RCCSE中文OA学术期刊类核心期刊
ISSN: 2161-8801 Volume 7, Number 4, April 2017

Hans汉斯

计算机科学与应用

Computer Science and Application

Ji Suan Ji Ke Xue Yu Ying Yong

2017年4月7卷4期

ISSN: 2161-8801

www.hanspub.org/journal/csa

Editorial Board
编委名单

ISSN: 2161-8801 (Print) ISSN: 2161-881X (Online)
http://www.hanspub.org/journal/csa

Computer Science and Application, Vol.7, No.4, 281-406
计算机科学与应用, 7 卷 4 期
Published Online April 2017 in Hans (http://www.hanspub.org/journal/csa)

Hans汉斯

TABLE OF CONTENTS
目　　录

Computer Science and Application, Vol.7, No.4, 281-406
计算机科学与应用, 7 卷 4 期
Published Online April 2017 in Hans (http://www.hanspub.org/journal/csa)

期刊信息
期刊中文名称：《计算机科学与应用》
期刊英文名称：**Computer Science and Application**
期刊缩写：**CSA**
出刊周期：月刊
语　　种：中文

出版机构：汉斯出版社(Hans Publishers, www.hanspub.org)
编辑单位：《计算机科学与应用》编辑部
主　　编：戴元顺，电子科技大学教授，千人计划入选者
　　黄廷祝，电子科技大学教授
　　叶培新，南开大学教授
网　　址：http://www.hanspub.org/journal/csa

订阅信息
通过中国教育图书进出口有限公司订购
订阅邮箱：sub@hanspub.org
订阅价格：240 美元每年

广告服务
联系邮箱：adv@hanspub.org

Computer Science and Application 计算机科学与应用, 2017, 7(4), 281-290
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74035

Hans汉斯

Extraction and Analysis of Location-Based Information in Android Forensics

Yi Zhang, Zelin Zhang, Yan Chen, Rui Mao, Lei Xu

Jiangsu Police Institute, Nanjing Jiangsu
Email: 464604236@qq.com

Received: Mar. 28th, 2017; accepted: Apr. 11th, 2017; published: Apr. 18th, 2017

Abstract

Android smartphones generate a lot of location-based information. In mobile forensics, location information, which effectively reflects users' behavior trace, shows very high investigative value and evidence value. This paper builds a forensics process model to extract, analyze and display location-based information in order to provide new ideas for case investigation.

Keywords

Android Forensics, Location-Based Information, Data Analysis, Data Visualization

Android取证中地理位置信息提取与分析研究

张　祎，张泽林，陈　岩，冒　睿，徐　雷

江苏警官学院，江苏　南京
Email: 464604236@qq.com

收稿日期：2017年3月28日；录用日期：2017年4月11日；发布日期：2017年4月18日

摘　要

Android手机在日常使用中产生了大量的地理位置信息。在手机取证中，这些信息能够有效反映用户的行为轨迹，具有重要的证据价值，同时也能为案件侦查提供线索。构建Android取证中地理位置信息提取分析的实现模型，探讨对Android取证中地理位置信息的提取、分析及证据可视化展示，为案件侦查取证提供新的思路。

文章引用: 张祎, 张泽林, 陈岩, 冒睿, 徐雷. Android 取证中地理位置信息提取与分析研究[J]. 计算机科学与应用, 2017, 7(4): 281-290. https://doi.org/10.12677/csa.2017.74035

关键词

Android取证，地理位置信息，数据分析，数据可视化

1. 引言

截止 2016 年第二季度，Android 手机的市场占比已经达到了 86.2% [1]。伴随着移动通讯技术的快速发展，可以预见手机取证技术，尤其是针对高普及率的 Android 系统的取证技术，将成为打击各类犯罪的重要技术手段。

Android 手机用户在日常使用中产生了大量的地理位置信息。在手机取证调查中，这些地理位置信息能够有效反映用户的行为轨迹，具有重要的证据价值，同时也为案件侦破提供重要线索。目前，公安机关在案件的侦破和审理中主要针对涉案人员移动终端设备中的通讯录、通话记录、聊天记录、照片等用户信息数据进行取证调查，对上述大量的地理位置信息并没有广泛利用。

本文围绕 Android 平台的地理位置信息，探讨在 Android 取证中的数据提取与分析，并辅助后续侦查和证据固定，以试图构建电子调查取证中地理位置信息提取分析的实现模型，为调查取证提供新的思路。全文分为四个部分：第一部分概述相关概念与实现模型，第二部分介绍各类地理位置信息及其提取，第三部分概述对获取信息的数据分析，第四部分介绍地理位置信息可视化展示的实现途径，第五部分总结已有成果并对未来研究提出展望。

2. 相关概念

2.1. 地理位置信息

本文所称地理位置信息，是指以各种形式表现且能够反映特定对象空间位置(通常还附带特定时刻或时间范围)的数据信息，不局限于 GPS 等定位系统产生的定位数据，还包括诸如表述地点位置信息的文本数据等其他各类形式。

2.2. 数据可视化

数据可视化(Data visualization)是关于数据之视觉表现形式的研究。数据可视化旨在借助图形化手段，清晰有效地传达与沟通信息[2]。目前可视化技术在商业金融、科学计算、气象海事、网络安全、犯罪学等领域得到了广泛而有效的运用，已成为各领域揭示数据集中数据之间的关系和背后隐匿信息的基本工具[3]。数据可视化虽然手段形式多样但实质都是将数据多维属性及其相互关系以二维或三维的图形图像进行展示的过程。

2.3. 数据取证分析的流程

根据取证实践将数据取证分析建立图 1 所示的流程模型。

第一部分是数据获取。在电子取证中，涉案电子数据主要来源于对涉案电子设备中数据的搜索和恢复。另外，在特定案件中，根据国家法律法规，调查人员有权依法从第三方机构调取涉案电子数据。

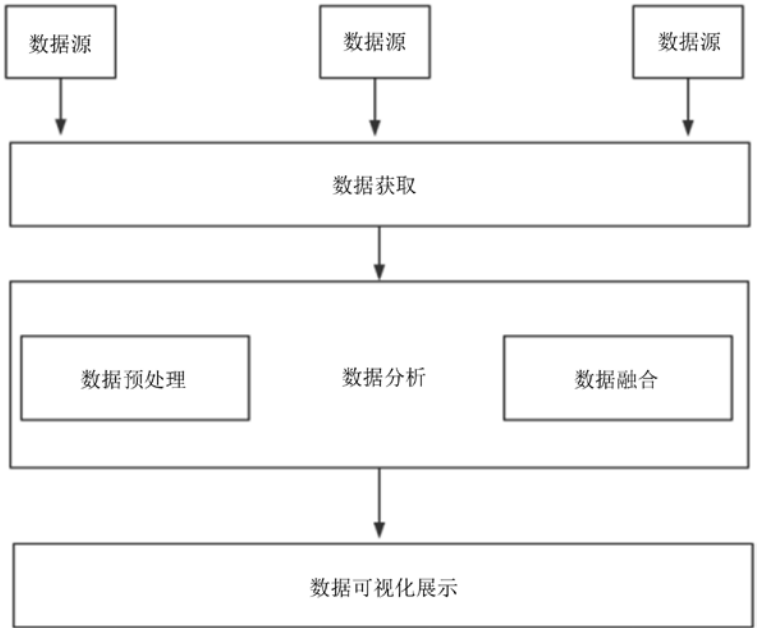


Figure 1. Procedure of forensics and data analysis
图 1. 数据取证分析流程

第二部分是数据分析。对于获取的原始数据往往并不能完全满足数据分析的要求，需要通过预处理对原始数据进行规整并集成。对于经过预处理的数据分析方法众多，本文主要介绍数据融合达到数据分析的目的。

第三部分是数据可视化。数据可视化要求根据数据特点和具体需求，综合利用多种手段工具，将数据分析结果进行图形化展示，以达到固定案件证据和挖掘涉案线索的目的。

3. Android 取证中地理位置信息及其提取

Android 手机用户在日常使用中产生了大量的地理位置信息，主要包括移动通信的基站信息，移动操作系统记录的位置信息，照片视频元数据中的附加信息，各类地图、打车、旅游、社交等移动应用的定位服务信息，WiFi 连接的位置信息等。

这些地理位置信息大部分都是由系统程序向用户声明且经用户允许下采集，存储于本地介质或者上传服务器云端。因此对地理位置信息的取证技术同时包含手机取证技术和网络取证技术。

Android 系统版本更新速度快，同时各个手机厂商对 Android 原生系统进行了定制，因此 Android 手机的系统平台具有多样性。本文以三星 S5 (SM-G9008V)作为实验平台，其系统环境为 Android 5.0 并已取得 ROOT 权限。数据的提取方法基于实际的数据提取，其方法对于基于镜像的提取也同样适用。

3.1. WiFi 信息

大量移动终端通过私人或者公共 WiFi 网络来访问互联网。Android 手机会自动存储 WiFi 网络的连接记录和网络的基本属性(例如 SSID 和 MAC 地址)。同时，WiFi 网络的无线接入点地理位置一般固定，根据 WiFi 接入信息可以得知关联的地理位置信息。

Android 系统使用一个修改版 wpa_supplicant 作为 daemon 来控制 WiFi，其配置文件/wpa_supplicant.conf 存储位置在/data/misc/Wi-Fi/目录下。该配置文件中定义的部分信息数据结构如表 1 所示。

3.2. 照片元数据

可交换图像文件格式(EXIF)通常附加于 JPEG、TIFF、RIFF 等文件之中，用于记录照片的属性信息和拍摄数据。其所记录的参数通常被称作照片元数据。照片元数据内容丰富，其中就包括 GPS 定位数据形式的地理位置信息。Android 智能手机在拍摄时，如果打开定位服务及相关选项会自动在 EXIF 中加入 GPS 定位数据。这一类地理位置信息配合拍摄时间较精确地反映了照片的拍摄时空信息。

3.2.1. EXIF 数据结构

对于 Android 智能手机默认存储照片格式 JPEG 文件，其以十六进制数值 FF D8 作为 SOI 标志(图像开始)，并以 FF D9 作为 EOI(图像结束)。文件头部有一系列“0xFF？？”格式的十六进制数值，称为 JPEG 标识，用来标记 JPEG 文件的信息段。EXIF 利用从“0xFFE0”-“0xFFD9”之间的应用标记(APPn)记录照片元数据[4]。EXIF 的官方文档关于元数据标签的定义较复杂，这里只列举常用的 GPSInfo 组件，见表 2。

3.2.2. 提取方法

针对 EXIF 的数据结构，我们使用 C#语言编写提取 JPEG 文件元数据的工具。以三星 SM-G9008V 手机拍摄的一张照片为例，读取其中 EXIF 数据。最后使用导出功能得到 html 格式的对应格式化摘要。摘要内容包括原照片浏览图、GPS 数据和 GPS 时间戳等，见图 2。

3.3. 即时通讯类应用

3.3.1. 地理位置信息存储形式

移动端常见即时通讯类应用有微信、手机 QQ、飞信、陌陌和米聊等，本文以较为广泛使用的微信为例。微信中的地理位置信息主要由用户聊天时地理位置分享产生，并保存在聊天记录数据中，Android 版聊天记录经加密存储在\data\data\com.tencent.mm\MicroMsg 路径下的 EnMicroMsg.db 数据库文件中，其中表 message 存储聊天内容，其表结构见表 3。

3.3.2. 提取方法

对于 Android 版微信，可以通过手机取证系统直接解密微信聊天记录，进一步获得地理位置信息。同时也可以借助免费的第三方软件实现微信聊天记录的找回操作。

Table 1. Data structure of Wi-Fi data
表 1. Wi-Fi 信息数据结构

字段名	ssid	Key_ngmt	autojoin	priority	psk
中文说明	接入点名称	认证密钥管理协议	自动加入	优先级	密钥

Table 2. Common GPSInfo components
表 2. 常见 GPSInfo 组件信息

字段名	标签名称	数据类型
GPSVersionID	GPS tag version	BYTE
GPSLatitudeRef	North or South Latitude	ASCII
GPSLatitude	Latitude	RATIONAL
GPSLongitudeRef	East or West Longitude	ASCII
GPSLongitude	Longitude	RATIONAL
GPSTimeStamp	GPS time	RATIONAL

Table 3. List structure of message (part)
表 3. message 表结构（部分）

字段	字段类型	说明
msgID	INTEGER	对话 ID/序号
msgSvrId	INTEGER	msg 服务号
type	INT	类型
status	INT	状态
isSend	INT	已发送
isShowTimer	INTEGER	计时器
createTime	INTEGER	创建时间
talker	TEXT	发起聊天者账号
content	TEXT	内容
imgPath	TEXT	图片路径
reserved	TEXT	存储位置
lvbuffer	BLOB	二进制标记
transContent	TEXT	传输内容
talkerId	INTEGER	发起聊天者 ID/序号

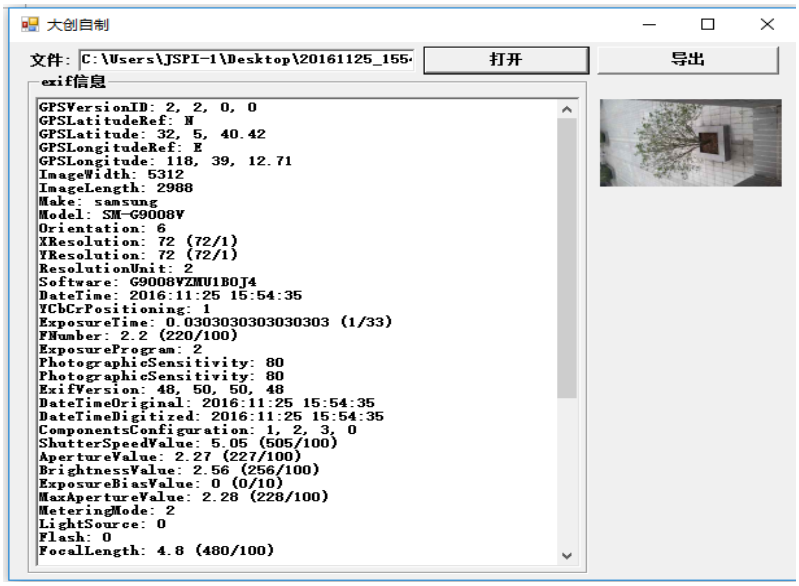


Figure 2. Program to read EXIF
图 2. EXIF 信息读取程序

如下是对微信(版本号：6.3.31) EnMicroMsg.db 数据库解密并提取位置分享信息的实验。首先将手机 IMEI(国际移动设备身份码)和微信 uin (微信用户信息识别码)字符串连接(见图 3)，其 MD5 值前七位即为数据库解密密码。利用工具 SQLCipher.exe 打开 EnMicroMsg.db 数据库文件，输入密码，即可成功打开文件。读取表“message”中记录的聊天记录，使用关键词“location” 执行 SQL 查询语句，从而获取所需的地理位置信息(见图 4)。

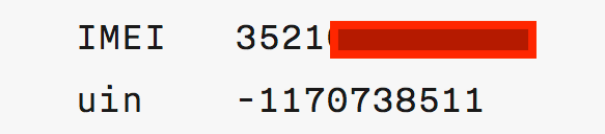


Figure 3. IMEI and uin
图 3. IMEI 与 uin

```
<?xml version="1.0"?>
<msg>
        <location x="32.092300" y="118.658653"
 scale="16" label="浦口区江苏警官学院(浦口校区)旁"
 maptype="0" poiname="[位置]" />
</msg>
```

Figure 4. Shared location data
图 4. 分享地理位置数据

3.4. 导航地图

经济社会高速发展的今天，复杂的路况已经成为户外出行的首要问题，由此百度地图、高德地图、腾讯地图等手机地图导航软件应运而生。本文以手机百度地图软件中地理位置信息的提取为例。

3.4.1. 相关信息存储形式

百度地图应用中地理位置信息主要包含用户定位记录、导航记录、步行记录等，这些数据存储在/data/data/com.baidu.BaiduMap/databases 路径下。在众多生成的数据文件中，名为“tracks.db”的数据库文件中各表详细记录使用者相应的地理位置信息，“tracks”数据库结构见表 4。

以存储用户定位记录的表 Track_Location 为例，其表结构如表 5 所示。

3.4.2. 提取方法

如图 5 用户定位记录所示，访问 data/data/com.baidu.BaiduMap/databases 目录，利用数据库查看软件 SQLite Expert Professional 即可查看数据库 tracks.db，并导出数据。

3.5. 网约车应用

网约车软件的诞生改变了传统打车市场格局。这里以滴滴出行为例进行相关地理位置信息的提取。

通常用户在使用应用时，会授权软件定位当前的地理位置，滴滴再根据司机的实时位置，向附近的车辆发送订单请求，一经接受即订单生效，以此来快速提供出行服务。这些位置信息包括常用地址信息(家和公司地址信息)，起点和终点信息。

由于滴滴出行应用将大量订单和用户信息数据存储于服务器，因此对此应用的数据提取需要可视化取证和介质取证相结合。滴滴出行应用的一部分订单信息和用户个人信息以数据库文件的形式存储在 data/data/com.sdu.didi.psnger/databases/路径下。

3.5.1. 数据库结构

在 datebases 中有许多数据库文件和 journal 文件，其中 im_database_281475098446424.db 中存储着有关订单信息，如图 6 所示。图中 Android_metadata 为 Android 数据库原文件数据，im_message_table 为会话信息，im_user_table 为用户相关的额表信息。

图 6 展示提取 im_user_table 表得到的用户与滴滴司机活动信息，然而具体位置信息并未包含于本地数据库中，而是存储在应用服务器上。因此需要使用可视化取证手段提取更多详细信息，见图 7。

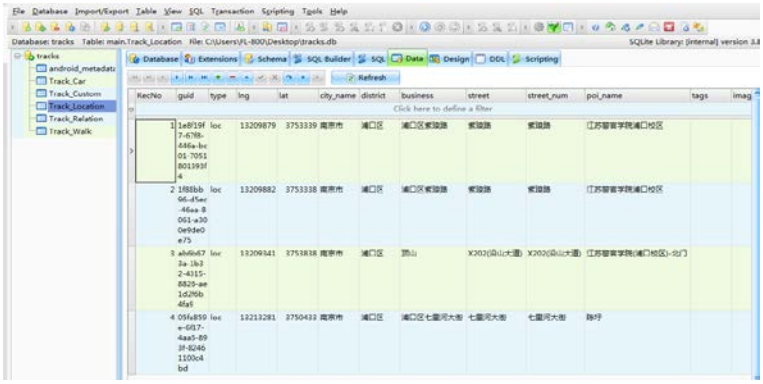


Figure 5. Record of user's tracks
图 5. 用户定位记录

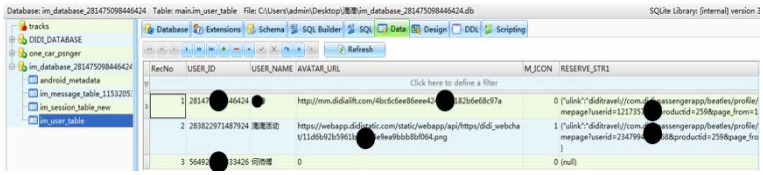


Figure 6. Orders of Didi Taxi
图 6. "滴滴出行"订单信息

Figure 7. Activities of users and drivers
图 7. 机主与司机进行的滴滴活动

Table 4. Database structure of tracks
表 4. tracks 数据库结构

表名称	说明
android_metadata	Android 元数据
Track_Car	用户导航数据
Track_Custum	用户常用地址
Track_Location	用户定位记录
Track_Relation	用户线路查询记录

Table 5. List structure of Track_Location
表 5. Track_Location 详细表结构

字段	字段类型	说明
guid	TEXT	数据库唯一标识符
type	TEXT	表的类型
lng	REAL	经度
lat	REAL	纬度
city_name	TEXT NOT NULL	城市名称
district	TEXT	行政区名称

从图 7 中可读出常用地址信息与订单的起终点信息，其中起终点信息以图针的形式在地图上标记出来，能更好地观察出行车路线与方位信息。

4. 数据处理

4.1. 数据预处理

地理位置信息数据来源途径较多，并且不同途径所获取的数据形式与结构不同，因此需要通过预处理对原始数据进行规整并集成统一。本研究提出可以使用基于 Javascript 对象表示法的地理空间信息数据交换格式——GeoJSON 格式进行统一编码。以 GeoJSON 数据结构为模板将各类信息调整为规范的数据结构，以规范地表示诸如点、线、面、多点、多线、多面等多种几何类型的地理位置信息，同时添加自定义属性。将地理位置信息规范为 GeoJSON 格式存储后，即可使用 Python 和 Javascript 等编程语言对其快捷读取与操作。

4.2. 数据融合

将来源不同的地理位置信息进行数据规整与集成，利于统一操作，从而进行数据融合和合成分析。数据融合的目的是合并或者综合多个数据集，成为一个完整的性能更好的数据集[5]。时间信息和空间信息是地理位置信息的两大基本要素。从这两要素出发，可以将地理位置信息的数据融合分为基于时间的数据融合和基于空间的数据融合，前者能体现信息对象某段时间的移动轨迹，后者能够反映某个位置或某个位置范围信息对象的惯性行为。

5. 数据可视化

5.1. 电子取证中的数据可视化

在电子数据取证中，数据可视化的实现目前主要有两种。第一种是使用商业软件产品，如 IBM i2 Analyst’s Notebook、美亚柏科可视化数据智能分析系统、智器云火眼金睛等。此类软件直接面向警务工作开发，操作简单，但价格不菲。第二种是利用开源工具和脚本语言实现，如 R 语言和 Python 语言及相关包，针对实际工作需求，开发扩展性较强的可视化工具，节约了购买商业工具授权的成本，然而具体实现的理论和技术基础要求高。可以说数据可视化的实现途径各有优势，在具体取证工作中应该根据实际需求，综合利用。

5.2. 地理位置信息可视化的实现途径

5.2.1. 基于百度地图 API 的展示

百度地图 URI API 是为开发者提供直接调用百度地图产品以满足特定业务场景下应用需求的程序接

口。按照接口规范构造一条标准的 URI，例如“http://api.map.baidu.com/geocoder?location=39.990912172420714,116.327158634486 07&coord_type=gcj02&output=html&src=test”(参数说明见表 6)，将前期提取的地理位置信息传递给百度地图-URI API 反地址解析接口，经过逆地理编码后，定位信息即可以标注形式展示出来，如图 8 所示。

5.2.2. 基于 Leaflet.js 的展示

Leaflet.js 是一个用于创建交互式地图的开源 Javascript 库，其允许载入 GeoJSON 格式的地理位置信息并创建带有特定标识图案的地图，再结合 HTML 便能以网页的形式进行可视化展示。Leaflet.js 实现的可视化图案要素有图钉，折线轨迹，圆圈和弹窗等。基于 Leaflet.js 能够有效地表达位置痕迹和移动轨迹。为了方便高效地完成大量地理位置信息输出，使用 Python folium 库读取 GeoJSON 规范数据并输出 HTML 文件完成位置信息可视化展示。图 9 所示为以简化实验数据为例的实现代码和输出效果。

6. 结语

在案件侦查中，地理位置信息是重要的电子证据之一，能够为案件分析、确定相关人员的行动轨迹

Figure 8. Visual display via Baidu Map
图 8. 百度地图可视化展示

```
#coding=utf-8
import folium
import json

map = folium.Map(location=[32.093,118.652],zoom_start=12)
marker_cluster = folium.MarkerCluster().add_to(map)

folium.CircleMarker([32.093,118.652],      #绘制已知地点范围
                    radius=500,
                    popup='江苏警官学院',
                    color='#3186cc',
                    fill_color='#3186cc',
                   ).add_to(map)

folium.GeoJson(open('test.json'),    #读取GeoJOSN文件
               name='json'
              ).add_to(map)

map.save('./test.html')
```

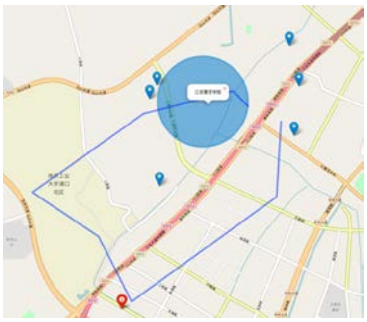


Figure 9. The code and examples of visual display
图 9. 实现代码与可视化效果示例

Table 6. Parameters of Baidu Map API
表 6. 百度地图 API 参数说明

参数名称	参数说明
location	lat<纬度>, lng<经度>
output	表示输出类型，web 上必须指定为 html 才能展现地图产品结果
coord_type	坐标类型，可选参数
zoom	展现地图的级别，默认为视觉最优级别
src	App 名称

提供帮助。本文构建 Android 取证中地理位置信息提取与分析的实现模型，通过数据可视化技术进行证据展示，为 Android 平台电子数据取证提供新的方法与思路。

另一方面，面对实际工作中海量复杂的数据，传统的数据分析手段已经捉襟见肘。因此，将机器学习和数据挖掘等技术运用于电子数据取证，实现对电子数据进一步有效和深度的分析，通过实战运用探讨手机取证获得的地理位置信息的应用价值，是未来研究和实践的方向。

参考文献 (References)

[1] Statista: 2011 年-2016 年全球智能手机出货量对比. http://www.199it.com/archives/509159.html

[2] Dunleavy, D. (2015) Data Visualization and Infographics. *Visual Communication Quarterly*, **22**, 68-68. https://doi.org/10.1080/15551393.2015.1029070

[3] Feng, Y.D., Wang, G.P. and Dong, S.H. (2001) Information Visualization. *Journal of Engineering Graphics*, 324-329.

[4] CIPA. DC-008-2016. Exchangeable Image File Format for Digital Still Cameras: EXIF Version 2.3.1[S]. CIPA, 2016.

[5] 魏彦飞, 等. 一种新的数据融合方法——广义融合法[J]. 天文研究与技术, 2016(3): 318-325.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 291-300
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74036

Hans汉斯

Key Technology of IQA Robot Based on Telecom Scene

Yaofeng Tu

Cloud Computing Institute, ZTE Corporation, Nanjing Jiangsu
Email: tu.yaofeng@zte.com.cn

Received: Apr. 1st, 2017; accepted: Apr. 14th, 2017; published: Apr. 19th, 2017

Abstract

Intelligent question answering system is a new type of information interaction for natural language understanding. With the development of intelligent question answering system, it will bring new human-computer interaction mode and new business pattern. Intelligent question answering system involves many fields such as Natural Language Processing, knowledge management, intelligent dialogue and so on. In this paper, take the scene of telecom service for instance, we discuss that the architecture of Intelligent Question Answering System based on natural language understanding is put forward, and the related technologies are analyzed deeply.

