是一种模块化的设计方式，分为FPGA板卡以及接口部分，其中FPGA板卡部分是K26 SoM，它采用Zynq UltraScale+ MPSoC架构，包含4核ARM Cortex-A53处理器，提供256 K个系统逻辑单元和1.2 K个DSP单元。在软件开发环境方面，使用赛灵思的统一软件平台Vitis以及高层次综合工具Vivado HLS作为编译测试环境。

神经网络模型方面使用的是MobileNetv2[5]。作为经典轻量化网络的MobileNet，自诞生就被广泛应用于工业界。它是一种构造体量小、低延时的网络结构，对于很多移动和嵌入式设备

收稿日期：2022-12-12 **修稿日期：**2022-12-25

作者简介：韦苏伦（1997—），男，四川自贡人，硕士研究生，研究方向为计算机应用与图像识别；*通信作者：陶青川（1972—），男，四川南充人，博士，副教授，研究方向为模式识别与智能系统，E-mail：451911031@qq.com

的图像应用都比较适合。MobileNet是由谷歌团队提出的应用于移动端或者嵌入式设备中的轻量级神经网络，在准确率只有极小幅降低的情况下，大量减少参数与运算量。MobileNet的特点是提出了深度可分离卷积，其还可被拆分成两个子模块：逐通道卷积(depthwise convolution)与逐点卷积层(pointwise convolution)。

表 1　KV260板卡资源

配值项	参数
参数	KV260
器件	Zynq® UltraScale+™ MPSoC
系统逻辑单元	256K
Block RAM	144
Ultra RAM	64
DSP单元	1.2 K
DDR内存	4 GB
一级启动内存	512 Mb QSPI
以太网接口	1 X 10/100/1000 Mb/s

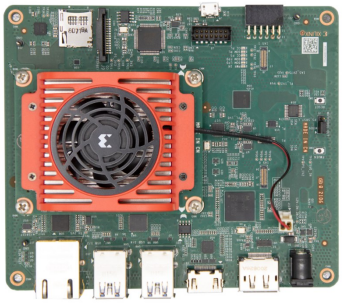


图 1　KV260开发板俯视图

对于传统卷积来说，若卷积核大小为D_k，数量为N，图像的尺寸为D_f，图像和卷积核的通道深度为M，那么对于N个卷积操作来说，总的计算量如公式(1)所示。

$$Cost_{Conv} = D_f \times D_f \times D_k \times D_k \times M \times N \quad (1)$$

而对于深度可分离卷积神经网络来说，其计算量包括逐通道卷积和逐点卷积两部分，如公式(2)、(3)、(4)所示。

$$Cost_{DSC} = Cost_{DW} + Cost_{PW} \quad (2)$$

其中，CostDW和CostPW分别为

$$Cost_{DW} = D_f \times D_f \times D_k \times D_k \times M \quad (3)$$

$$Cost_{PW} = D_k \times D_k \times M \times N \quad (4)$$

所以可计算出传统卷积与深度可分离卷积的计算量比值，由公式(5)可知，后者的计算效率明显高于传统卷积。

$$\frac{Cost_{Conv}}{Cost_{DSC}} = \frac{1}{N} + \frac{1}{D_f^2} \quad (5)$$

MobileNetv2是MobileNetv1的升级版，在MobileNetv1的深度可分离卷积基础上，新增加了线性瓶颈和倒残差结构，其中倒残差结构如图2所示。

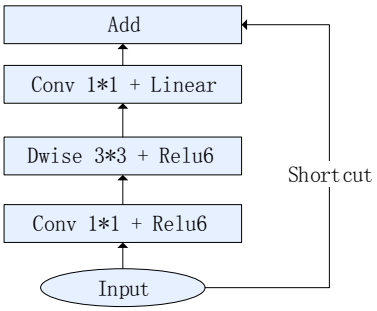


图 2　MobileNetv2倒残差结构

在倒残差结构中，先用小的1*1卷积升维，通过3*3的逐通道卷积提取特征，最后再使用1*1卷积降维，呈两头小、中间大的梭型结构。为了提升精度，在倒残差结构中，前两个激活函数使用ReLU6来代替ReLU，最后使用线性激活函数。

1.2　加速核的整体设计

本文的总体设计基于vivado HLS高层次综合[6]和PYNQ平台[7]。由HLS工具通过C++语言实现卷积神经网络的IP核，通过仿真和综合之后可以打包zip核文件，通过vivado平台导入HLS的zip文件进行块设计、布线和整体编译得到硬件平台的.bit和.hwh文件，这是PL部分的设计。PS端可以使用C++语言在Vitis中进行主机端的代码编写或者使用Python语言在PYNQ中进行编写，进而控制数据的输入与输出并与PL部分进行计算与交互，其数据传输的结构如图3所示。

首先通过S_AXI_lite接口将主机端与FPGA相连接，通过读取和写入S_AXI_lite端的控制寄存器来控制PL端的行为，一般用来传输控制和状态寄存器以及读写的地址等。对于输入的图像数据、权重和偏置等数据则使用存储器映射接口M_AXI来进行传输，M_AXI存储器映射

接口可以支持最高4 K字节的突发量，并且具有独立的读取和写入的通道，因此单个接口也可以同时执行读取和写入的操作。在传输读写地址之后，通过M_AXI接口将经过主机端映射存到全局存储器中的数据读取至PL内核中进行计算。

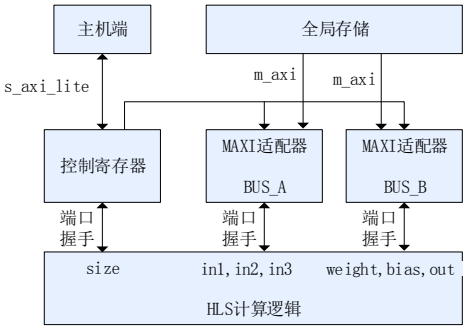
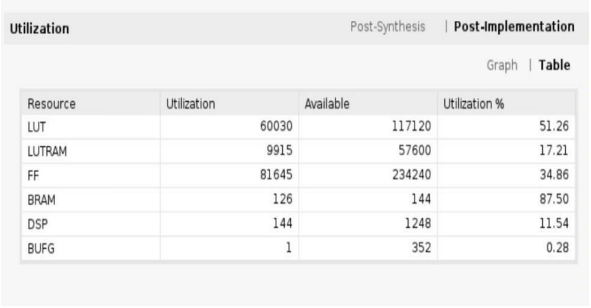


图3　数据传输结构

为了加快推理速度，可以进行BN融合，即将BN层参数和卷积层参数融合在一起，这样在模型训练之后的推理阶段就可省去BN层的计算[8]。

考虑到MobileNetv2结构的重复性，将整个网络分成三个部分来设计：Bottleneck部分、Bottleneck之前以及之后的部分。在每个部分也有基础的计算单元，如普通二维卷积、深度卷积、逐点卷积、残差结构以及全剧平均池化层等，本文将这些基础单元单独设计成IP计算核。

加速器总的系统结构块设计如图4所示，左边是HLS设计导出的5个IP核，中间从上往下分别是ZYNQ的中心处理模块和时钟与复位信号模块，右边的两列则是传输总线AXI的相关模块，对块设计进行综合之后得到总的资源使用情况，如图5所示。



Utilization　　Post-Synthesis | Post-Implementation

Graph | Table

Resource	Utilization	Available	Utilization %
LUT	60030	117120	51.26
LUTRAM	9915	57600	17.21
FF	81645	234240	34.86
BRAM	126	144	87.50
DSP	144	1248	11.54
BUFG	1	352	0.28

图5　综合资源占用情况

1.3　加速器优化策略

1.3.1　pingpong操作

在两个计算模块之间传递数据时，由于前一个模块需要等待下一个模块计算完成才能交付数据，造成了一定的性能损失，所谓pingpong操作是指设置缓冲来进行交替存储，进而实现读、算、写的同步。

数据的分块计算使得多个模块之间的计算满足上述情况，所以设置两个buffer来实现数据的pingpong操作，buffer的大小根据不同IP核进行

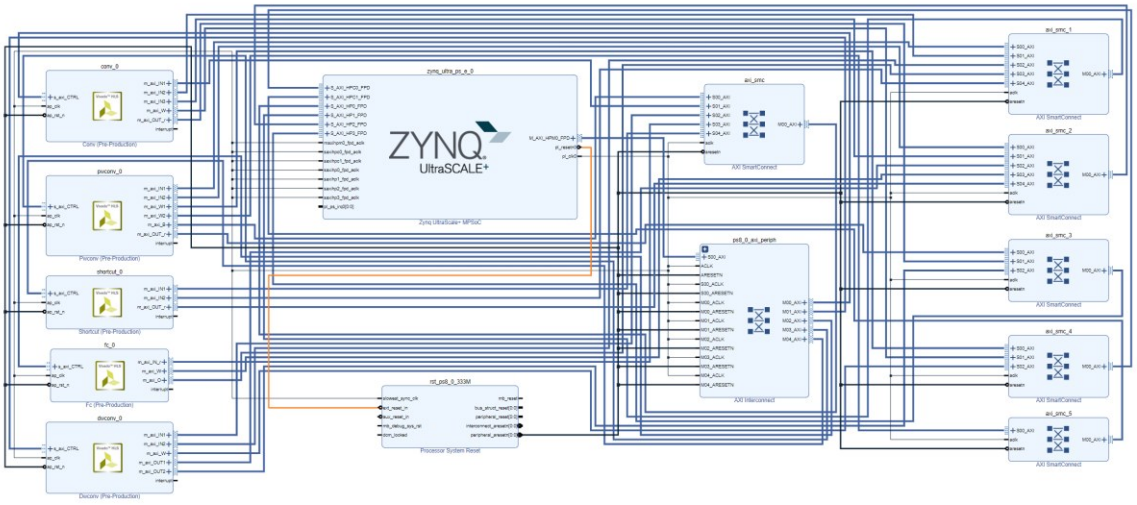


图4　加速器整体块设计

独立设置。pingpong操作的计算原理如图6所示。

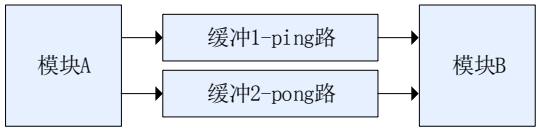


图 6　pingpong操作原理

1.3.2　for循环展开

HLS可以使用Unroll指令对for循环进行展开，for循环在默认状态下是折叠状态的，即在电路里中每一次循环都会分时地使用同一套电路。使用Unroll可以对for循环的代码区进行循环体展开，将之前的电路复制多份，实现以资源换取并行的计算逻辑，如图7所示。Unroll允许完全展开或部分展开，其中部分展开需要指定循环因子factor=N，即把循环展开N倍来减少循环迭代。

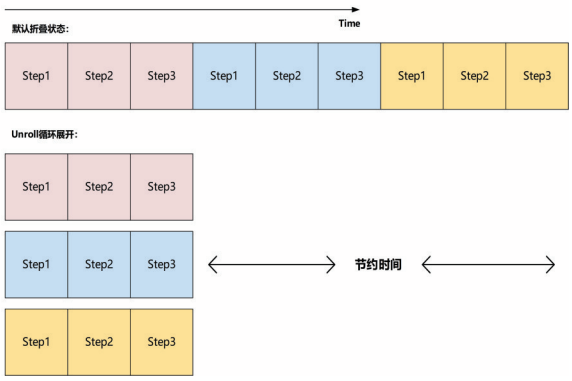


图 7　Unroll循环展开原理

1.3.3　多精度计算优化

在IP核中采用的是AP_FIX16的精度进行存储和运算，这就有可能会因为数据精度过低以及数值溢出而产生一些计算误差，并有可能在深度神经网络中持续积累这种误差。于是本文在考虑资源占用的同时，采用多精度数据的方式来避免这种精度的溢出：将输入与输出设置为AP_FIX16的数据类型，然后在计算卷积的过程中，扩大精度来保存临时的数值，并在最后输出时恢复为最开始的精度。

1.3.4　计算分组分块

由于网络模型的参数量比较大，直接使用FPGA中片上资源来保存每一层所有的数据并不是一个好的方法，所以本文对网络的各个模块用数据分组分块的方式来计算，以降低片内资源的占用，并配合pingpong操作等优化方式进行并行计算，其过程如图8所示。

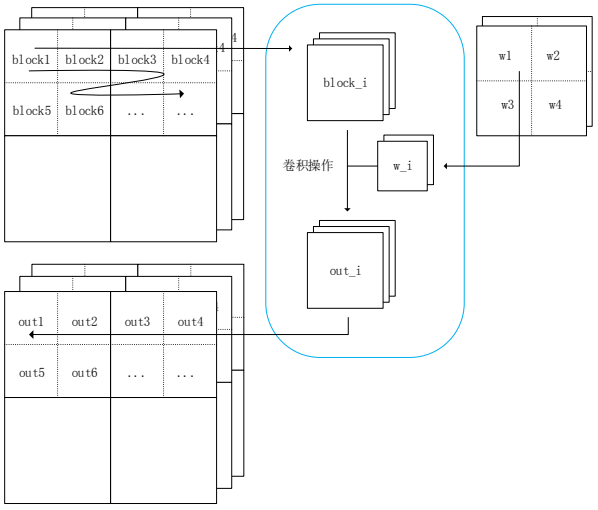


图 8　分组分块计算

1.3.5　流水线优化

流水线的设计思想是新的输入数据在前面的数据计算完成之前就能提前处理，例如有一个复杂电路，它需要固定的时间周期才能得出最后的稳定结果，我们将其拆解为N个步骤，第一个步骤计算完成后将结果存起来传送给第二个步骤，然后紧接着继续往第一个步骤输入数据，以此类推，用这种方式实现流水线优化的并行处理，如图9所示。