Keywords

Intelligent Question Answering, Natural Language Understanding, Machine Learning, Deep Earning

基于电信业务场景的智能问答机器人关键技术

屠要峰

中兴通讯，云计算研究院，江苏 南京
Email: tu.yaofeng@zte.com.cn

收稿日期：2017年4月1日；录用日期：2017年4月14日；发布日期：2017年4月19日

摘　要

智能问答系统[1]是一种针对自然语言理解的新型的信息交互方式。它的发展将带来新的人机交互模式，带来新的业务形态。智能问答系统涉及自然语言处理、知识管理、智能对话等多领域技术，本文以电信业务场景为例，提出了基于自然语言理解的智能问答系统架构，并对相关关键技术进行了深入分析。

文章引用: 屠要峰. 基于电信业务场景的智能问答机器人关键技术[J]. 计算机科学与应用, 2017, 7(4): 291-300.
https://doi.org/10.12677/csa.2017.74036

关键词

智能问答，自然语言理解，机器学习，深度学习

1. 引言

传统的客服中心以人工服务的呼叫中心为主，但随着移动互联网与智能手机的不断普及，社交渠道多元化和应用软件功能的不断丰富，传统客服面临着急剧增加的服务需求和更为碎片、多元化的客户服务场景，这导致传统客服中心陷入到现实的多方困境：第一，人工客服受时间影响，难以保证 7 * 24 小时全天候服务。第二，人工客服业务操作繁琐，响应速度和服务质量的一致性难以保证，容易影响客户满意度。第三，业务知识的更新速度加快，增加了对服务人员的培训成本。第四，为满足增大的业务量和提高并发度，需要增加服务人员或者加班时间，导致成本高昂，而且近年来人力成本不断的上升。同时，服务信息难以保存，难以进行及时有效的分析，导致知识无法共享，服务灵活性不足。

2. 智能问答系统发展机遇

艾媒咨询在 2015 年发布的报告显示：中国客服市场整体规模超过千亿人民币[2]。但在庞大的市场规模下，难掩的是多数传统企业客服部门正在陷入上述之高成本、低效率、需求碎片化、满意度低、订单转化率低等矛盾。传统客服面临的现实困境和巨大的市场容量，迫切需要智能化的客服产品来突破发展瓶颈。如何构建可商用的智能问答系统是运营商及客服厂家亟需解决的问题。

当前智能机器人技术不断发展和成熟，智能机器人被应用于金融、财务、客服工作等领域，其中，智能机器人在客服工作中的应用效果最为显著。它通过自动客服、智能营销、内容导航、智能语音控制等功能提高了企业客服服务水平[3]。

智能问答应用在客服工作中有着显而易见的优势。一是提高用户感知，为企业在线客服、新媒体客服等提供统一智能的自助服务支撑，减少了用户问题得到解决的难度和复杂度；二是提升服务效率，缩短咨询处理时限，分流传统人工客服压力，节省服务成本(据统计：智能机器人投入是人工座席成本的 10%)；三是收集用户诉求和行为数据，支撑产品迭代优化[3]。

目前国际上的智能问答技术主要采用检索技术、知识网络、深度学习这三大技术，代表性平台是苹果的 Siri、谷歌的 GoogleNow 和微软的 Cortana。国内的智能应答技术发展较晚，这和中文的语法、语义复杂性等多种因素有关,目前主要是以人工模板和智能检索技术为主，典型代表有小 i 机器人、百度度秘等。

为了促进传统客服形态向自动化、智能化、人性化、多渠道的方向演进，更好的支撑电信域的业务发展，推动智能客服在电信领域落地，本项目根据电信业务场景的特性与需求，提出了一种基于自然语言的交互模式，可通过引导式应答或者反问使得问题加以确认，分领域建立领域本体的知识库和问题应答库，采用推理技术进行问题的自动识别与应答，形成基于语义的智能问答机器人 IQA(下文简称 IQA)。

本文介绍了 IQA 系统的设计思想及架构，并对其中的关键技术进行了分析。IQA 已在电信、移动现网多个局点运营和实证，能快速准确识别用户意图，自组答案，大幅提升客服工作效率，在实践中检验

了技术路线的成效，并建立了丰富的电信领域语料库。

3. 智能问答机器人 IQA 设计思想

问答系统的目标是给定一个问题，能够得到简短、精确的答案。

IQA 是基于自然语言理解技术和知识图谱技术，并配合使用语音识别(ASR)、语音合成(TTS)等智能人机交互技术，通过微信、APP、网页、短信、电话等渠道，以文字、语音等方式提供智能问答交互服务的信息服务系统。

图 1 展示了 IQA 的设计和架构图，该系统分为应用层、接入层、分析层和数据层。各层之间相对独立，耦合度小，易于扩展。

智能应答系统首先要有数据，数据可以来自于互联网爬取，也可以是现有的知识库，或者特定的语料库，这就涉及到数据的来源、获取、挖掘、存储等设计。

- 数据源：大致可以分为三类：1、百度知道等社区问答对；2、电信领域内专业数据及网站语料；3、第三方提供的数据接口，比如天气、笑话等。
- 数据获取：针对以上三类数据源，对应的获取方式可以是垂直爬虫爬取、人工维护录入、数据厂家提供，以及第三方开放平台提供，比如中国气象网等。
- 数据挖掘：数据挖掘就是对所获取的数据进行挖掘成有用的信息或结构化信息，按一定的结构和规则来组织有用的信息和知识，最终形成各种语料库，比如问答对 FAQ、聊天对话库、各种分词、领域词等词典，以及通用和行业等知识库，以便于问答直接使用或进一步处理。
- 数据存储：存储方式至少包括四类：1、像各种词库等，可直接文本形式存储，为了管理方便也可以存在数据库中；2、对于问答对、训练数据等可采用关系数据库进行存储；3、对于知识库，可以采

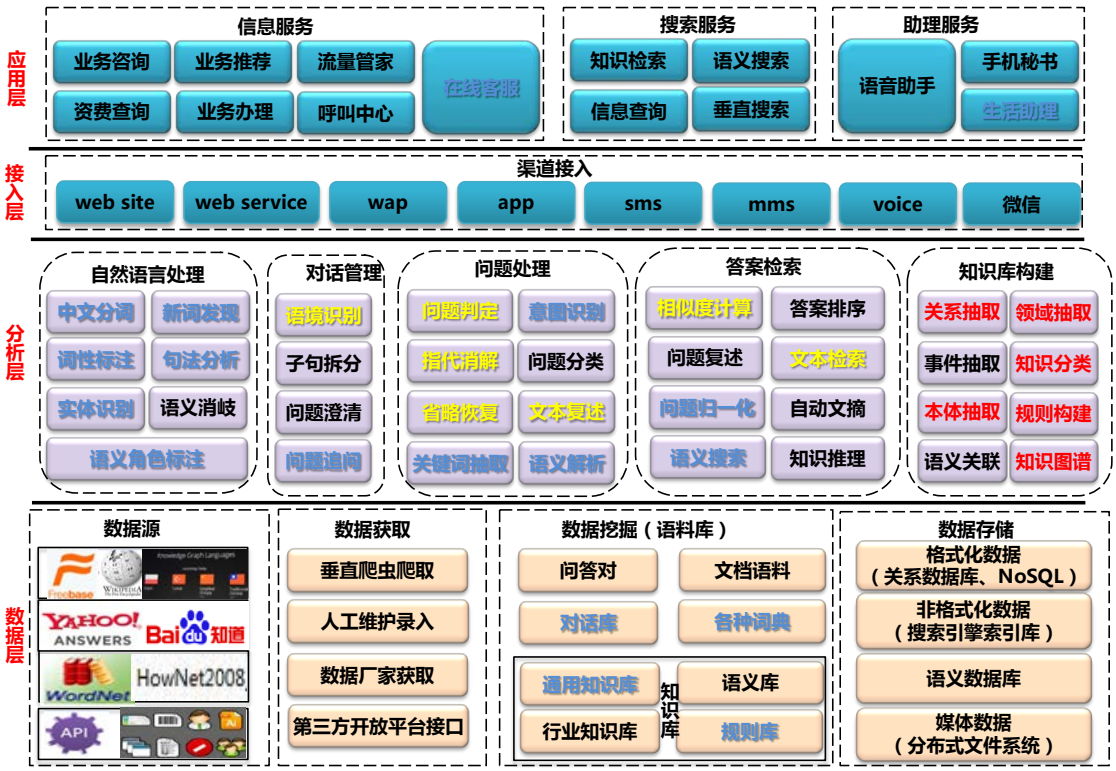


Figure 1. Robot IQA system architecture.
图 1. 智能应答机器人 IQA 系统架构

用语义数据库来存储；4、对于大数据量，可以采用分布式文件系统和 NoSQL 等。

有了结构化或者清洗后的数据支持，需要对数据进行分析，这里的分析主要分为两部分，一部分是对用户输入问题的分析，即问题语义理解，一部分是线下数据的分析，如知识库构建等。分析层是智能问答系统 IQA 的核心引擎，分为自然语言处理(预处理)、对话管理、问题语义理解、答案检索获取，以及知识库构建更新 5 大部分，在知识数据已具备的情况下，由这 5 个模块即可以组成一个问答系统。

- 自然语言处理 NLP：该模块属于预处理模块，主要对用户输入问题和知识库进行预处理，比如中文分词、词性标注，供后面的关键词提取使用；抽取语料库的实体和新词发现，获取领域词典；通过句法分析和语义角色标注，获取用户问题的主谓宾、施事受事等。
- 对话管理：该模块属于上下文的复杂问题处理：如问题一次输入多个问题，需要进行子句拆分；问句模糊无法理解，需要问题再次澄清；问题缺少必须元素，需要追问达到回答目的；问题涉及上下文语境的，需要语境识别等。
- 问题语义理解：首先对单个问题进行语义理解：如判定用户的是问题还是聊天；问的是哪类业务或类别问题，便于快速定位；识别用户问的意图，究竟是咨询还是购买等；对缺少的问题成分，根据上下文恢复成语义齐全的问题，便于检索答案。深层语义分析主要是理解问题的真正语义并处理复杂问题，将多个问题拆分，根据上下文进行缺省句恢复和意图理解，对于多种问法可以抽取语义规则，进行规则匹配，对检索的结果进行相似度计算，找出最佳答案。
- 答案检索获取：该模块主要为答案获取模块，如知识库为 FAQ,则需要对问题进行复述，或将问题归一化到 FAQ 库标准问题，在没有的情况下，可以直接根据关键词检索，返回的结果进行相似度计算，答案排序，最终返回答案；如为知识库，则需要进行语义检索，并进行一定的推理；如为文档，则需要进行自动文摘，找到答案。
- 知识库构建更新：构建知识库可以更好的组织知识，更快速准备检索答案，和 FAQ 相结合，使问答系统适用各种语料，而不仅仅局限于 FAQ，需要包括本体提取、领域词提取、关系提取、推理规则构建等功能。

IQA 为了方便用户使用，满足用户不同的使用习惯，需要有多种接入方式，主要包括：

- 语音：为了解放用户双手，支持语音接入交互，使问答更加智能。
- 微信：支持和微信平台接入，成为微信智能客服和问答机器人，便于用户低成本使用。
- 短信：在传统运营商有些业务依然会使用短信来提醒，为此，用户可以直接在此基础上进行回复，而无需记忆传统的短信代码，使用自然语言，提高用户使用体验。
- App：支持 app 接入，比如生活助理、吃喝玩乐等 app 应用中的智能问答。
- 平台：在设计时，需要考虑各个模块、算法和接口的松耦合性，做到可插可拔，黑盒复用，因此中间的分析层可以独立成问答平台，以便开放给第三方调用。

应用层是在智能问答核心技术的支撑下产生的各种应用和服务，不仅仅是问答，还可以是信息服务、搜索服务和助理等服务。

- 信息服务：可面向运营商或政企等在线客服场景，做业务咨询、业务办理、业务推荐等，也可和传统人工坐席相结合，在智能问答无法回答的情况下，再呼叫人工坐席。
- 搜索服务：我们的知识库支持专业的知识工程构建，因此可以提供知识检索、语义检索等，也可以作为智能检索，嵌入企业管理系统，作为垂直企业检索。
- 助理服务：生活助理、个人助理、语音助理也是智能问答的应用场景，为用户提供助理等服务。

此外，IQA 涉及大数据量的语料处理，需要支持离线分析。IQA 系统又是在线的高并发服务系统，需要支持实时访问、分布式扩展。

4. 智能问答系统关键技术

4.1. 知识图谱构建

智能问答的智能核心来自于强大的知识资源，这需要对大规模数据资源进行理解和抽取，转换成计算机可以处理的形式来表示和存储。

早期的知识库是把专家知识通过人工进行构建的，需要大量的人力和物力，这些知识库资源最典型的代表包括英文词汇知识库 WordNet、FrameNet、中文词汇知识库 HowNet 等。这些知识库基本属于通用领域的知识库。

通用领域知识库资源不能满足智能问答系统对知识资源的需求。针对电信业务场景，IQA 还需要构建专用领域知识库。

知识图谱构建主要分成三部分：核心信息抽取、知识库构建与更新、知识库可视化。其中核心信息抽取指从语料中抽取核心信息，如实体、实体类别、实体关系、实体属性等；知识库构建与更新指将核心信息抽取的结果以知识库的方式进行构建与更新；知识库可视化指知识库的展示，并提供一定的管理功能。

由于 IQA 的设计目标是电信场景的智能应答，所以使用了基于知识库和 FAQ 相融合的智能交互服务技术，将知识与知识的关联整合到系统和数据里面，使 IQA 更加智能。知识管理既包含了知识库，还有 FAQ 库等：

- 知识库：主要是进行知识库构建和更新，比如本体提取、本体关系提取，将领域知识构建成知识库专家系统，另外能进行一定的推理。
- FAQ 库：FAQ 主要是面向客服类语料，不能光靠检索和相似计算，也需要对 FAQ 库进行处理，如同一问题不同问答的语义规则提取、意图类别提取、问题类别提取等。
- 对话库：任何问答都离不开正常的对话交流，因此需要提供对话寒暄库来提高问答的用户体验性。
- 各种词典：存放问答需要的各种词典，如分词、同义词、领域词等。

4.1.1. 实体识别

命名实体识别包括对实体的识别及属性的抽取。实体识别是把文本中的实体划为某一语义类型。主要有三种方法，基于字典、基于统计与基于规则的方法。基于字典的方法这一方法较为简单，主要是通过字符串匹配找寻词库中命名实体，但是通常没有一个全面的实体库，而且比对费时。基于规则算法主要在实体识别过程中加入词法规则、语法规则、语义规则，然后通过规则匹配的方法识别各种类型的命名实体。基于规则方法受限于人工添加规则。基于统计的方法先建立语言模型，利用人工标注或原始语料进行训练，然后在训练数据上估算模型参数，这有利于移植到不同的语言及新领域。基于统计的方法主要利用一些统计模型如隐马尔可夫模型、最大熵模型、支持向量机、条件随机场(Conditional Random Fields，CRFs)等。

在电信领域，NER(命名实体)识别的目的为从语料中抽取出电信领域相关实体。例：“如何办理酒店**留言**优惠**套餐**？答：…”，其中加粗部分即为电信领域相关实体。

IQA 的实体识别结合了基于字典和基于统计的方法。一方面查找字典，一方面采用基于统计的方法，通过标注语料获得一定数量的已标注的 NER 数据，用于训练 NER 模型。然后对于给定的生文本语料，先进行文本预处理(分词、词性标注等)，然后使用训练好的 NER 模型进行 NER 识别，最终得到 NER 识别结果，如图 2 所示。

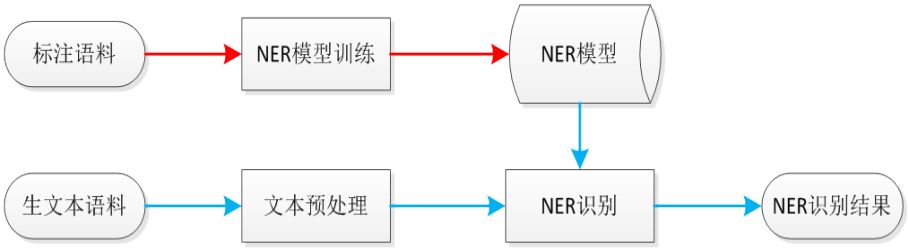


Figure 2. The process of Named Entity Recognition
图 2. 命名实体识别流程图

构建电信领域命名实体库是一个动态的过程。开始时针对电信业务手动建立较小规模的电信实体及属性词典，然后从电信文本及期刊网上获得电信领域的文献，通过建立的词典信息提取特征，使用统计模型进行训练，从而使模型在样本有限的情况下学习到新知识，将筛选出的元素加入词表中。随着项目的深入，数据的增多，则通过对大量数据的学习自动识别电信实体从而扩大命名实体库的规模。

属性抽取的任务是为每个实体语义类构造属性表并抽取出属性值。一方面对于一些数据如电信套餐业务类表格可以通过解析原始数据中的半结构化信息来获得属性和属性值，这种方法绕开了自然语言理解的阅读句子。而更多的知识隐藏在自然语言句子中，当前从句子中提取属性和属性值的基本手段是模式匹配[4]和对自然语言的浅层处理。

4.1.2. 关系抽取

由于知识库采用实体－属性－值的方式进行构建，知识库中的实体之间的关系有以下三种：

- 顺序关系：

在电信领域的问答中，用户对于实体的提问具有先后的次序关系，比如先问手机话费业务，再问短信业务。实体之间的顺序关系抓住用户提问的这种先后关系，建立实体间的顺序关系。利用这种顺序关系，解决自动问答系统中当用户提到一个实体时，为其自动推荐下一个实体。

- 共现关系：

在电信领域的文本中，不同的实体经常一起出现，具有比较强的关联关系，我们称之为共现关系。与顺序关系不同的是，共现关系不仅在用户提问中获取，也可以在其它文本中获取(比如业务说明文档等)，没有先后顺序关系，从某种意义上来讲，顺序关系是共现关系的一种。利用这种共现关系，当用户提到某一实体时，为其推荐关联度较大的若干个实体。

- 父子关系：

从电信领域的目录结构中获取实体之间的上下级关系。

关系抽取一般通过人工定义各类关系模型，然后通过模式匹配或者机器学习的方法进行抽取。一般领域知识内，关系模式相对比较固化，可以通过统计算法找到关系模式，再基于关系模板进行抽取。基于深度学习的抽取算法近来比较火热，不仅能减少人工标定的工作，还能够返现尚未人工定义的规则。

4.2. 问题理解

智能问答系统只有知识远远不够，还需要理解人提出的问题，将自然语言表述的问题转化为计算机可以理解的形式化语言。让计算机理解自然语言是非常困难的，也是人工智能领域最难处理的问题之一。

4.2.1. 领域分类

领域分类主要解决的是把用户提出的问题对应到相应的领域中，减少问题答案的范围，提升对问题的理解力，这是智能问答系统要解决的一个关键问题。在实际的问答系统中，领域知识可能会包含很多

种，如果分类错误，答案基本就是错误的。领域分类有词匹配法、规则方法、统计学方法等几种常见方法。

最早的词匹配法仅仅根据文档中是否出现了与类名相同的词来判断文档是否属于某个类别。很显然，这种过于简单的方法无法带来良好的分类效果。

规则方法是为每个类别定义大量的推理规则，如果一篇文档能满足这些推理规则，则可以判定属于该类别。由于在系统中加入了人为判断的因素，准确度比词匹配法大为提高。但是分类的质量严重依赖于规则的好坏，也就是依赖于制定规则的“人”的水平。规则方法可推广性差，一个针对金融领域构建的分类系统，则无法扩充到医疗或社会保险等相关领域，造成巨大的知识和资金浪费。

统计学习方法进行文本分类的一个重要前提是认为文档的内容与其中所包含的词有着必然的联系，同一类文档之间总存在多个共同的词，而不同类的文档所包含的词之间差异很大。而且不光是包含哪些词很重要，这些词出现的次数对分类也很重要。这一前提使得向量模型(VSM)成了适合文本分类问题的文档表示模型。在这种模型中，一篇文章被看作特征项集合来看，利用加权特征项构成向量进行文本表示，利用词频信息对文本特征进行加权。它实现起来比较简单，并且分类准确度也高，能够满足一般应用的要求。

然后根据 VSM 生成的描述文本特征向量，用例如朴素贝叶斯算法、kNN 算法或者 SVM 算法进行分类。上述方法各有各的优势[5]。kNN 算法具有简单、稳定、有效的特点，但时间复杂度较高，适用于文本训练集合规模较小的文本分类系统。朴素贝叶斯算法可应用到大规模文本集合中，具有方法简单、速度快、分类准确率高等优点，但由于朴素贝叶斯算法所基于的假设太过于严格，在实际应用并不完全符合理论中假设条件的情况下，其准确率会有一定程度的下降。因此，在类别数目较多或者类别之间相关性较小的情况下，该算法模型的分类性能才能达到最佳。SVM 是以 VC 维理论和结构风险最小化原则为基础的，克服了特征空间中的维数灾难问题，解决了小样本学习问题，可得到小样本条件下的全局最优解。相对于其它分类算法，稀疏、高维的数据对 SVM 算法基本没有影响，能够更好地体现文本数据的类别特征。

统计学习模型在实现文本分类中也存在明显的缺陷。统计学习模型没有利用上下文的关系来理解文本信息进行分类。VSM 中生成的描述文本特征向量，需要人为的统计并人为的做特征提取工作。统计学习模型泛化能力薄弱，在不同的场景下要进行人为的调整模型的特征。

为了解决上述统计学习模型的缺陷，我们设计了深度学习中的 RNN [6]模型来进行分类模型的实现，RNN 很好的利用了整个文本结构的信息并减少了人为固定特征提取的工作。加强了泛化能力的同时，在应用在类似领域的分类只需增加几个参数而已，可以大大减少人为的工作。

使用 RNN 模型实现我们要做的领域分类过程如图 3 所示，首先用对大量广泛语料和特定领域的语料进行分词处理，将分词后的语料代入 word2vec [7]中，得到词向量的训练模型。然后将我们要验证的特定领域语料代入词向量的训练模型，继而生成特定领域中每个句子所有词的词向量。最后将每个句子的词向量作为输入数据输入到 RNN 模型当中，根据我们要求的分类情况给出输出向量，进而对整个模型进行训练得到 RNN 的分类模型。将我们要验证的语句代入到 RNN 分类模型当中，判断出该语句是否属于特定领域。其中，word2vec 开源工具直接从网上下载获得。大量广泛的训练语料从网站上爬虫获得。

我们以聊天语料作为特定领域为例，根据设计的 RNN 神经网络，对聊天语料进行训练，然后用训练好的 RNN 分类模型对验证语料进行测试，判断出该语句是否属于聊天类别。

整个问题的难点在于如何用词向量来训练 RNN 模型，如何确定 RNN 模型中各个参数，如何根据分类的结果对句子是否为聊天语料进行判别。

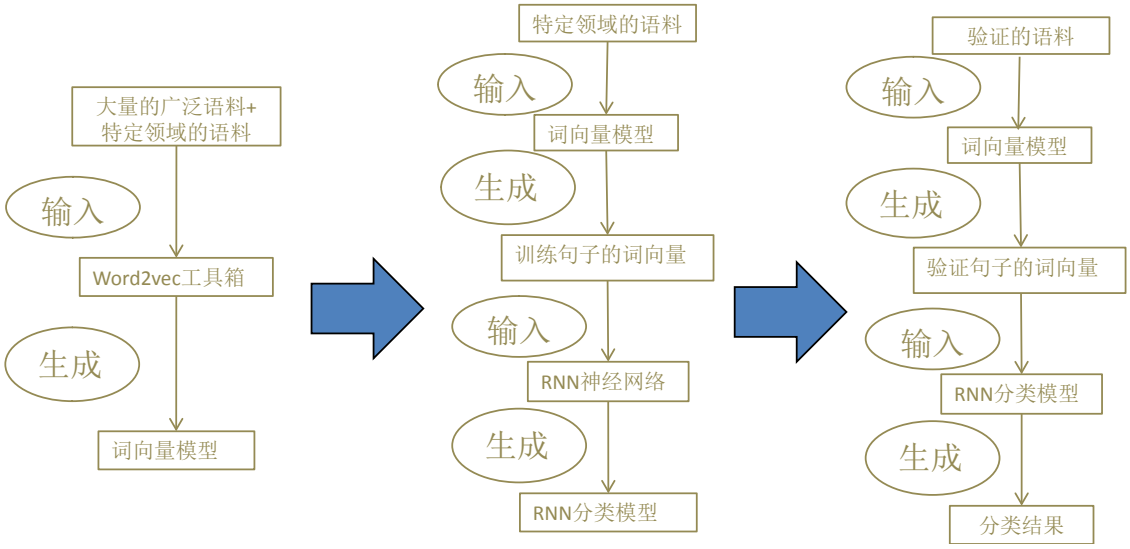


Figure 3. The process of domain classification with RNN
图 3. RNN 实现领域分类的过程

下面给出 RNN 的整个 BPTT 算法[8]：

输入：训练集 $D=\left\{\left(x_i^t,z_j\right)\right\},i=1,2,\cdots,I,j=1,2,\cdots,J,t=1,2,\cdots,T$

学习速率 η

过程：

1. 在(0, 1)内随机初始化网络中的所有连接权和偏移量。
2. Repeat。
3. for all $\left\{\left(x_i^t,z_j\right)\right\}$ do。
4. 求解中间变量 $\frac{\partial L}{\partial y_j^t}=\frac{\sum_{t=1}^{T}\frac{y_j^t}{T}-z_j}{T}$。
5. 更新输出层偏移量 $\Delta\theta_j=\eta\sum_{z=0}^{T}\frac{\partial L}{\partial y_j^{t-z}}$。
6. 更新隐藏层到输出层的权值 $\Delta w_{hj}=-\eta\sum_{z=0}^{T}\frac{\partial L}{\partial y_j^{t-z}}b_h^t$。
7. 更新隐藏层到隐藏层的权值 $\Delta p_{ih}=-\eta\sum_{z=0}^{T}b_i^{t-z-1}\left[1-\left(b_h^{t-z}\right)^2\right]\sum_{h=1}^{H}w_{hj}\frac{\partial L}{\partial y_j^{t-z}}$。
8. 更新隐藏层的偏移量 $\Delta r_h=\eta\sum_{z=0}^{T}\left[1-\left(b_h^t\right)^2\right]\sum_{h=1}^{H}w_{hj}\frac{\partial L}{\partial y_j^{t-z}}$。
9. 更新输入层到隐藏层的偏移量 $\Delta v_{ih}=-\eta\sum_{z=0}^{T}x_i^{t-z}\left[1-\left(b_h^{t-z}\right)^2\right]\sum_{h=1}^{H}w_{hj}\frac{\partial L}{\partial y_j^{t-z}}$。
10. until 停止条件。
11. end for。
12. 输出：各层的连接权和偏移量。

一般的停止条件设定为确定误差 $err=\sum_{z=0}^{T}\left(\frac{e^{y_j^t}}{\sum_{j=1}^{J}e^{y_j^t}}-z_j\right)^2<\varepsilon$，其中 ε 根据要求去人为给定。

当然上述给出的 BPTT 算法是基于 SGD(随机梯度下降法)去实现的，即每输入一个句子语料的词向量进行一次权值的更新。这样在以后对模型进行泛化能力的加强有更好的作用，也可以作为增强模型去训练。

4.2.2. 语境识别

交互式问答中问题语境检测可以看作一个二元分类问题，即判别用户提出的问题是与之前问题相关还是不相关。如果相关就可以认为是属于同一个语境，否则认为其属于不同语境。构造二元分类器主要问题就是找出能够有效进行问题相关检测的特征。目前研究来看，对分类有效的特征有：指代成分特征、线索词特征、最长公共词序列特征、名词特征等。利用以上特征就可以利用二元分类算法对当前的问题进行语境识别了。

指代成分特征：如果在问题中包含第三人称代词和指示代词，例如“他”“它的”“这些”等，并且这些指代成分不是指代本问题中出现的实体，那么指代成分只能指代之前问题或答案中出现的词语，所以问题是后继问题。这个特征通过构造的指代词表过滤分词后的问题获得，是布尔值特征，即当问句中包含指代成分时，特征值为 True，否则为 False。

线索词特征：线索词是指问题句中包含的如“其他的”“综上所述”“总之”这样的词汇，这些词汇提示了问题是与之前问题相关的。我们发现，尽管包含线索词的后继问题数量不是很多，但是线索词可以很准确地识别出后继问题。这个特征是通过线索词表过滤分词后的问题获得的，是布尔值特征，即当问句中包含线索词成分时，特征值为 True，否则为 False。

最长公共词序列特征：如果当前问题与之前问题中，包含有相同顺序出现的内容词，这样的问题通常为后继问题。例如：“问题 1：布什和戈尔的第一场辩论在哪所大学举行？问题 2：辩论是在什么时候举行？”，在问题 2 和问题 1 有相同的公共词序列“辩论举行”。最长公共词序列特征通过计算两个问题分词之后序列的最长公共字串获得，为减少计算复杂度，可以采用动态规划算法。最长公共词序列特征值是一个自然数集合。

名词特征：如果问题在去除了停用词、疑问词以及问题常用词后，没有包含任何名词，这样问题中没有提及任何具体信息，所以之前的问题一定提供了相关信息，则问题应与之前问题相关。通过对问题分词和词性标注处理，我们就可以获得这个特征值。这个特征值是一个布尔值，即当问句包含名词时，特征值为 True，否则为 False。

利用构建的基于以上 4 个特征二元分类器来进行交互式问答的问题相关检测，分类器采用标注了相关和不相关的 TRECQA 任务翻译成中文的问题集训练获得。训练二元分类器的方法是对于每个问题获取 7 维向量 $v_i=\left\langle v_i^1,v_i^2,v_i^3,v_i^4\right\rangle$，其中每个向量分量对应于指代成分特征、线索词特征、最长公共词序列特征、名词特征、相同实体特征、句法结构特征和内容词相关性特征。首先对问题集中每个问题进行分词、词性标注、句法分析以及命名实体识别的预处理，然后根据问题句预处理的结果，就可以获得指代成分特征、线索词特征、名词特征以及句法结构特征。而对于最长公共词序列特征、实体相同特征以及内容词相关性特征，采用经过预处理后的问题 q_i' 和之前的 n 个问题 $Q_n'=\{q_{i=1}',q_{i=2}',q_{i=3}',\cdots,q_{i=n}'\}$ 分别按照特征中描述的方法计算，取其最大值作为问题向量的相应分量，即问题 q_i 的 m 特征分量：$v_i^m=\max f\left(q_i',q_{i-1}'\right),1\le i\le n,m\in\{3,5,7\}$。根据训练集问题的问题向量以及问题相关和不相关标注，采用

C4.5 决策树的方法通过训练就可以获得问题相关检测的二元分类器。

智能问答系统涉及的关键技术点不限于以上所述，还有省略恢复[9]与指代消解[10]、相似度计算、智能对话管理等等，鉴于篇幅有限，不再展述。

5. 总结与展望

智能问答系统是目前人工智能和自然语言处理领域中一个备受关注并具有广泛发展前景的研究方向[1]。从行业发展方向来看，融入了人工智能技术的智能客服机器人成为了推动传统客服转型升级的变革力量。从技术现状和我们项目的实际应用经验来看，目前智能问答机器人 IQA 还需要借助于工程化的方法达到商用要求，例如关键词、等价句等配置使用等。通用的、高质量的智能应答系统，仍然是较长期的努力目标。随着大数据和深度学习技术的发展，基于知识和推理的深层方法和基于统计等“浅层”方法有机结合，未来智能问答系统将向着更智能化和人性化的方向发展。随着项目语料库的扩充和丰富，IQA 将在省略恢复技术点上使用深度学习的 LSTM 网络进行训练，尝试通过自动学习得到隐藏信息的网络结构；在领域分类技术点上，当 LSTM 训练模型的准确率积累到一定的程度时，摒弃线上分类规则，直接使用 LSTM 进行领域分类处理；问题归一化能力的强弱决定问答系统智能化程度的强弱，现在研究解决通用领域的问题归一化还有很多困难，但研究解决垂直领域的归一化还有很大的可能性。2016 年 9 月，谷歌提出了一种使用单个神经机器翻译(NMT)模型在多种语言之间进行翻译的简洁优雅的解决方案[11]，实现了机器翻译领域的重大突破。而问题归一化本质上也是一种机器翻译，可以尝试使用神经网络训练得到口语化到标准问题转换的模型网络。

参考文献 (References)

[1] 史忠植. 人工智能[M]. 北京: 机械工业出版社, 2001。

[2] http://money.163.com/16/1130/10/C745HHAB002580S6.html#from=keyscan

[3] http://finance.qq.com/a/20160302/004820.html

[4] 杜永萍, 黄萱菁, 吴立德. 模式学习在 QA 系统中的有效实现[J]. 计算机研究与发展, 2015, 43(3): 449-455.

[5] Russell, S.J., Norvig, P. 人工智能, 一种现代的方法[M]. 第 3 版. 北京: 清华大学出版社, 2013.

[6] http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns

[7] https://code.google.com/p/word2vec

[8] Graves, A. (2012) Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin.

[9] 李旺, 李绍滋. 基于 DRT 理论的汉语省略恢复研究[J]. 计算机工程, 2004, 30(17): 39-41.

[10] 宋洋, 王厚峰. 基于马尔可夫逻辑的中文零指代消解[J]. 计算机研究与发展,2015, 52(9): 2114-2122.

[11] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., *et al.* (2016) Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 301-309
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74037

Hans汉斯

Hierarchical ICA Encoding Combined with Motion-Appearance Information for Anomaly Detection

Shilei Duan, Kewei Wu, Shen Tang, Huan Liang, Zhao Xie

School of Computer and Information, Hefei University of Technology, Hefei Anhui
Email: shilei_duan@yeah.net, wu_kewei1984@163.com

Received: Apr. 2nd, 2017; accepted: Apr. 14th, 2017; published: Apr. 19th, 2017

Abstract

Due to the shortcomings of the existing anomaly detection methods in terms of the representation of visual hierarchical perception, inspired by the regularities in perception encoding of bio-mimetic vision, a hierarchical ICA encoding approach combined with motion-appearance is presented for abnormal events detection. First of all, this method extends the Existing biological visual hierarchy coding framework with three-level layer-wise learning, and uses ICA statistical method to extract intra-layer visual perceptual coding patterns, and utilizes the HMAX mechanism to transmit the Hierarchical information. In addition, with the processing theories of double channels in the visual system, Each channel is proposed to separately complete three-layer coding pattern learning. Then, these double features are fused to represent the Anomaly patterns. Based on the joint representations, the one-class SVM model is used to predict the abnormal score for each input. The proposed method, whose properties of motion perceptual coding and the performance of anomaly detection, is evaluated on UCSD datasets, and the results demonstrate that the learned feature representations in this paper for anomalous patterns are superior to other traditional hand-crafted features and deep learning features.

Keywords

Hierarchical ICA Encoding, Motion-Appearance Representations, One-Class SVM, Anomaly Detection

基于运动外观多通道层级ICA编码的异常检测

段士雷，吴克伟，唐　燊，梁　欢，谢　昭

合肥工业大学计算机与信息学院，安徽　合肥
Email: shilei_duan@yeah.net, wu_kewei1984@163.com

收稿日期：2017年4月2日；录用日期：2017年4月14日；发布日期：2017年4月19日

文章引用: 段士雷, 吴克伟, 唐燊, 梁欢, 谢昭. 基于运动外观多通道层级 ICA 编码的异常检测[J]. 计算机科学与应用, 2017, 7(4): 301-309. https://doi.org/10.12677/csa.2017.74037

摘　要

针对现有异常表示方法对视觉感知层级关系描述能力的不足，基于生物视觉感知编码特性启发，本文提出一种基于运动外观多通道层级ICA编码模型，实现复杂场景中的异常检测任务。首先，对现有的生物视觉层级编码框架，进行三级逐层学习拓展，采用ICA统计方法提取层内视觉感知编码模式，利用HMAX机制实现层级信息传递。其次，借助视觉双通道处理机制，各通道独立完成三层编码模式学习，随后联合双通道特征构建异常模式表达，最终，利用单类支持向量机模型对正常和异常情况进行判定。在UCSD数据集上，分别验证了本文方法的运动感知编码特性和异常检测的性能，实验结果能够说明本文异常模式表达优于现有的手工设计特征，以及深度学习特征。

关键词

层级独立成分分析编码，运动外观联合表达，单类支持向量机，异常检测

1. 引言

视频场景分析与理解研究已经吸引了来自计算机视觉领域众多研究者的关注，其致力于研究新技术、新方法去更精确快速地分析、理解场景内容，从而更有效地协助监控人员获取准确信息以及处理突发事件，并最大限度地降低误报漏报，起到监督管理的作用[1]。视频场景中的异常事件检测是其中一项重要的研究内容，同时也是研究的热点和难点。

异常检测最经典的做法通常是基于训练数据学习描述正常运动模式的模型，然后通过评估新视频中的运动模式偏离该模型的程度来判断其是否属于异常运动。如 Hu 等人[2]采用多目标追踪算法提取正常运动轨迹特征，然后学习其统计分布，充分考虑时空信息用于异常检测。Mehran 等人[3]首次采用表达行人交互力信息的社会力模型(Social Force Model，SFM)来检测视频中的异常行为。此外，基于多尺度光流直方图的稀疏编码模型[4]也成功用于异常检测，该模型采用稀疏重构代价(Sparse Reconstruction Cost，SRC)为判断准则。Li 等人[5]采用混合动态纹理模型(Mixture Dynamic Texture，MDT)对外观、运动以及空间尺度特征进行建模，提出了时空异常的联合检测器。上述方法虽然能够实现异常检测，但是其采用的是手工设计特征，该类特征需要专业的先验知识，而这在复杂的视频场景下难以实现，也限制了检测性能的进一步提升[6]。

近年来，深度学习方法被成功应用于各项视觉任务，证明了其强大的编码表达能力。深度学习本质思想也是以人类大脑对视觉信息的层次处理方式为基础，构建多层次学习模型进行学习。如蔡瑞初等人[7]提出了一种基于多尺度时间递归神经网络的人群异常检测和定位方法。Xu 等人[6]则提出外观和运动深度网络(Appearance and Motion Deep Net，AMDN)学习运动、外观以及联合信息的特征表达用于异常检测。该方法采用堆栈去噪自动编码器网络进行特征学习，并提出双融合策略，分别是输入端的运动外观信息融合和输出端的异常得分融合，其中，前者融合是为了充分考虑外观和运动信息之间的互补性，后者融合是为了得到运动区域的最终异常得分。但是，在异常检测任务中，这些深层框架仅仅被看作是输

入直接至输出的黑盒子模式的学习过程，即图 1 中红框部分，忽略了运动感知层次编码中潜在的编码特性，而且其层级较多，结构较为复杂，在学习过程中易产生过度拟合现象，从而导致结果不准确。

针对现有手工设计特征和端对端深度学习特征，在特征编码的层级关系描述能力的不足，即缺少对图 1 中红色框中的编码机制的深入研究，本文启发于仿生物视觉计算模型的(Hierarchical Max-pooling，HMAX)机制[8]，对现有层级模型进行了扩展，采用具有稀疏特性的独立成分分析(Independent Component Analysis，ICA)统计方法，分析了运动感知编码中潜在的三层编码特性，并模仿人类视觉的双通道处理机制，采用独立的外观通道和运动通道分别学习高层结构特征，通过特征层融合来进行正常运动模式表示和异常检测。本文的异常检测框架如图 1 所示。与现有的异常检测方法相比，本文主要贡献如下：

1. 针对异常检测任务中的特征编码难题，对现有的生物视觉层级编码框架，进行三级逐层学习拓展，采用 ICA 统计方法提取层内视觉感知编码模式，利用 HMAX 机制实现层级信息传递，如图 1 中的层级 ICA 基元学习部分。

2. 借助视觉双通道处理机制，运动外观各通道独立完成三层编码模式学习，随后联合多通道信息构建异常模式表达，采用 One-Class SVM 对正常和异常情况进行判定，多通道特征层融合效果能够提升现有层级编码表达的准确性。

3. 在 UCSD 数据库上，与其他公认手工设计特征[3] [4] [5]和 AMDN[6]深度学习特征进行了对比，实验表明本文方法的检测性能要优于对比方法，同时也直观解释了层级框架的三层运动感知编码特性。

2. 基于高层结构特征的异常检测

现有异常检测方法主要采用手工设计特征，如轨迹特征，光流直方图特征和 3D 时空梯度特征等等。这类特征不仅属于低层特征，也缺乏强大的编码表达能力，从而导致不佳的检测性能。AMDN 方法首次用深层结构分别学习外观、运动及两者联合的编码特征用于异常检测。本文方法是生物视觉启发式的计算模型，采用层级学习结构，不同于传统手工设计特征提取，与 AMDN 方法在层级处理上也存在明显不同：1) 本文层级编码框架是模拟人类视觉系统的双通道处理机制，AMDN 方法采用复杂的三通道学习机制；2) 本文层级框架包含三层学习结构，AMDN 方法的深度更为复杂，层级与视觉通路的对应关系不明确；3) 运动外观通道融合策略不同：本文采用特征融合方式，AMDN 方法采用得分融合方式；4) 本文能够有效解释层级结构的运动感知编码特性，AMDN 方法只是当作输入至输出的黑盒子模式；5) 本文采用具备稀疏特性的 ICA 学习算法，而且在卷积之后增加了归一化和校正线性单元[9]等非线性操作以提高层与层之间的信息传输能力，AMDN 方法采用传统稀疏编码学习算法。

2.1. 层级编码理论

人类视觉系统中存在结构和功能相对分离的两条通路[10]：一条是腹侧通路，主要感知目标形状和纹理等信息；另一条是背侧通路，主要感知目标位置和运动方向信息等。两条通路均呈现出明显的层次特性，层与层之间存在信息相互传递。图 1 中红框部分体现类似的双通道信息处理机制。两条通路拥有相似的层级学习过程[11]，静态视觉信息具有层级编码过程[12]，启发于此，本文的双通道机制采用相同的算法学习方式，如算法 1 所示。

层级编码特性中不同层次具有不同的感受野。感受野理论[13]是支持视觉信息分层串行处理的生物学基础理论，其主要是指视野某特定区域受到光线等刺激时，该区域可以引起相应神经元细胞的响应。研究表明，运动感知层级编码过程中，随着层级的不断提升，即算法 1 中参数 l 的增加，感受野区域也在不断地增大，所提取的特征表达复杂度在不断增加。因此，根据感受野特征的不同，本文模型的编码特性可以分为以下三层：低层主要编码运动模式边缘和方向等信息，中层则主要编码运动模式的局部结构

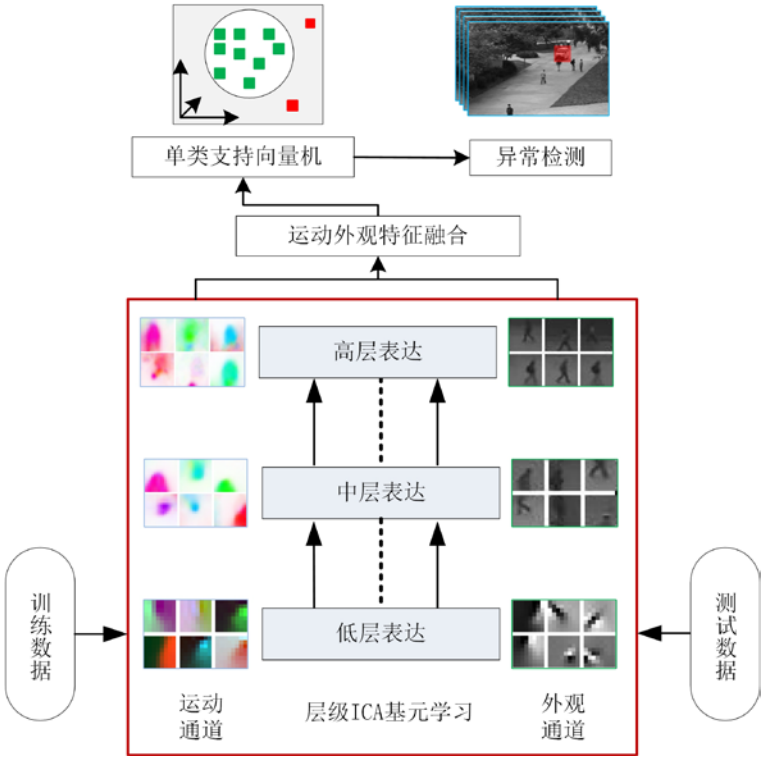


Figure 1. Framework of this paper for anomaly detection.
图 1. 本文异常检测框架

Algorithm 1. Multi-channels hierarchical ICA encoding algorithm
算法 1. 多通道层级 ICA 编码算法

输入：帧序列 $\{F_i^0\}_{i=1}^{N_F}$

输出：高层特征集合 $P=\{\boldsymbol{p}_i\}_{i=1}^{N_F}$

1：对帧序列 $\{F_i^0\}_{i=1}^{N_F}$ 提取光流特征 $\{Fm_i^0\}_{i=1}^{N_F}$

2：运动通道层级 ICA 基元学习：For $l=1:N_l$ (l 表示是层级框架的第 l 层)

3：对 $\{Fm_i^{l-1}\}_{i=1}^{N_F}$ 随机采样，构建训练集 Xm_{l-1}

4：利用 ICA 进行层内基元模式学习 $\min\|Bm_l*\Psi(Xm_{l-1})\|_1, s.t.\ Bm_l Bm_l^T=I.$ $\Rightarrow Bm_l$ 表示基元模式

5：利用基元模式实现层内编码 $Sm_l=\varphi(Bm_l*\Psi(Xm_{l-1}))$，其中 $\varphi(\cdot)$ 表示归一化和校正线性单元等非线性操作，增强层与层之间的信息传输能力

6：层间信息传递，$Fm_i^l=\max\text{-}pooling(Sm_i)$，通过局部邻域最大值池化操作增加局部平移不变特性

7：Endfor

8：利用运动通道高层基元，提取运动基元响应 $\boldsymbol{pm}_i=\bigcup_{c=1}^{Nm_c}{}_{\oplus}\left(dm_{(h,z)}^c\right)$，$dm_{(h,z)}^c\in Fm_i^l$ 表示坐标 (h,z) 处像素点，$\bigcup_{\oplus}$ 表示像素点的层间级联串行操作，Nm_c 表示 Fm_i^l 的通道数

9：对运动区域 $\{F_i^{l-1}\}_{i=1}^{N_F}$ 提取亮度特征 $\{Fa_i^{l-1}\}_{i=1}^{N_F}$，重复步骤 1-7，学习外观通道的三层 ICA 基元 Ba_l

10：利用外观基元 Ba_l 重复步骤 8，提取外观基元响应 $\boldsymbol{pa}_i$，两通道特征融合形成 $P=[\lambda\boldsymbol{pm}_i,(1-\lambda)\boldsymbol{pa}_i]$

11：所有运动区域的运动外观多通道进行高层表达的矢量集合，形成 $P=\{\boldsymbol{p}_i\}_{i=1}^{N_F}$

信息，而高层则编码运动模式的全局结构信息，更趋于类别信息。

为了在实际数据集上充分验证上述编码特性，本文将采用 ICA 学习算法进行运动模式的学习，算法

1 描述了模型的学习过程。其中， Bm_l 表示第 l 层的神经元卷积核。 $\Psi(\cdot)$ 表示白化操作，此外，受上述启发可知， $N_l=3$，表示低中高三层编码。该方法与传统稀疏编码方法相比，具有更高效的计算速度，也具备很好的稀疏特性，被广泛应用于生物视觉模拟。层级框架通过逐层非线性变换学习，提取运动模式的高层全局特征。

2.2 独立成分分析

独立成分分析(ICA)方法是利用统计独立作为目标来分离独立成分的信号分解技术，有利于小目标以及小类别信息的保留，其目的是把观测到的混合信号分解为相互之间统计独立的成分，然后利用该独立成分信息对其他信号进行线性组合表达，即 ICA 通过在特征空间上寻找能使数据互相独立的空间，将 n 维随机信号分解成一组统计独立的随机变量的线性组合。此外，相互独立的各个分量之间需要满足正交的约束限制[14]。假定一组观测信号 $X=\{x_1,\cdots,x_{nn}\}$ 是源信号 $S=\{s_1,\cdots,s_n\}$ 的观测值，假设第 i 个观测信号是由 n 个独立分量 S 线性混合而成，公式表达如下：

$$x_i=a_{i1}s_1+a_{i2}s_2+\cdots+a_{in}s_n,i=1,\cdots,nn \tag{1}$$

上式可以用矢量表达为 $\boldsymbol{X}=\boldsymbol{AS}$ ，其中 $\boldsymbol{A}=[\boldsymbol{a}_1,\cdots,\boldsymbol{a}_n]$ 表示混合矩阵， $\boldsymbol{a}_i$ 表示混合矩阵的基向量。而 ICA 方法就是仅通过数据观测信号估计出独立源信号或混合矩阵数据。

稀疏编码模型学习到的基矩阵属于超完备基，维数较大，基与基之间也存在相关性。而 ICA 算法学习到的基不仅线性无关，还会保持相互正交的特性。ICA 算法的目标函数为 $J(W)=\|Wx\|_1$ ，其加入正交约束后的优化策略如下：

$$\min\text{imize}\|Wx\|_1 \quad s.t.\ WW^{\mathrm{T}}=I \tag{2}$$

此外，在迭代优化过程中，如果每次用梯度下降法更新权值 W 后，都需要对其进行正交约束，更新方式如下面公式所示：

$$W\leftarrow\left(WW^{\mathrm{T}}\right)^{-\frac{1}{2}}W \tag{3}$$

其中权值 W 矩阵中基的维数低于输入数据的维数，并且在采用 ICA 算法进行学习时，要对输入数据进行白化预处理操作，目的是为了去除数据维度之间的相关性，保证特征间的协方差矩阵为单位矩阵。

2.3. 异常判断

本文把异常检测问题看成是基于数据块的二分类问题，为了降低计算代价及提高计算效率，异常检测过程中只考虑包含有运动区域的数据块，忽略背景信息块，即给定含有运动区域的测试块数据，采用训练学习所构建的模型来预测该数据块包含的运动模式是属于正常模式还是异常模式。综上，本文模拟人类视觉系统中的双通道机制，采用外观和运动两个独立编码框架，分别学习运动模式的高层表达。为了充分考虑视频场景中具有互补性的外观和运动两类信息，需要对运动和外观特征进行融合操作共同表达运动模式。本文采用线性加权的融合方式进行整合，即按照公式 $P=\left[\lambda\boldsymbol{p}m_i,(1-\lambda)\boldsymbol{p}a_i\right]_{i=1}^{N_F}$ 进行融合操作获得最终的特征矢量，描述运动模式信息，框架如图 1 所示。基于该高层融合特征，采用单类支持向量机[15]对正常运动模式进行学习建模，评估其分布，建模优化策略如下：

$$\begin{aligned}&\min_{w,\rho}\ \frac{1}{2}\|\boldsymbol{w}\|^2+\frac{1}{uN}\sum_{i=1}^{N}\xi_i-\rho\\&s.t.\ \boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{p}_i)\geq\rho-\xi_i;\xi_i\geq0\end{aligned} \tag{4}$$

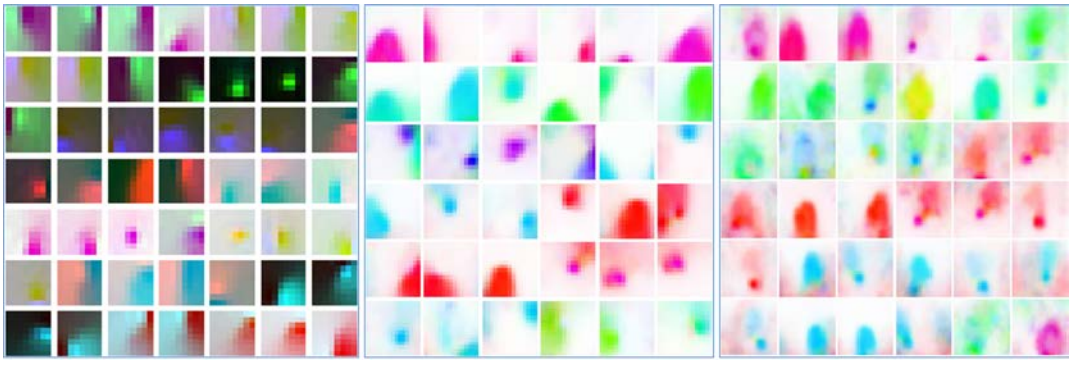
其中 $\boldsymbol{w}$ 表示权重矢量，ρ 表示偏差，$u \in (0,1]$ 为平衡控制参数，$\Phi(\cdot)$ 表示特征映射函数。然后对测试样例判断其是否符合正常运动模式的分布并采用已训练好的权重和偏差计算相应异常得分 $As(\boldsymbol{p}_i) = \rho - \boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{p}_i)$。此外，根据接收器操作特性曲线计算取得最大曲线围成的面积时的阈值作为判断的最终阈值。当异常得分 $As(\boldsymbol{p}_i)$ 大于阈值时，表示该样例是异常运动模式，反之，其属于正常模式。

3. 实验结果及分析

本文验证实验是基于 UCSD 数据集[16]，该数据集包含有两部分：Ped1 和 Ped2，分别描述两个不同的场景。两个子集中异常运动模式包括骑自行车的人、踩滑板的人、汽车等等，而正常运动模式都是人行道上正在行走的行人。为了提高计算效率，本文保持分辨率长宽比不变，统一把视频分辨率缩小为 210 × 140 像素。三层卷积核个数设置为 100、100 和 210，卷积核尺寸分别为 8 × 8，4 × 4 及 3 × 3。此外，本文所采用的学习框架是采用无监督方式进行学习。测试过程中以帧层和像素层的接受器操作特性曲线(Receiver Operating Characteristic，ROC)以及等错误率(Equal Error Rate，EER)作为最终的评价标准来验证本文方法的检测性能。其中，EER 表示在漏检率等于假阳率情况下的检测错误率。该评价标准沿用了 Mahadevan 等人[15]提出的评估方式，即当某个像素被认为是异常的时候就判定该帧是异常帧；而如果引起异常的对象有 40%的像素被检测到，就认为异常定位准确。

为了更有效评估方法的检测性能，本文采用以下四种对比方法：社会力模型[3]，混合动态纹理模型[4]，稀疏重构代价模型[5]及深度学习模型 AMDN[6]。该四类方法按其所采用特征可以分为两大类：1) 非深度学习方法，采用不同的传统手工设计特征来检测异常。其中，社会力模型是受物理学启发的用于异常检测的经典模型，主要通过交互力和环境因素来描述场景动态；混合动态纹理模型充分考虑场景外观和运动信息，基于局部特征表达来检测异常；稀疏重构代价模型是基于光流特征的稀疏表达思想，采用重构误差来评估运动的异常程度。2) 深度模型。AMDN 方法是第一个用于异常检测的深度模型，采用自动去噪编码器层级结构。这四种方法代表不同的特征编码方法，也能够全面地反映出本文框架高效性。

实验一：验证层级编码框架的运动感知编码特性。图 2 表示层级学习框架中三层编码特征的可视化。针对图 2，可以得出以下结论：1)低层主要编码运动边缘和方向等信息；2)中层则主要表征运动模式的局部结构信息，具有一定的类别信息，如头部运动、脚部运动等；3)高层更倾向于表达不同的运动模式类别信息，包含运动模式的全局结构信息。此外，深度学习的本质思想也是模拟人类视觉系统的视觉信息层次处理机制，从海量的数据中进行学习提取有用的特征。因此，无论其他检测方法使用了几层的处理机制，本文实验揭示的三层编码规律都应该能够去解释其编码特性。



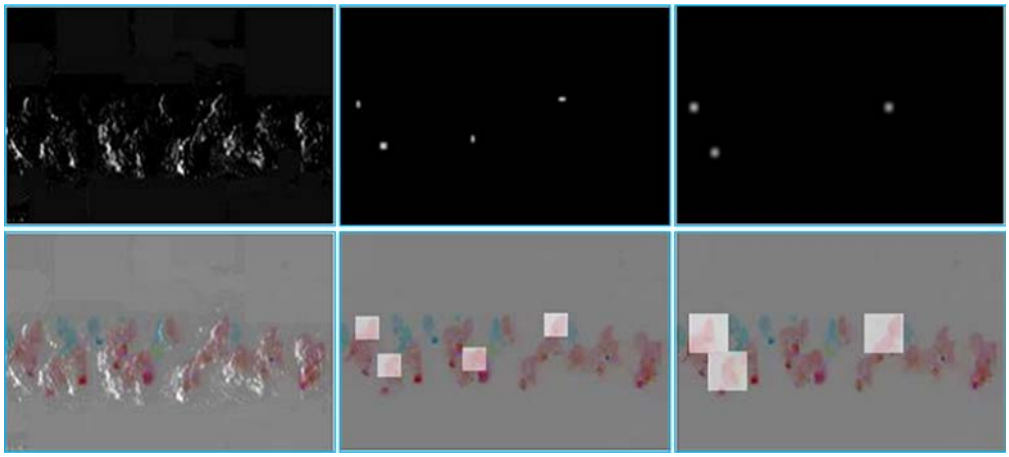
(a) (b) (c)

Figure 2. Visualization of feature representations for three-level neurons, (a) Low-level; (b) Mid-level; (c) High-level
图 2. 三层神经元的特征表达可视化，(a) 低层；(b) 中层；(c) 高层

图 3 表示某卷积核特征响应图及其在原始数据上的感受野区域。针对图 3，可以看到：1) 低层编码的边缘等信息与运动模式类别无关，具有共性，上述三层编码规律再次得到了充分验证；2) 中层局部结构信息的特征表达适用于分析某些拥挤场景下局部运动模式；而高层编码特征，可以有效用于运动模式检测和行为识别等视觉任务。基于此，本文将高层全局结构特征来建模场景中的运动模式用于异常检测。

实验二：评估三层学习框架在异常检测任务中的检测性能。图 4 和表 1 给出了不同方法在 UCSD 数据集上的异常检测和异常区域定位的定量结果。由于现有大多数方法都没有给出 Ped2 子数据集上的异常区域定位结果，因此，图 4 中也没有给出异常区域定位曲线。针对定量结果，可以得到：1) 异常帧检测层面：本文方法在 Ped1 子集上取得了与最好方法相当的性能，但明显优于其他对比方法；在 Ped2 子集上本文方法的性能是最好的，相比于其他方法，有了较大的性能提升。2) 异常区域定位层面：本文方法在 EER 和 AUC 两个标准下都获得最好的性能。此外在 UCSD 数据集上的一些异常检测实例如图 5 所示。

针对上述获得的优异定量结果进行分析：1) 相比于采用手工设计特征的传统方法，本文方法是采用学习机制提取高层全局特征，不需要专业的先验知识，并且特征具有更强的表达能力，因此本文方法的异常帧检测和异常区域定位性能都优于该类方法。2) 相比于基于深度学习的检测方法，AMDN 是第一个用于异常检测的采用深度框架的方法，从结果不难发现，其取得了优异的检测性能。但是该方法采用的去噪自动编码器网络结构较为复杂，层级较多，忽略了运动感知层级编码特性。而本文方法采用结构较为简单的三层学习框架，模拟生物视觉的双通道机制，充分考虑运动感知层级编码规律。尽管 AMDN 方法在 Ped1 子集上的异常帧检测取得最好的性能，但是本文方法同样也取得了相当的性能，而且异常区域定位和 Ped2 子集上异常帧检测性能上都有一定提升，综合来看，本文方法的异常检测优势更明显。



(a) (b) (c)

Figure 3. Responsive feature maps and associated receptive field region
图 3. 响应特征图及其对应的感受野区域

Table 1. Quantitative comparison for anomaly detection on UCSD dataset
表 1. UCSD 数据集上的异常检测定量评比结果

检测方法	Ped1 帧层	Ped1 像素层	Ped2 帧层
	AUC (%)	AUC (%)	AUC (%)
Social Force [3]	67.5	19.7	55.6
MDT [4]	81.8	44.1	82.9
Sparse Reconstruction [5]	86.0	46.1	—
AMDN [6]	**92.1**	67.2	90.8
本文方法	89.8	**72.5**	**92.3**

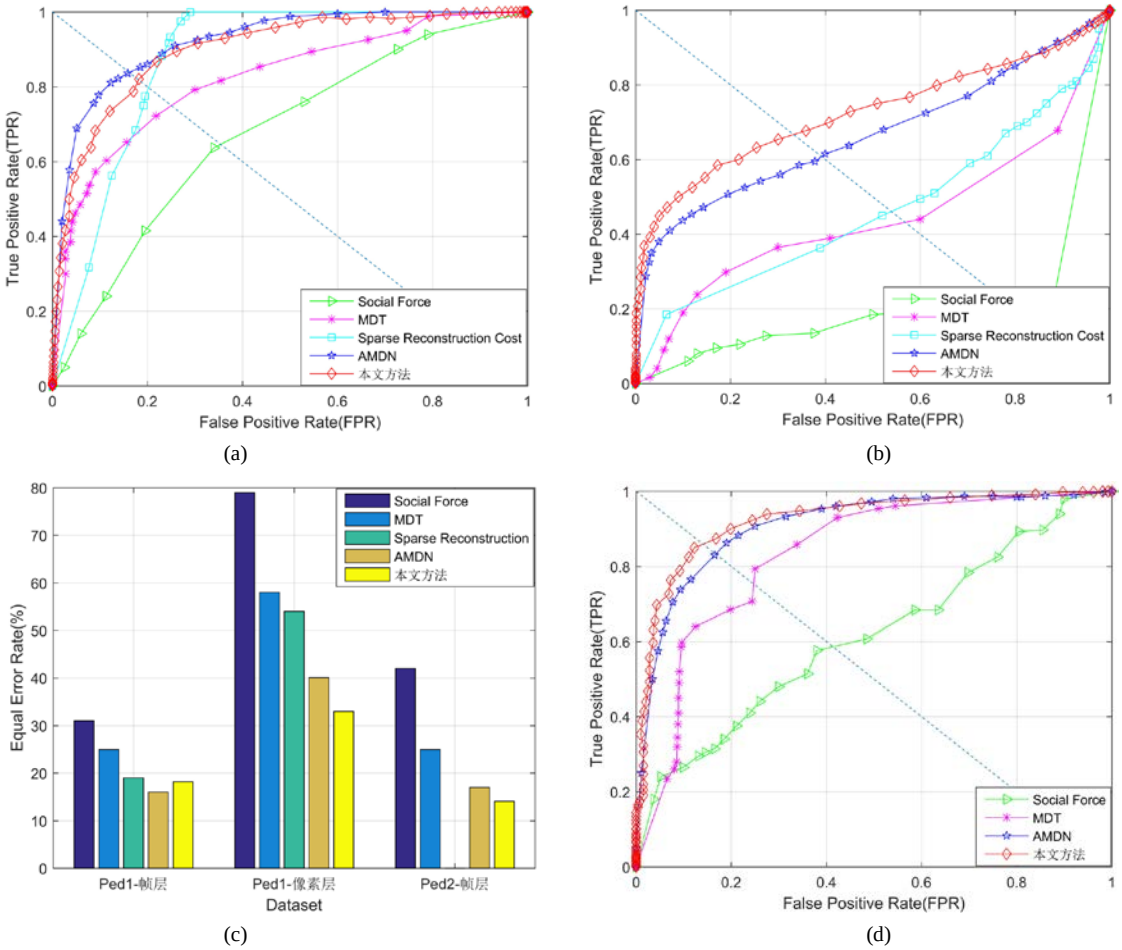


Figure 4. Quantitative curves for anomaly detection on UCSD dataset。(a) ROC curves of frame-level for detection results on Ped1；(b) ROC curves of pixel-level for detection results on Ped1；(c) EER results for UCSD dataset；(d) ROC curves of frame-level for detection results on Ped2

图 4. UCSD 数据集上异常检测结果定量曲线。 (a) Ped1 帧层检测结果 ROC 曲线；(b) Ped1 像素层检测结果的 ROC 曲线；(c) UCSD 数据集上 EER 结果；(d) Ped2 帧层检测结果 ROC 曲线

Figure 5. Examples for anomaly detection on UCSD dataset

图 5. UCSD 数据集上异常检测实例

4. 结束语

现有异常检测方法在至关重要的特征编码层面存在不足，忽略了运动感知编码特性。针对该问题，本文模拟生物视觉系统的信息处理机制，通过采用具备稀疏特性的 ICA 算法构建 HMAX 计算模型，研究其运动感知层级编码特性，揭示了低中高三层编码规律，然后，采用高层运动全局结构特征来对视频场景中的正常运动模式进行建模，用于异常运动模式发现。UCSD 数据集实验已经充分证明了本文方法的有效性，本文的异常检测方法获得的性能也优于其他传统方法和 AMDN 方法。

致　谢

感谢国家自然科学基金(No. 61273237；No. 61503111；No. 61501467)的支持。

参考文献 (References)

[1] 余昊, 孙锬锋, 蒋兴浩. 基于光流块统计特征的视频异常行为检测算法[J]. 上海交通大学学报, 2015, 49(8): 1199-1204.