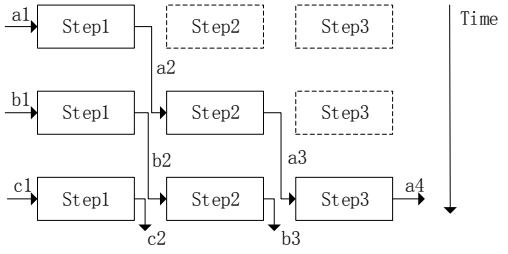


图 9　流水线加速时序

1.3.6　加法器树优化

在一次卷积计算过程中涉及到多次加法的运算，这个时候可以采用加法器树的设计方式，

将卷积运算单元中的相加过程设计为树形的并行相加模式，进而提高运算效率，可以把时钟周期为N的工作缩小到$\log_2 N$个时钟周期。

1.4　实验对比

通过实验对比，在不使用并行加速策略时，单核的Arm设备平均耗时为35.6 s，而本文使用上述的并行优化设计，在KV260的ZYNQ平台中的平均计算耗时降低到0.046 s，且功耗并无明显变化，证明该加速器在保证功耗基本不变的情况下，还能充分发挥FPGA的并行计算优势，提高计算速度。

表2　加速器优化策略性能对比

测试设备	频率/MHz	平均耗时/s	平均功耗/W
Arm A53	33	35.600	4.65～5.63
KV260(ZYNQ)	**333**	**0.046**	**5.59～6.35**

2　卷积神经网络加速器的设计与实现

2.1　普通卷积层的设计与实现

该部分是神经网络最上层卷积，输入为224×224×3的图像数据与该层的权重数据。为了提升计算速度，我们对数据流进行了控制，将输入分为三个通道来处理输入数据，使用三条128位AXI总线进行传输。在实际传输时数据为16位的定点数，这里将24个16位定点数打包，同时在AXI总线上进行并行传输。对于权重、偏置和输出则分别使用一条64位AXI总线进行传输。

为进一步提升计算效率，对计算分为两路的pingpong缓存和计算，如图10所示。在pingpong操作中进行后续的循环展开计算卷积模块，将输入数据进行如图8的分块卷积操作，对于权重来说，同样在for循环中每次读取出8个filter来进行卷积操作，以达到数据并行计算的目的。

```
if(pingpong){    // 存入缓冲ping路，计算缓冲pong路
    load(data1, data2, data3,ping_buff2);
    compute(weight,out,ping_buff1,bias_buff);
    pingpong =false;
}
else{           //  存入缓冲pong路，计算缓冲ping路
    load(data1, data2, data3,ping_buff1);
    compute(weight,out,ping_buff2,bias_buff);
    pingpong =true;
}
```

图10　pingpong模块伪代码

经过HLS仿真测试和综合之后，得到其资源使用报告，如图11所示。

Utilization Estimates

Summary

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	320	-
FIFO	-	-	-	-	-
Instance	28	27	9696	16367	-
Memory	24	-	0	0	-
Multiplexer	-	-	-	829	-
Register	-	-	1700	-	-
Total	52	27	11396	17516	0
Available	4032	9024	2607360	1303680	960
Available SLR	1344	3008	869120	434560	320
Utilization (%)	1	~0	~0	1	0
Utilization SLR (%)	3	~0	1	4	0

图11　普通二维卷积IP核综合报告

2.2　逐通道卷积与逐点卷积层的设计与实现

Depthwise是逐通道卷积，是将特征图的每个通道单独使用卷积核来进行卷积操作，获得特征图每个通道的空间特征，并使得到的特征图和输入的通道数保持一致。逐通道卷积的卷积计算采用PIPELINE指令进行流水优化以及UNROLL进行循环展开，同时利用pingpong模块对缓存和计算进行并行处理。在接口方面，将输入设置为两个不同通道来处理输入数据，使用两条64位AXI总线进行传输，同样将8个16位定点数据打包到AXI总线上并行传输，权重和偏置则采用一条32位AXI总线进行传输。

由于每次不同层的逐通道卷积尺寸不同，我们将数据分块来实现统一的单元化卷积计算，具体是将数据切分为若干个8×32的小尺寸，且进一步将32划分为4×8的多通道数据，使得不同尺寸的卷积转换为多个固定尺寸的小的卷积，有利用逐通道卷积的加速核的统一化设计。对于不同步长的层，可通过传入的stride值判断输出特征尺寸是否需要减半。分块卷积的伪代码如图12所示，W为卷积核的大小，B_r、B_c、B_ch分别为分块的尺寸以及通道大小，在B_ch处进行PIPELINE循环展开，使用B_ch个并行的乘法器和深度为$\lceil \log_2 B_ch \rceil$的加法器树，每时钟将输入缓存的ch个通道的特征值与对应的权重值进行乘法计算，然后将计算的数据进行累加，并使用加法器树优化计算，最后通过输出缓存存储结果。

```
for(int w_x=0;w_x<W;w_x++)
    for(int w_y=0;w_y<W;w_y++)
        for(int m=0;m<B_r;m++)
            for(int n=0;n<B_c;n++){
HLS PIPELINE
                for(int ch=0;ch<B_ch;ch++){
                    if (w_x == 0 && w_y == 0) {
                        buff_out[ch][m][n] = buff_b[ch];
                    }
                    buff_out[ch][m][n] +=buff_in[ch][m*s+w_x][n*s+w_y]*buff_w[ch][w_x][w_y];
                }
            }
```

图 12　分块卷积实现伪代码

在HLS中进行仿真和综合后的报告如图13所示。

Utilization Estimates

Summary

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	494	-
FIFO	-	-	-	-	-
Instance	11	9	3148	7506	-
Memory	24	-	512	544	-
Multiplexer	-	-	-	2570	-
Register	-	-	1273	-	-
Total	35	9	4933	11114	0
Available	4032	9024	2607360	1303680	960
Available SLR	1344	3008	869120	434560	320
Utilization (%)	~0	~0	~0	~0	0
Utilization SLR (%)	2	~0	~0	2	0

图 13　逐通道卷积层IP核综合报告

Pointwise是逐点卷积，使用和输入特征图通道相同数量的1*1的卷积核，对特征图深度方面做了加权组合，相当于获得每个点的特征信息，大大减小了总体的计算量。

在具体的实现中，使用两条64位AXI总线进行传输，同样将8个16位定点数据打包到AXI总线上并行传输，权重和偏置则分别采用两条64位AXI总线和用一条32位AXI总线进行传输。计算部分由于该模块存在许多1*1的卷积，计算得到的通道数相对较多，同样采用了数据分块的思想，利用两层pingpong嵌套操作来读取权重和特征数据并进行卷积计算。

同样经过仿真以及HLS综合之后获得资源使用情况的报告，如图14所示。

Utilization Estimates

Summary

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	1	-	-	-
Expression	-	-	0	781	-
FIFO	-	-	-	-	-
Instance	40	265	43248	102403	-
Memory	48	-	0	0	-
Multiplexer	-	-	-	2867	-
Register	-	-	1023	96	-
Total	88	266	44271	106147	0
Available	4032	9024	2607360	1303680	960
Available SLR	1344	3008	869120	434560	320
Utilization (%)	2	2	1	8	0
Utilization SLR (%)	6	8	5	24	0

图 14　逐点卷积层IP核综合报告

2.3　残差层、全连接模块的设计与实现

由于已经单独实现了逐通道卷积和逐点卷积，所以这里残差层是指最后的原始数据与经过了残差之后的输出相加的操作，并通过流水展开来进行加速。

全连接模块由全局平均池化层和全连接层组成，本文将其实现为一个IP核，输入大小为7×7×Channel，输出为最后的分类结果。接口方面直接使用一条16位AXI总线进行传输，具体的实现则是将卷积核的值当做分数来模拟求平均的计算，使用加法器树对齐进行并行加速，加载数据时同样利用pingpong模块来进行缓存和计算优化。仿真和综合之后得到报告，如图15所示。

Utilization Estimates

Summary

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	690	-
FIFO	-	-	-	-	-
Instance	4	-	850	1254	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	421	-
Register	-	-	930	32	-
Total	4	0	1780	2397	0
Available	4032	9024	2607360	1303680	960
Available SLR	1344	3008	869120	434560	320
Utilization (%)	~0	0	~0	~0	0
Utilization SLR (%)	~0	0	~0	~0	0

Utilization Estimates

Summary

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	5263	-
FIFO	-	-	-	-	-
Instance	30	53	19836	39836	-
Memory	2	-	64	16	-
Multiplexer	-	-	-	1733	-
Register	-	-	5962	128	-
Total	32	53	25862	46976	0
Available	4032	9024	2607360	1303680	960
Available SLR	1344	3008	869120	434560	320
Utilization (%)	~0	~0	~0	3	0
Utilization SLR (%)	2	1	2	10	0

图 15　残差层、全连接层IP核综合报告

3　实验与结果

经过上述的仿真和综合之后，从Vivado HLS 2020.2中导出IP核的压缩文件，在Vivado 2020.2中导入IP文件并在块设计中进行连线。对块设计进行仿真和综合后得到.bit和.hwh文件，利用Python语言调用PYNQ软件层的接口进行主机端编程。

数据集采用公开的DeepFashion服装数据集来进行训练和测试，其中包括运动夹克、毛衣、

连衣裙等46种类别，并使用2000张测试图片进行上板验证。分别从计算时间、识别准确率、芯片功耗几个方面来展示在KV260的FPGA上推理MobileNetv2网络的实验结果，并且将该结果分别与网络在CPU和GPU计算平台上的推理结果作为对比来分析基于FPGA方法的优势。

3.1　实验测试平台

对于使用FPGA加速器的方案，使用赛灵思的Kria KV260作为硬件计算平台；对于直接使用CPU的方案，使用i7-12th作为硬件计算平台；而对于使用GPU的方案，则使用RTX3060作为硬件计算平台。

3.2　实验性能对比

表3　计算速度对比

测试设备	平均计算时间/s	平均准确率/%	平均功耗/W
i7 12th^{laptop}	0.094	73.57	35～44
RTX3060^{laptop}	0.012	74.32	42～63
KV260	**0.046**	**72.55**	**5.59～6.35**

4　结语

本文基于赛灵思提供的Kria KV260开发板，使用高层次综合工具通过C++语言进行MobileNetv2的加速核设计，并在使用Python在PYNQ平台中对主机端进行编程以及验证推理。实验表明，相比单核的Arm芯片，利用FPGA的并行计算设计以及HLS相关指令对加速核进行计算优化效果明显，且该种方法设计的卷积神经网络加速器在推理Top1上并无明显下降，在计算速度上相较于CPU来说提升了2倍左右，虽然与GPU相比还是有一定差距，但在功耗方面降低了10倍左右，有着较大的优势。

参考文献：

[1] AMODEI D, ANANTHANARAYANAN S, ANUBHAI R, et al. Deep speech 2: end-to-end speech recognition in English and Mandarin [EB/OL]. arXiv:1512.02595,2015.

[2] 张丽丽. 基于HLS的Tiny-YOLO卷积神经网络加速研究[D]. 重庆:重庆大学,2017.

[3] 李振宇. 基于PL-PS架构的图像处理系统的实现与算法应用[D]. 济南:山东大学,2016.

[4] CORNU A, DERRIEN S, LAVENIER D. HLS tools for FPGA: faster development with better performance [C]//Proceedings of the 7th International Conference on Applied Reconfigurable Computing. Springer, Berlin, Heidelberg, 2011:67-78.

[5] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: inverted residuals and linear bottlenecks [C]//Proceedings of 2018 IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018:4510-4520.

[6] 张志超,王剑,章隆兵,等. 基于FPGA的浮点可分离卷积神经网络加速方法[J]. 高技术通讯,2022,32(5):441-453.

[7] 陈禹,谷文成,渠吉庆,等. 基于PYNQ的图像分类识别技术研究与实现[J]. 计算机技术与发展,2021,31(12):73-77.

[8] 龚豪杰,周海,冯水春. 基于FPGA的卷积神经网络并行加速设计[J]. 计算机工程与设计,2022,43(7):1872-1878.

Realization of MobileNet accelerator based on HLS

Wei Sulun, Tao Qingchuan*

(School of Electronic Information, Sichuan University, Chengdu 610065, China)

Abstract: With the development of convolutional neural networks, more and more modern intelligent applications appear in our daily life. Among them, the application of embedded scenarios gives more consideration to the design of low power consumption and resource control. This paper proposes a convolutional neural network accelerator based on FPGA, and uses HLS high-level synthesis to design and verify. The experimental results show that, compared to i712th, the speed has improved by about 3 times, while compared to RTX3060 in accuracy, the power consumption has decreased by about 10 times.

Keywords: FPGA acceleration; convolutional neural network; high level integration; low power consumption

文章编号: 1007-1423(2023)08-0098-06 DOI: 10.3969/j.issn.1007-1423.2023.07.016

基于Vue和SpringBoot的乡村文旅平台设计与实现

李晟曈[1,2]，刘 哲[2,3]，俞定国[2,3*]，方申国[4,5]，孙学敏[1,2]

（1.浙江传媒学院新闻与传播学院，杭州 310018；2.浙江省影视媒体技术研究重点实验室，杭州 310018；
3.浙江传媒学院媒体工程学院，杭州 310018；4.浙江省工业和信息化研究院，杭州 310012；
5.浙江理工大学纺织科学与工程学院，杭州 310018）

摘要： 乡村文旅是新时期乡村振兴战略实施的重要抓手。立足于经济社会发展背景及乡村文旅发展需要，设计并实现了基于Vue和SpringBoot的乡村文旅平台。平台采用Redux模块分层结构，以Vue-cli进行脚手架搭建，运用SpringBoot微框架和Docker应用容器，实现以二维码为标识的乡村文旅信息集成，旨在通过赋能特色文化传播、催生并推广特色产业、优化文旅多主体统筹管理的方式，创新传播优秀传统文化，完整呈现乡村旅游图景，多方受益提高旅行效率。

关键词： 乡村文旅；乡村振兴；Vue；SpringBoot；二维码

0 引言

《“十四五”文化发展规划》指出要坚持以文塑旅、以旅彰文，推动文化和旅游在更广范围、更深层次、更高水平上融合发展，打造独具魅力的中华文化旅游体验[1]。文旅融合发展成为未来文化产业、旅游产业的重要趋势。乡村文旅作为一种将中华民族的优秀传统文化与旅游经济结合起来的产业形态，是新时期我国落实文化强国战略和乡村振兴战略，缩小城乡差距，助力共同富裕的重要抓手。然而新冠疫情爆发以来，我国旅游业受到极大冲击，至今仍处于“休眠”状态。同时，由于我国乡村发展的不充分和不平衡特点，乡村文旅面临资源整合难、信息化水平低、管理措施落后等问题。在保障社会安全的前提下，如何提升乡村文化旅游中的游客体验，助力乡村优秀传统文化的传播，并拉动旅游的经济效益增长，成为当下亟需解决的问题。

基于上述背景，针对当代乡村文化旅游发展的痛点和要点，本文设计并实现了一种基于Vue和SpringBoot的乡村文旅平台设计，旨在通过数字化平台进行资源整合，集游客、商家、景区等多元主体协同管理为一体，实现扫码进入、码上导游、码上惠客、码上消费、码上评价、码上合作、码上监管等应用，创新乡村文化传播、激发特色产业活力、优化文旅管理，促进乡村振兴和旅游业可持续发展。

1 平台建设目标及功能设计

1.1 建设目标

本平台立足乡村文旅实际需求，依托支付宝平台进行小程序的设计和开发。平台采用Vue

收稿日期： 2022-12-09 **修稿日期：** 2022-12-27

基金项目： 浙江省重点研发计划项目(2021C03138)：文旅融合支撑平台关键技术研究与应用示范——基于‘一码一图一网’(QDI)的乡村文旅综合服务平台关键技术研究与应用示范；2022年浙江省大学生科技创新活动计划(新苗人才计划)项目(2022R416C032)：数字‘画’乡愁，文旅在‘浙’游——文旅码助力美丽乡村文化传播

作者简介： 李晟曈(1998—)，男，浙江嘉兴人，在读硕士研究生，研究方向为文化与科技融合；刘哲(1985—)，女，山东菏泽人，博士，讲师，研究方向为大数据挖掘与分析；*通信作者：俞定国(1976—)，男，浙江新昌人，博士，教授，研究方向为媒体融合，E-mail：yudg@cuz.edu.cn；方申国(1983—)，男，浙江温州人，博士研究生在读，助理研究员，研究方向为文化与科技融合、数字经济；孙学敏(1995—)，女，山西大同人，在读硕士研究生，研究方向为数字媒体与智能传播

的渐进式框架进行“微前端”的开发、基于SpringBoot微框架的后端开发以及数据库管理系统，最终实现以二维码(QR code, QRC)为标识的乡村文旅融合功能，实现创新传播优秀传统文化，完整呈现乡村旅游图景，多方受益提升旅游体验。

1.2　功能设计

1.2.1　用户登录

乡村文旅平台是依托于支付宝开发和搭建的小程序应用，也是以智能手机移动端为主的轻量化程序。安装支付宝的用户可以利用支付宝、浏览器等“扫一扫”功能，扫描小程序二维码进入程序主体。用户在进行文旅信息的订阅、平台互动、地点打卡等操作时需要登陆个人账号。个人账号通过用户手机号进行关联性注册，达到账号个性化和实名认证的效果。用户登录界面如图1(a)所示。

1.2.2　乡村文旅信息推荐

平台小程序首页呈现乡村文旅信息，实现自动拉取并推荐，包括乡村景点名称、乡村景点略缩图等。用户可根据推送的信息点击进入相应的景点详情页，也可以在首页滑动浏览多个景点得到简要信息。乡村文旅平台首页如图1(b)所示。

1.2.3　景点详情页面

景点详情页用以展示每个乡村景点的详细信息和具体介绍，包括景点简介、文化元素、宣传视频、具体景色以及地图导航功能。用户可以通过主页乡村景点列表、搜索等方式进入页面，并可以通过滑动浏览详情页面查看景点详情信息。景点详情页面如图1(c)所示。

1.2.4　精确搜索

平台的精确搜索功能模块主要用于提升用户体验，满足用户的个性化需求。平台为每个景区、景点设置标签，提供筛选和精确查找功能。用户可以通过平台内置的搜索功能精确查找相应的景区，通过筛选功能查找特定类别的景区。精确搜索页面如图1(d)所示。

1.2.5　用户信息页面

平台按照用户的个性化设置和旅游动态，搭建用户个人信息页面，如图1(e)所示，包括用户头像、个性签名、积分、隐私、收藏、设置等功能模块，用以满足用户的个性化使用需求和个人使用痕迹记录。

图1　功能设计页面

2 关键技术实现

2.1 前端设计

2.1.1 技术选取

前端设计选用Vue 3.0作为开发工具，完成本平台的轻量小程序的开发。作为一套构建用户界面的渐进式框架，相比其他重量级框架(如angular)，Vue在设计上采用自底向上的增量开发设计[2]。这种框架采用响应式的方式对平台内的乡村文旅相关数据、用户数据进行双向绑定，使用虚拟DOM执行异步更新，具有轻量、数据双向绑定、指令简单、插件丰富等特点[3]。虚拟DOM的核心思想是只要页面依赖数据有所修改，对应的视图就会更新，其优势在于能够将多次操作保存起来，进行合并计算，减少真实DOM的渲染计算次数，提升用户体验[4]。这样一方面能够实现乡村文旅数据的实时更新和真实变化，另一方面通过队列缓冲中去除重复数据的方式，避免了系统不必要的计算，符合本平台“微前端”的思想，有利于轻量化程序的跨平台应用。此外，Vue使用门槛上的便利性，利于乡村文旅平台系统架构的迭代升级，便于后期修改和使用，符合乡村文旅数字化的可持续发展理念。

2.1.2 项目结构

本平台的项目结构主要分为三大模块：状态模块的定义、Redux模块分层与Vue-cli脚手架的搭建。本平台将用户登录、乡村文旅信息推荐、景点详情、精确搜索、用户信息定义为领域实体。领域实体是一系列具有相同目的与相同功能的资源的集合，如景点详情中封装了乡村各种旅游景点相关的信息资源，它们共同构成一个整体，即领域实体。前端架构采用Redux模块分层，平台前端设计中的多个容器组件设为共享状态，平台内组件、页面、领域的状态能够及时改变，乡村文旅信息能够实现及时的互联共享。其工作原理如图2所示，容器组件的信息变动能够及时引起页面状态、通用前端状态的变动，最后实现程序领域状态的变化。

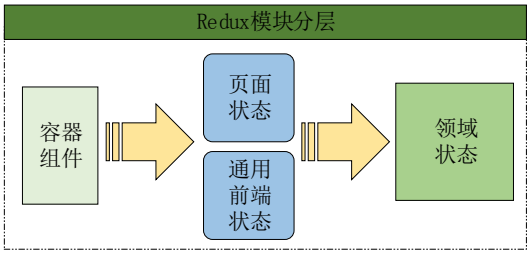


图2 Redux模块分层原理示意

平台前端中Redux模块分层由中间件(middleware)和模块(moudules)组成，redux通过api来加载middleware，每一个中间件middleware能够处理一个相对独立任务，经compose函数串联不同的middleware，便于平台内部复杂状态的管理和处理。

项目需求与技术实现相辅相成，平台项目树中包含node、public、src等模块，展示了项目前端的架构与各模块内部构造。其项目结构详情如图3所示。其中，index.js用于聚合所有的领域状态和页面UI状态。

项目结构采用Vue-cli进行脚手架的搭建。首先使用npm install webpack-g命令全局安装webpack；其次创建脚手架工程，命令vue create mydemo创建mydemo作为项目名，如图3所示，表示Vue-cli脚手架搭建成功。

Vue-cli具有相对成熟的Vue项目架构设计，其本地测试服务器能够满足乡村文旅平台轻量化小程序的测试版运行、测试、改进需求。同

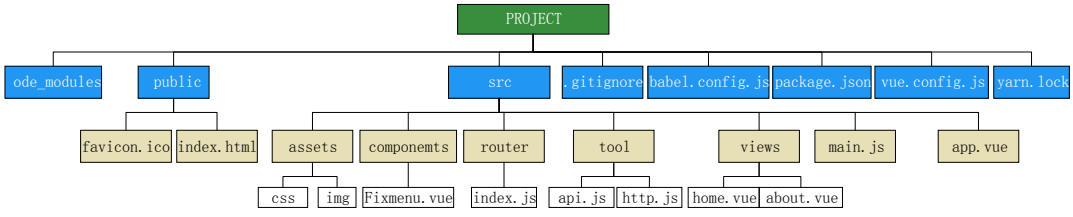


图3 项目文件结构层级

时，其集成化的打包上线方案能够实现平台版本的快速更新，便于平台的迭代升级。

项目前端设计采用Ddd（领域驱动设计），domain域充分增强，多个模块开发互相支持，使得平台中的乡村文旅信息、用户互动模式能够基于小程序的有机整体呈现快速响应、及时联动、实时同步的模式，有利于乡村文旅产业、乡村文化信息的高效运行和传播。

2.2　后端设计

2.2.1　技术选取

平台后端技术选用SpringBoot微框架和Docker应用容器。SpringBoot的约定大于配置的核心思想，默认帮用户做了许多基础设置，很多SpringBoot应用只需要很少的Spring配置就可以正常运行[5]。同时，SpringBoot能够集成大量的框架，解决了重要的项目之间包的版本依赖和稳定性问题[6]。这一特点能够满足本平台微服务的核心数据开发。依托Java平台和Spring框架的可扩展性和可定制性，SpringBoot支撑了本平台的边界工具应用，例如小程序中布局管理器BorderLayout可以将容器的布局分为五个位置，经控件设置窗口大小和窗口显示的位置。

Docker是PaaS提供商dotCloud开源的一个基于LXC（Linux container）的应用容器引擎[7]，其满足了开发者开发应用过程中的应用主体和依赖包的移植镜像需求。平台基于Docker的沙箱机制进行后台开发，能够将相关应用信息的前端+后端+自动化部署的代码全部放入docker并打造镜像，后续可以在服务器中通过docer-ps和docker-run命令直接启动服务，简单便捷，也有较好的封闭性，保障了乡村文旅业态中商家、游客、景区之间多个主体数据和程序移植的安全性。

2.2.2　数据库设计

乡村文旅平台的程序中包含了数以万计的景点信息、用户信息以及周边自定义标签（tag）和复杂的关系（connection，con），因此，数据库的设计和管理不便采用单独的MySQL，而采用关系型数据库MySQL+和key-value型数据库MongoDB。在较为简单的信息存储时使用MySQL，在相对复杂的信息存储时使用MongoDB，即双数据源的数据流管理。这有利于文旅信息能够实现高效的储存、提取、修改、更新，从数据储存层面保障了轻量化小程序使用过程中维护成本的节省和用户体验的提升。目前数据库中已收录浙江省内500个乡村小镇详细数据，部分数据内容如表1所示。

表1　500个乡村小镇数据部分内容

序号	名称	详情介绍	经纬度坐标	地名标签
1	法云古村	法云古村坐落于灵隐景区旁，与灵隐禅寺相接，是杭州最早的……	（120.22，30.25）	幽静；佛教；灵隐寺……
2	芹川古村	典型的徽派古村落，整个村庄以小桥流水古民居为主要特色……	（120.22，30.25）	千岛湖；小众；古樟树……
3	新叶古村	已有800多年历史，是国内目前最大的叶氏聚居村，还保留着……	（119.34，29.34）	建德；爸爸去哪儿；明清建筑……
4	荻浦古村	荻浦古村距今已有900多年，文化底蕴丰厚，孝义文化、古戏曲……	（119.83，29.87）	宋代；孝义文化；徽派建筑……
5	塘栖古镇	塘稻名胜古迹遍布，依景称“栖溪二十四景”。迄今尚存……	（120.19，30.48）	广济桥；京杭大运河；知名景区……
6	龙门古镇	龙门90%以上的村民是三国东吴大帝孙权家族的后裔，据家谱记载……	（119.96，29.91）	孙权；三国；八卦阵……
…	…	…	…	…

注：经纬度坐标是关联地图显示的村镇空间点位，见景点详情界面（图1（c））；地名标签是互联网用户对村镇的特色描述，关联精确搜索页面（图1（d））

2.2.3 自动化运行及部署

乡村文旅平台的自动化运行和部署基于Docker容器进行设计，使用虚拟机的托管环境，相对简化了程序运行和部署的资源，同时也便于后期程序的维护和使用。首先构建应用环境，封装Docker容器的添加、删除等修改操作；其次创建具有集群唯一虚拟IP的Docker容器，实现集群管理；最后构建应用功能划分，运行Docker注册表，实现自动化运行、部署及IT运维。

3 应用前景分析

3.1 以文化为驱动力:具有特色文化的乡村景点

乡村文旅平台构建了一个共享、开放的数字化平台，以数字呈现的方式赋能优秀文化传播，增强特色文化的知名度和影响力。这种面向大众的传播方式为具有一定文化特色的乡村提供了新的平台和机会。这类乡村即使知名度相对欠缺，也能够依托平台实现特色文化元素宣传资料的投放，如宣传片、文物展示等，从而突出自身别具一格的文化内容。例如主打历史名人文化乡村，可以借助平台突出历史名人的介绍以吸引游客；主打特色建筑文化的乡村，可以借助平台生成VR建筑、全景地图等，使游客能够远程体验乡村文化，推动当地优秀文化的传播。

3.2 以产业为纽带:具有特色产业的乡村景点

数字化是乡村经济发展的新阶段与新方向，可以加快乡村经济信息化转型，催生新兴的乡村产业形态[8]。乡村文旅平台能够通过数字化方式赋能乡村景点产业发展，尤其是具有特色的产业。平台一方面能够为乡村特色产业的产品开拓新的销售、变现渠道；另一方面能够通过引流的方式，将游客的目光聚焦于某一产业领域，促进产业的进一步发展或催生产业的新型业态。例如以林业、果业为特色产业的乡村，能够借助平台推广林、果产品，并发展农家乐、果园采摘等体验经济，以此为纽带促进乡村经济、文化、基础设施建设的多维发展。

3.3 以优化管理为方式:具有多主体统筹管理需求的乡村景区

疫情常态化背景下，如何使原本数字化、信息化水平相对落后的乡村景区实现多主体协同管理的优化，成为亟需解决的难点。乡村文旅平台基于QRC的数据信息集成方式，打破各主体、各地区之间标准不对接、接口不统一、设施不共用、信息不互通、资源不共享的问题，能够将游客、商家、景区等多元主体的信息集于一体、实时共享，实现行业监管、流量控制、旅游攻略、特色产品销售等多方面的功能应用，助力具有规模化、规范化的多主体统筹管理需求的乡村景区资源整合、管理优化。

4 结语

基于Vue和SpringBoot框架的乡村文旅平台，能够实现用户登录、乡村文旅信息推荐、景点详情页面、精确搜索等功能。平台采用Redux模块分层结构，以Vue-cli进行脚手架搭建，运用SpringBoot微框架和Docker应用容器，通过二维码的标识实现了乡村文化旅游过程中的游客、商家、景区等多元主体的信息集成。平台以开放、共享的发展理念，面向文旅行业监管、景区运营、游客服务等应用场景为突破口。未来，平台在乡村文旅的文化传播、产业赋能、旅游管理方面具有应用前景。

参考文献：

[1] 中办国办印发《“十四五”文化发展规划》[N]. 人民日报,2022-8-17.