[2] Hu, W., Xiao, X., Fu, Z., *et al*. (2006) A System for Learning Statistical Motion Patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **28**, 1450-1464. https://doi.org/10.1109/TPAMI.2006.176

[3] Mehran, R., Oyama, A. and Shah, M. (2009) Abnormal Crowd Behavior Detection Using Social Force Model. *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 935-942.

[4] Cong, Y., Yuan, J. and Liu, J. (2013) Abnormal Event Detection in Crowded Scenes Using Sparse Representation. *Pattern Recognition*, **46**, 1851-1864.

[5] Li, W., Mahadevan, V. and Vasconcelos, N. (2014) Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **36**, 18-32. https://doi.org/10.1109/TPAMI.2013.111

[6] Xu, D., Yan, Y., Ricci, E., *et al*. (2017) Detecting Anomalous Events in Videos by Learning Deep Representations of Appearance and Motion. *Computer Vision & Image Understanding*, **156**, 117-127.

[7] 蔡瑞初, 谢伟浩, 郝志峰, 等. 基于多尺度时间递归神经网络的人群异常检测[J]. 软件学报, 2015, 26(11): 2884-2896.

[8] Riesenhuber, M. and Poggio, T. (1999) Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, **2**, 1019. https://doi.org/10.1038/14819

[9] Nair, V. and Hinton, G.E. (2010) Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair. *International Conference on Machine Learning*, Haifa, 21-24 June 2010, 807-814.

[10] Ungerleider, M.L.G. (1982) Two Cortical Visual Systems. *Analysis of Visual Behavior*, **35**, 549-586.

[11] Giese, M.A. and Poggio, T. (2003) Neural Mechanisms for the Recognition of Biological Movements. *Nature Review Neuroscience*, **4**, 179. https://doi.org/10.1038/nrn1057

[12] Hu, X., Zhang, J., Li, J., *et al*. (2014) Sparsity-Regularized HMAX for Visual Recognition. *PLoS ONE*, **9**, e81813. https://doi.org/10.1371/journal.pone.0081813

[13] Hubel, D.H. and Wiesel, T.N. (1968) Receptive Fields and Functional Architecture of the Monkey Striate Cortex. *Journal of Physiology*, **195**, 215-243. https://doi.org/10.1113/jphysiol.1968.sp008455

[14] 冯燕, 何明一, 宋江红, 等. 基于独立成分分析的高光谱图像数据降维及压缩[J]. 电子与信息学报, 2007, 29(12): 2871-2875.

[15] Schölkopf, B., Platt, J.C., Shawetaylor, J., *et al*. (2001) Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, **13**, 1443. https://doi.org/10.1162/089976601750264965

[16] Mahadevan, V., Li, W., Bhalodia, V., *et al*. (2010) Anomaly Detection in Crowded Scenes. *Computer Vision and Pattern Recognition*, San Francisco, 13-18 June 2010, 1975-1981.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 310-319
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74038

Hans汉斯

Operating System Integrity Measurement Method Based on UEFI

Yihua Zhou[1], Hui An[1,2], Guan Wang[1,2], Liang Sun[3]

[1]College of Computer Science, Beijing University of Technology, Beijing
[2]Key Laboratory of Trustworthy Computing in Beijing, Beijing
[3]ZD Technologies (Beijing), Beijing
Email: 1216589268@qq.com

Received: Apr. 2nd, 2017; accepted: Apr. 14th, 2017; published: Apr. 19th, 2017

Abstract

If the operating system kernel is under attack, it can pose a significant threat to operating systems and applications. In order to ensure the integrity of the operating system kernel, this paper presents an operating system integrity measurement method based on UEFI firmware. In the scheme, we measure the integrity of operating system mainly in UEFI BIOS boot process, using TCM chip's encryption, authentication functions and Hash algorithm. The scheme can effectively protect the kernel and the safety of the operating system.

Keywords

UEFI, OS Kernel, TCM, Integrity Measurement

基于UEFI的操作系统完整性度量方法

周艺华[1]，安　会[1,2]，王　冠[1,2]，孙　亮[3]

[1]北京工业大学计算机学院，北京
[2]可信计算北京市重点实验室，北京
[3]中电科技(北京)有限公司，北京
Email: 1216589268@qq.com

收稿日期：2017年4月2日；录用日期：2017年4月14日；发布日期：2017年4月19日

摘　要

操作系统内核受到攻击，会对操作系统以及应用程序造成重大的威胁，为了保证操作系统内核的完整性，

文章引用: 周艺华, 安会, 王冠, 孙亮. 基于 UEFI 的操作系统完整性度量方法[J]. 计算机科学与应用, 2017, 7(4): 310-319. https://doi.org/10.12677/csa.2017.74038

本文提出了一种基于UEFI固件的操作系统完整性度量机制，该方案主要在UEFI BIOS启动过程中，利用TCM芯片的加密，认证和Hash运算等技术，对操作系统内核进行完整性度量，能够有效保护内核以及操作系统的安全。

关键词

UEFI，操作系统内核，TCM，完整性度量

1. 引言

操作系统是计算机硬件和软件的纽带，是应用软件运行的基础环境，而内核的安全性是操作系统安全的核心问题，若操作系统内核受到攻击，会对操作系统以及应用程序造成重大的威胁。

文献[1]中指出了可信验证包括启动验证和运行时验证两种类型，启动验证来源于认证启动[2]，认证启动保证了操作系统开机启动时是安全可信的，如果开机启动时不能够保证操作系统的安全性，则开机启动之后的验证将不再有意义。如何在开机启动时保护操作系统的完整性，防止恶意程序攻击内核系统是可信计算[3] [4]亟待解决的问题。本文通过在 UEFI BIOS 启动的过程中，基于 TCM [5]芯片的加密和认证等功能，完成对操作系统内核文件的完整性度量，从而保证了操作系统内核的完整性和安全性。

IMA [6] (IBM integrity measurement)提出了一种有效的度量方法，该方法是在加载动态链接库或内核模块时，对用到的一些关键数据和信息进行度量，并将度量的结果扩展到通过 TPM 来保护的信任链中。IMA 架构要求任何软件在加载的时候都要被度量，而用户可能只关心某一部分的完整性，对于这些用户来说 IMA 度量会有冗余，会降低系统的效率[2]。

而本文提出的基于 UEFI [7]的操作系统完整性度量方法只是对操作系统的内核进行完整性度量，而且适用于一台计算机装一个或者多个操作系统的情况。该方法是以硬件芯片作为起点，度量在 BIOS 中执行，增加了其自身的安全性，相对于那些在应用层做度量的方法更安全。而且将被检测的对象分离开，更能够有效的防止篡改和不可旁路性，在实现技术上，通过一台安全管理服务器来管理度量初始值的安全性，更简单方便。

UEFI (United Extensible Firmware Interface)是 Intel 联合业界采用开源方式共同制定推出的规范[8]。UEFI BIOS 定义了操作系统和平台固件之间的接口标准，是运行在操作系统和硬件之间的一个新的模型[9]。UEFI 较 BIOS 有可编程性好，可扩展性好，安全性高等优点，所以能够很快速的取代传统 BIOS，成为新一代的趋势。

2. 基于 UEFI 的操作系统完整性度量的总体设计

基于 UEFI 的操作系统完整性度量主要包括两部分，一部分是需要度量的客户机，另一部分是进行安全管理的服务器。整个系统的流程图如图 1 所示，客户机首先在 UEFI BIOS 启动的时候获取操作系统的内核以及版本，对操作系统内核做 Hash 运算，然后判断本地存储的 OsKernalFile 文件是否被篡改，如果被篡改，则将 OsKernalFile 文件里的内容清空后向服务器发送请求，要求服务器发送该版本的安全内核文件；如果没有被篡改，再查看本机中是否存储该操作系统的安全内核 Hash 值以及该 Hash 值是否失

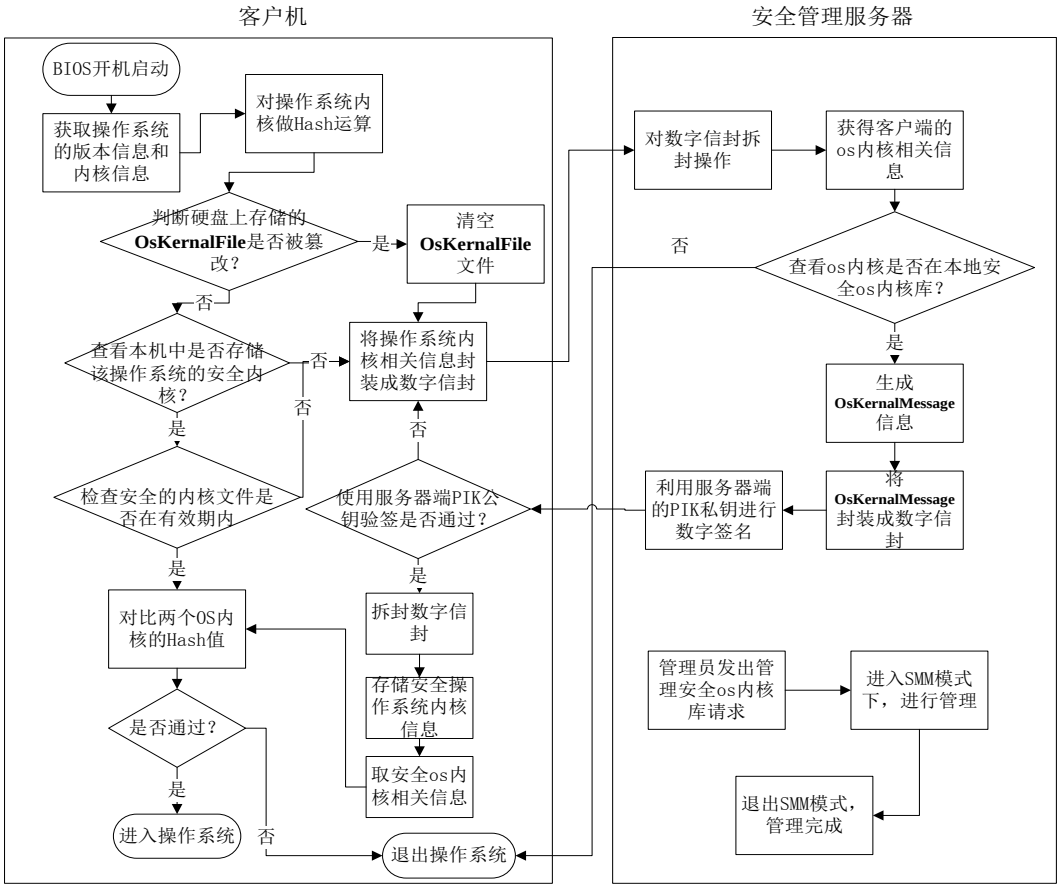


Figure 1. System flow diagram
图 1. 系统流程图

效，如果存在该操作系统的安全内核 Hash 值并且该 Hash 值在有效期内，则取出该安全内核 Hash 值，与开机启动时获取的操作系统内核 Hash 值进行对比，对比通过，进入到操作系统，对比不通过，直接关机；如果不存在该操作系统的安全内核 Hash 值或者该 Hash 值已经失效，则向服务器发送请求，要求服务器发送该版本的安全内核文件。客户机向服务器发送的内容主要包括操作系统的内核 Hash 值和版本信息，安全管理服务器通过通讯模块接收请求，对客户机进行身份的认证，然后得到操作系统内核镜像文件 Hash 值以及版本信息，进一步跟 DB 里存放的安全的操作系统内核镜像 Hash 值进行对比，对比通过之后，将该版本的操作系统内核相关的信息(内核 Hash 值，版本信息和有效期)发送给客户机，客户机接收到该信息之后，将该条信息更新到 OsKernalFile 文件，将整个 OsKernalFile 文件的 Hash 值更新到 TCM 上。

OsKernalFile 文件存储的是客户端装的所有的操作系统内核相关的信息，每一条记录是一个操作系统相关的信息，包括内核的有效期，内核的版本信息和内核的 Hash 值。每次开机，要先对整个文件做 Hash 运算，再跟 TCM 中存储的整个文件的 Hash 做对比，对比通过，说明 OsKernalFile 文件没有被篡改，对比不通过，首先清空该文件，然后再请求服务器重新发送安全内核的 Hash 值。

本系统使用 SM3 杂凑算法[10]产生 Hash 值，SM3 算法是国家密码管理局编制的商用算法，适用于商用密码应用中的数字签名和验证、消息认证码的生成与验证以及随机数的生成，可满足多种密码应用的安全需求。SM3 密码杂凑算法的设计主要遵循的原则有 3 点，第一点是能够有效抵抗比特追踪法及其他分析方法；第二点是软硬件实现需求合理；第三点是在保障安全性的前提下综合性能指标与 SHA-256

同等条件下相当[11]。

安全管理服务器中的 DB 是专门用于存储所有安全的操作系统内核镜像文件的相关信息，包括系统的版本和内核的 Hash 值。在接收验证请求时，首先会对客户端发过来的操作系统内核镜像进行对比，如果在 DB 里能够找到该镜像文件，说明该镜像文件是安全的，将该安全的内核镜像文件发送给客户机，所以要想保证整个过程的一个安全性，首先要确保存放在安全管理服务器上的 DB 的安全性，如果用户攻击了 DB，修改了部分操作系统内核镜像，那么在运行验证内核安全性模块的时候就不会通过，整个过程将会终止，针对这个问题，本系统采取的措施是在硬盘上存储加密的 DB，秘钥存放到 TCM 上，每次修改该 DB 时，都要通过 SMI 中断进入 SMM 模式，在 BIOS 层修改。

整个系统的框架图如图 2 所示，客户机主要包括发送请求模块，安全内核管理模块，信息处理和度量模块，安全管理服务器包括通讯模块，内核安全性验证模块，数据管理模块，发送安全内核相关信息模块。客户机是在操作系统启动之前的 UEFI BIOS 启动的时候进行的一系列的操作，安全管理服务器是在操作系统起来之后的应用层进行的，在 UEFI BIOS 启动的 DXE 阶段，就已经能够进行网络传输，所以向安全管理服务器发送请求模块和安全管理服务器向客户机发送安全内核模块都具有可行性。

3. 基于 UEFI 固件的操作系统完整性度量的模块设计

3.1. 客户端模块的设计

客户机从本地获取安全的系统内核 Hash 值，通过跟自己机器中的内核 Hash 值对比，决定是否开关机，在本地没有内核 Hash 值时需要向服务器发出请求，要求服务器将安全的内核 Hash 值存储到本机中。如图 3 所示，客户机包括四个模块，分别是请求发送模块，安全内核文件管理模块，信息处理模块，度量模块，其中安全内核管理模块又包括文件的存储管理，失效检测和读取管理 3 个模块。

1) 请求发送模块

该模块主要是用来向服务器端发送请求，请求服务器端将某种版本的操作系统内核文件传输给自己。本地存储的 OsKernalFile 文件被篡改，或不存在该内核 Hash 值或存在但是不在有效期内时调用该模块。

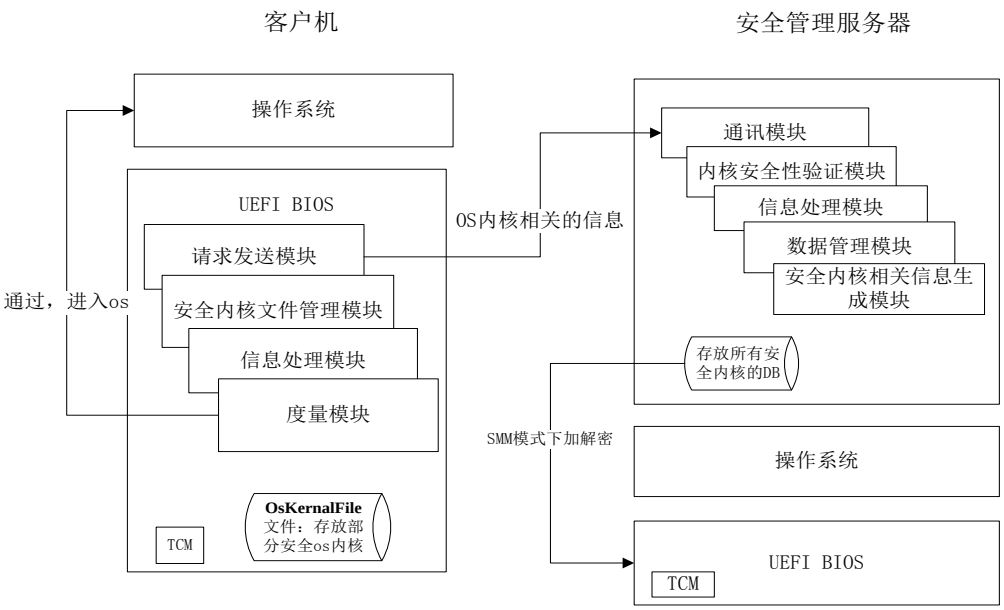


Figure 2. The overall system architecture diagram
图 2. 系统的总体框架图

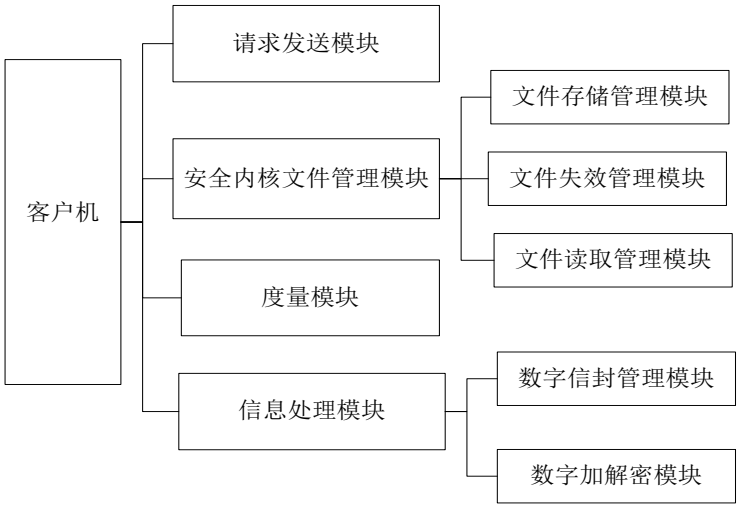


Figure 3. The client organization chart
图 3. 客户端的组织结构图

2) 安全内核管理模块

文件存储管理模块：主要用来存储安全管理服务器端发送过来的安全操作系统内核相关信息的模块。在存储该模块的时候，首先将该条消息存储到硬盘上，然后对整个文件进行杂凑运算得到新的文件 Hash 值，将整个文件 Hash 值更新到 TCM 中。

文件失效检测模块：该模块主要实现检测本机中已存在的安全操作系统内核 Hash 值是否失效，如果失效，调用请求发送模块。主要是通过有效期来判断，首先获得当前的日期，再跟服务器端发来的文件的有效期进行对比。

文件读取管理模块：该模块是用来实现对安全操作系统内核 Hash 值的读取，但是在读取信息之前，需要先确定该 Hash 值是不是被篡改过，需要将存储到本机硬盘上的安全操作系统内核 Hash 值文件在进行杂凑运算，将生成的整个文件的 Hash 值与 TCM 中已存的文件 Hash 值进行对比，对比通过之后，在对硬盘上的安全操作系统内核 Hash 值进行读取操作。

3) 信息处理模块

数据加解密模块：该模块主要实现的是 TCM 相关的一些算法，包括 SM3 杂凑算法，平台身份秘钥 PIK (本质上是一对 SM2 秘钥)。

数字信封解封管理模块：该模块主要实现的功能主要包括两个：一个是对客户端向安全管理服务器端发送信息的封装操作，另一个是对服务器端向客户端发送数据的解封操作。

4) 度量模块

在 BIOS 启动的时候，首先获取 os 内核镜像文件，使用 SM3 杂凑算法产生 Hash 值，调用文件读取管理模块读取文件中的安全 os 内核 Hash 值，将这两个 Hash 值对比，从而决定操作系统的启动与关闭。

3.2. 安全服务器端的设计

服务器端接收到请求之后，将客户端发送过来的内核 Hash 值与本地安全内核 DB 中的同一版本的 Hash 值进行对比，通过之后给客户端发送安全的操作系统内核 Hash 值相关信息。如图 4 所示，服务器端主要包括通讯模块，内核安全性验证模块，信息处理模块，数据管理模块，安全内核相关信息生成模块 5 个模块，其中数据管理模块又包括数据增加模块，数据查找模块和数据删除模块。

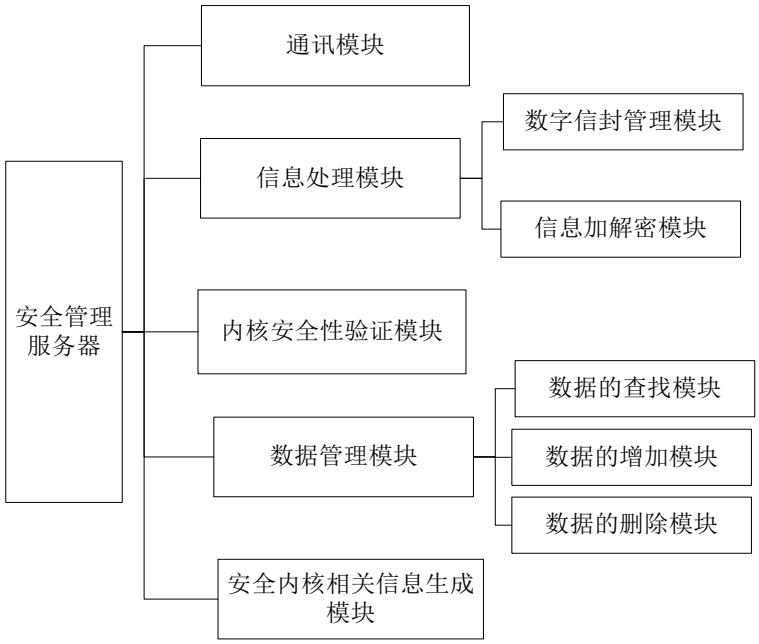


Figure 4. The server module diagram
图 4. 服务器端模块图

1) 通讯模块

主要利用 UEFI 中的网络协议相关的知识实现接收验证请求，以及与客户端的通讯功能。

2) 内核安全性验证模块

该模块主要实现的是将客户端发送的操作系统内核 Hash 值与本地存储的同一版本 os 内核 Hash 值做对比，如果两个 Hash 值一致，说明该客户端发送的内核信息是安全的，如果不一致，则返回一个错误的信息。由于在安全服务器端存放的是已经加密的所有安全操作系统内核 Hash 值文件，所以需要先对其解密。

3) 数据管理模块

该模块是供安全服务器管理员使用的模块，主要实现的是对服务器中的安全操作系统内核 Hash 值文件的管理，包括数据的查找，增加和删除。在服务器端存放的是使用 SM2 公钥对整个文件加密过的，而秘钥存放在 TCM 中。

数据查找模块：该模块主要实现的是内核 Hash 值的获取，首先利用 SMI 中断进入到 SMM 模式，然后在 SMM 模式下获取 SM2 私钥信息，在 BIOS 下对文件进行解密，退出 SMM 模式，找到想要的内核 Hash 值，然后进入到 SMM 模式，在 BIOS 层对文件使用 SM2 公钥加密，最后退出 SMM 模式[9]。

数据增加模块：该模块主要实现的是管理员在 SMM 模式下增加一条或者多条安全的操作系统内核 Hash 值记录，同样需要对安全的操作系统内核 Hash 值整个文件在 UEFI BIOS 层进行解密和加密。

数据删除模块：该模块实现的过程与上述两个模块的实现过程类似，只不过是在对文件的操作不同。在 UEFI BIOS 层进行解密，删除一条信息之后，在对整个存放内核信息的文件进行加密。

4) 安全内核相关信息生成模块

该模块是在跟客户机发送的内核信息比较通过之后调用的模块，主要是用来生成安全内核文件的一些信息，在将这些信息发送到客户机端，生成的安全操作系统内核信息(OsKernalMessage)的格式如图 5 所示，设置分隔符的作用主要是为了便于程序的读取。

内核有效期	分隔符	内核版本信息	分隔符	内核的 hash 值

Figure 5. File format diagram from the server
图 5. 服务器端发送文件格式图

5) 信息处理模块

数字信封封装模块：该模块主要实现的是对客户端发过来的数字信封进行拆封以及将要发送给客户机的信息进行封装。

信息加解密模块：同客户端的信息加解密一样，调用 TCM 相关的算法完成对数据的加解密操作。

4. 基于 UEFI 固件的操作系统完整性度量流程设计及算法实现

4.1. 客户端存储度量初始值的流程及算法实现

Client：

1. 首先获得操作系统内核信息 OsKernal 和版本信息 OsVersion；

2. 对操作系统内核信息做 SM3 杂凑运算：OsKernalHash = SM3 (OsKernal)；

3. 对本地存储的 OsKernalFile 文件做 hash 运算得到 OsKernalFileHash，跟 TCM 中存储的 OsKernalFileHashTCM 值对比看是否一致，如果不一致，或者一致，但是本地没有存储该操作系统相关的信息，或者存储但是已经失效，拼接系统内核信息和版本信息：OsKernalMsg = OsVersion|OsKernal；

4. 将操作系统内核相关的信息封装成数字信封，封装数字信封的过程如下：

1) 产生一个对称的秘钥，用对称的秘钥对 OsKernalMsg 信息进行加密，利用 TCM 中的随机数生成器生成一个 128 位的数据，将这个数据作为 SMS4 的对称秘钥，利用 SMS4 对称算法对 OsKernalMessage 加密。

encOsKernalMessage = SMS4_Encrypt (SMS4Key，OsKernalMsg)；

2) 利用服务器的 PIK 公钥 Server_PIK_public 对 SMS4Key 对称秘钥进行加密，得到加密的对称秘钥 EncryptSMS4Key，EncryptSMS4Key = SM2_Encrypt (Server_PIK_public，SMS4Key)；

3) 将加密的操作系统内核信息 encOsKernalMsg 和加密秘钥 EncryptSMS4Key 封装成数字信封。

DigitalEnvelop = encOsKernalMsg|EncryptSMS4Key。

5. 将封装好的 DigitalEnvelop 发送给安全管理服务器。

Server。

6. 安全管理服务器接收到 DigitalEnvelop 信息之后，对 DigitalEnvelop 进行拆封运算，具体的流程如下：

1) 拆封数据信封，分别得到 encOsKernalMsg 信息和 EncryptSMS4Key 信息。

2) 使用自己的 PIK 私钥 Server_PIK_private 解密，得到对称秘钥 SMS4Key。

SMS4Key = SM2_Decode (Server_PIK_private，EncryptSMS4Key)。

3) 使用对称秘钥 SMS4Key 解密加密的操作系统内核信息 encOsKernalMsg 得到 OsKernalMsg。

OsKernalMsg = SMS4_Decoode (SMS4Key,encOsKernalMsg)。

7. 查看 os 内核信息 OsKernalMsg 是否在本地库中，String SerchMessage (File f，String str)函数实现的是查找文件中是否存在 s 功能，查找成功返回数据库中的这条记录，并进行下一步信息的封装操作；查找失败返回 NOT_EXIT,并将该消息发送给客户端，客户端收到该消息之后，执行关机命令。

8. 将整条 os 内核记录与有效期进行拼接，得到 OsKernalMessage。

OsKernalMessage = indate|OsVersion|OsKernalHash。

9. 对 OsKernalMessage 进行封装，封装的过程如下，其中 Client_PIK_public 是客户端 PIK 公钥，SMS4Key 是一个 128 位的随机数：

encOsKernalMessage = SMS4_Encrypt (SMS4Key，OsKernalMessage)；

EncryptSMS4Key = SM2_Encrypt(Client_PIK_public, SMS4Key)；

DigitalEnvelop = OsKernalMessage|EncryptSMS4Key。

10. 使用服务器端的 PIK 私钥数字签名，并将签名信息发送给客户端。

SignatureDigitalEnvelop = SM2_Signature (Server_PIK_private，DigitalEnvelop)；

Client。

11. 客户端用服务器端的 PIK 公钥验证签名，对安全管理服务器进行身份认证。

DigitalEnvelop = SM2_VerifySignature (Server_PIK_public，SignatureDigitalEnvelop)。