[2] 朱二华. 基于Vue.js的Web前端应用研究[J]. 科技与创新,2017(20):119-121.

[3] 王宏杰. 一种与课堂网络无关的高校考勤系统的设计与实现[J]. 现代计算机,2021,27(29):97-103.

[4] 张浩洋,顾丹鹏,陈肖勇. 基于Vue的数据管理平台实践与应用[J]. 计算机时代,2022(7):66-67,72.

[5] 闫四洋,胡昌平,卞德志,等. 基于SpringBoot+MongoDB的微服务日志系统的实现[J]. 计算机时代,2020(8):69-71,74.

[6] 张峰. 应用SpringBoot改变Web应用开发模式[J]. 科技创新与应用,2017(23):193-194.

[7] 刘思尧,李强,李斌. 基于Docker技术的容器隔离性研究[J]. 软件,2015,36(4):110-113.

[8] 李翔,宗祖盼. 数字文化产业:一种乡村经济振兴的产业模式与路径[J]. 深圳大学学报(人文社会科学版),2020,37(2):74-81.

Design and implementation of rural cultural tourism platform based on Vue and SpringBoot

Li Shengtong[1,2], Liu Zhe[2,3], Yu Dingguo[2,3*], Fang Shenguo[4,5], Sun Xuemin[1,2]

(1. School of Journalism and Communication, Communication University of Zhejiang, Hangzhou 310018, China;
2. Zhejiang Provincial Key Laboratory of Film and Television Media, Hangzhou 310018, China;
3. College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China;
4. Zhejiang Institute of Industry and Information Technology, Hangzhou 310012, China;
5. College of Textile Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Rural cultural tourism is an important step to implement the strategy of rural revitalization in the new era. Based on the background of economic and social development and the needs of rural cultural tourism development, this paper designs and implements a rural cultural and tourism platform based on Vue and SpringBoot framework. The platform adopts the hierarchical structure of Redux module, constructs scaffolding with Vue-cli, and uses SpringBoot microframework and Docker application container to realize the information integration of rural cultural travel marked by QRC, aiming to innovate and spread excellent traditional culture, present a complete picture of rural tourism and improve the efficiency of travel through enabling characteristic cultural transmission, promoting characteristic industries and optimizing the multi-subject overall management of cultural tourism.

Keywords: rural cultural tourism; rural revitalization; Vue; SpringBoot; Qr code

(上接第81页)

Research on application of high-precision vison technology in intelligent manufacturing

Zhang Guangyu*, Xu Han, Wei Yafeng, Zhao Yukun, Li Shouwei, Ji Yong

(Wuxi Zhongwei Hi-tech Electronics Co., Wuxi 214035, China)

Abstract: With the development of science and technology, the highly efficient automatic inspection function of intelligent manufacturing equipment has put forward higher requirements for machine vision 3D reconstruction technology and its reconstruction accuracy and efficiency. Here, the measurement principle and key technologies of high-precision 3D vision reconstruction algorithm were analyzed to research the possible problems in the intelligent manufacturing equipment application of line laser triangulation ranging method, phase corridor measurement method and photometric stereoscopic vision method. Simultaneously, the development direction of 3D vision technology for intelligent manufacturing equipment was prospected.

Keywords: 3D imaging; high-precision; measurement; intelligent manufacturing; stereo vision

文章编号: 1007-1423(2023)08-0104-05 DOI: 10.3969/j.issn.1007-1423.2023.08.017

基于微信小程序的计算机类课程教学平台的设计及应用

黄寿孟[1,2]，刘小飞[1,3]，韩　强[4*]，陆娇娇[1,2]，焦萍萍[1,3]

（1. 三亚学院信息与智能工程学院，三亚 572022；2. 三亚学院陈国良院士团队创新中心，三亚 572022；
3. 三亚学院容淳铭院士工作站，三亚572022；4. 琼台师范学院信息科学技术学院，海口 571100）

摘要：近年来，线上线下混合教学应用广泛，教学平台不断更新，基于微信小程序的教学平台也深受广大师生的青睐。充分利用手机微信进行教学是当前高校教学改革的热点问题，通过学生前台的学习、教师后台的课程管理实现碎片化的教学模式，优化线上教学的效果，提高学生的学习兴趣。解锁基于微信小程序的计算机类课程的教学平台设计与开发，同时验证其教学效果的有效性，对其他学科的教学具有一定的借鉴意义。

关键词：微信小程序；计算机类课程；教学平台；系统设计

0　引言

中国互联网络信息中心发布截至2022年6月，使用手机上网的比例达99.6%，手机在线教育用户规模已超过3亿人，微信日活帐户数已超过4亿[1]。疫情以来，各高校师生进行线上教学的混合学习模式成为教育行业的教学新方向，学生采用手机、平板电脑等电子设备拓展学习资源的获取，在教育行业得到了广泛的应用[2]，同时各高校开展“一师一优课、一课一名师”活动，利用信息化手段扩大优质教育资源覆盖面的有效机制基本形成[3-5]。然而计算机类课程的网络教学平台不多，满足不了高校IT类专业的公开使用，无法为学生提供碎片化课后的学习环境[6-9]。因此本文设计了基于微信小程序的高校计算机类课程教学平台，结合线上线下教学的理念和模式，充分发挥手机微信小程序的教学平台优势，帮助计算机类专业的学生获取学习资源，为师生开展混合教学提供便捷的网络平台。

1　相关工作

1.1　微信小程序

微信小程序是一种基于微信平台的简易手机应用，其设计理念就是方便快捷。只要用户动动手扫描二维码、搜一搜，就能在手机上直接使用。微信小程序的种类繁多，其功能也极其丰富，最突出的还是“随用随走”的理念，使用户不再担心安装太多应用，耗费手机内存资源的问题。微信小程序的优点有：一是加载速度快；二是不消耗手机内存；三是不需要下载特定软件，既不费流量，又不占内存空间；四是丰富的设备访问能力，微信平台为小程序提供了一系列的控件，让微信小程序拥有了极其丰富的设备访问能力(比如GPS、相机等)；五是接口方便，微信平台搜索、微信产品链接、公众号内链、自定义菜单等均可进入小程序；六是开发成本低、维护简便。缺点是无法开发大型的应用程序，无法跳转外网，间接影响了其开放性等。

收稿日期：2022-12-21　　**修稿日期：**2023-01-06

基金项目：三亚学院校级课程考核改革试点项目(SYJGKH2022110)；三亚学院重大专项课题(USY22XK-04)

作者简介：黄寿孟(1975—)，男，广东湛江人，硕士，副教授，主要研究方向为信息技术、现代教育技术；刘小飞(1984—)，女，吉林松原人，硕士，副教授，研究方向为软件应用、图像处理；*通信作者：韩强(1982—)，男，海南海口人，硕士，讲师，研究方向为软件应用、数据安全、人工智能，E-mail：huang123888@126.com；陆娇娇(1984—)，女，湖北随州人，硕士，讲师，研究方向为软件应用、信息处理；焦萍萍(1983—)，女，江西南昌人，硕士，讲师，研究方向为软件应用、信息处理

1.2　线上教学平台

随着教育赋能的需求越来越高，线上教学平台的种类和规模也越来越多，有综合性的教学模式，也有单一学科的教学模式、有手机端也有电脑用户端，张茵茵教授[10]认为在高等教育中教师将手机学习引进课堂，开展丰富的教学活动是大势所趋。韩锡斌等[11]学者也支持教师个性化、多模式的课程教学，充分利用网络教育资源开展手机移动终端的沉浸式、强交互、重体验的学习环境。绝大多数的国内外专家认同线上教学平台主要是以手机学习为主的网络资源服务。基于从学习者这种手机学习方式出发，线上教学平台务必提供协作手机学习环境，支持手机学习的新型协作方式。

本文所指的是基于微信小程序的计算机类课程教学平台，它是一种基于手机学习的计算机类课程资源的“互联网+”教学平台，它既支持方便又快捷的学习模式，又支持强交互的教学模式，是教师和学生互动教学的一种网络平台，它的主要功能包括创建管理课程、题库管理、课程学习、互动答疑等，支持提问、讨论、错题、评分、作业等多种功能。

2　教学平台设计

教学平台是师生线上学习的交流媒体，也是师生共享资料、体验知识的学习界面。在新时代的教学手段下，教师开展教学活动应为学生提供新的学习环境，以此更加方便学生进行个性化学习，更加有效地提高学生的学习兴趣。

2.1　总体架构设计

本教学平台，通过课题组成员一年多的调研，总结出计算机类课程教学平台的功能模块主要包括由师生使用的前台模块和后台教师管理模块，其中前台主要功能有课程学习、课后管理以及帐户管理，而后台主要功能包括课程管理、学习管理、用户信息管理，其功能架构设计如图1所示。

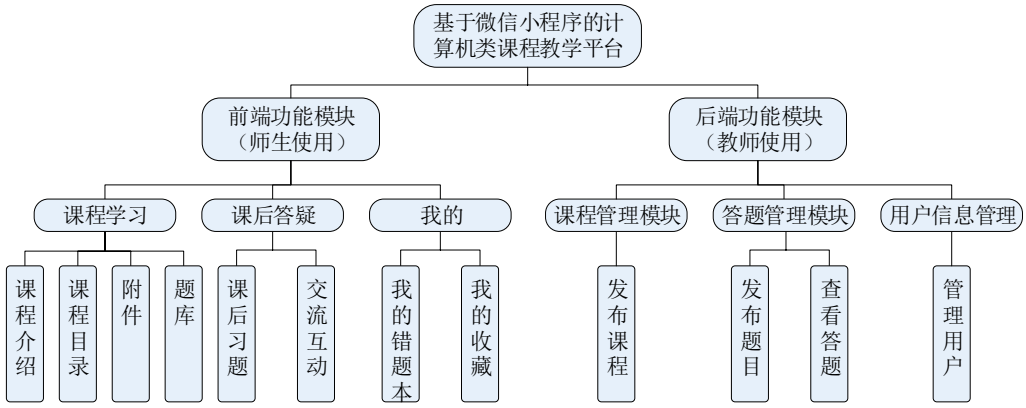


图1　教学平台的总体设计框架

2.2　设计原则

本教学平台是基于手机微信小程序而开发的，目标是解决计算机专业学生在预习或复习时资料缺乏的困扰，充分利用小程序特点与优势，弥补计算机类教学资源上的不足，让教学效果更有效。因此本教学平台的设计需满足以下几点原则：

（1）平台界面友好：第一层界面重点突出，方便师生快速理解重点内容，第二层界面流程明确，减少不必要的干扰内容。

（2）链接清晰明确：第一层导航明确，有页面导航、标签分页、安全跳转；第二层愉悦使用，回应及时，减少等待；第三层提醒明确，异常返回。

（3）操作便捷：手机学习，操作简单，减少输入，流畅便捷。

（4）画面统一：页面风格统一，具有延续性，学习成本极低。

2.3 界面设计

2.3.1 前台界面

师生共享资料的界面，要求简易明了，所以只有课程、课后答疑、个人信息共三个模块，比如本课题组的成员设计的计算机课程中的《信息技术》教学平台，其前台界面如图2所示，点击界面中的“课程学习”即有“课件”视频和课程讲解视频；点击界面中的“课后答疑”即有课后作业以及相关要求，师生互动解答相关的资料；还有“我的”模块是用户学习过程中的记录内容，包括学习课程的收藏、答题记录、互动记录、观看记录，等等。

图 2 教学平台的前台界面

2.3.2 后台界面

后台功能主要是教师课程管理功能，包括课程资源上传、课后答疑回复记录、课程评价分析、课程设置管理等内容。其中教学进度或学习情况在每章节中可以体现出来，让教师及时掌握知识点的学习进度，还有评价考试的结果等内容，后台界面如图4所示。