12. 拆封数字信封，得到 OsKernalMessage 信息，并将该信息存储到硬盘上。

SMS4Key = SM2_Decode (Client_PIK_private，EncryptSMS4Key)；

OsKernalMessage = SMS4_Decoode (SMS4Key，encOsKernalMessage)。

13. 将存储 OsKernalMessage 的整个文件 OsKernalFile 做 hash 运算，得到文件 Hash 值，在将该信息更新到 TCM 中。

OsKernalFileHash = SM3 (OsKernalMessage)。

4.2. 客户机完整性度量的流程及算法实现

1. 客户机开机启动获取内核相关信息并对内核信息进行 hash 运算，得到内核的 hash 值。

OsKernalHash = SM3 (OsKernal)。

2. 对本地存储的 OsKernalFile 文件做 hash 运算得到 OsKernalFileHash，跟 TCM 中存储的 OsKernalFileHashTCM 值是否一致，如果不一致，则清空 OsKernalFile 文件里的内容，然后调用客户端存储度量初始值模块，具体的步骤在上一小节已详细介绍；如果一致，则说明硬盘上存储的没有被篡改。

OsKernalFileHash = SM3 (OsKernalFile)。

3. 在本地 OsKernalFile 文件中取出该操作系统的 OsKernalHash’，将 OsKernalHash 信息与 OsKernalHash’做对比，如果一致，说明内核没有篡改，进入到操作系统；如果不一致，说明内核被篡改，执行关机命令。

5. 系统安全性的核心问题

本系统是在 UEFI BIOS 启动的时候根据本地存放的安全内核的文件，对操作系统内核文件进行完整性度量，度量通过开机启动，要想保证整个过程的安全性，需要做到以下 3 点：

1) 客户端 OsKernalFile 文件的安全存储问题。OsKernalFile 文件中存储的是本计算机中装的所有操作系统内核 Hash 相关的信息，是度量的初始值，如果 OsKernalFile 文件被篡改，度量会失去意义。为了保证该文件的安全性，需要将整个的 OsKernalFile 文件的 Hash 值存储到 TCM 中。在每次读取该文件时，都需要将 OsKernalFile 文件的 Hash 值与 TCM 中的 Hash 值做对比，对比通过，说明存储到该硬盘上的 OsKernalFile 文件是安全的。

2) 服务器端 DB 的安全存储问题。在安全管理服务器端存放着所有安全可信的操作系统内核镜像的文件，每次都要判断客户机发过来的操作系统内核文件是否在该 DB 中，如果有，说明该操作系统内核镜像文件是安全的。如何保障 DB 的安全性是一件极其重要的工作，该系统采用的是硬件防护，将存储

和数据加解密进行有效的隔离，在操作系统上存放的是已加密的文件，在加密和解密的时候在可信的UEFI BIOS 完成。

3) 传输数据的安全性。数据在网络中传输，为了保证数据的安全性，一般采取的是对需要传输的数据进行加密，常见的一些传输加密技术包括数字信封技术，SSL 协议等技术，由于客户机的固件只是在开机启动的时候工作，而 SSL 在传输数据的时候需要三次握手才能够确认其身份，所以在该系统中 SSL 协议不适用，所以本系统采用了数字信封的方式来保证数据传输的安全性。

6. 实验及结果

该系统的安全性在上一章节已经论述，下面我们将对整个系统的功能进行验证：

1) 在一台计算机中装一个 win8 的操作系统，调用该系统，能够正确的对操作系统的完整性度量，计算机正常开机。

2) 在重新开机启动该计算机，也能够实现对操作系统的完整性度量，计算机正常开机。

3) 修改操作系统内核文件，在开机时进行度量，该操作系统的内核摘要跟本地存储的内核度量初始值不一致，度量不通过，计算机不会开机。

4) 修改客户机本地硬盘中存储的内核度量初始值，在开机启动时度量，对 OsKernalFile 文件做 SM3 杂凑，跟 TCM 中存储的 OsKernalFile 文件 hash 不一致，会请求服务器重新发送该操作系统内核信息。

5) 伪装成服务器的身份，发送给客户机不安全的内核信息，客户机收到该消息后，对服务器进行身份认证不通过，会再次请求服务器发送安全内核信息。

7. 总结

OS 的安全性是计算机安全的核心，而内核的安全性是操作系统安全的核心，本文提出了一种基于UEFI BIOS 的操作系统完整性度量机制，实现了在 UEFI BIOS 的启动过程中对操作系统内核镜像文件的完整性度量，保证了操作系统内核的安全性。在实现的整个过程中最主要的是要保证操作系统度量初始值的完整性和安全性，这是该度量机制的难点也是关键点，如果不能够保证度量初始值的正确性，该系统的度量将不再有意义。

该机制会对电脑的开机速度有较大的影响，在后续的工作中，会对该问题进行改进。本机制只是实现对操作系统完整性的度量，但是不能够保证电脑安装操作系统时操作系统的安全性，下一步可以增加操作系统的远程安装功能，进一步保证电脑中操作系统的安全性。

参考文献 (References)

[1] 胡浩, 张敏, 冯登国. 基于信息流的可信操作系统度量架构[J]. 中国科学院大学学报, 2009, 26(4): 522-529.

[2] Smith, S.W. (2004) Outbound Authentication for Programmable Secure Coprocessors. *International Journal of Information Security*, **3**, 28-41. https://doi.org/10.1007/s10207-004-0033-0

[3] Trusted, B.G. (2010) Computing Group. Trusted Platform Module (TPM) Specifications.

[4] Parno, B., Mccune, J.M. and Perrig, A. (2010) Bootstrapping Trust in Commodity Computers. Security and Privacy. IEEE,:414-429.

[5] 国家密码管理局. 可信计算密码支撑平台功能与接口规范[S]. 国家密码管理局, 2007.

[6] Sailer, R., Zhang, X., Jaeger, T., *et al*. (2004) Design and Implementation of a TCG-Based Integrity Measurement Architecture. *Conference on USENIX Security Symposium*, San Diego, 9-13 August 2004, 16.

[7] 戴正华. UEFI 原理与编程[M]. 北京: 机械工业出版社, 2015.

[8] The Unified EFI Forum (2011) Unified Extensible Firmware Interface Specification Version 2.3.1. http://www.uefi.org

[9] 周艺华, 王伟, 王冠, 等. 基于固件的远程身份认证[J]. 信息安全与技术, 2016, 7(3): 35-39.

[10] 国家密码管理局. SM3 密码杂凑算法(SM3 Cryptographic Hash Algorithm)[M]. 国家密码管理局, 2010. http://www.oscca.gov.cn/UpFile/20101222141857786.pdf

[11] 王小云, 于红波. SM3 密码杂凑算法[J]. 信息安全研究, 2016, 2(11): 983-994.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 320-328
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74039

Hans汉斯

Design and Implementation of Attack Detection System Based on UEFI Firmware

Yihua Zhou[1], Yang Liu[1,2], Guan Wang[1,2], Liang Sun[3]

[1]College of Computer Science, Beijing University of Technology, Beijing
[2]Key Laboratory of Trustworthy Computing in Beijing, Beijing
[3]ZD Technologies Ltd. (Beijing), Beijing
Email: zhouyh@bjut.edu.cn, 1005415619@qq.com

Received: Apr. 4th, 2017; accepted: Apr. 17th, 2017; published: Apr. 25th, 2017

Abstract

With the continuous development of network attack technology, attack against the firmware has emerged, causing a huge threat to the computer. This paper analyzes the function, characteristics and security threat of UEFI BIOS in detail, explaining the importance of firmware security. The malicious codes in the firmware are hard to be found and eliminated by the antivirus software. This paper has designed and implemented an attack detection system based on UEFI firmware by getting and analyzing the information about the firmware.

Keywords

UEFI, BIOS, Attack Detection

基于UEFI固件的攻击检测系统的设计与实现

周艺华[1]，刘　阳[1,2]，王　冠[1,2]，孙　亮[3]

[1]北京工业大学 计算机学院，北京
[2]可信计算北京市重点实验室，北京
[3]中电科技(北京)有限公司，北京
Email: zhouyh@bjut.edu.cn, 1005415619@qq.com

收稿日期：2017年4月4日；录用日期：2017年4月17日；发布日期：2017年4月25日

摘　要

随着网络攻击技术的不断发展，针对固件的攻击已经出现，对计算机造成了巨大的威胁。本文详细分析

文章引用: 周艺华, 刘阳, 王冠, 孙亮. 基于UEFI固件的攻击检测系统的设计与实现[J]. 计算机科学与应用, 2017, 7(4): 320-328. https://doi.org/10.12677/csa.2017.74039

了UEFI BIOS的功能、特点及安全威胁，诠释了固件安全的重要性。固件中的恶意代码难以被杀毒软件发现和消除。本文通过对固件层信息的获取和解析，设计与实现了一种基于UEFI固件的攻击检测系统。

关键词

UEFI，BIOS，攻击检测

1. 引言

固件(Firmware)是计算机上电后首先执行的一组程序，运行在计算机底层，固化在 Flash 芯片中，常用于完成配置硬件设备，为操作系统提供硬件操作接口，引导操作系统的启动等功能，被业界称为是连接计算机基础硬件和系统软件的“桥梁”。UEFI (Unified Extensible Firmware Interface，统一可扩展固件接口)是新的固件标准，目前被业界广泛使用。UEFI 为固件和操作系统之间的接口制定了新的标准和规范，并为启动操作系统提供了所需的标准环境[1]。

BIOS (Basic Input/Output System)的主要功能[2]包括：1)开机自检，计算机启动时，BIOS 掌握着硬件设备的控制权，首先检查 CPU 是否工作正常，还包括定时器、可编程中断控制器、基本内存、显卡等的自检工作。2)系统初始化，初始化相应的硬件设备，并对外围设备进行驱动。3)BIOS 提供了系统设置功能，用户在进入操作系统之前可以进入 BIOS 设置界面，对系统的一些参数进行设置，如光盘/硬盘/USB 引导顺序、端口使能/禁用等。4)引导操作系统启动，BIOS 按照设置中保存的启动顺序读入操作系统引导代码，从而完成操作系统的加载启动。作为连接硬件和操作系统的枢纽，BIOS 的安全对计算机至关重要。Flash 芯片已经允许在操作系统运行过程中，通过特定的软件技术直接对 BIOS 中的内容进行刷新写入[3]，给用户和系统维护带来了便利，但与此同时也引入了很多的安全风险。

早期，众所周知的 CIH 病毒[4]就是直接向 BIOS 固件芯片和硬盘中写入乱码，篡改了 BIOS 中的正常代码，造成计算机无法启动，使用户数据严重破坏并无法还原。1999 年，Phoenix 公司提出并实现了通过在 BIOS 固件中嵌入一个代码模块 ILS [5] (Internet Launch System)，为用户自动连接网络，下载服务软件，启动 web 服务，后来此项目遭到了抵制，被终止。后来出现了 ACPI BIOS rootkit [6]和 PCI rootkit [7]，分别利用 ACPI 和 PCI 在固件中植入恶意代码。文献[8]根据 UEFI 自身和启动过程中存在的安全隐患，提出了基于 UEFI 存储攻击设备和劫持操作系统内核的两种攻击方式。BIOS 攻击的特点：不同于操作系统攻击和网络攻击，针对 BIOS 的攻击，在操作系统中是找不到的，在硬盘中也是看不到的，传统杀毒软件杀不掉，格式化硬盘或更换硬盘清不了，重装系统也无法去除。BIOS 恶意代码驻留在被攻击终端的计算机主板固件芯片中，难以检测和清除。随着信息技术的飞速发展，针对固件的攻击手段越来越多，计算机安全受到严重的威胁，BIOS 一旦被植入恶意代码，可能致使整个计算机瘫痪。因此，BIOS 攻击检测和防护成为当前亟待解决的问题。目前，对于 BIOS 攻击检测的研究和分析，国际国内的信息安全领域所做的相关工作还较少，固件安全领域的相关资料也较少，对开展固件攻击检测的研究和实现增加了一定的难度。

文献[9]提出了一种 BIOS 安全检测方法，即进行 BIOS 完整性度量。文献[10]提出对 BIOS 安全隐患

进行扫描，通过匹配安全隐患库特征码，判断 BIOS 是否存在隐患。本文研究 UEFI 固件技术并借鉴文献[9] [10]的检测方法，设计与实现了一个新的基于 UEFI 固件的攻击检测系统。本文针对 BIOS 镜像解析模块所采用的实现方式与以往不同。传统的 BIOS，对于不同的类型如 Award BIOS，AMI BIOS，Phoenix BIOS 等，各厂商的规范不尽相同，BIOS 镜像文件的格式处理和模块组成有很大差异，因此以往的 BIOS 安全检测系统通常是针对一种特定类型 BIOS 的研究与实现，有一定的局限性。目前，市面上新主板的 BIOS 普遍支持 UEFI 规范，不同的 BIOS 厂商都实现了 UEFI 规范所定义的接口。UEFI 规范定义了新的固件结构，BIOS 文件格式遵循 UEFI 组织发布的 PI 规范里定义的 Firmware Storage Specification。本文系统根据 UEFI 规范对 BIOS 镜像文件进行解析，设计与实现基于 UEFI 固件的攻击检测系统，增加了系统的通用性。而且，传统 BIOS 检测系统通常采用静态检测技术，本文系统客户端在 Windows 系统下开发，可以被广泛使用，不仅可以完成 BIOS 攻击检测工作，也可以使用户对 BIOS 有更深入的了解和认识。

2. 基于 UEFI 固件的攻击检测系统总体设计

本系统利用实验室已经实现的基于 UEFI 的固件木马攻击系统[11]来实施向 BIOS 中植入木马。固件攻击开发系统以原 X86 平台 BIOS 为基础，通过利用计算机 BIOS 中系统管理模式(SMM)的基本原理、接口设计以及 SMI 中断处理实现，可模拟对计算机固件的攻击。固件木马攻击系统网络拓扑图如图 1 所示。

固件木马攻击系统验证了木马能够被植入到 BIOS 中。基于固件木马攻击系统平台，本文所设计的系统主要通过获取 BIOS 镜像文件并对 BIOS 镜像进行解析，将 BIOS 采样和标准模块进行对比，对 BIOS 进行完整性度量，以此检测 BIOS 是否被篡改、是否受到木马攻击，并且检测出 BIOS 的哪一模块受到破坏。被篡改的模块可能隐藏着一定的安全漏洞，需要深入研究，提出防护措施。此外，本文系统利用 SMBIOS 实现了从固件层获取到计算机系统各组件的信息，包括 BIOS、主板、内存、CPU 及其他计算机设备的详细信息，对计算机硬件进行状态检测，便于用户监测系统状态，更好地进行攻击检测工作。整个固件攻击检测系统的总体架构如图 2 所示。

基于 UEFI 固件的攻击检测系统由客户端和服务器端组成，需要客户端和服务器端进行数据的通信。客户端按照功能划分总体上包含四个模块：BIOS 镜像解析模块，基准信息记录模块，计算机系统信息模块和检测分析模块。其中 BIOS 镜像解析模块是系统的核心模块和系统实现的关键环节，该模块包含的任务有：获取固件层 BIOS 镜像文件，对 BIOS 镜像进行模块分解和解压，并获取 BIOS 各个模块的 MD5

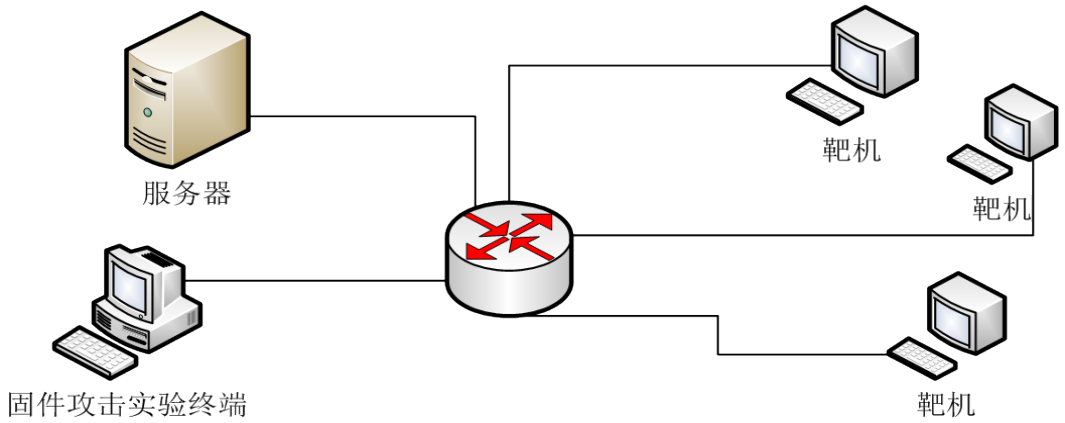


Figure 1. Firmware trojan attack system network topology
图 1. 固件木马攻击系统网络拓扑图

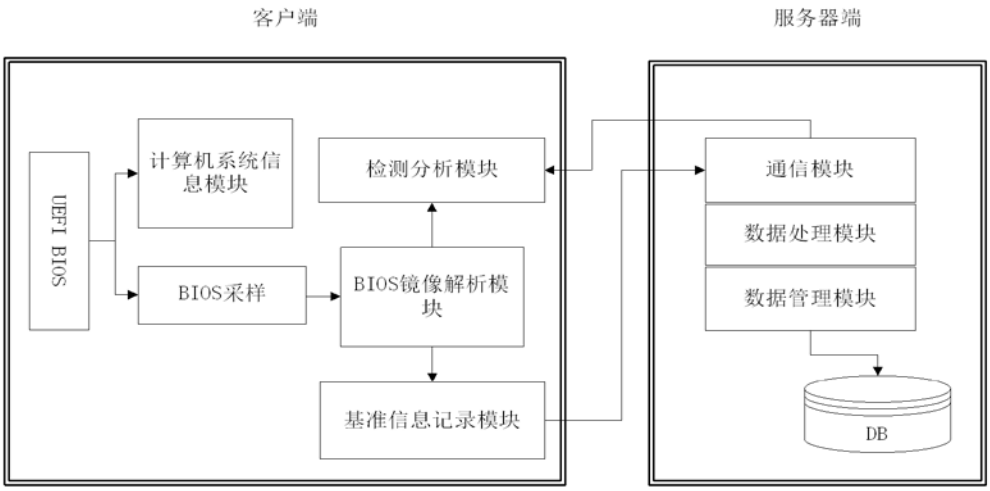


Figure 2. The overall framework of the firmware attack detection system
图 2. 固件攻击检测系统总体框架图

值。计算机系统信息模块显示了 BIOS、CPU、内存、缓存等计算机组件的详细信息。检测分析这一模块的功能是将解析后的被测 BIOS 文件与标准代码样本进行模块 MD5 值对比，检测 BIOS 是否被篡改，并具体显示出 BIOS 的哪一模块发生变化。服务器端利用通信模块与客户端进行通信，接收客户端发出的请求，并给出响应。数据处理模块是对客户端发出的请求进行处理，并对通信数据做加解密工作。木马攻击前的原 BIOS 叫做标准代码样本，当前不知道是否被攻击的 BIOS 是被测代码样本，被测代码与标准代码进行对比、检验，才能实现检测功能。因此，标准代码样本库需要安全地备份和存储，显然存储在本地并不安全，我们将 BIOS 标准代码远程存储在服务器上。服务器端通过数据管理模块对用户的 BIOS 标准样本和模块 MD5 进行统一的管理与存储。客户端记录基准信息时将标准 BIOS 文件和模块 MD5 值上传至服务器上。当系统调用检测分析模块时，服务器端会返回相应 BIOS 数据给客户端。

3. 基于 UEFI 固件的攻击检测系统流程设计

本系统的流程主要是获取到固件层 BIOS 镜像文件之后，将固件木马攻击之前的原 BIOS 镜像文件和当前被测 BIOS 镜像文件进行模块分解，由于 BIOS 实际是经过压缩、加密后存储在 Flash 芯片上的，所以需要对压缩存储的模块按照模块长度、压缩算法等进行解压缩，还原为最原始的 BIOS 二进制文件。对解析出的各个模块计算 MD5 消息摘要，将目标被测 BIOS 代码样本的各个模块分别与相对应的标准代码样本模块进行 MD5 值比较。如果全部一致，说明所有模块都没有变化，与原来代码一样，BIOS 并未受到攻击；如果有一个或多个模块对比不一致，说明 BIOS 受到攻击，MD5 消息摘要不一致的模块被破坏。固件攻击检测系统流程图如图 3 所示。

本系统实现的关键技术在于固件 BIOS 和操作系统之间的通信。UEFI BIOS 系统主要包括 8 个模块：UEFI BIOS 基础代码模块，硬件相关代码包括 CPU 代码、芯片组代码和设备代码，端口控制代码，兼容支持模块，UEFI OS 加载器，固件应用程序，文件系统驱动模块，SMI 中断处理模块。其中 SMI 中断处理模块以及文件系统驱动模块是 OS 和 BIOS 通信的桥梁，是本文系统实现的基础。

4. 基于 UEFI 固件的攻击检测系统的模块设计与实现

4.1. BIOS 镜像解析模块的设计与实现

BIOS 镜像解析模块需要完成的任务包括获取 BIOS 镜像文件，对 BIOS 进行模块分解，将压缩的部

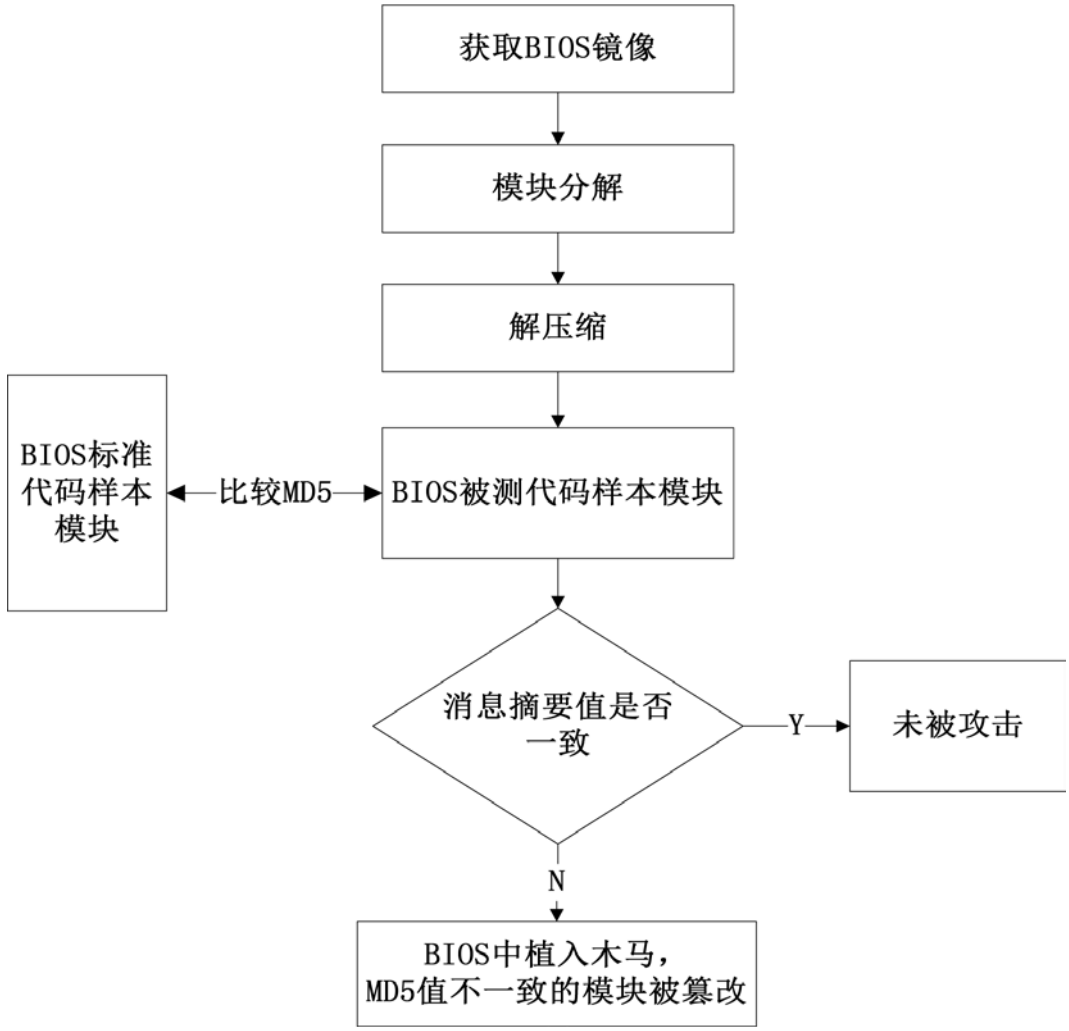


Figure 3. Firmware attack detection system flow chart
图 3. 固件攻击检测系统流程图

分进行解压缩得到最原始的 BIOS 二进制代码，并计算每个模块的 MD5 消息摘要。这一模块是整个 BIOS 攻击检测系统的最重要的部分，是实现检测、达到检测目标的基础。BIOS 在 Flash 芯片中是根据规范按照一定固件结构存储的，所有模块保存在固件结构单元中，此外，BIOS 厂商为了降低 BIOS 所需的存储空间，并增强代码的安全性，对 BIOS 的许多模块进行了压缩。因此，对 BIOS 模块进行攻击检测之前，首先需要将 BIOS 镜像文件分解，并对压缩的模块进行解压缩得到原始代码。

UEFI 作为广泛使用的新一代 BIOS 规范，为了方便 UEFI BIOS 的开发与扩展，为 Flash 芯片中 BIOS 镜像定义了新的存储结构。这种固件结构类似于一种文件系统，对固件设备文件的组织和存储进行了明确的规定，UEFI BIOS 的每个功能模块都存储在固件结构单元中。在这种规范中，整个 BIOS 被视为一个 FD (Firmware Device)，BIOS 镜像文件就是一个固件设备文件。FD (固件设备)是一片连续的存储区域，一个 FD 又分成多个逻辑区块，被称作固件卷。FV (Firmware Volume)是连续的格式化的存储空间，一个 FV 种又包含一个或多个 FF (Firmware File)，固件文件保存着 FV 中存储的数据和代码，是固件文件系统最重要的部分。而一个固件文件中又包含着固件文件段，Firmware File Sections 是不连续的段，是特定 File Type 里的独立部分。UEFI BIOS 镜像文件结构层次如图 4 所示。

固件文件系统(Firmware File System，FFS)描述了在固件卷FV上固件文件FF和空闲空间的组织关系，每个 Firmware Volume Image 包含 Header，FFS Image 和 Free Space，每个 FFS Image 包括 Header 和 File Sections。FFS 首部的格式如图 5 所示。其中 Type 字段描述了固件文件所属类型，固件文件类型有：PEI 核心固件文件、DXE 核心固件文件、UEFI 驱动文件等多种类型文件。

基于 UEFI 的固件攻击检测系统，通过读取 BIOS 所对应的内存地址获取 BIOS 镜像文件，得到 BIOS 镜像文件之后，利用 UEFI 固件文件规范对 BIOS 文件进行模块分解，对于压缩存储的 FV，根据固件文件段中描述的压缩方式，反向操作对其进行解压缩，便可得到 BIOS 二进制文件，完成 BIOS 镜像文件的解析。系统 BIOS 镜像解析效果如图 6 所示。

4.2. 计算机系统信息模块的设计与实现

计算机系统信息这一模块是对计算机中重要组件的详细信息进行展示，用来查看和检测硬件状态，便于用户监视系统状态。从此模块中可以获取到的信息有：BIOS 基本信息，系统信息，系统外围，处理器，缓存，端口连接器，系统插槽，物理存储阵列，内存阵列映射地址以及系统引导的详细信息。这些硬件信息是通过编程从 SMBIOS 中得到的，并以直观、易读的文本格式显示出来。SMBIOS (System

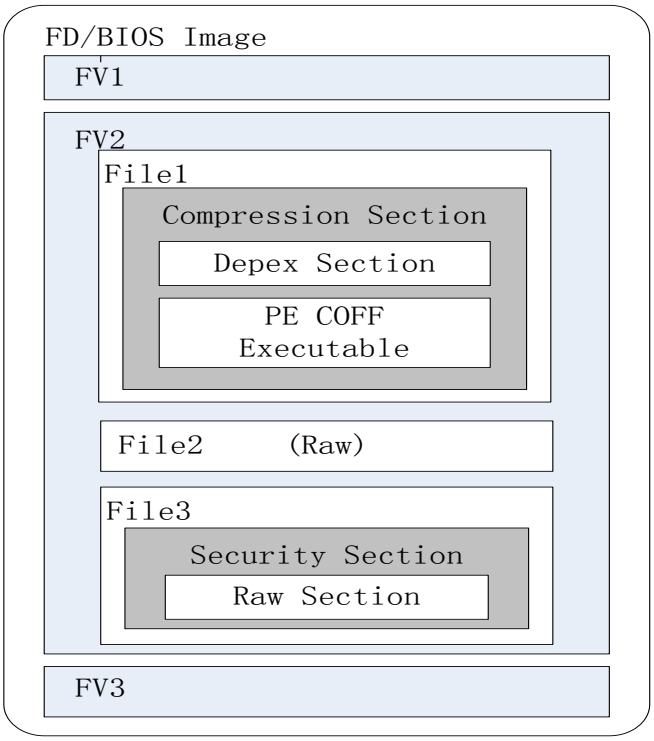


Figure 4. UEFI BIOS image file hierarchy diagram
图 4. UEFI BIOS 镜像文件结构层次图

Name		
IntergrityCheck	Type	Attributes
Size		State

Figure 5. FFS header format
图 5. FFS 首部格式

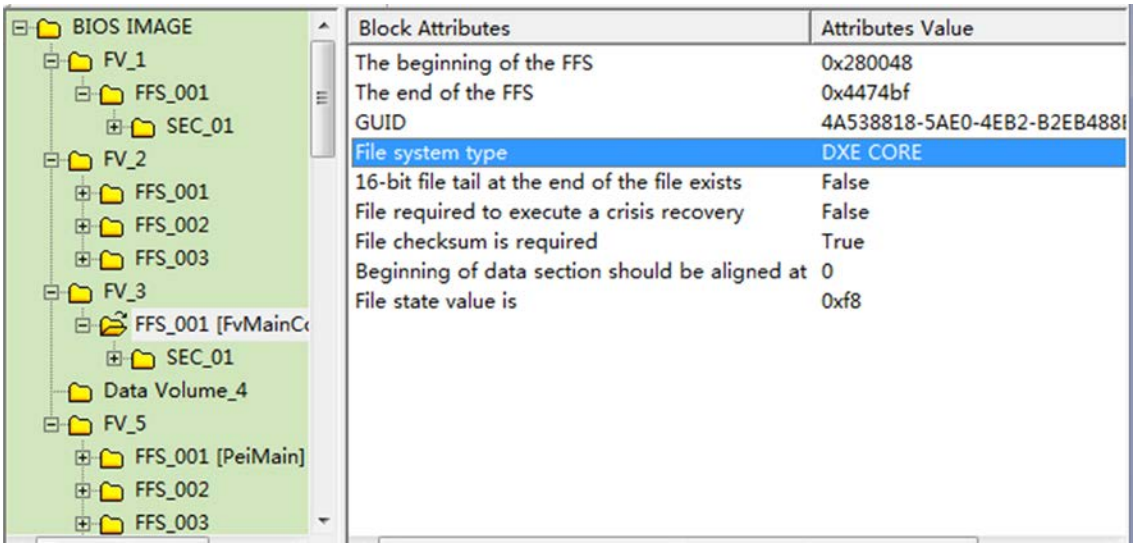


Figure 6. The BIOS image analytical results
图 6. BIOS 镜像解析效果

Management BIOS)参考规范描述了主板和系统供应商通过扩展英特尔架构系统上的 BIOS 接口如何将其产品的管理信息以标准格式呈现。DMI (Desktop Management Interface)是用来收集计算机系统自身以及其外围设备相关数据信息的管理系统，遵循 SMBIOS 规范，并将收集到的数据信息存在 BIOS 中。因此，我们可以从 SMBIOS 中获取到想要的计算机组件信息。在 SMBIOS 2.7 版本规范中为 SMBIOS 信息定义的唯一访问方法是基于表结构的方法。

编程实现从 SMBIOS 读取计算机系统信息的步骤：首先寻找 EPS 表，获取 SMBIOS 表的入口地址。SMBIOS 中数据的存储结构是由 EPS 表和 SMBIOS 表一起来描述的，EPS 表结构如表 1 所示。

在 EPS 表的 16H 处存储的是 SMBIOS 表的总长度，表的 18H 处存储了 SMBIOS 表的起始物理内存地址。根据 EPS 表结构，可以看出，找到“_SM_”并找到“_DMI_”，便找到了 EPS 表。而“_SM_”和“_DMI_”是用固定的 ASCII 码代表，很容易找到。找到 EPS 表后，根据表的 18H 处就可以得到 SMBIOS 的内存地址。SMBIOS 的首地址是 Type 0 的存储地址，Type 0 描述的是 BIOS 的基本信息。对 Type 0 表进行解析并以一定格式读出来，便获得了想要的 BIOS 信息。SMBIOS 表是由多个不同类型的表结构构成的，每一个表结构代表一种计算机组件的信息或者其他类型的系统信息。SMBIOS 2.7 版本描述了 Type 0 - 42，126 和 127 的表结构。本文所设计的系统只选取其中部分常见的计算机组件信息进行展示。Type 0 存储的是 BIOS 信息，Type 1 是系统信息，Type 3 是系统外围信息，Type 4 是处理器信息，Type 7 是缓存信息，Type 8 是端口连接器，Type 9 是系统插槽，Type 16 是物理内存阵列。Type 19 是内存阵列映射地址，Type 32 是系统引导信息。每个 SMBIOS 表都具有相同的头结构，每个表又划分为格式区域和字符串区域，字符串区域紧跟在格式区域之后。每个 Type 表的区域内容有所不同，部分 BIOS 表的格式区域见表 2 所示。

在 BIOS 表的 00H 处代表 Type 号，01H 中描述了其格式区域的长度，在格式区域之后是其字符串区域。根据表结构得知，字符串区域的第一个字符串描述的是 BIOS 厂商信息，第二个字符串代表的是 BIOS 版本，第三个字符串是 BIOS 发布日期。每一个字符串以 00H 结束，字符串区域以 0000H 结束。接着是 Type 1 的内容，获取其他 Type 信息的方法和 BIOS 信息类似，本文不再赘述。本文以一台实验机为例，直接获取到的 BIOS 信息的展示比较混乱，本文系统通过解析，将 BIOS 的基本信息以直观、易读的格式化文本形式显示出来，效果如图 7 所示。

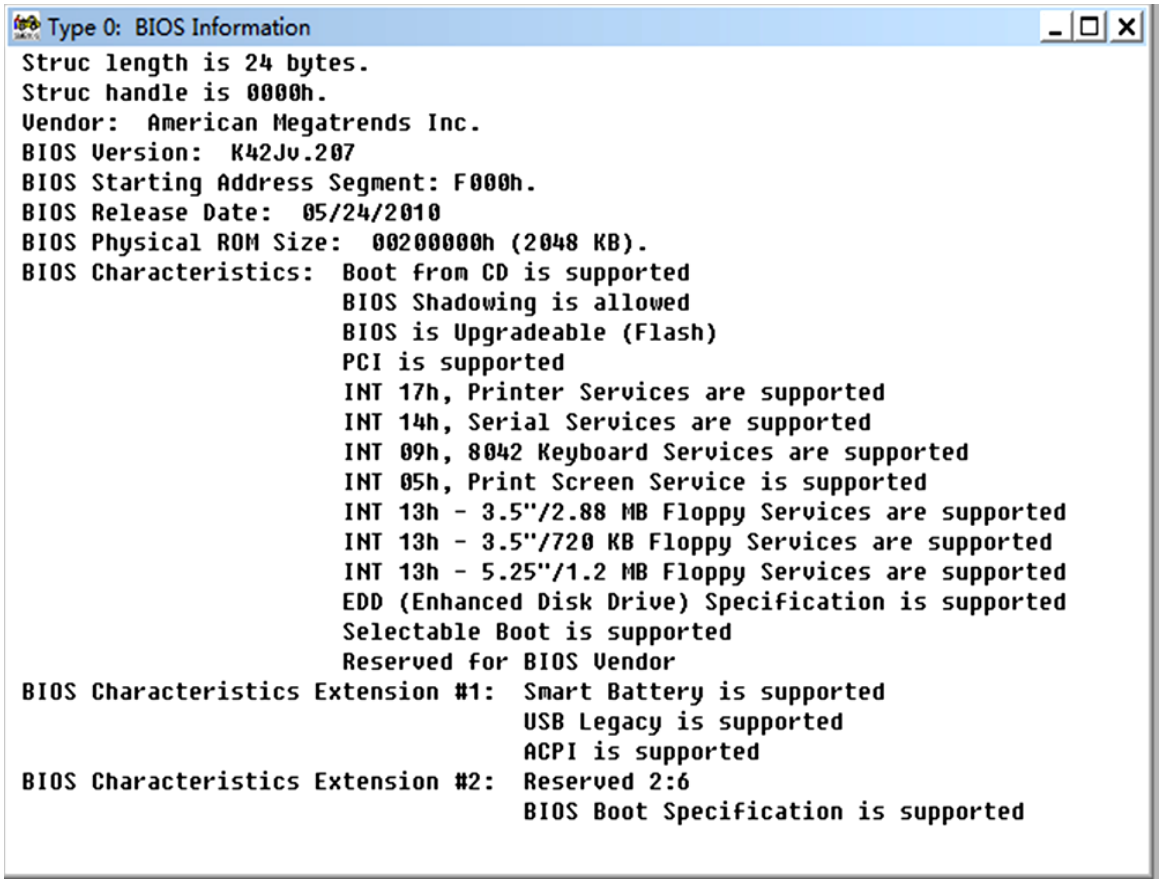


Figure 7. System BIOS information display effect
图 7. 系统 BIOS 信息显示效果

Table 1. EPS table structure
表 1. EPS 表结构

offset	name	length	description
00H	字串锚	4 BYTEs	“_SM_”,指定为四个 ACSSII 码(5F 53 4D 5F)
04H	EPS 校验和	BYTE	
05H	EPS 长度	BYTE	EPS 表的长度
06H	SMBIOS 主版本	BYTE	
07H	SMBIOS 次版本	BYTE	
08H	最大结构大小	WORD	最大的 SMBIOS 结构的大小
0AH	EPS 修正	BYTE	
0BH-0FH	格式化区域	5 BYTEs	
10H	中间锚串	5 BYTEs	“_DMI_”,指定为五个 ACSSII 码(5F 44 4D 49 5F)
15H	中间校验和	BYTE	
16H	结构表长度	WORD	SMBIOS 结构表的总长度
18H	结构表地址	DWORD	SMBIOS 表的起始物理地址
1CH	结构表数量	WORD	SMBIOS 结构表的总个数
1EH	SMBIOS BCD 修正	BYTE	

Table 2. Part of the BIOS table format area
表 2. 部分 BIOS 表格式区域

offset	name	length	value	description
00H	Type	BYTE	0	BIOS 信息指示
01H	长度	BYTE	Varies	Type0 格式区域的长度
02H	句柄	WORD	Varies	EPS 表的长度
04H	厂商	BYTE	01H	BIOS 供应商的信息， 一般为 01H 代表在字符串域中的第一个字符串
05H	版本	BYTE	02H	BIOS 的版本信息， 一般为 02H 代表在字符串域中的第一个字符串
06H	BIOS 起始地址段	WORD	Varies	
08H	BIOS 发布日期	BYTE	03H	一般为 03H，代表在字符串域中的第三个字符串
09H	BIOS ROM 大小	BYTE	Varies	

BIOS 镜像解析模块和计算机系统信息模块是系统通过固件层实现的主要模块，其他模块的实现相对简单，由于篇幅问题，在此暂时不作介绍。

5. 结束语

当今社会普遍将更多的关注点放在操作系统层的安全上，容易忽略计算机固件层的安全。而针对固件的攻击手段已经越来越多，BIOS 作为连接计算机硬件和操作系统的中间桥梁，在计算机中的地位极其重要，一旦受到攻击，将可能致使整个计算机瘫痪。固件攻击造成计算机安全面临严重的威胁，因此，BIOS 攻击检测和防护成为当前亟待解决的问题。针对传统 BIOS 安全检测系统的局限性，本文根据 UEFI 中固件存储结构规范设计和实现了一种基于 UEFI 固件的攻击检测系统，增强了系统的通用性。而且，传统 BIOS 检测系统通常采用静态检测技术，市面上还没有广泛投入使用的 BIOS 检测工具，本文系统客户端在 Windows 系统下开发，可以被广泛使用。系统根据 UEFI 固件文件系统实现了对 BIOS 镜像文件的解析，检测分析模块将被测 BIOS 的各个模块的 MD5 值与服务器中存储的标准 BIOS 的相应模块进行对比，可以检测出 BIOS 是否受到攻击，并显示出具体哪个模块受到攻击，定位攻击者的具体行为。此外，系统通过 SMBIOS 实现了对计算机硬件信息的获取和格式化展示，方便用户检测计算机硬件的状态。

参考文献 (References)

[1] 刘挺. UEFI BIOS 实现原理与结构分析[J]. 电子技术与软件工程, 2015(20): 171.

[2] 周振柳. 计算机固件安全技术[M]. 北京: 清华大学出版社, 2012: 13-14.

[3] 赵丽娜, 陈小春, 张超, 肖思莹. BIOS 安全更新及保护系统设计[J]. 微型机与应用, 2015(8): 2-4.

[4] 李越, 黄春雷. CIH 病毒的分析与清除[J]. 计算机科学, 2000(5): 104-105.

[5] Phoenix Technologies Ltd. (2000) Phoenix Net 1.4 PRD Revision 0.6.2000.6.

[6] Heasman, J. (2006) Implementing and Detecting an ACPI BIOS Rootkit. Next Generation Security Software Ltd., Manchester.

[7] Heasman, J. (2007) Implementing and Detecting a PCI Rootkit. Next Generation Security Software Ltd., Manchester.

[8] 何宛宛. 基于 UEFI BIOS 攻击方式的研究[D]: [硕士学位论文]. 北京: 北京工业大学, 2014.

[9] 王晓箴, 周振柳, 刘宝旭. BIOS 采样分析系统的设计与实现[J]. 计算机工程, 2011(11): 7-9.

[10] 张智, 袁庆霓. BIOS 安全检查系统设计与实现[J]. 计算机技术与发展,2012(2): 172-175 + 180.

[11] 孙亮, 陈小春, 王冠, 等. 基于 UEFI 固件的攻击验证技术研究[J]. 信息安全与通信保密, 2016(7): 89-93.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 329-335
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74040

Hans汉斯

Intelligent Irrigation Control System Using Internet of Things

Yuxuan Feng, Shengyue Wang, Dandan Yang, Renchun Guo, Lijie Zhao, Jie Xing

College of Information Engineering, Shenyang University of Chemical Engineering, Shenyang Liaoning
Email: zlj_lunlun@163.com

Received: Apr. 4th, 2017; accepted: Apr. 17th, 2017; published: Apr. 27th, 2017

Abstract

Traditional orchard cultivation is inefficient and heavy work, and the Internet of Things technology + traditional orchard cultivation mode is conducive to improving the efficiency of the orchard management. In this paper, with STM32 series of single-chip microcomputer, 2.4 G wireless module, and Unity3D engine mobile development platform, we design and develop an orchard planting remote monitoring and control system of Internet of Things + Unity3D interactive intelligent virtual reality. The system consists of the bottom part and the top part of the composition. The bottom part of the design uses soil moisture sensors and air temperature and humidity sensors to detect the soil temperature and outdoor environment temperature and humidity information. According to different fruit soil moisture settings, the controller adjusts the solenoid valve and controls the amount of irrigation. The top part of the design establishes three-dimensional virtual scene to achieve roaming, real-time monitoring, and information display. The bottom part establishes protocols with the top part, then we can investigate fruit tree farming professional information to set the intelligent watering, and establish remote manual control watering, which facilitate the management staff at any time to view the data and remotely control watering, thus reducing the difficulty of orchards maintenance.

Keywords

Smart Orchards, Remote Control and Detection, Internet of Things, Virtual Reality

物联网智能浇灌控制系统

冯雨轩，王圣玥，杨丹丹，郭仁春，赵立杰，邢　杰

沈阳化工大学信息工程学院，辽宁 沈阳
Email: zlj_lunlun@163.com

*通讯作者。

文章引用: 冯雨轩, 王圣玥, 杨丹丹, 郭仁春, 赵立杰, 邢杰. 物联网智能浇灌控制系统[J]. 计算机科学与应用, 2017, 7(4): 329-335. https://doi.org/10.12677/csa.2017.74040

收稿日期：2017年4月4日；录用日期：2017年4月17日；发布日期：2017年4月27日

摘　要

传统果园种植低效且工作繁重，物联网技术+传统果园种植的模式有利于提高果园管理效率。本文采用STM32系类单片机、2.4 G无线模块，结合Unity3D引擎移动开发平台，设计和开发了一种物联网+Unity3D可交互智能化虚拟现实果园种植远程监控控制系统。该系统由底层部分和顶层部分组成，底层部分设计使用土壤湿度传感器和空气温湿度传感器检测果园土壤温度和外部环境温度和湿度信息，控制器根据不同果树土壤湿度设定值，调节电磁阀，控制浇水量。顶层部分设计建立三维虚拟场景，实现场景漫游、实时监视、信息显示功能。底层部分与顶层部分建立协议，通过查询果树养殖专业信息设定智能浇灌，同时也建立远程手动控制浇灌，方便管理人员随时查看数据和远程控制浇灌，降低果园养护难度。

关键词

智慧果园，远程控制与监测，物联网，虚拟现实

1. 引言

传统果园种植手工劳动方式造成果园养护效率低下，果农劳动强度大。由于专业养护人员的缺乏、果树养护不及时，常常导致果树营养不良甚至死亡，给果农造成极大的损失。随着信息时代的高速发展，传统产业迎来了物联网时代。如何让电脑客户端和手机 APP 应用程序自动检测控制果树的生长情况，在果树缺失水分时自动通知果农，并进行自动浇水。通过远程监控控制功能实现智能管理，越来越受到关注，因此迫切需要一种可交互智能化果园种植虚拟现实监控控制系统。

文[1]设计的基于 PC 机和单片机智能灌溉系统无智能终端即智能手机 APP 操控，极大的限制了远程操作的距离。文[2]设计开发了一种基于 ZigBee 技术实现农田节水灌溉、施肥以及信息采集与处理的系统数据只能通过路由器传输，限制了数据传输和控制的范围。文[3]发明的基于无线传感器网络智能灌溉系统仅仅可通过无线网络对果园进行浇水控制，但无法检测到果树周围环境的具体信息，同时需要大量的接线不利于在果园安装和维护。虚拟现实作为一种高度逼真的交互式视景仿真技术，在军事、医学、设计和娱乐等领域得到广泛应用[4]。但是，基于虚拟现实技术的可交互远程智能果园浇灌系统尚未见报道。

针对以上已有研究中出现的问题，本文设计和开发了一种物联网 + Unity3D 可交互智能化虚拟现实果园种植远程监控控制系统。该系统具有数字化、网络化、虚拟与现实的深度融合的特点。系统硬件采用 STM32 系类单片机、2.4 G 无线模块，虚拟仿真应用程序的开发选用 Unity3D 引擎移动开发平台。整个系统包括底层部分和顶层部分。其中，底层部分的功能是：对果树进行实时信息采集和控制，顶层部分的功能是实现虚拟漫游、实时信息显示、远程控制功能。顶层部分和顶层部分通过 GPRS 模块进行数据交换和传输。

2. 系统总体结构和功能设计

2.1. 总体结构设计

本文提出一种可交互智能化果园种植虚拟现实监控控制系统，该系统由底层部分和顶层部分组成，如图 1 所示。底层部分包括：核心控制器 STM32 系类单片机、2.4 G 无线模块、土壤湿度传感器、空气温湿度传感器、485 通信模块、GPRS 模块、继电器模块、水泵、水管、喷头。顶层部分包括：PC 端和手机 APP。

土壤和空气环境信息采集部分采集部分：采用主—从机模式，主从机均采用 STM32F103 系列单片机，主机采集空气环境的温湿度和其所在区域的土壤湿度信息，从机负责采集其他区域的土壤湿度信息，并且通过 2.4 G 通信模块[5]与主机通信，实现一主机多从机的模式，主机收集到各区域的环境信息后将数据发送给上位机。

底层控制系统的设计与开发：电磁阀一端通过水管连接水泵，另一端通过水管连接到土壤，将单片机信号线与继电器接口相连，继电器触点和电磁阀连接，通过改变 I/O 口的高低电平就可完成对浇灌动作的控制。

下位机与上位机之间通信系统的设计开发：STM32 通过串口将数据发送到 485 模块上，再传输到 USR-GPRS-730 上，上位机通过 TCP/IP 协议与 GPRS 进行通信，使得上位机与下位机可通过物联网相互传送数据。同时，上位机与下位机建立协议，上位机可根据下位机发出的信息进行处理并反馈数据到下位机，下位机根据反馈数据后进行对应的控制处理。

客户端三维虚拟人机交互 APP 设计与开发[6]：采用 3D Max 进行场景建模、渲染和加工，生成 3D 模型文件后导入 Unity3D，后台 c#.net 脚本语言进行场景漫游、信息显示和远程控制实现[7]。

底层部分封装进一个独立设计的包装中，底层部分都是无线进行相互连接，方便安装和使用；顶层部分开发 APP，可以让使用者方便操作和远程监测。最终结合成一个完善的 3D 数字化智慧果园管理系统。

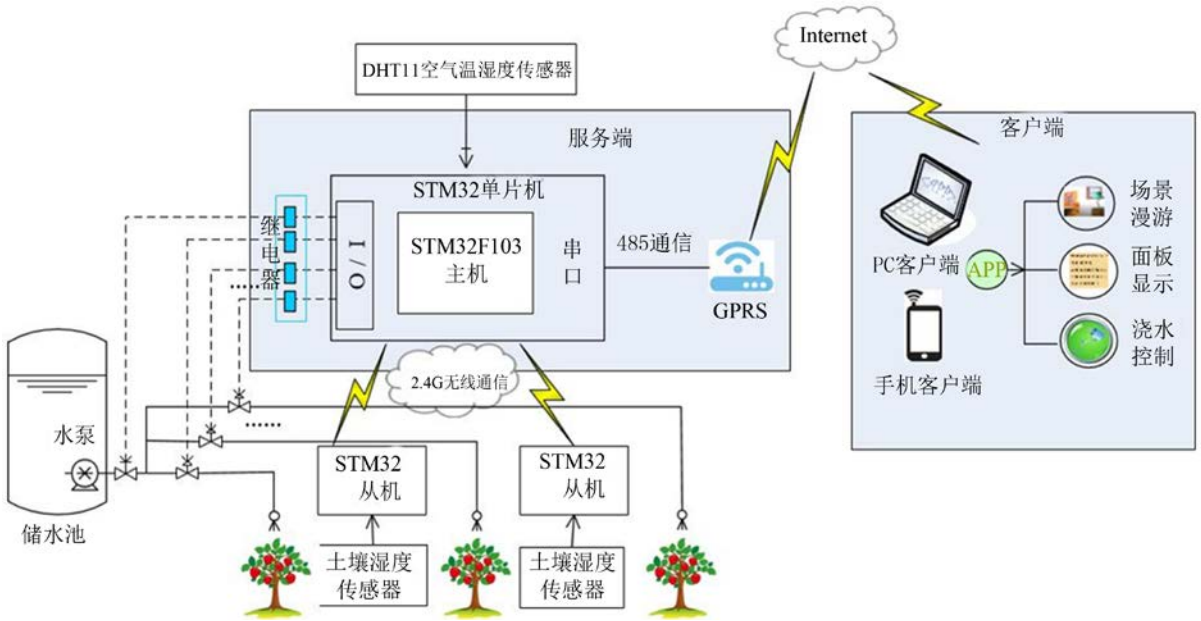


Figure 1. System global structure chart
图 1. 系统总体结构图

2.2. 系统功能设计

本系统将传感器采集到的信息通过 2.4 G 无线网络同意发送至主核心控制器 STM32 系类单片机，对果树信息进行实时采集，采用 Unity3D 引擎开发移动平台实现虚拟现实应用程序开发。通过三维虚拟场景漫游，与果树浇灌设备交互实现远程开启停止控制，应用虚拟现实 VR 技术，实现智能浇灌控制，手自动工作切换模式功能。

具体包括：

1) 信息监测：PC 端和手机 APP 信息显示面板自动显示果树品种、当前环境温湿度、土壤温度和浇灌湿度控制系统设定值。

2) 远程控制：设置了手动浇灌模式和自动浇灌模式。开启自动模式，系统根据实时采集的土壤湿度信息，比较温湿度传感器采样值和控制系统设定值之间的大小，根据差值调整浇水量。手动模式时，可通过 PC 端或者 APP 进行对果园的管理，由于对果园进行了三维场景的建立，因此，果农可以直接看到果园的全部场景，利用鼠标点击电脑屏幕或者使用手机 APP 直接点击手机屏幕对果树进行浇灌。

3) 场景漫游：通过场景前后左右移动和旋转漫游，查看果园场景漫游。

3. 系统开发关键技术

3D 数字化智慧果园系统开发包括底层控制系统开发和顶层远程站客户端应用程序开发。底层部分核心控制器采用 STM32F103 系列单片机最小系统板将收集的土壤和环境空气信息的模拟数据转为二进制数据，STM32F103 系列单片机成本低、体积小、开发简单、系统开发灵活，并且 STM32F103 系列单片机中断、定时器等外设多，非常适用于控制。虚拟现实开发选择 Unity3D 平台[8]，它具有大型场景支持和在线控制功能，模块资源丰富，编程周期短，脚本强大，渲染高速的优点。

3.1. 传感器数据采集和单片机的数据处理

通过湿度传感器采集土壤的湿度信号和通过温度传感器采集空气、土壤、水的温度信号，通过单片机的 AD 模块采样并转换成单片机能识别的信号后送入核心控制单元，核心控制单元对信号进行判断决策后，实时对喷灌设备进行控制，调整喷灌时间、喷灌水量及温度上下限报警，以达到节水节能的目的。

根据实时采集土壤水分、温湿度等数据和电磁阀、泵等驱动执行设备参数，远程设置和修改现场的环境参数(温湿度、土壤水分等)以及现场设备的控制模式，实现智能自动、定时自动、手机遥控、手动灌溉等多种灌溉模式。

3.2. 客户端与单片机之间无线数据传输

上位与下位机之间通信系统的设计开发：STM32 通过串口将数据发送到 485 模块上，再传输到 USR-GPRS-730 上，上位机通过 TCP/IP 协议与 GPRS 进行通信，使得客户端与单片机可通过物联网相互传送数据。同时，上位机与下位机建立协议，上位机可根据下位机发出的信息进行处理并反馈数据到下位机，下位机根据反馈数据后进行对应的控制处理。

3.3. 三维交互系统 APP 的开发

客户端三纬虚拟人机交互 APP 设计与开发：采用 3D Max 进行场景建模、渲染和加工，生成 3D 模型文件后导入 Unity3D，后台 c#.net 脚本语言进行场景漫游、信息显示和远程控制实现，虚拟场景实时与现实交互，使操作者更加直观的感受到果园的现实场景，具有强烈、逼真的感官冲击，给人身临其境的视觉享受。3D 数字化智慧果园系统人机交互界面的设计与开发如图 2 所示。

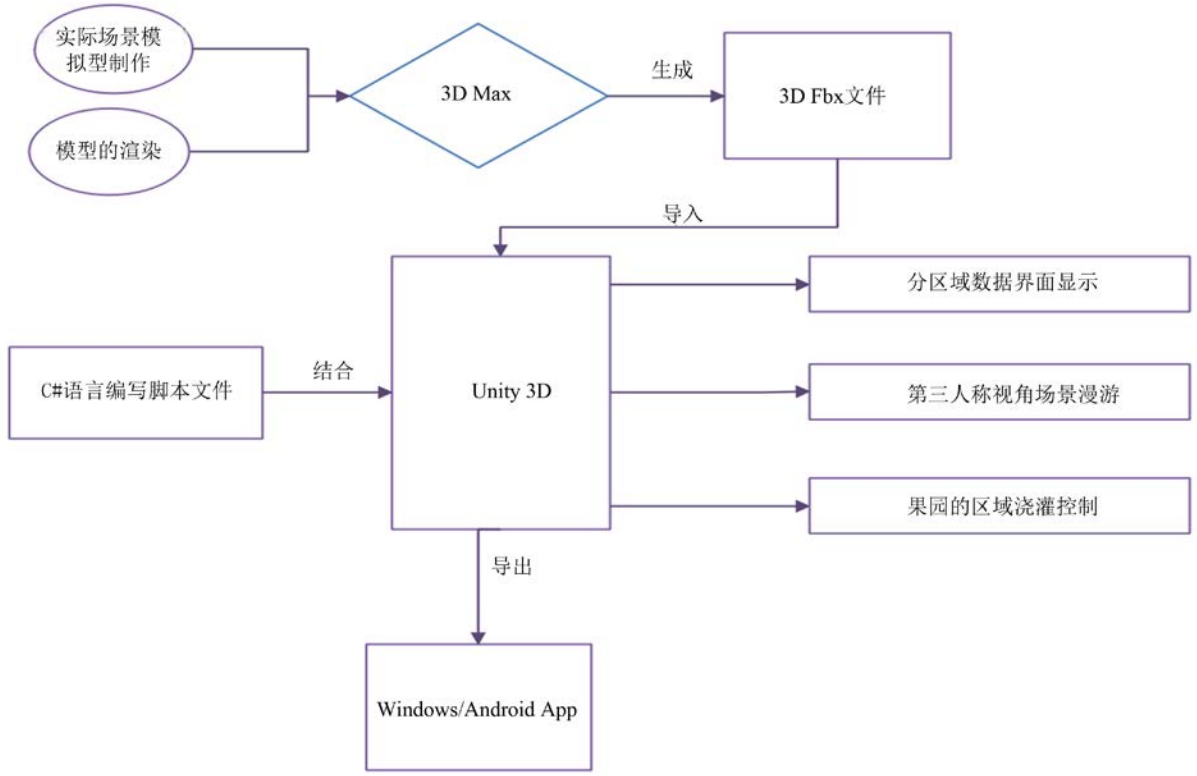


Figure 2. The human-computer interaction interface flow chart of design and development
图 2. 人机交互界面的设计与开发流程图

具体开发过程包括：

1) 果园全景取景，建模。使用 3D Studio Max 软件进行对果园果树等场景的建模，并对模型进行加工和渲染，再利用 Photoshop 对模型进行精细加工、美化，最后得到逼真的三维模型，将制作好的模型生成 Fbx 文件导入 Unity3D 中。

2) 基于 Unity3D 实现果园场景漫游、实物可视化和实时三维动画。本系统采用 c#.net 语言作为开发的脚本语言，利用 Unity3D 软件对微软 Visual Studio 开发软件提供 API，然后通过控制第三视角前后左右、放大、缩小、旋转开发实现三维场景漫游以及手动灌溉模式。

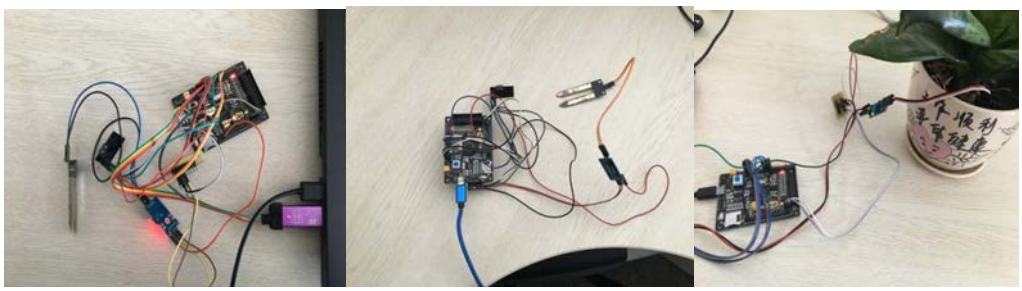
3) Unity 3D 应用程序多平台导出。根据 Unity3D 提供的导出功能，添加导出的场景到列表里，根据应用程序运行在 PC 端还是移动端，选择切换平台，生成应用程序。