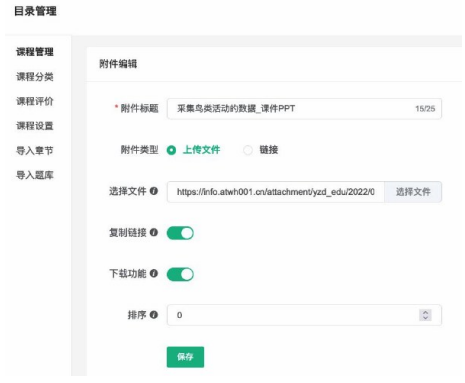


图 3 教学平台的后台管理界面

2.4 功能设计

2.4.1 课程管理功能

课程管理功能主要是课程资源的管理，在前台显示是“课程资源”，在后台显示是“课程管理”，首先由老师上传学习课件、附件、题库等课程资源，并开通学习权限，学生即可查看课程进入学习，老师在学生结束课程时即可观看学习数据，其功能示意图如图4所示。

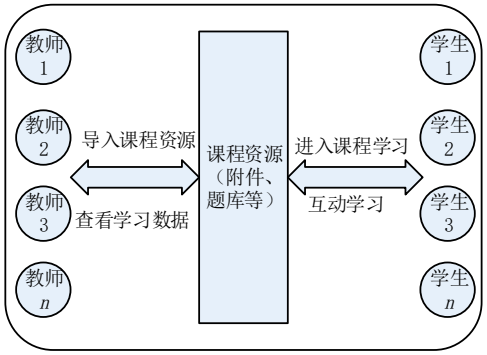


图 4 课程管理功能的示意图

2.4.2 学习管理功能

学习管理功能是本教学平台最重要的功能模块，主要是教师发布学习资源、题库练习及作业，学生即可访问相关资源，学生答题、上交作业后平台自动评定结果，提供结果解析，学生即可了解学习结果，教师也可了解学习分析，其功能流程如图5所示。

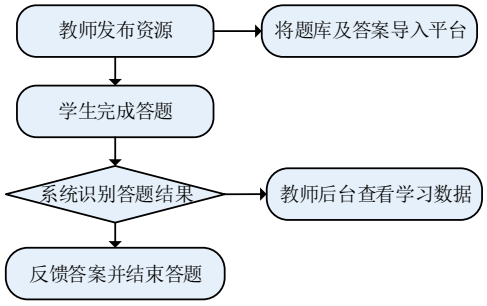


图 5 学习管理功能的流程

3 教学平台的应用

3.1 应用环境及对象

微信小程序开发时有系统架构的渲染层与

逻辑层，渲染层接收用户数据，并触发其内容的更新，同时传递给逻辑层，让其进行相关事务的处理。本教学平台选用阿里云服务平台作为后台服务器ECS(1 M出口的带宽，1核CPU 1 GB，64位CentOS操作系统，含宝塔服务器管理软件)，使用对象是大一信息技术专业的学生(信息技术专业有2101班作为实验班，2102班为对照班)，学生在大学之前并没有使用过此类基于微信的教学平台。当然为了让学生更主动自学，增强学习效果，课前预习时会发布如图6所示的教学阶段图给学生，为学生创造条件，提高教学效果。

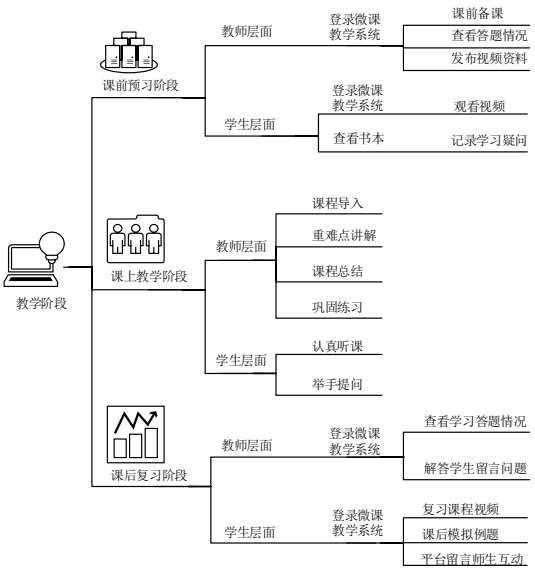


图6　教学平台的学习阶段图

3.2　应用效果

本课题成员对使用教学平台上课的学生进行教学满意度的调查(采用问卷调查形式)，学习的教学平台包括计算机导论、程序设计C、大数据可视化、网络与信息安全、信息技术等课程，调查内容主要有教学平台对学习的帮助情况、学习兴趣情况、学习主动性效果等方面，然后对有效的问卷进行统计，通过软件SPSS分析出可信度Crobach Alpha的系数为0.938，高于标准值0.8，验证调查问卷的有效性。同时本课题成员对《信息技术》教学平台的实验班和对照班的考试成绩进行分析，结果如图7所示，实验班考试成绩的平均分均高于对照班10分之多，这也说明了教学平台有益于学生学习的体现。

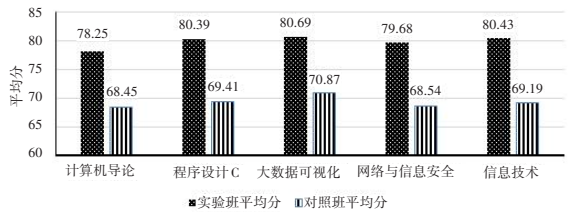


图7　考试成绩统计

4　结语

微信小程序与教学平台相结合，为教育改革提供新的动力，推动混合教学的不断发展，本课题充分说明了教学平台对教学效果有显著的积极作用，深受师生的认可。当然教学平台的教学资源有待补充和更新，学生个性化学习的能力也有待加强，还有应用课程的科目较少，希望今后继续开发出更强更好更多的教学平台。

参考文献：

[1] 陆禹文. 基于微信小程序的移动学习平台的设计开发及应用[D]. 兰州:兰州大学,2020.
[2] 王斌. 基于微信小程序的信息技术课程教学平台的设计与应用[D]. 广州:广东技术师范大学,2022.
[3] 黄寿孟. 基于超网络模型的混合教学知识传播研究[J]. 信息与电脑(理论版),2019(5):37-39.
[4] 马莲姑,黄寿孟,张洁. 高校微课评价指标体系构建研究[J]. 中国现代教育装备,2022(13):127-129.
[5] 黄寿孟,马莲姑. 基于深度学习的混合教学评价体系研究[J]. 甘肃广播电视大学学报,2022,32(1):38-44.
[6] 马莲姑,黄寿孟,吴坤明. 我国高校微课教学有效性现状分析研究[J]. 中国教育技术装备,2020(23):77-79.
[7] 马莲姑,黄寿孟,贾晓婷. 微课资源在高校教学中的应用效果评价[J]. 现代计算机,2020(25):60-63.
[8] 程亚恒,黄寿孟,韩凤娟,等. 基于微信小程序的出行导航预警系统设计[J]. 信息与电脑(理论版),2020,32(12):90-91.
[9] 冯淑娟,黄寿孟. 微课教学下的高校教师产学研培养模式研究[J]. 和田师范专科学校学报,2019,38(6):47-50.
[10] 张茵茵. 移动教学平台在高等教育教学中的应用[J]. 软件导刊(教育技术),2017,16(30):23-24.
[11] 韩锡斌,葛文双. 中国高校教师信息化教学能力调查研究[J]. 中国高教研究,2018(7):53-59.

(下转第112页)

文章编号: 1007-1423（2023）08-0108-05 DOI: 10.3969/j.issn.1007-1423.2023.08.018

基于云平台的批量话单快速解码方法

武小波[1]，李建林[2*]

（1.中国电信山西分公司，太原 030006；2.山西应用科技学院，太原 030006）

摘要：针对当前运营商业务支撑系统的处理话单能力，特别是对用户批量漫游话单解码处理慢的现状，探索了利用分布式计算框架搭建的云计算能力解码的方法。在服务器内存中建立分层处理逻辑，将解码需要的配置信息、索引信息等关键数据放入内存中，同时建立多节点部署的集群处理架构，采用云计算方式对批量话单进行并行解码处理，实时输出处理结果。实验表明该方法可有效提升批量话单解码的效率，有效提升运营商对外服务的能力。

关键词：云计算；分布式计算；解码；批量话单

0 引言

漫游话单是用户离开本省使用移动终端通话产生的使用记录，因记录落在漫游省，需要通过编码处理后，利用特殊途径回传给用户归属省进行处理。随着用户数量和出省使用量的增加，漫游话单量持续增加，对本地处理系统的性能考验越来越高，目前的漫游话单处理系统以集中物理数据库为基础，对漫游话单数据进行统一解码处理，需要将配置文件预先设置在数据库，处理话单时批量读取配置到缓存进行读取，整体处理时效在分钟级，其性能会随着数据量的增加而逐渐降低，且系统扩充涉及到的修改量和扩容费用庞大，影响了系统对漫游话单的解码处理速度[1]。

本文提供了一种基于云平台的快速处理批量编码话单的方法，利用分布式计算技术和文件存储技术，取消传统物理库加缓存机制，达到快速处理编码话单的能力。

1 技术应用

1.1 云平台

云平台是应用廉价的服务器，在不同机房搭建的分布式处理集群，云平台搭建时采用分层架构，即Iaas、Paas、Saas三层。Iaas层属于硬件资源层，使用统一的管控平台进行资源管理；Paas层是组件层，统一搭建数据库、缓存、中间件等，进行组件资源的统一分配；Saas层部署应用，采用主备方式进行应用的搭建，确保应用使用的安全[2]。

本文讨论的云平台是使用内网方式组网，通过内部交换机安全、可靠地传输数据[3]。云平台分为三层四个部分，最下面的一层为Iaas的硬件资源层，包括基础硬件资源层和资源调度层，基础硬件资源层提供服务器资源、网络资料、存储资源等系统运行的基础硬件环境，资源调度层实现对基础硬件的管理和控制，可以实现对硬件资源使用率的弹性计算，能根据应用的需求实现资源的灵活分配，可以对资源进行动态的扩缩容管理；第二层为组件运行的Paas层，有进行数据存储的分布式数据库，文件存储的分布式文件系统，用于处理消息的分布式消息中间件，实现数据临时访问的分布式缓存，进行任务调度的组件，还有用于平衡应用处理的负载均衡组件；第三层为应用部署的Saas层，应用有采集中心、预处理中心、剔重中心、批价中心四个，其构成了系统对外的统

收稿日期：2022-12-27 **修稿日期：**2023-01-14

作者简介：武小波（1979—），男，山西太原人，硕士研究生，高级工程师，研究方向为IT项目管理；*通信作者：李建林（1981—），女，山西太原人，硕士研究生，高级工程师，研究方向为数据挖掘，E-mail：pc_ljl@126.com

一应用能力；除了三层功能外，系统架构还设计了运维管理部分，用于对各层的监控，建立规范约束，进行日常的安全管理，实现运维操作。其结构如图1所示。

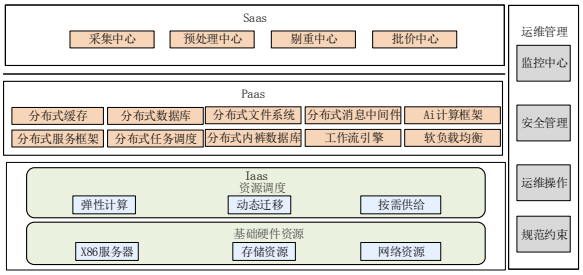


图1　云平台总体结构

1.2　分布式云计算

传统运营商的话单存放在集中搭建的物理数据库上，数据统一存放在磁盘上，集中部署的资源以共享为主，包括CPU处理器、内存资源、存储设备等，资源依赖于一个商用的操作系统进行管理。虽然已经出现的固态硬盘、闪存硬盘等高性能硬件，带来了数据存储、读取的速度提升，但是依然存在海量数据读写性能瓶颈、数据灾备恢复时间长等问题[4]。随着对外服务质量要求的提升，利用云资源实现从集中式到分布式的变革成为新的要求，云资源可以部署在集中或分布式的数据中心上面，由物理集群或虚拟计算资源组成，实现集中式的磁阵计算向分布式云计算发展成为趋势[5]。在分布式计算中，处理器资源和内存资源，可以相互紧耦合，也可以虚拟化实现松耦合，资源通过管理工具实现共享，分布式计算由一个分布式管理系统和众多独立自治的服务器组成，各自拥有独立运行的内存，相互通过计算机网络实现信息交互[6]。

分布式的云计算系统建立在大量自治的服务器节点上，节点之间通过SAN、LAN或WAN网络，实现跨网络层级方式的互连。本系统采用内存存取的分布式云计算系统，基于物理内存存储和读取数据，访问性能高，将数据统一存放在物理内存中，可以充分利用内存访问的超强速度优势，实现快速的I/O交互。分布式部署，可以承载的数据量大，应用部署在互通网络的多个节点上，对外可以提供统一的访问入口[7]。采用文件形式组织数据结构，文件存储是将话单依据业务类型，直接解析放到文件中，不再放到传统物理数据库中，数据会以单条信息的形式存储在文件夹中，正如将几张纸放入一个文件袋中一样。当需要访问该数据时，计算机通过相应的路径查找[8]。存储在文件中的数据会根据数量有限的元数据来进行整理和检索，这些元数据会告诉计算机文件所在的确切位置，它就像是数据文件的库卡目录[9]。

依据摩尔定律的发展，服务器的硬件费用逐年降低，其运算速度逐年提高，应用系统充分利用分布式云计算的快速部署、高速计算、动态扩展能力，加上文件存储的丰富多样性，将极大提升对漫游话单的处理效率[10]。

1.3　话单编码技术

为了确保话单的安全传输，运营商的话单普遍采用ASN.1编码技术，ASN.1(abstract syntax notation one)是一种抽象语法标记，它定义了抽象数据类型形式标准，是一种通用的用于表示数据层次的数据结构。抽象语法让使用者可以根据实际需要定义数据类型，并指明这些数据类型的值。ASN.1使用一整套正规的格式来描述对象的结构，实际使用过程中不管语法上的指代，也不管如何执行，这种语法标记不关心执行的应用程序。