4. 系统测试

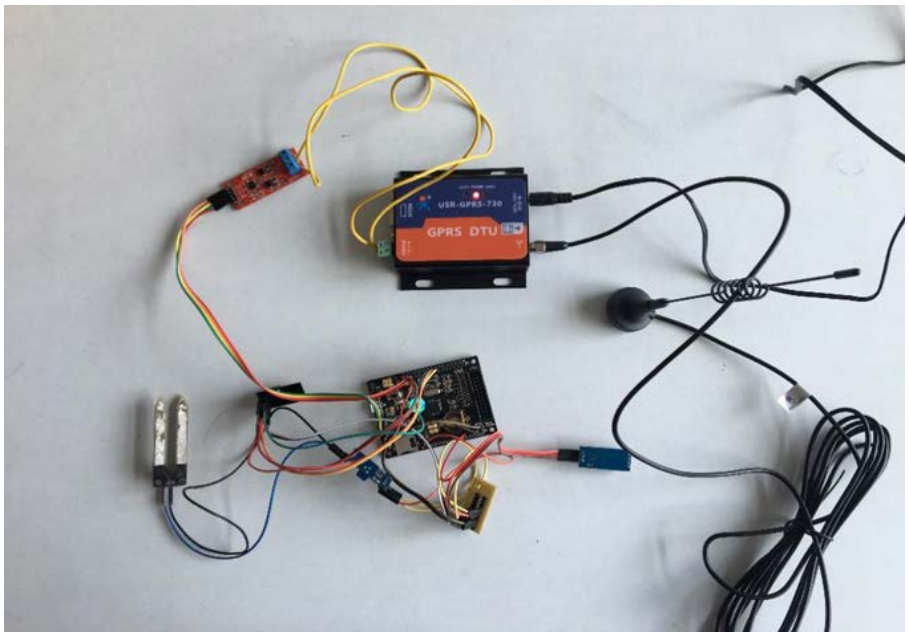
3D 物联网智能浇灌控制系统的底层部分如图 3 所示，底层部分包括机械部分和控制部分。机械部分包括电磁阀、储水箱、水泵、水管；控制部分包括土壤湿度传感器、空气温湿度传感器和继电器。实物图如图所示。

土壤湿度传感器和空气温湿度传感器与核心控制器采用 STM32F103 系列单片机最小系统板相连，土壤湿度传感器探头插入果树周围土壤中，空气温湿度传感器置于空气中，单片机将收集的土壤和环境空气信息的模拟数据转为二进制数据。果园划分区域进行监控，每块区域放置一个独立的核心控制器，所有区域的信息都将通过 2.4 G 无线网络传送到一个主核心控制器，主核心控制器与 GPRS 模块相连，将信息通过物联网发送到上位机，并在 APP 应用中显示。每个区域各设置一个电池阀，一端通过水管连接

(a)



(b)



(c)

Figure 3. Physical diagram of bottom. (a) Information acquisition and control systems; (b) Master-slave STM32 series microcontrollers; (c) GPRS wireless communication
图 3. 底层实物图。(a) 信息采集和控制系统；(b) 主从机 STM32 系列单片机；(c) GPRS 无线通信部分

水泵，另一端通过水管连接到喷头，每块区域中的核心控制器 STM32F103 系列单片机信号线与继电器接口相连，继电器触点和电磁阀连接，通过改变 I/O 口的高低电平就可改变电池阀的开关，对果树浇灌进行控制。

APP 建立的三维虚拟场景下每个区域都有独立的显示土壤和空气环境信息界面，可以结合科学知识设定果树生长所需环境进行智能浇灌，也可以点击喷头进行手动浇灌。通过界面显示与提示对果园进行管理。基于虚拟现实技术的应用程序可以在 PC 端使用，也可以在手机上使用。

5. 结论

本文设计开发了一种物联网 + Unity3D 的虚拟现实果园种植远程监控控制系统，该系统由底层部分和顶层部分组成，底层部分主核心控制器采用 STM32 系类单片机，顶层部分应用程序采用 Unity3D 和 3D Studio Max 开发。底层部分和顶层部分通过 GPRS 模块进行数据交换和传输。系统测试表明该系统可以实时监测采集果树的基本信息，监测果树的生长情况，融合采集到的果园信息，实时自动控制浇灌，达到节水、智能、高效的目的。操作者还可以选择手动浇灌，更加精准地控制果园的浇灌。此系统克服了果园现场有线布线的不便，实现了全网络无线传感器信息传输和控制，系统实用性增强，市场应用前景广阔。

参考文献 (References)

[1] 吴爱萍, 李嘉琪. 智能灌溉控制系统设计[J]. 工业控制计算机, 2015, 28(6): 142-143.

[2] 温宗周, 豆朋达, 钱佳佳, 等. 基于 ZigBee 的智能灌溉系统设计[J]. 单片机与嵌入式系统应用, 2016, 16(11): 38-42.

[3] 杨柏楠. 基于无线传感器网络智能灌溉系统的设计[J]. 农业科技与信息, 2016(10): 95-95.

[4] 姜学智, 李忠华. 国内外虚拟现实技术的研究现状[J]. 辽宁工程技术大学学报, 2004, 23(2): 238-240.

[5] 邱林, 覃江峰. 智能灌溉系统中无线传感网络 2.4GHz 无线信道传播特性[J]. 湖北农业科学, 2015(9): 2242-2244.

[6] 王圣霖, 朱世范, 胡海辉. 基于移动设备的虚拟实境技术在景观设计中的应用[J]. 中国园林, 2015, 31(11).

[7] 王洪源. Unity3D 人工智能编程精粹[M]. 北京: 清华大学出版社, 2014.

[8] 朱惠娟. 基于 Unity3D 的虚拟漫游系统[J]. 计算机系统应用, 2012, 21(10): 36-39.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 336-343
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74041

Hans汉斯

Architecture Design of Private Cloud Computing Platform Based on OpenStack

Cheng Guo[1], Haojun Zhang[1], Shougang Yuan[2], Ruyi Guo[2], Xiaotao Ju[1]

[1]China Xidian Group Corporation, Xi'an Shaanxi
[2]School of Electronic & Information Engineering, Xi'an Jiaotong University, Xi'an Shaanxi
Email: 854222409@qq.com

Received: Apr. 7th, 2017; accepted: Apr. 22nd, 2017; published: Apr. 27th, 2017

Abstract

In this paper, firstly we design an Architecture Design of Private Cloud Computing Platform based on OpenStack, then a prototype system is installed in the actual application environment based on the architecture. Finally using the benchmark tool Rally of OpenStack community, we simulate the actual load environment and test the platform synthetically. The result shows that the cloud computing platform still performs well in high load environment.

Keywords

Private Cloud, OpenStack, Performance Evaluation, Rally

基于OpenStack的企业私有云架构设计

郭　诚[1]，张豪俊[1]，袁守刚[2]，郭如意[2]，琚晓涛[1]

[1]中国西电集团公司，陕西 西安
[2]西安交通大学电子与信息工程学院，陕西 西安
Email: 854222409@qq.com

收稿日期：2017年4月7日；录用日期：2017年4月22日；发布日期：2017年4月27日

摘　要

本文基于开源云平台框架OpenStack设计一个企业私有云平台的架构，根据该架构在实际应用环境中搭建原型系统，并利用OpenStack社区的基准测试工具Rally，模拟真实云环境中的实际负载场景对该平台进行全面测试。结果表明：在高压环境下，基于该架构的企业私有云平台在高负载环境中仍表现出优异

文章引用: 郭诚, 张豪俊, 袁守刚, 郭如意, 琚晓涛. 基于 OpenStack 的企业私有云架构设计[J]. 计算机科学与应用, 2017, 7(4): 336-343. https://doi.org/10.12677/csa.2017.74041

性能。

关键词

私有云，OpenStack，性能评估，Rally

1. 引言

近年来，云计算技术不断发展，所谓云计算是一种按使用量付费的模式，可以提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池(资源包括网络，服务器，存储，应用软件，服务)，这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。对于应用程序开发人员和 IT 运营商来说，云计算可以开发、部署、运营更容易扩展的、极少失效的应用程序，无须关心底层基础结构或所处位置。随着云计算技术的不断发展与成熟，各种云计算管理平台应运而生，目前较为著名的产品有 Eucalyptus、OpenStack、CloudStack 以及 OpenNebula 等[1]。

按照计算提供者与使用者的所属关系为划分标准，将云计算分为三类，即公有云、私有云和混合云。本文主要讨论的私有云是指某个企业独立构建和使用的云环境。不同企业的需求不同，其部署的云平台的架构也不同。私有云部署模式赋予企业对于云资源使用情况极高水平的控制能力，可提供对数据、安全性和服务质量的充分保证，因而近年来得到飞速发展，大量企业开始部署自己的 IaaS 开源私有云平台。私有云是为一个客户单独使用而构建的，云基础设施特定为某个组织运行服务，其可部署在企业数据中心的防火墙内，也可以将它们部署在一个安全的主机托管场所，私有云的核心属性是专有资源[2] [3]。

本文详细分析了开源云平台框架 OpenStack，设计了一个私有云方案的整体架构，并在实际应用环境中搭建原型系统，借助于 OpenStack 社区的测试工具 Rally，规划了详细的测试步骤，对该平台进行性能测试与评估，证明了该方案的可行性。

2. 相关主要技术工具

IaaS 开源私有云平台包括加州大学建立的开源项目 Eucalyptus，Rackspace 和美国国家航天局共同开发的 OpenStack，新加入到 Apache 基金会中的 CloudStack，还有一个是一个虚拟化企业数据中心和云基础设施建设和管理的行业开源解决方案，具有开放性、模块化和可扩展的架构 OpenNebula。这四种典型的开源 IaaS 云平台的特点对比如表 1 所示。

通过以上比对，本设计中最后选择 OpenStack 作为云计算开发平台。

2.1. OpenStack

OpenStack 是由 NASA 和 Rackspace 公司共同开发支持的云计算平台，是一个自由开发的源代码的软件项目，为任何一个组织提供可靠的云计算服务，大幅提高数据中心的运营效率。

OpenStack 经过多年发展，现在主要包含 7 个核心项目，包括计算服务 Nova、对象存储服务 Swift、块存储服务 Cinder、网络服务 Neutron、界面展示 Dashboard、身份认证服务 Keystone 以及镜像管理模块 Glance [4]。

Table 1. The comparison of main IaaS platform
表 1. IaaS 典型开源云平台对比

	Eucalyptus	OpenStack	CloudStack	OpenNebula
发布时间	2008 年 5 月	2010 年 7 月	2010 年 5 月	2008 年 7 月
基本架构	主要包括 CLC、CC、NC、Walrus、SC 五个组件	主要有 Compute、Storage、Image、Identity 和 Dashboard 等几个子项目	主要包括管理服务、云基础设施和网络三大部分	主要包括接口与 API、用户与组、主机、网络、存储、集群 6 个部分
虚拟化技术支持	XEN、KVM、VMware	XEN、KVM、VMware、LXC、QEMU、UML	XEN、KVM、VMWare	XEN、KVM、VMWare
用户界面	命令行、简单的浏览器用户界面	命令行、简单的浏览器用户界面	命令行、较好的浏览器用户界面	命令行、简单的浏览器用户界面
社区活跃程度	总人数最多，但活跃用户较少	总人数较多，活跃用户最多	总人数最少，活跃用户数较多	总人数较少，活跃用户最少
公共云平台支持	兼容 Amazon 公共云平台	兼容 Amazon 公共云平台	兼容 Amazon 公共云平台	兼容 Amazon 公共云平台
官方网站	Eucalyptus.com	Openstack.org	Cloudstack.org	opennebula.org

其中，计算服务 Nova 提供 CPU 计算资源，是 NASA 开发的虚拟服务器部署和业务计算模块，负责虚拟机的调度与管理；对象存储服务 Swift 提供对象存储服务，是一个可扩展并且提供了冗余的存储系统；块存储服务 Cinder 提供块存储服务，为云环境中的计算实际提供块设备的创建、添加和卸载；网络服务 Neutron 负责云计算环境中的虚拟网络功能，可以管理网络地址、虚拟路由器；界面展示 Dashboard 是 OpenStack 的图形化接口，用户通过 Dashboard 可以在浏览器上登陆私有云平台，查看管理资料的各种资源；身份认证服务 Keystone 为系统提供统一的授权和身份验证服务，管理员可以添加用户、角色等；镜像管理模块 Glance 提供镜像服务，包括磁盘和服务器虚拟镜像的查询、注册和传输的功能。

如图 1 所示，为 OpenStack 的整体架构图，箭头表示各模块之间的依赖关系。例如一个创建虚拟机操作，需要用户登陆 Dashboard，通过 Keystone 的认证后，用户可以操作 Nova 服务发布创建虚拟机命令，Nova 模块调用从镜像管理模块 Glance 中获得存储于对象存储服务器上的镜像文件，从而实现虚拟机的创建。

2.2. Rally

Rally 是 OpenStack 社区开发的基准测试工具，用于对 OpenStack 进行性能测试。Rally 是通过插件的形式在实际部署的 OpenStack 云平台中运行，通过用户自定义的参数，来模拟真实云环境中的负载场景[5]。其架构如图 2 所示。

其主要包含以下四个组件：

1) Server Providers 用来提供虚拟服务器并可通过 ssh 访问这些服务器；

2) Deploy Engines 在 Server Providers 提供的虚拟服务器上部署 OpenStack 云平台；

3) Verification 在部署的云平台上运行 Tempest，收集 Tempest 运行结果并显示，而 Tempest 是一个旨在为云计算平台 OpenStack 提供集成测试的开源项目。它是基于 unittest2 和 nose 建立的灵活且易于扩展及维护的自动化测试框架，使得 OpenStack 相关测试效率得到大幅度提升；

4) Benchmark engine 允许我们通过具体参数来模拟真实负载场景，并在部署的 OpenStack 云平台上运行该场景。

Rally 提供两种部署方式，一种以服务的形式(as-a-service)运行，该方式通过一系列的守护进程来运

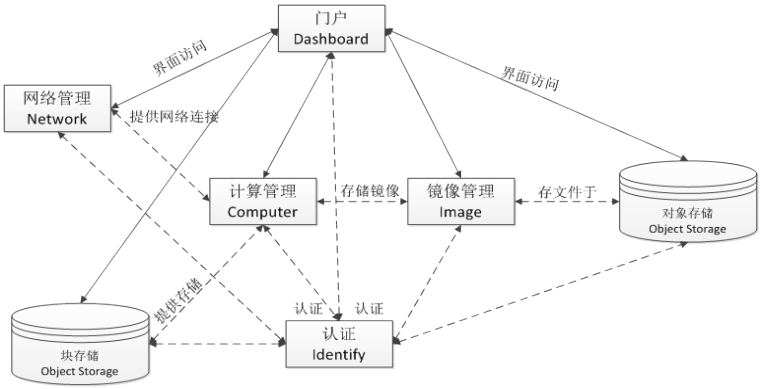


Figure 1. OpenStack architectural diagram
图 1. OpenStack 架构

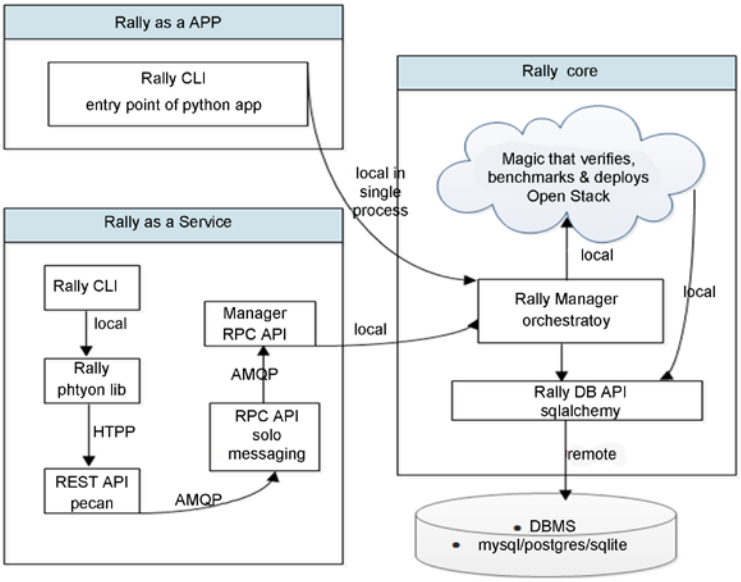


Figure 2. Rally architectural diagram
图 2. Rally 架构

行 Rally 工具；另一种以应用形式(as-a-app)运行，这种方式不需要守护进程，而是通过命令行界面(CLI)的形式运行。

3. 私有云平台系统架构

本文立足于企业现有软硬件资源，经过调研发现，部分企业基于业务的信息系统各自为政，相对比较独立，由此产生的大量数据信息也比较分散，IT 资源异构性强，不同设备的计算能力差异大，增加了维护成本，严重阻碍了集团业务的发展，因此需要建立一个具有资源共享能力的云计算平台[6]。

通过充分论证，基于 OpenStack 开源框架设计私有云平台方案，实现资源的虚拟化，最终部署企业业务应用并支持企业业务发展。云平台的规划和部署，需要进行 OpenStack 技术与抽象的部署模型的映射，然后对应到具体的物理设施中去。本文设计的电力设备远程运维私有云平台系统架构如图 3 所示。

平台最底层是基础设施层，包括服务器、网络、储存等物理资源，利用虚拟化软件(KVM/XEN)实现资源抽象，形成资源池。中间为 OpenStack 管理平台，通过部署 OpenStack 的计算服务、认证服务、网

络服务等相关组件用户可以方便的管理虚拟资源。最上层为应用层，可以部署企业业务应用软件。后台管理可以部署操作和监控应用，负责对整个平台进行管理和控制，实现对服务器的统一管理。

本文搭建的私有云平台基于 OpenStack 的开源框架，包括控制节点、存储节点以及计算节点。控制节点部署了 Nova、Glance、Swift、Cinder、keystone 及 Horizon 等组件；计算节点主要用于运行虚拟化实例，提供计算服务。平台的部署基于实验环境，采用三台浪潮通用服务器和一台存储服务器搭建原型系统，各节点的配置情况如表 2 所示。

其中，一台节点作为控制节点 controller，一台节点作为计算节点 compute，一台作为 block 块存储节点。部署过程中，每个节点使用两块网卡，eth0 用作 Openstack 内部管理网络，eth1 则可访问外部网络。云计算平台的网络主要分为两类：管理网络与外部网络。这样有效的隔离不同功能的网络，增强系统的安全性，当某些网络出现故障时不影响其他网络。管理网络包括宿主物理主机、镜像管理、存储管理、网络管理、虚拟机管理控制器等。从安全性角度来说，管理网与数据网(虚拟网络)的隔离是非常有必要的，确保系统各部分良好运行，隔离故障。外部网络指的是管理节点可以连接外网，实现外部网络访问。

在部署私有云平台过程中，首先需要配置操作系统以及网络环境，选用 Ubuntu 或者 CentOS。其次部署各个节点，控制节点部署 Nova、Glance、Neutron、Cinder、keystone 及 Horizon 等组件；计算节点部署 Nova 以及 Neutron 模块；存储节点上部署 Cinder 模块。最后在平台上部署企业相关应用。

4. 云平台架构测试

测试与验证是及时发现云计算平台安全隐患与缺陷的有效手段之一，为证明本文涉及的电力设备远程运维私有云平台的设计合理性，本文基于 OpenStack 社区的基准测试工具 Rally 设计了详细的负载参数，对原型系统的性能进行评估。Rally 的 benchmark 工具主要是模拟真实云环境中的平台负载，并将其应用到被测试的云平台，收集运行结果并以易于用户理解的方式呈现结果[7]。

Figure 3. Enterprise private cloud platform system architecture
图 3. 企业私有云平台系统架构

Table 2. Prototype system hardware configuration
表 2. 原型系统硬件配置

角色	内存配置/数量	CPU 配置/数量	硬盘容量
控制节点	16G ECC Registered DDR4/	Intel Xeon E5-2620v4(2.1GHz)/2	1TB
存储节点	4G DDR3/2	Intel Xeon E5310(1.6GHz)/2	21TB
计算节点	16G ECC Registered DDR4/	Intel Xeon E5-2620v4(2.1GHz)/2	1TB

Rally 提供两种方式模拟真实云环境负载，一种是通过创建临时用户进行，另一种通过已有用户进行模拟。相比于创建临时用户的方式，通过已有用户进行模拟可以在不同用户组中进行，一旦发生错误，不会影响其他组的用户，这种方式更安全。

通过云平台中已存在的用户向 Rally 注册，并通过脚本设置负载参数，用于测试场景。其具体内容如下：

用户注册：

```
{
    "type": "ExistingCloud",
    "auth_url": "http://example.net:5000/v2.0/",
    "region_name": "RegionOne",
    "endpoint_type": "public",
    "admin": {
        "username": "username",
        "password": "password",
        "tenant_name": "tenant_name"
    },
    "users": [
        {
            "username": "username",
            "password": "password",
            "tenant_name": "tenant_name"
        },
        {
            "username": "username",
            "password": "password",
            "tenant_name": "tenant_name"
        }
]}
```

场景参数设置：

```
 NovaServers.boot_and_delete_server:
      args:
        flavor:
            name: "name"
        image:
            name: "name"
      runner:
        type: "constant"
        times: 2
        concurrency: 1
      context:
        users:
          tenants: 1
          users_per_tenant: 1
```

在单一场景测试中，本文进行 nova 实例测试，通过指定 OpenStack 用户启动虚拟机并列出虚拟机，并将这两个操作迭代 200 次，其结果如图 4 所示，由图可见随着迭代次数的增加，用户创建的虚拟机越来越多，但系统认可正常运行，在高压场景下并未崩溃。

如图 5 所示，在复杂场景测试中，本文设置的复杂场景主要包括以下几个操作：启动虚拟机、虚拟机快照、删除虚拟机、利用快照启动虚拟机、删除虚拟机、删除虚拟机快照；并将该系列操作迭代 100 次，系统仍可正常运行。

5. 结束语

本文立足于企业的应用需求，详细分析了云计算、私有云有关理论基础，提出了一种基于开源框

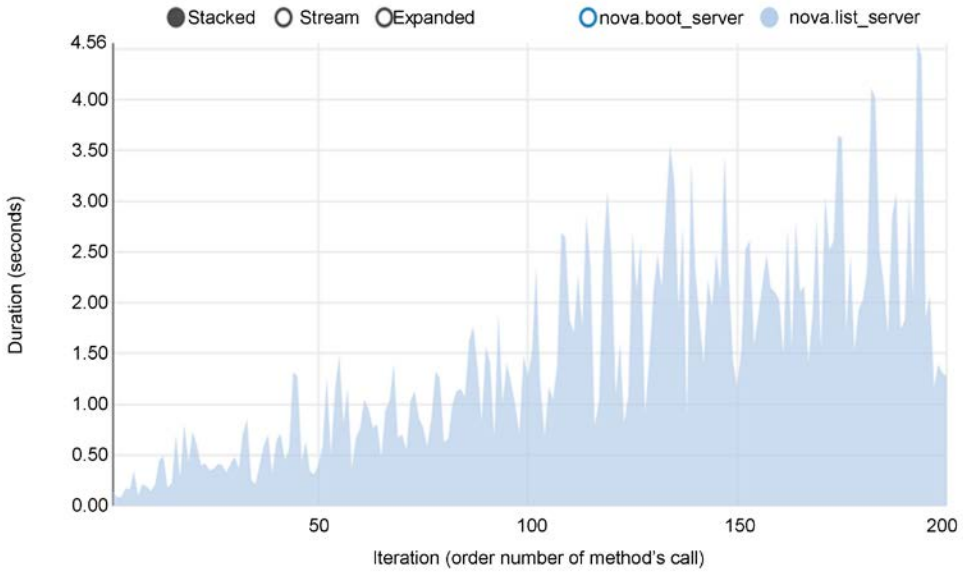


Figure 4. Nova experiment result
图 4. Nova 实例测试

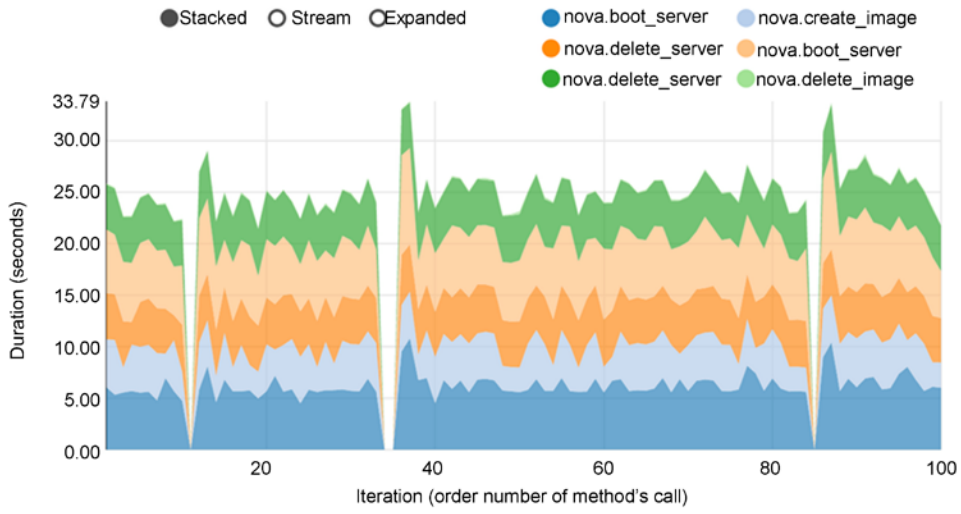


Figure 5. The complex scenes experiment result
图 5. 复杂场景测试

架 OpenStack 构建私有云计算平台的可行性方案，依据该方案搭建了原型系统，并基于 OpenStack 社区的基准测试工具 Rally 设计了详细的测试程序，对原型系统的性能进行评估，证明了该方案的合理性和可应用型，对于企业私有云平台的建设和性能评估具有重大意义。

参考文献 (References)

[1] 中国电子技术标准化研究院. 云计算标准化白皮书[M]. 北京市: 中国电子技术标准化研究院, 2014.

[2] 李乔, 郑啸. 云计算研究现状综述[J]. 计算机科学, 2011, 38(4): 32-37.

[3] 李志军, 孔朋朋, 雷振伍. 基于 OpenStack 的私有云平台设计[J]. 微型机与应用, 2016, 35(9): 24-26.

[4] OpenStack 社区. http://docs.OpenStack.org/#

[5] Rally. https://wiki.OpenStack.org/wiki/Rally#Architecture

[6] 赵少卡, 李立耀, 凌晓, 等. 基于 OpenStack 的清华云平台构建与调度方案设计[J]. 计算机应用, 2013, 33(12): 3335-3338.

[7] 林闯, 苏文博, 孟坤, 等. 云计算安全: 架构, 机制与模型评价[J]. 计算机学报, 2013, 36(9): 1765-1784.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 344-350
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74042

Hans汉斯

Design and Implementation of Fiber Link Monitoring System

Jianjun Guo

Tianjin Port Information Technology Development Co., Ltd., Tianjin
Email: 269596957@qq.com

Received: Apr. 10th, 2017; accepted: Apr. 23rd, 2017; published: Apr. 27th, 2017

Abstract

Optical fiber link monitoring system using advanced wavelength division multiplexing technology, computer communication technology, fiber measurement technology, machine learning TensorFlow framework, etc., achieved the status of the optical fiber link automatic monitoring, fault analysis, positioning, fault management, maintenance arrangements, line status prediction. This system can reduce maintenance cost, and significantly reduce the time required for maintenance.