2　快速解码设计

漫游话单处理需要经过采集、预处理、剔重等环节，各个环节需要读取配置信息，才能正确对话单进行解析。本文采用内存分层多单元并行处理方法，改变传统漫游话单采集、预处理、剔重信息置于物理库的方式，直接将采集、预处理、剔重的解析信息部署在云平台，利用分布式计算速度快的优势并行计算解码程序，将各个环节的操作时间压缩至秒内。

2.1　设计思想

将服务器的内存块逻辑分层，一层存放内存文件数据，将解码信息存放此单元，用于批量话单的解码检索、解码信息存放。一层存管理网络文件数据，用于记录通过网络传输的文件信

息，保障文件处理不重复。一层进行数据处理，用于根据上两层信息快速进行文件中的话单解码。处理过程中充分利用分布式计算的快速、安全特性，对话单文件实现类似流水操作。

2.2 整体架构

本方法快速解码话单原理图，如图2所示，该系统包括20个处理模块M1,M2,…,M20，每个模块中都包括内存文件数据库单元A，网络文件管理单元B以及相连的数据处理单元C，其中A1表示第一内存文件数据检索单元，B1表示第二网络文件数据管理单元，C1表示第三数据处理单元。处理主机分为1主2备模式，一台主机进行处理操作时，另外两台主机进行同时备份操作，出现主机宕机问题影响话单的解码流程时，两台备份主机没有主次备份区别，直接接管进行操作。

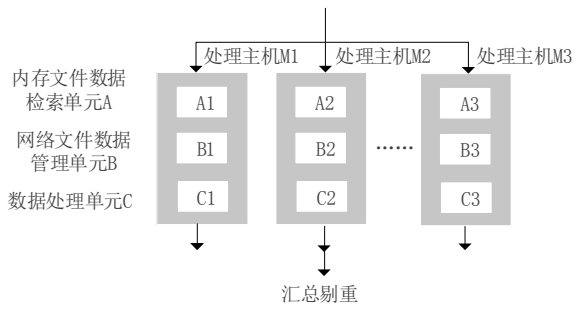


图2 整体架构

处理流程为，通过两级查询处理生成当前索引记录的查询结果，将所有查询结果发送查询请求的网络文件数据管理单元B。网络文件数据管理单元B，用于存储话单文件，以及为各数据处理单元C分配目录文件。数据处理单元C，用于进行解码程序的操作和运行，以及执行分配给自身的任务，收集自身及本系统其他正常工作的数据处理单元C的执行结果。

具体处理步骤如下：

步骤1：将漫游话单文件中的通话记录的索引存储到云平台的内存文件单元A中。

步骤2：将需要处理的话单文件存储到网络文件管理单元B中。

步骤3：数据处理单元收到查询请求后，转发给内存文件数据库单元A。

步骤4：内存文件数据库单元查找符合条件的索引记录，返回查询结果。

步骤5：数据处理单元C将查询结果作为一批次任务，分配给自身及系统中其他正常工作的数据处理单元。

最后将上述执行结果进行汇总处理后返回，结果落地为文件，之后进行剔重等环节操作。

2.3 集群部署

本系统部署在高可用的集群上，采用间隔5公里的双中心双机房搭建集群。集群搭建需要利用一组服务器，服务器作为一个整体向用户提供所需资源，这些单独的服务器系统就是集群的节点。集群的部署需要具备高可用，高可用集群的部署是为了使集群的整体服务的质量高，能保障应用的连续运行，以便减少因服务器硬件和软件的运行故障所带来的损失。如果应用运行的某个节点故障，它的备用节点将在几秒钟的时间内接管业务。高可用集群通过智能调度机制，保障业务程序对外提供服务的不间断，把因为软件、硬件等因素出现的故障对业务的影响降低到最小程度。集群部署架构如图3所示。

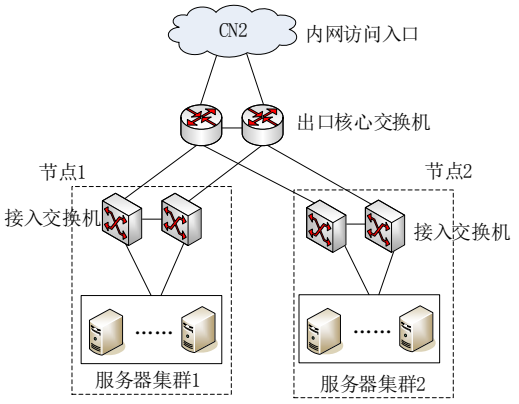


图3 集群部署架构

3 实验

3.1 实验环境

本实验搭建集群测试框架整体效果，实验基于Python3.6.8，应用Zookeeper进行集群资源调度，话单解码使用C++完成撰写，使用Python作为测试编程语言。配置参数见表1。

表 1　集群配置

配置项	参数
操作系统	CentOS 7.4.1708
处理器型号	Intel Xeon E5-2630v4 25M 2.20GH
处理器个数	2
处理器核数	20
内存(GB)	256
Zookeeper	V3.4

3.2　实验数据

在实验中，为规避数据异常丢失的风险，提高数据的安全性，服务器集群采用一主两从部署，主服务器存储磁盘中的文件X，另外两个文件副本将同时存储在备服务器。当主服务器所在物理设备发生故障，文件X在主服务器上不能被访问，但是存储在另外两个服务器上的文件副本可以被正常访问，文件系统仍能正常对外提供文件X的数据。由此可以增加系统的可靠性，数据的安全性。

内存文件数据检索单元A用于存储话单文件中的通话记录的索引，用于查找符合查询条件的索引记录，并逐条读取符合查询条件的索引记录中的特征字段和文件名字段，使用地市ID和用户ID组成的两层查询条件，精确分配给各个正常工作的网络文件数据管理单元，以及执行分配给自身的任务，收集自身及上述其他正常工作的网络文件数据管理单元B的执行结果。

根据当前任务，网络文件数据管理单元中查找相应的话单文件，并在找到的话单文件中查找对应的通话记录，即当前网络文件数据管理单元根据收到的位置信息，读取其中的漫游话单记录内容并汇总，将汇总结果返回给数据处理单元C。上述漫游话单记录的索引包括主被叫号码、通话起止时间、通话记录在话单文件中的偏移量以及上述话单文件在上述网络文件管理单元中的文件名等。

数据处理单元C通过解码文件在本地索引记录中查找符合查询条件的索引记录，逐条读取符合查询条件的索引记录中的偏移量字段和文件名字段，生成解码结果。最后判断本次收到的任务是否执行完毕，若是，则汇总上述目录文件的位置信息。

分布式系统运行的性能有多个衡量的指标，常用的系统吞吐量用TPS(transactions per second，每秒事务数)测量，其它度量指标还包括网络延迟、任务响应时间等。系统自身运行的消耗通常包括指令编译的时间、操作系统的启动时间、I/O读取数据速率和程序运行时指令交互系统消耗，业务运行时还需要考虑服务的QoS、可靠性和系统可用性，以及系统安全运行的能力，这些指标都需要综合分析。

3.3　实验结果和分析

以10万漫游话单处理进行对比，分布式计算采用双节点部署，每个节点5台服务器组成集群。对比现有整体程序单节点10台服务器处理，新架构因为采用分布式调度，直接读取内存库数据，利用内存分层进行漫游话单解码，10万条话单的处理效率从之前的5分钟缩短到了30秒。

表 1　实验结果

方式	效率(10万话单)	TPS	节点数量
单节点	5分钟	300条/秒	1
分布式计算	30秒	2000条/秒	2

4　结语

本文探索了利用云平台的分布式计算，解决运营商支撑系统处理批量漫游话单解码效率低的问题。本方法通过分层并行的处理模式，快速将漫游数据进行分层处理解码，各个处理单元并行处理其各自独立信息，优化加速整体处理流程和速度。实验表明：利用云计算快速部署、快速计算的特性，将话单全程存放在文件中，批量处理话单的效率从分钟级提升到秒级，提升了运营商对外服务的感知。

参考文献：

[1] 刘岩,柳乐怡,王冬,等.基于深度学习算法的调度自动化云平台任务优化策略研究[J].机械与电子,2022,40(10):32-36.

[2] 姚利侠,付萍华,周红艳.云平台的大数据信息安全机制的几点探讨[J].网络安全技术与应用,2022(8):68-70.

[3] 张春,方瑞,程宇,等.电信运营商多样性存储应用实践[J].电信工程技术与标准化,2022,35(6):14-21.

[4] 刘天成,凌书平,罗平. 政企OA文件存储的新方法[J]. 现代计算机,2022,28(15):91-95.
[5] 赵泓尧,赵展浩,杨皖晴,等. 内存数据库并发控制算法的实验研究[J]. 软件学报,2022,33(3):867-890.
[6] 严逸,徐伟. 一种基于Redis数据库的海量动目标存储方法[J]. 电子质量,2022(8):112-115.
[7] 高益. ASN.1的PER分层运行库系统的设计和实现[J]. 微处理机,2020,41(5):26-29.
[8] 王丽,杜鹏程,许一鸣,等. 基于分层架构模式识别的软件架构重构技术[J]. 电子学报,2021,49(1):201-208.
[9] 莫凌飞,胡书铭. 用于目标检测的邻域融合与分层并行特征金字塔网络(英文)[J]. Journal of Southeast University (English Edition), 2020, 36(3):252-263.
[10] 杨朝树,诸葛晴凤,沙行勉,等. 持久化内存文件系统的磨损攻击与防御机制[J]. 软件学报,2020,31(6):1909-1929.

Fast decoding method for batch call detail record based on cloud platform

Wu Xiaobo[1], Li Jianlin[2*]

(1. Shanxi Branch of China Telecom, Taiyuan 030006, China;
2. Shanxi College of Applied Science and Technology, Taiyuan 030006, China)

Abstract: In view of the CDR processing capability of the current carrier service support system, especially the slow processing of batch roaming CDR decoding, the method of cloud computing capability based on the distributed computing framework is explored. The hierarchical processing logic is built in the server memory to store key data such as configuration information and index information required for decoding in the memory. In addition, a multi-node cluster processing architecture is built to concurrently decode batch CDRS in cloud computing mode and output processing results in real time. The experimental results show that this method can effectively improve the efficiency of batch CDR decoding and enhance the external service capability of operators.

Keywords: cloud computing; distributed computation; decoding; batch CDRS

(上接第107页)

Design and application of computer courses teaching platform based on WeChat mini program

Huang Shoumeng[1,2], Liu Xiaofei[1,3], Han Qiang[4*], Lu Jiaojiao[1,2], Jiao Pingping[1,3]

(1. School of Information & Intelligence Engineering, University of Sanya, Sanya 572022, China;
2. Academician Guoliang Chen Team Innovation Center, University of Sanya, Sanya 572022, China;
3. Academician Chunming Rong Workstation, University of Sanya, Sanya 572022, China;
4. College of Information Science and Technology, Qiongtai Normal University, Haikou 571100, China)

Abstract: In recent years, online and offline hybrid teaching is widely used, and the teaching platform is constantly updated, and the teaching platform based on WeChat applet is also used by teachers and students. Making full use of mobile phone Wechat for teaching is a hot issue in the current teaching reform in colleges and universities. Through students' learning at the front desk and teachers' curriculum management at the back desk, the fragmented teaching mode can be realized, the effect of online teaching can be optimized, and students' interest in learning can be improved. It unlocks the design and development of the teaching platform of computer courses based on Wechat applet, and verifies the effectiveness of the teaching effect, which has certain reference significance for the teaching of other disciplines.

Keywords: WeChat mini program; computer courses; teaching platform; system design

文章编号: 1007-1423（2023）08-0113-04 DOI: 10.3969/j.issn.1007-1423.2023.08.019

高校智能化实训管理系统的研究与设计

管 彤*

（贵州经贸职业技术学院信息工程系，贵阳 580001）

摘要：随着国家对职业教育的重视程度加大，培养更多高素质技能人才是高校职业教育的重要目标，而实训室则是培养学生动手实践能力，提高专业技能的重要场所。随着职业院校招生规模的日益增大，传统的实训管理模式已不能满足目前工作的需要，当前大部分实训工作都是采用手工和半自动化的方式进行，不仅工作效益低，也容易出现一些人为因素的失误。为了减轻实训室工作人员的负担，依据贵州经贸职业技术学院实训管理的现况，采用软件手段，实现实训管理的自动化、信息化，提高实训管理效率，提升实训教学的效果。

关键词：实训管理；系统设计；研究与设计

0 引言

为便于实训室工作人员的实训工作管理，并切实提高学校实训教育水平和实训室管理质量，各院校相继开发了不少实训管理系统，这些系统各具特色，但也有不足，比如不少系统还不能完全满足各个学校的实际实训状况的需要。这些实训系统，仅在实训工作的整个业务处理的过程中向各学院提供片面支持，而不能与实训相关的工作部门进行资源共享，因此未能达到实训资源利用率的最大化要求[1]。

目前，贵州经贸职业技术学院的大部分实训工作都是采用手工或半自动化的方法进行的。由于实训课程科目繁多、管理复杂，目前的管理模式工作效率低，且容易出现主观原因的失误。本文将运用科学合理的分析方法，拟提出一个解决方案，期望较为全面地解决现有系统存在的问题，探讨如何开发一套采用UML建模、基于B/S架构的实训管理系统，使之适用于贵州经贸职业技术学院的智能化实训管理工作。

在系统分析设计过程中，可用多种开发技术，包括UML统一建模语言、B/S三层架构技术、.NET框架技术及MySQL数据库等[2]。根据系统需求需要实现考勤管理模块、预约管理模块、实训管理模块、上课管理模块、基础信息管理模块、系统管理模块等功能模块。

面向对象技术[3]是目前软件项目中的主流技术，它能有效降低项目开发本身和项目管理的难度，开发出稳定性强、易于升级和维护的程序。UML是对软件密集型系统进行可视化建模的一种通用语言，其通过标准的、统一的图符构成图形来描述软件模型。UML的主要目标[4]是以面向对象图的方式来描述任何类型的系统。UML常用于建立软件系统的模型，在非软件领域也有不少应用，如机械系统、企业机构、业务过程、工业系统等。在本项目的分析、设计中都采用了UML和面向对象的技术。

1 需求分析

1.1 业务需求分析

1.1.1 业务问题定义

随着学院对实训强度的加大，更要求实训室加强现代化管理[4]。在传统方式下，教师对实训室预约主要以口头通知为主，容易造成实训室使用混乱、实训设备安排不当等问题，提

收稿日期：2022-12-15 **修稿日期：**2023-01-04

作者简介：*通信作者：管彤（1983—），女，山东菏泽人，硕士，讲师，主要研究方向为计算机技术应用、计算机教育，E-mail：409522569@qq.com

前预约也存在管理人员遗忘或安排错乱的情况，降低了实训室的使用效率。传统上课考勤记录以点名为主，浪费教学的时间。教师上完实训课后，一般都不对实训的效果进行考评，导致学生主动性及积极性差，对实训课程不重视，且教师对实训成绩一般是凭印象进行评定，这些情况都极大降低了实训的效果。

1.1.2 组织职能分析

贵州经贸职业技术学院实训管理的组织机构如图1所示。实习实训科是在主管院长及教务处长领导下负责全校学生实习和实训工作的职能部门。其下属分为六个系部、十多个专业的核心实训室。实习实训科主要负责督促、检查各教研室各专业实训基地建设情况，实训设备的选购与备案情况；并组织各专业制定好学生实训的计划、大纲；负责各专业学生进行实训课的排课安排，实训室的预约管理，解决实训课中存在的问题和突发事件等。

1.2 功能需求分析

一般来说，功能需求即确定了系统必须要实现的功能，用户通过对系统内各项功能的使用来完成各自的工作任务，即达到了其业务需要[4]。

1.2.1 角色分析

以下从贵州经贸职业技术学院实训管理系统的实际进行需求分析，系统涉及到的角色有：管理员、教师、学生，具体角色划分如表1所示。

表1 用户角色划分

角色	职责或功能
学生	进行上课考勤签到、查看实训成绩等
教师	进行实训室预约、上课课程登记、学生实训成绩登记等
管理人员	实训预约回复、实训管理、用户管理等

1.2.2 系统用例分析

（1）总体用例分析从功能分析的角度出发，给出了贵州经贸职业技术学院实训管理系统的总体用例图，包含系统管理、基础信息管理、预约管理、上课管理、实训管理、考勤管理用例。

（2）预约管理用例描述了教师必须是教师角色登录系统进行预约登记和查询预约，管理员根据预约的情况回复预约信息的过程。

（3）上课管理用例描述了教师使用系统进行上课登记和上课查询的过程。

（4）考勤管理用例描述了学生进行考勤登记的过程，教师对考勤情况进行查询的过程，前置条件是学生角色登录系统验证无误；上课登记记录；学生所属上课登记的班级，操作后生成考勤信息记录。

（5）实训管理用例图描述了教师使用本系统实训管理模块进行实训成绩记录的过程，生成实训成绩信息记录，已签到的学生可以在系统中进行成绩查询。

（6）基础信息管理用例描述了管理员使用本

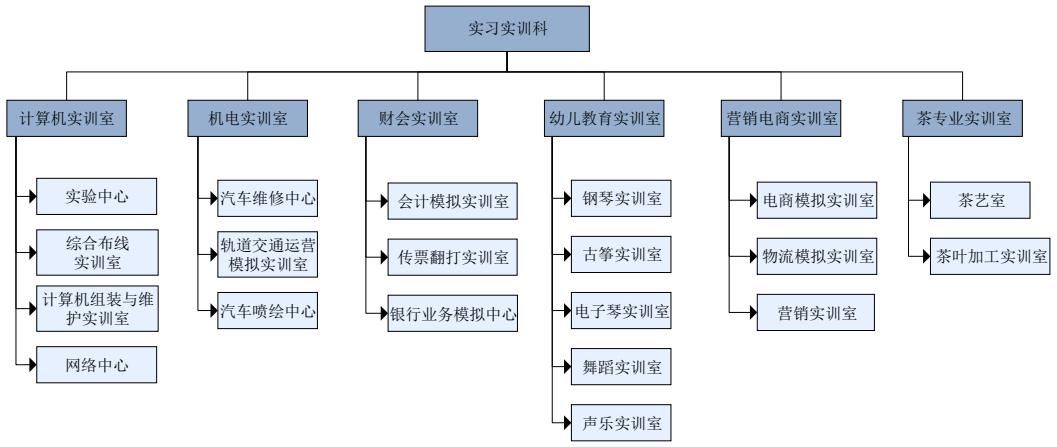


图1 贵州经贸职业技术学院实训管理组织机构图

系统对基础管理模块进行班级信息、学生信息、教师信息、课程信息及实训室信息基础数据维护的过程。

（7）系统管理用例描述了管理员使用本系统对系统管理模块进行用户管理、模块管理、角色管理、权限管理等基础信息模块数据维护的过程。

2　系统设计

任何软件工程项目，在编码前必须要作软件设计，软件设计是软件开发工程的关键步骤，直接影响项目的质量。在软件需求分析阶段已经了解了软件的各种需求，那就意味着解决了本软件中“做什么”的问题，并且在软件需求说明书中应充分地阐明这些需求，接下来就开始设计系统体系结构和各个功能模块的结构。即软件设计阶段要解决“如何做”的问题，最终的结果以“设计模型图”来反映[5]。

2.1　总体设计

2.1.1　系统设计原则

只有在一定的原则指导下，系统设计工作才能做的更好，本系统设计过程中遵循以下原则[6]：

（1）安全性可靠性。本系统中数据资源涉及到一些敏感数据，对系统资源管理机制上需要提供一定的权限管理，系统管理员对不同角色用户分配不同权限，以此来确保系统的可靠性。

（2）实用性。系统提供的功能可以满足用户需求，在实际工作中能真实有效地减轻实训工作人员的负担。

（3）先进性。在系统设计时要考虑到系统的先进性。运用当前较先进的开发技术以保证系统不会在短时间内淘汰。

（4）易操作性。系统界面能够易于理解和易于操作，对系统界面设计要美观友好，能够让用户得心应手地使用。

（5）可扩展性原则。系统的开发过程应考虑到系统的各个功能模块尽量独立，增加预留了可扩展接口，以便后期系统的维护和功能的扩展工作。

2.1.2　系统体系结构设计

对于本系统来说，为了便于实训教师能够及时对实训室进行预约，同时也便于学校实训管理人员对系统进行维护，并以节约成本为前提，本系统采用基于B/S的体系结构来进行设计，这样用户可以方便快捷地通过网络来访问本系统，更进一步简化了系统的开发、维护和使用。

2.2　功能模块设计

如图2所示，贵州经贸职业技术学院实训管理系统一共分为六个功能模块：预约管理、上课管理、考勤管理、实训管理、基础信息管理、系统管理[4]。

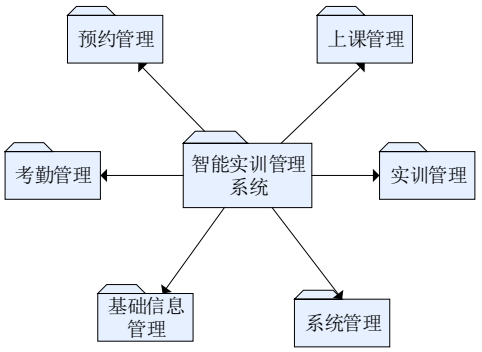


图2　贵州经贸职业技术学院实训管理系统

2.2.1　预约管理

预约管理功能模块是由预约登记和回复预约两个子模块组成。预约登记功能主要是教师向管理员预约实训室的使用时间、使用班级、课程等；回复预约是管理员回复教师收到预约并确定的过程。

2.2.2　上课管理

上课管理功能模块是由上课登记和上课查询两个子模块组成。上课登记主要为教师提供上课记录的功能：教师进行上课登记，记录上课班级、上课课程、上课时间、上课内容等信息；同时教师可通过上课查询功能对历史上课记录进行查看。

2.2.3　考勤管理

考勤管理功能模块是由考勤登记和考勤查询两个子模块组成。考勤登记是给学生提供上课考勤签到，学生进行考勤登记时，选择上课数据即可签到；考勤查询给教师和学生提供考勤查询的功能。

2.2.4 实训管理

实训管理功能模块是由实训登记和成绩查询两个子模块组成。实训登记是教师对学生的实训效果进行记录；成绩查询为教师和学生提供成绩查询的功能。

2.2.5 基础信息管理

基础信息管理功能模块是由学生信息、教师信息、班级信息、课程信息和实训室信息五个子模块组成。管理员使用各模块对其信息进行维护，如使用学生信息模块对学生基础信息进行维护等。

2.3 数据库设计

完整的软件必须要有数据库的设计，它负责存储系统中所有数据及信息，因此，它对数据库的安全性要求较高[7]。本文设计的系统主要应用于贵州经贸职业技术学院实训管理工作中，存储的数据较多，这对数据库要求就比较高。

在对数据表设计时要注意一个关键问题，由于该系统要与贵州经贸职业技术学院学生教学管理信息系统进行数据交互，所以为了满足这个需求，设计数据库时在数据库字段按照贵州经贸职业技术学院学生学籍数据库中表字段的类型和长度来设计本系统数据库数据类型和字段长度[8]。

3 结语

通过认真研究与分析后开发了智能化实训管理系统，大大提高了实训管理工作的便利性、可操作性以及科学性，这使得实训室的各类资源管理得到了大大增强[9]。从课程的预约、实施、结束到实训室的各类管理，从学生到老师、实训管理员，都从功能和职责上进行了统一分工和协作。从一定程度上来说，实训室的智能化管理水平，体现了学校的管理水平以及科研水平。本系统的设计，将使得实训管理水平、实训资源的最大化利用以及学校的管理水平大大提高，从而更好地为学生和教职工服务[10]。

参考文献：

[1] 陈窕. 永康职校实训室设备管理系统的设计与实现[D]. 成都:电子科技大学,2012.

[2] 田黎莉. 西南工程学校实验室管理系统的设计与实现[D]. 成都:电子科技大学,2010.

[3] 林仙芝. 闽西职业技术学院考试管理系统 的研究与分析[D]. 昆明:云南大学,2015.

[4] 冯宇. 宁波工程学院排课管理系统功能及数据分析[J]. 科技视界,2016(4):165-166.

[5] 杨梅. 计算机系网络教学实训管理系统的研究与分析[D]. 昆明:云南大学,2015.

[6] 徐颖芳. 高校实训室建设与管理的探讨[J]. 世界华商经济年鉴·高校教育研究,2009(9):64.

[7] 曾惠萍. 软件学院实训管理系统的研究与分析[D]. 昆明:云南大学,2016.

[8] 罗平,胥光辉. 软件工程方法与实践[M]. 北京:机械工业出版社,2009.

[9] 傅峰. 基于移动平台的学生实训实习管理系统的设计[J]. 电子设计工程,2016,24(9):66-68.

[10] 邓宏伟. 基于Web的实训管理系统的设计与实现[J]. 电子技术与软件工程,2016,(7):50.

Research and design of intelligent training management system in colleges and universities

Guan Tong*

(Department of Information Engineering, Guizhou Vocational and Technical College of Economy and Trade, Guiyang 580001, China)

Abstract: With the increasing emphasis of the state on vocational education, it is an important goal of vocational education in colleges and universities to cultivate more high-quality skilled personnel, and the training room is an important place to cultivate students' hands-on ability and improve professional skills. With the increasing enrollment of vocational colleges, the traditional training management model can no longer meet the needs of the current work. At present, most of the training work is carried out manually and semi automatically, which not only has low work efficiency, but also is prone to some human errors. In order to reduce the burden of the staff in the training room, this paper, based on the current situation of the training management of Guizhou Vocational and Technical College of Economics and Trade, aims to use software means to realize the automation and informatization of the training management, improve the efficiency of the training management and the effect of the training teaching.

Keywords: training management; system design; research and design

文章编号: 1007-1423(2023)08-0117-04 DOI: 10.3969/j.issn.1007-1423.2023.08.020

基于模糊逻辑的踝足矫形器设计

蒲 文，华 伟*

（四川大学电子信息学院，成都 610065）

摘要： 中风、脑瘫患者因对下肢运动控制存在缺陷导致步态异常。设计一款气动踝足矫形器来预防足下垂等行动障碍。将踝关节位置和足底压力作为输入信号，采用模糊逻辑控制器(FLC)控制输出电压。监测和记录步行测试实验中控制器的输入和输出，结果表明基于模糊逻辑的踝足矫形器能在行走时提供辅助，有助于防止足下垂。

关键词： 踝足矫形器；足下垂；模糊逻辑控制

0 引言

瘫痪、中风会使患者运动控制存在缺陷导致步态异常，如足下垂是卒中(中风)患者十分常见的后果之一[1]。由于足下垂患者行走时，脚尖拖地或者过度屈膝导致脚尖离地，类似上楼梯时的状态，所以足下垂患者会表现出较短的步长、较长的站立期、较慢的步态速度、更有限的功能性运动和更高的跌倒风险。若下肢长期不进行活动，会导致下肢肌肉萎缩等问题。

目前，踝足矫形器(ankle-foot orthosis，AFO)是中风相关偏瘫患者改善步态的常用装置，例如2020年加拿大研究机构研制出由电机驱动的踝足矫形器可以实现踝关节的跖屈和背屈运动[2]，周聪[3]设计了一款由绳索驱动的踝关节康复机器，患者穿戴AFO后，在站立和行走时能更好地控制脚踝关节来辅助患者行走，有助于下肢血液循环，通过康复训练来恢复踝关节活动能力，使患者能重新回归正常步态，回归正常生活。

1 踝足矫形器系统设计

1.1 系统组成

基于模糊逻辑的踝足矫形器系统设计如图1所示，主要是由矫形鞋、控制单元、执行单元、数据采集单元构成。

矫形鞋采用超高分子聚乙烯(UHMWPE)材料制成，采用二连杆模型，与气动肌腱组合后实物图如图2所示。

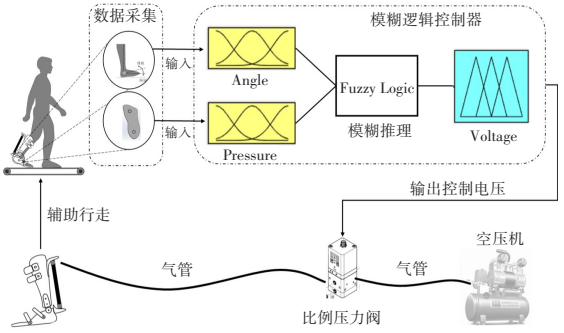


图 1 系统概览图

图 2 矫形鞋和气动肌腱组装图

系统所用的核心控制器是意法半导体公司开发的STM32F1系列微处理器产品。执行单元是由Festo公司生产的比例减压阀和气动肌腱构成，气动肌腱具有功率重量比高、柔韧性好、重量轻等优点，非常适合应用于医疗领域[4]。比

收稿日期： 2022-12-13 **修稿日期：** 2023-01-12

作者简介： 蒲文(1994—)，男，四川南充人，在读硕士研究生，研究方向为嵌入式系统设计；*通信作者：华伟(1967—)，男，山东聊城人，副教授，研究方向为电子与通信工程、等离子体，E-mail:hua23557@163.com

例减压阀能够把电压值转为气压值，从而控制气动肌腱充气和放气。数据采集单元主要由传感器构成，其中足底压力传感器用于采集脚后跟和脚前掌的压力值，脚踝位置传感器MPU6050用来采集踝关节俯仰角数值。

1.2 工作原理

人行走时，踝关节主要进行跖屈和背屈活动[5]，如图3所示。本文设计的踝足矫形器通过电压来控制比例减压阀开度，比例减压阀又称比例阀，其作为先导控制的调压阀，可以控制0～10 V的电压信号，实现气动肌肉进出气。具体功能如下：当比例阀显示气压增加，使气动肌腱充气，其轴向压缩、径向膨胀，气动肌腱长度缩短，矫形鞋呈现跖屈状；当比例阀显示气压减小，气动肌腱放气，长度由短变长，为穿戴者提供一个背屈的支撑力。因此矫形鞋可以弥补病人缺失或减少的肌腱力量，达到改善步行的目的。

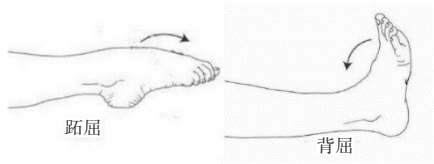


图3 跖屈与背屈

目前已有控制策略是根据脚踝所处位置来输出不同的电压，从而控制矫形鞋跖屈和背屈。但该控制策略只考虑脚踝传感器数据，对于步态周期的判断会因人而异，适合定制开发。模糊逻辑控制属于智能控制的一种，对于被控对象的控制不依赖精确的数学表达式来完成[6]，模糊逻辑控制被应用于各种系统中[7-8]，都取得了不错的效果。针对踝足矫形器的主动控制策略非线性、时变的特点，本文提出一种模糊逻辑控制策略，以足底压力和踝关节位置作为控制器输入，经过模糊规则计算后，由控制器输出确定的电压。

2 模糊逻辑控制器设计

模糊逻辑控制器的控制目标是控制气动肌腱充气或放气，使得矫形鞋配合穿戴者行走，并且在摆动提供支撑，预防足下垂情况的发生。本系统尽可能地优化隶属度函数和规则的数量，以使得系统简单有效。

2.1 输入、输出量的模糊子集与隶属度函数的选取

本文设计的模糊逻辑控制器为双输入、单输出的控制器，输入为踝关节角度(ankle)和足底压力(pressure)，输出为电压值。

在穿戴矫形器行走的过程中，位于踝关节处的位置传感器的论域为[-25, 25]。将其所在论域的模糊子集分为3个，分别为BWD(向后)、STD(站立)、FWD(向前)。其隶属度函数如图4所示。选择广义钟型隶属函数，是因为该函数比三角形或梯形等隶属函数更平滑。

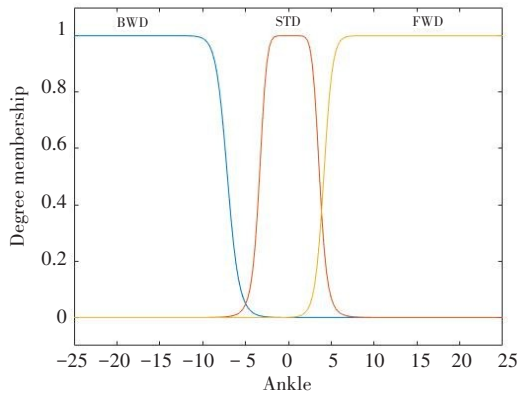


图4 踝关节角度(Ankle)隶属度函数

足底压力的论域为[0, 3]，输入模糊逻辑控制器的足底压力数据经过处理，能判断出压力点受力情况。因此将其论域的模糊子集分为4个，分别是ZZ(前后脚掌都无压力)，ZO(前无后有)，OZ(前有后无)，OO(前有后有)。其隶属度函数如图5所示。

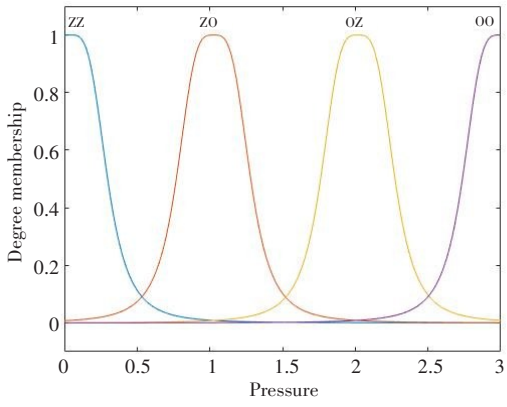


图5 足底压力(Pressure)隶属度函数

输出电压的论域是[0, 10]，将其论域的模糊子集分为5个，分别为L(低)，ML(中低)，M(中)，MH(中高)，H(高)。其隶属度函数如图6所示。

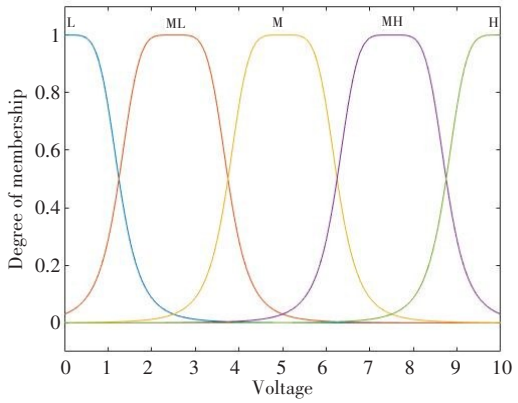


图6　输出电压(Voltage)隶属度函数

2.2　模糊逻辑控制规则

本系统想要达到的目的是，在行走时，当踝足处于正常步态时，矫形器协同行走；当踝足处于不正常步态，如足下垂时，矫形器将限制过度跖屈，保持踝足所处位置以进入下一步态周期。根据先验知识，设置了6条模糊逻辑控制规则，如表1所示。

表1　模糊逻辑控制规则

踝关节位置(Ankle)	足底压力(Pressure)	输出电压(Voltage)
BWD	ZZ	MH
BWD	ZO	L
BWD	OZ	H
STD	ZZ	M
FWD	ZZ	ML
NULL	OO	M

2.3　输出模糊量的清晰化

利用Matlab软件中的模糊系统设计器对模糊逻辑控制器的行为进行了仿真。在模糊系统设计器中插入隶属函数和规则设计后，图7所示的模糊表面查看器的形式观察输入和输出之间的关系，显示了输入和输出的所有可能的组合。

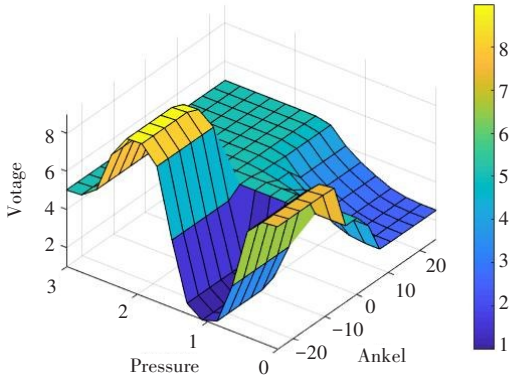


图7　输出电压Voltage随输入变化的曲面

当足底压力为3(前后脚掌都有压力)时，此时电压输出为5左右，气动肌腱充气量适中，矫形器保持站立姿态；当足底压力从3到1，角度为负数时，代表从站立阶段到摆动阶段，此时输出电压由大变小，刚好配合穿戴者行走；当摆动阶段结束时，输出电压达到最小，矫形器呈背屈姿态，使脚后跟先着地，刚好预防足下垂，避免了脚尖先着地的情况。

3　系统调试

利用本文设计的模糊逻辑控制器，移植到系统后，让试穿者穿戴踝足矫形器进行测试。记录测试过程中的传感器数据和输出电压。如图8所示。当踝关节角度在0附近时，穿戴者处于站立姿态，此时输出电压为5 V左右，气动肌腱充气使得矫形鞋也处于站立位置。当踝关节角度持续下降到−20甚至−30时，穿戴者脚尖蹬地，即将进入摆动阶段，此时输出电压增大，气动肌腱持续缩短，使得矫形鞋呈现跖屈状，完全配合穿戴者行走意图。当踝关节角度逐渐增加时，穿戴者处于摆动阶段，角度增加达到极限时，穿戴者理应脚后跟着地，为进入站立阶段做准备，从图8可见，此时输出电压达到该步态周期最小值，表明气动肌腱放气，其长度增加，使得矫形鞋呈背屈状，从而保证穿戴者脚后跟先着地。当穿戴者再次恢复站立姿态时，输出电压又变成5 V左右。整个步态周期内，矫形鞋先是配合穿戴者跖屈意图，然后从摆动阶段到站立阶段前，又使得矫形鞋背屈状，防止穿戴者脚尖拍打地面，避免了足下垂情况的发生。

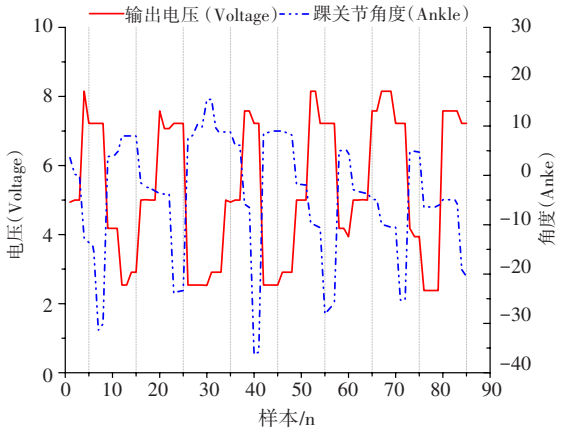


图 8 行走测试中的踝关节角度和电压

4 结语

本文将模糊逻辑算法应用于主动型踝足矫形器的控制策略中，适应性较强，以此来辅助足下垂、肌无力等踝足损伤的患者进行训练康复。经过实际测试，表明该装置能够有效地预防脚尖拍打地面，减少了患者摔倒的可能性，规避了一些风险，有助于提高患者主动步行的能力。

参考文献：

[1] POURHOSEINGHOLI E, SAEEDI H. Role of the newly designed ankle foot orthosis on balance related parametersin drop foot post stroke patients[J]. Journal of Bodywork and Movement Therapies, 2021, 26: 501-504.

[2] GHOSH S, ROBSON N P, MCCARTHY J M. Kinematic design and evaluation of a six-bar knee-ankle-foot orthosis[J]. Journal of Engineering and Science in Medical Diagnostics and Therapy, 2020, 3(2): 21.

[3] 周聪. 可穿戴式绳驱动踝关节康复机器人设计研究[D]. 成都: 电子科技大学, 2020.

[4] 周彬滨, 邹任玲. 气动人工肌肉在康复器械中的应用现状[J]. 中国康复理论与实践, 2020, 26(4): 463-466.

[5] 任媛媛, 蒲文. 基于主被动的脚踝康复系统[J]. 现代计算机, 2021, 27(28): 101-105.

[6] CHERROUN L, NADOUR M, KOUZOU A. Type-1 and type-2 fuzzy logic controllers for autonomous robotic motion[C]//Proceedings of the 2019 3rd International Conference on Applied Automation and Industrial Diagnostics, Elazig, Turkey, 2019.

[7] 刘晋霞, 王莉, 刘宗锋. 四驱电动轮汽车模糊逻辑控制的再生制动系统[J]. 机械设计与制造, 2021(12): 164-168.

[8] 黄森, 周安妮, 何镇琦, 等. 基于天气预报的智能灌溉模糊控制系统设计[J]. 物联网技术, 2021, 11(12): 67-70.

Design of ankle-foot orthosis based on fuzzy logic

Pu Wen, Hua Wei*

(College of Electronic Information, Sichuan University, Chengdu 610065, China)

Abstract: Patients with stroke and cerebral palsy have abnormal gait due to deficiencies in motor control of the lower extremities, and a pneumatic ankle-foot orthosis is designed to prevent mobility disorders such as foot drop. Based on the ankle joint position and plantar pressure as input signals, a fuzzy logic controller(FLC) was used to control the output voltage. The inputs and outputs of the controller were monitored and recorded during walking test experiments, and the results showed that the fuzzy logic-based ankle-foot orthosis can provide assistance during walking and help prevent foot drop.

Keywords: ankle-foot orthosis; foot drop; fuzzy logic controller