Keywords

Fiber Link, Wavelength Division Multiplexing, TensorFlow, State Prediction

光纤链路监测系统的设计和实现

郭建军

天津港信息技术发展有限公司，天津
Email: 269596957@qq.com

收稿日期：2017年4月10日；录用日期：2017年4月23日；发布日期：2017年4月27日

摘　要

论文利用先进的波分复用技术、计算机通信技术、光纤测量技术和机器学习TensorFlow框架等设计实现了光纤链路监测系统，该系统实现了对光纤链路的状态自动监测、故障分析和定位、故障管理和维护安排、线路状态预测，该系统投入运行后降低了维护成本、减少了运维时间。

文章引用: 郭建军. 光纤链路监测系统的设计和实现[J]. 计算机科学与应用, 2017, 7(4): 344-350.
https://doi.org/10.12677/csa.2017.74042

关键词

光纤链路，波分复用，TensorFlow，状态预测

1. 引言

光纤通信具有传输容大、抗干扰能力强、传输衰耗小等优点，光纤作为一种优良的通信载体，已在天津港广泛用，但光缆数量的增加，早期铺设的光缆也已有了一定的年限，各种隐患，各种危机随时存在，光缆线路故障次数的连年增加也说明光缆的维护与管理问题也日渐突出。

为此，光缆监测系统是集测试、告警、信息处理和业务管理于一体的光纤网络综合维护系统，它综合运用地理信息系统(GIS)、全球定位系统(GPS)、光时域反射仪(OTDR)、光波分复用(WDM)、数据挖掘等多种技术，能及时掌握网络运行情况，预测发现光纤劣化趋势，做到基于运行情况的状态检修，并在故障发生时可以迅速、准确地定位故障，从而缩短障碍历时，与此同时，它还具有资料管理功能，可以保存一条光缆的设计、施工、维护等相关数据信息，以及 ODF 架等站端设备的数据信息，从而方便维护人员的管理和使用。

如果采用监测系统进行光功率监测，可以达到秒级的水平，系统一刻不停的监视着光缆线路上所有发生的一切细微变化。光缆线路监测系统不仅仅是用自动替代人工，它根本上提高了维护水平，也就是把以往一年的一两次测试提高到一年几十万、几百万次。监测系统是新时期光缆线路维护发展的需要，它把光缆线路纳到实时集中的监测维护当中，不管线路上发生了多大的变化，监测都是秒级的进行着。光传输设备监控不能取代光缆线路监测：传输监控能够发现设备故障，但它是事后行为；而光缆监测不仅可以进行障碍报擎定位，而且可以发现尚未影响通信的故障隐患和精确的故障位置，进行预警预防预维。由于两种监测方式目的不同、手段不同，所以光传输监测不能取代光缆线路监测。为保证光缆不断、不坏、不换，必须进行预防性线路维护和监测。

2. 功能分析

光缆自动监测系统是针对广电光纤网络管理和维护的智慧型系统，具有功能强大、资源结合紧密、智能化操作、维护便捷的特点[1]。不仅可以实现对光纤网络状况实时的监测，而且结合资源系统更加快讯准确的提供光纤故障点信息，以缩短故障历时，实现对光缆线路的监测与管理，动态地观察光缆线路传输性能的劣化情况，及时发现和预报光缆隐患，以降低光缆阻断的发生率，由人工检测与管理模式迅速过渡到集中监控与管理的模式。光缆监测系统可将其核心的测试功能和其它外部网管系统(如设备网管系统，资源管理系统，故障派单系统，SDH 网管告警接口等)结合，从而进一步提高系统功效，切实地加强维护管理工作，减少损失，可以为广电运营商带来不可低估的经济效益[2]。主要组成如下：

1) 远程、实时、在线地进行光缆线路中被监测光纤运行状况的监测，预防光缆线路的故障隐患；

2) 按规定的周期，向网管中心(TSC)传报被监测光缆线路运行状况的数据文件；

3) 系统可以根据不同时期，不同线路的情况设置不同的障碍分析参数，查询 RTU 的光缆运行数据文件，自动分析所监测光缆线路劣化的趋势；

4) 多种测试功能，实现了点名测试、定期测试、障碍告警测试；

5) 多种监测模式，具有在线监测、备纤监测、跨段监测等功能；

6) 通信功能，本部分实现的监测站(RTU)具有 RS-232 物理接口和 TELNET 网络接口控制方式，方便工程人员查询完整的设备状态；监测站(RTU)可通过 IP 网与网管中心(TSC)联网；进行数据通信时，采用 TCP/IP 协议，确保监控网络设备互通；

7) 设备管，本部分实现了 RTU 设备的运行状态及其板卡信息通过设备回传可以显示在操作界面上，以便工作人员实时了解该设备的运行情况，方便维护；设定 RTU 的运行时间；

8) 多曲线显示功能，系统通过曲线管理，可以在曲线界面中查看到多条曲线，以便对当前光缆状态信息更加直观的呈现在操作界面上；

9) 链路状态预测，系统使用 TensorFlow 框架搭建一个链路状态学习系统，能通过当前链路相应状态值，准确预测未来可能出现的链路问题。

3. 难点技术的解决方案

光纤的劣化，手工测试难以发现，因为光纤的劣化过程是一个长期而渐近的过程，若没有长时间测试数据的积累分析是很难发现的[3]。工程中常用的光时域反射损耗测试仪(OTDR)，可以检测出已经发生故障的光纤所在地点[4]。但从发现故障到发生故障的地点并完成维修会花大量的时间。这将会对集团或区域经济建设带来许多的直接或间接的经济损失。一个具有预测和自动化处理能力的光纤链路检测系统是一个非常必要的需求。下面分别从数据的自动化收集和利用数据实现动态预测两方面叙述系统的难点技术。

3.1. 光纤链路数监测

3.1.1. 链路监测数据获取

系统利用相关设备收集光缆损耗、链路属性、信号衰减、信号量等数据。设备采用在线监测方式，该测试方式与工作波长不同的测试波长通过 WDM (wavelength division multiplexer 分波合波器)设备合波在同一根光纤中测试，在对端利用滤波器将测试波长滤掉让工作波长通过[5]；在线监测采用分光器将工作光路上 3%~5%的光分给 OPM (光功率仪)进行实时监测(其中 3%~5%的分光模块集成在 OPM 模块中)此模式能真实反映业务所占纤芯工作状态。

图 1 是链路监测的原理图，WDM 是一种在一根光纤中能同时传输多个波长光信号的一种技术，在不影响光纤链路的正常使用情况下，使用检测设备监测链路的运行情况，获取监测系统需要的各种数据。

3.1.2. 监测信息与系统的设计

监测站(RTU)的设备服务进程用于接受和处理 RTU 上报的告警信息、设备信息以及周期测试信息将链路监控信息、报警等信息发送到综合监控平台进行实时监看，并对告警、解警信息进行及时响应，该设备支持 SNMP 协议可与综合监控平台对接[6]。系统的主要组成部分如图 2 所示。

系统主要由数据监测中心 MC、监测站 MS、通信网络 3 部分组成[3]。数据监测中心(MC)负责对本管区的各监测站进行控制和管理，是收集和处理数据的中心。监测站(MS)在监测中心控制下，对光纤传输损耗的变化进行监测。通信网络主要为监测站与各级监测中心之间的通信提供通道。数据中心获得数据后的显示效果如图 3 所示。

3.2. 光纤链路状态预测

3.2.1. TensorFlow 框架

TensorFlow 是一个采用数据流图(data flow graphs)，用于数值计算的软件库。主要用于机器学习和深

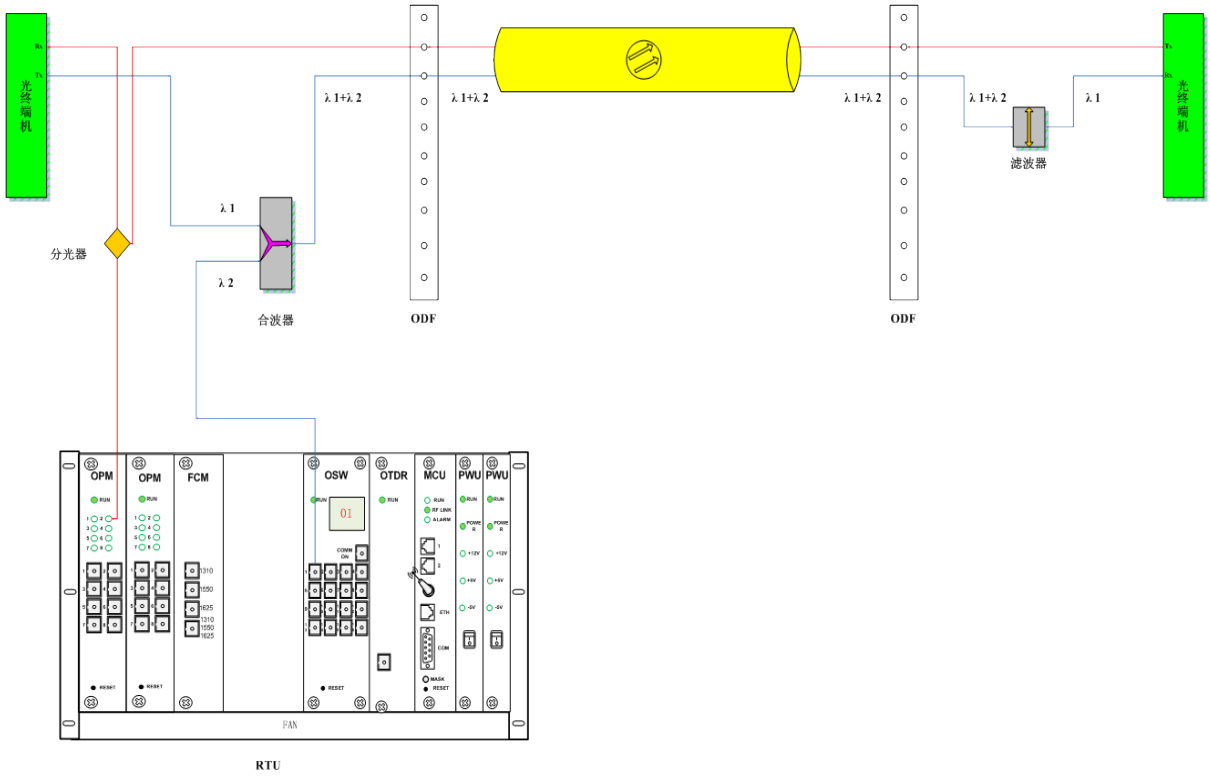


Figure 1. Link monitoring basic schematics
图 1. 链路监测基本原理图

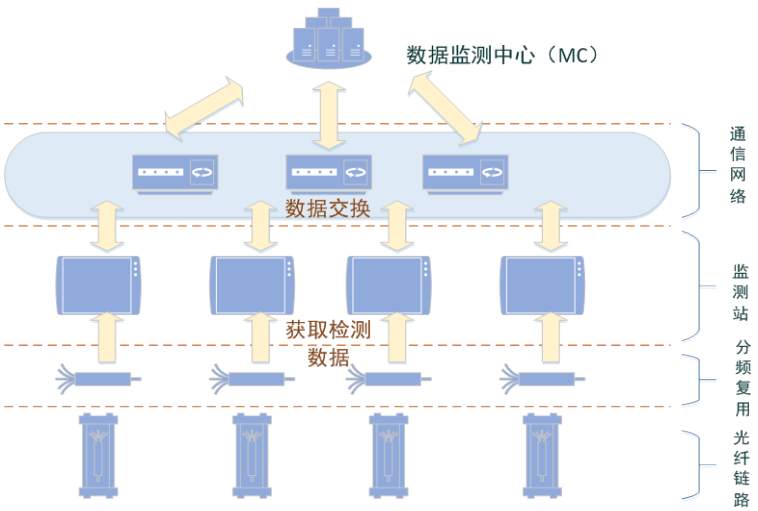


Figure 2. System general framework
图 2. 系统整体框架

度神经网络方面的研究[7]。它基于是一种基于图结构的计算，使用 SGD 优化中间参数，最终得到可以使用的模型，同时它支持 GPU 加速，甚至可以运行在移动设备上。TensorFlow 的基本原理示例图 4 所示。

图 4 中的输入数据需要经过变形后称为一维向量，组合多输入变量可以形成一个基础矩阵。这个矩阵可以看作是神经网络的输入层数据。上图是一个两层网络图的示例，中间层最常用的两个激活函数，

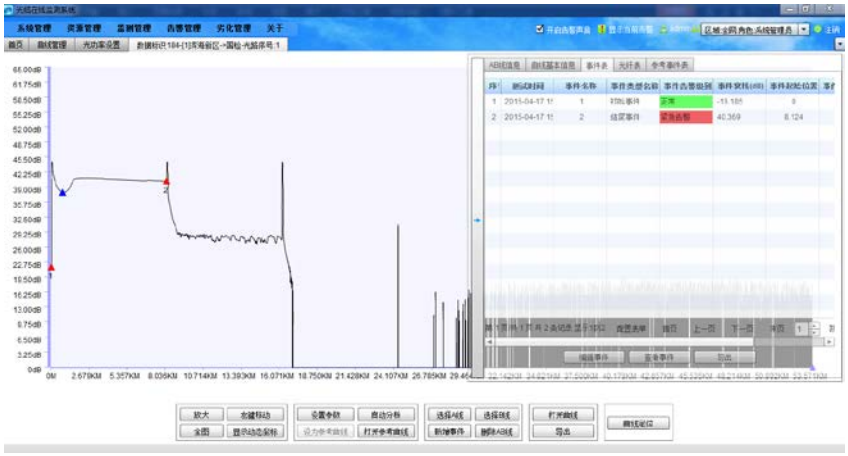


Figure 3. Platform information display
图 3. 平台信息展示

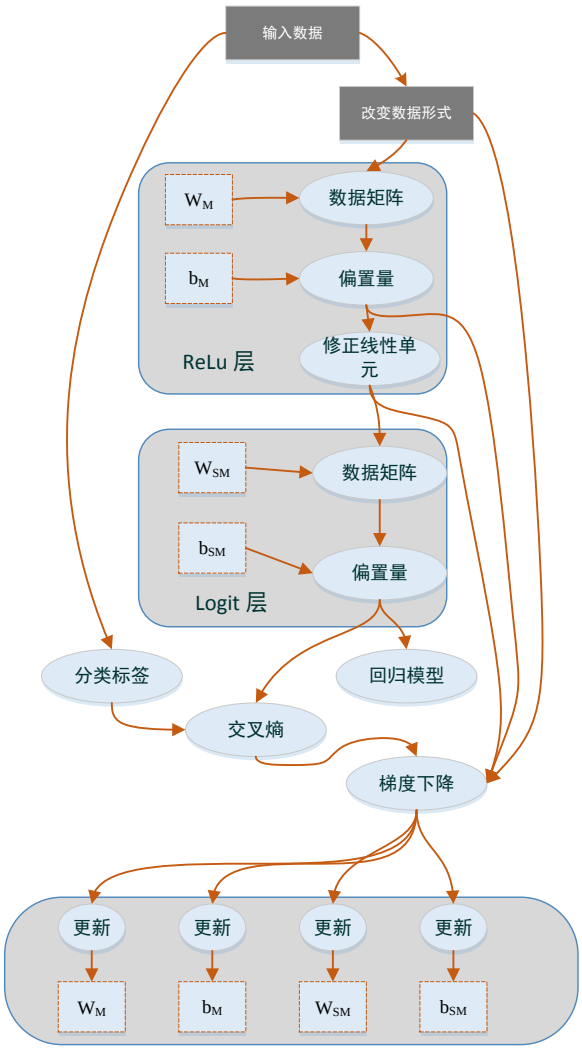


Figure 4. TensorFlow sample graph
图 4. TensorFlow 示例

Sigmoid 系(Logistic-Sigmoid、Tanh-Sigmoid)被视为神经网络的核心所在[8]。使用梯度下降和交叉熵来优化权值W_M和偏置量b_M，最终得到优化的参数及为模型。

3.2.2. 数据分析预测

为了达到预测任务需要使用上文的 TensorFlow 框架构建计算图(graph)，定义操作(ops)和数据类型(主要分为张量(tensor)、变量(variable)和常量(constant))，最后执行阶段(execution phase)执行会话(Session)。

使用由监测站传来的数据如光缆损耗值、链路属性值、信号衰减值、信号量值、光纤使用年份、是否曾有维护等多字段组成的向量作为输入。具体的设置如图 5 所示。

图 5 中 Layer 中 conv1 实现卷积以及 rectified linear activation、pool1 (max pooling)、norm1 局部响应归一化、conv2 卷积和 rectified linear activation、norm2 局部响应归一化、pool2 (max pooling)、local3 基于修正线性激活的全连接层、local4 基于修正线性激活的全连接层、softmax_linear 进行线性变换以输出 logits。

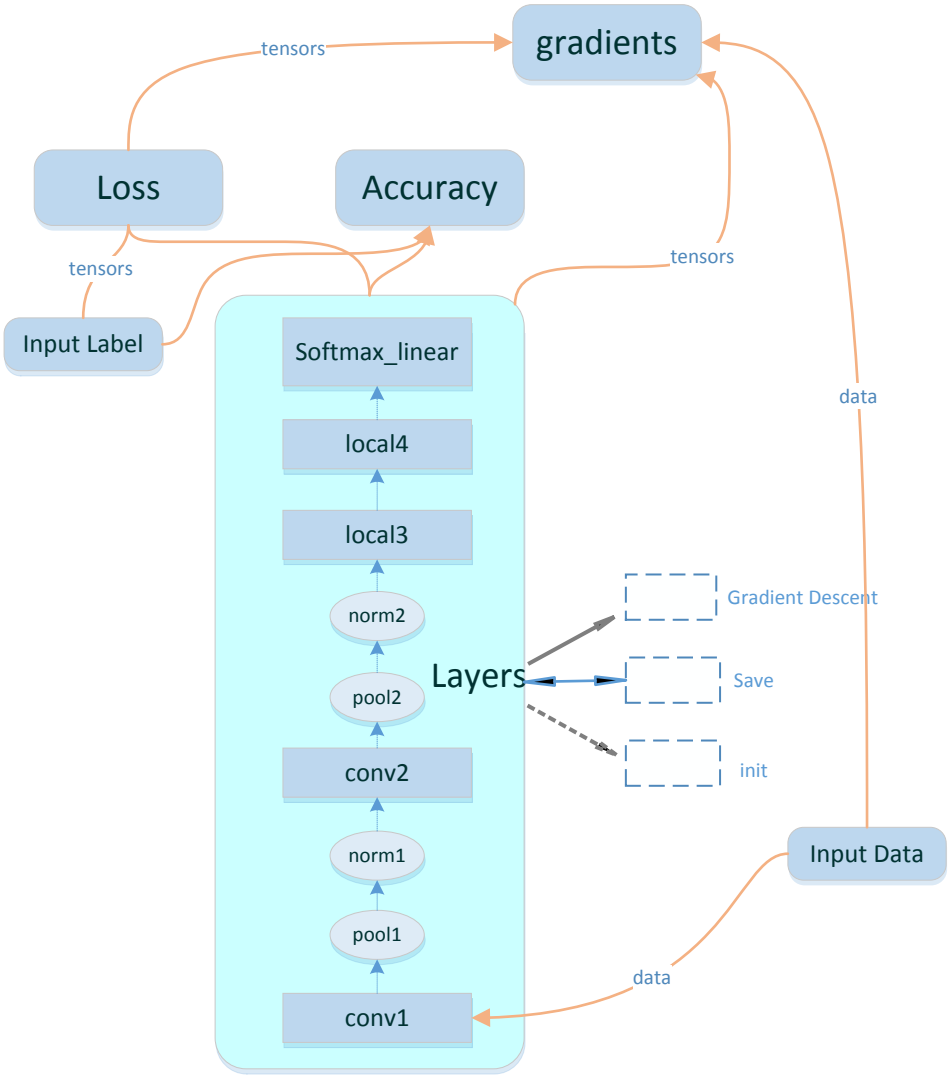


Figure 5. Apply the model structure diagram
图 5. 应用模型结构图

将经过处理带有标签的光纤链路数据输入模型后可以得到各数据的权值和偏差值，初始是模型的准确度不会很高，但这是一个可以不断改进权值和偏差的系统，因此经过一段时间的数据运行后，预测的准确度将不断提高。

4. 总结

港内交通、作业码头等视频监控信号众多，为便于日常维护管理，建立一套统一集中的光缆监测平台接入不仅可以进行障碍报警定位，而且可以发现尚未影响通信的故障隐患和精确的故障位置，进行预警预防预维。系统经投入使用，系统运行稳定，起到了预期的作用。

参考文献 (References)

[1] 朱磊, 张盛武. 光缆自动监测系统的实现[J]. 电信科学, 2000, 16(9): 52-54.

[2] 孟嗣仪. 电力系统光缆自动监测系统的设计及实现[J]. 北京交通大学学报自然科学版, 2002, 26(6): 56-58.

[3] 姜彬. 光纤自动监测系统的应用[J]. 电气化铁道, 2010, 21(1): 48-50.

[4] 肖平平, 袁睿. OTDR 波形分析及在光纤测量中的应用[J]. 光通信技术, 2010, 34(4): 42-44.

[5] 吴海西. WDM 技术的原理及其应用与发展[J]. 现代电信科技, 2000(10): 6-10.

[6] 徐明, 陈奇, 王凌武. SNMP 协议分析与协议栈的实现[J]. 计算机工程与设计, 2006, 27(14): 2669-2672.

[7] Abadi, M., Agarwal, A., Barham, P., *et al*. (2016) TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

[8] Abadi, M., Barham, P., Chen, J., *et al*. (2016) TensorFlow: A System for Large-Scale Machine Learning.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 351-358
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74043

Hans汉斯

Flying Streaming: A Platform for Real Time Multidimensional Statistical Analytics of Large-Scale Log Data

Hongbo Zhao[1], Hua Qin[1], Jianbo Zhao[2]

[1]Department of Information, Beijing University of Technology, Beijing
[2]Beijing 58 Information Technology Co., Ltd., Beijing
Email: zhaohongbo@emails.bjut.edu.cn

Received: Apr. 9th, 2017; accepted: Apr. 22nd, 2017; published: Apr. 27th, 2017

Abstract

At present, it is common that daily increment of log data reaches TB level in domestic internet companies, and the real-time multidimensional statistical analysis of large-scale log data is becoming more and more important for enterprise operation, management and decision-making. However, the current large-scale log data analysis and processing technology is very professional, and business departments and operation and maintenance departments whose demand of data processing is most urgent are difficult to have such capacity. This paper designed a real-time multidimensional statistical analysis platform for large-scale log data through integrating Flume, Kafka, Storm, HBase and so on. The platform is named Flying Streaming. It solves some key technical issues, such as manifold log data access, real-time multidimensional statistical analysis, submitting, updating and deleting tasks by configuration instead of programming. Flying Streaming provides users with the ability of real-time multidimensional statistical analysis without programming. The application effect of Flying Streaming in the Internet enterprise is good, and it can meet the needs for Multidimensional Statistical Analytics of most log Data of business departments and operation and maintenance departments.

Keywords

Storm, Large-Scale Log Data, Real-Time Analytics, Multidimensional Statistical Analytics, General Platform

飞流：基于Storm的大规模日志数据实时多维统计分析平台

赵宏博[1]，秦　华[1]，赵健博[2]

[1]北京工业大学信息学部，北京

文章引用: 赵宏博, 秦华, 赵健博. 飞流: 基于Storm的大规模日志数据实时多维统计分析平台[J]. 计算机科学与应用, 2017, 7(4): 351-358. https://doi.org/10.12677/csa.2017.74043

[2]北京五八信息技术有限公司，北京
Email: zhaohongbo@emails.bjut.edu.cn

收稿日期：2017年4月9日；录用日期：2017年4月22日；发布日期：2017年4月27日

摘　要

目前国内互联网企业单日日志数据增量达到TB级已很常见，大规模日志数据实时多维统计分析对于企业运行、管理和决策越来越重要。但目前大规模日志数据分析处理技术专业性强，企业中数据处理需求最为急迫的业务部门和运维部门都难有这样的技术能力。本论文整合Flume、Kafka、Storm、HBase等开源系统设计了飞流大规模日志数据实时多维统计分析平台，解决了多种日志数据接入、实时多维度统计分析、用户通过提交配置代替大数据编程来提交、更新和删除任务等关键问题，提供了飞流平台上用户不需要编程就能方便使用的大规模日志数据实时多维统计分析的功能。飞流平台在互联网企业中实际应用效果较好，满足了业务部门和运维部门的大部分日志数据多维统计分析需求。

关键词

Storm，大规模日志数据，实时统计分析，多维统计分析，统一平台

1. 引言

互联网企业的主要数据来源是散落在各个业务服务器上的半结构化日志，比如系统日志、程序日志、访问日志、审计日志等。目前国内互联网企业单日日志数据增量达到 TB 级很常见。互联网企业间的竞争非常激烈，实时统计分析日志数据并将结果指导决策能够提高企业的竞争力。因此最原始的日志数据记录具有丰富和巨大的价值。

目前很多开源的系统，如 Flume、Kafka、Storm、HBase 等可以对日志数据进行处理，但是这些系统相互独立，均庞大复杂，需要专门的数据处理人员根据需求编程来使用，而企业的业务和运维部门一般没有专业从事大数据实时处理的人员，因此设计大数据实时处理平台，为用户提供不需要编程就能方便使用的大规模日志数据实时多维统计分析功能，是各个互联网企业的迫切需求，飞流应运而生。

统一的大规模日志数据实时多维统计分析平台需要接入多种来源的日志数据，每种日志记录了各种不同维度的运行数据，用户需要的是灵活的、方便的、多维度的统计分析。因此，飞流平台的设计目标主要是实现大规模日志数据的多源采集和聚合、实时多维度统计分析、用户可以在线通过配置代替大数据编程实现统计分析任务的热提交、热更新和热删除、统计分析结果通过 WebUI 进行展示。

2. 相关工作

目前国内外工业界和学术界在大规模日志数据分析领域做了很多研究和实践，尤其是工业界推动的开源社区非常活跃。面对数据规模大这个问题，广泛采用的解决办法是使用分布式的方案。

多源采集和聚合。Flume [1]是 Apache 基金会下面的一个分布式、可靠和高可用的，从不同源采集、

聚合和传输大量日志数据到一个集中的数据仓库的系统。文献[2]的数据表明 Flume 满足大规模数据实时处理需求。直接将 Flume 收集的数据传递给数据分析系统会带来数据在传递和分析阶段可能丢失，数据量剧增时没有缓冲等问题，所以在 Flume 和数据分析系统的中间加一层消息系统就很有必要。实时的生产日志是流式数据，Kafka [3]是 Apache 基金会下面的一个分布式的、基于发布/订阅的，主要用于流式数据的消息系统，满足大规模数据实时处理需求[4]。

实时计算。满足低延迟大规模流式数据分析的分布式计算系统主要有 Apache 基金会下的 Spark Streaming [5]和 Storm [6]，按照它们提供的编程模型就可以写出分布式的低延迟数据处理程序。Spark Streaming 的处理模型类似批处理，需要收集数据的时间，造成几秒到几分钟的延迟，只能做到准实时的数据处理。Storm 是消息流处理模型，采用服务型作业，数据流实时采集，省去了作业调度时间，满足大规模数据实时处理需求[7] [8]。日志数据分析完后，如果直接做展示，分析结果只能被利用一次。HBase [9]是 Apache 基金会下的分布式的，支持随机实时读/写访问的数据库，可用于存储大规模数据并实时存取[10] [11]。分布式计算系统将分析结果写到 HBase，数据可视化系统随机读取 HBase 存储的分析结果并以某种方式展示出来。

依靠用户大数据编程才能进行多维统计分析。北邮陈任飞等[12]设计并实现的日志数据处理系统，通过整合 Flume、Kafka、Spark Streaming 这一方法，解决了多源采集和聚合、大规模数据的处理两个问题。延迟在秒级，为准实时级别。该系统完全依靠各个部门的各个用户分别使用 Spark 的 API 进行大数据编程来进行日志数据的多维统计分析。北邮薛瑞等[13]设计并实现的日志数据处理平台，通过整合 Kafka、Spark Streaming、HBase 这一方法解决了大规模数据的处理这一个问题。延迟在秒级，为准实时级别。该平台还通过提供一些日志数据统计分析的编程接口简化了用户的编程工作，但仍需要依靠各个部门的各个用户分别进行大数据编程来进行日志数据的多维统计分析。以上两个平台或系统，都只是提供了依靠用户分别进行大数据编程来进行日志数据准实时处理的功能，没有实现实时多维度统计分析、用户可以在线通过配置实现统计分析任务的热提交、热更新和热删除、统计分析结果通过 WebUI 进行展示这些目标。

飞流通过整合 Flume、Kafka、Storm 和 HBase，并在此基础上实现了一个 Storm topology，解决了多种日志数据接入、实时多维度统计分析、用户简单配置即可提交、更新、删除统计分析任务等问题。

3. 飞流

3.1. 飞流架构

首先从整体上看下飞流的架构，如图 1。

为了支持大规模日志数据的集中多源采集和聚合，飞流架构采用 Flume 将不同源的日志数据采集、聚合和传输到一个集中的数据仓库，飞流选择 Kafka 作为这个集中的数据仓库暂时缓存数据，可以有效避免数据在传递和分析阶段丢失和数据量剧增时没有缓冲的问题。大规模日志数据的实时统计分析计算的工作以 Storm 分布式计算系统为基础。飞流的核心就是多个 Storm topology。一个 Kafka topic 对应一个类型的所有日志。一个 Storm topology 订阅一个 Kafka topic。启动一个 Storm topology，就可以承载一个 Kafka topic 对应的日志数据的所有多维统计分析任务。多个 Kafka topic，对应启动多个 Storm topology，就可以承载所有日志数据的多维统计分析任务。启动 topology 是由数据平台部门做的，对于用户是透明的。用户看到的，是一个统一的大规模日志数据实时多维统计分析平台。每个 topology 从 Kafka 拉取各自订阅的 topic 的日志数据。

用户通过 Web UI 输入任务的配置信息(配置信息包含该任务分析的日志数据属于的 Kafka topic 信息、维度信息、度量信息和度量周期 4 部分，飞流目前可配置的统计分析功能包括求和、取平均、最小、最

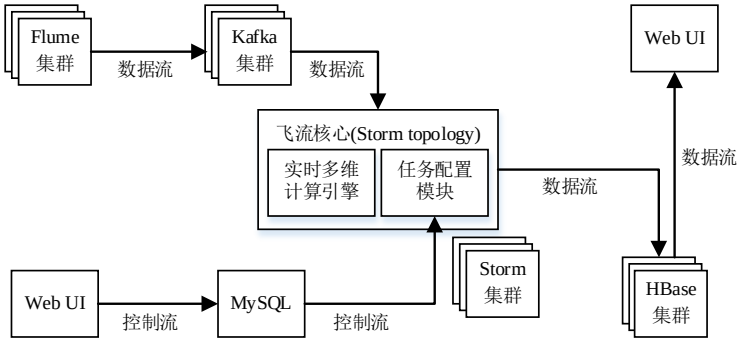


Figure 1. Architecture of Flying Streaming
图 1. 飞流架构

大、计数、唯一计数)等，此外还提供包括提交新任务、更新旧任务和删除旧任务在内的控制信息配置。用户通过 WebUI 配置的任务提交后被写到 MySQL 数据库的任务配置表中，MySQL 会自动为每一个配置信息生成一个 id，该 id 作为配置信息对应的任务的 id。

每个 topology 动态地从任务配置表中读取与自己订阅的 topic 对应的任务配置信息，按照配置好的任务需求分析日志数据，topology 将分析的结果实时持久化存储到 HBase。Web UI 动态地从 HBase 读取最新的分析结果，刷新展示结果的图表。

3.2. 关键技术

3.2.1. 多维统计分析

飞流的可配置的多维统计分析功能是多维分析理论[14]在互联网企业大规模日志数据实时处理场景中的新实践，包括数据源接入、多维数据的提取、多维聚合计算、多维聚合结果持久化等过程，其主要机制如下：

一个用户任务的配置信息中的度量信息由一个或多个<维度名，维度的值数据提取方式，黑/白名单>三元组表示(设置黑/白名单是为了解决任务只处理用户关心的该维度的所有维度值组成的集合的一个子集对应的日志数据的问题)。

假设任务 i 的配置信息中的维度信息里有 n 个<维度名，维度的值数据提取方式，黑/白名单>三元组，以第 j (1 <= j <= n)个三元组为例，维度名对应的所有维度值构成的集合记作 Aj。如果是黑名单，黑名单包含的所有维度值构成的集合记作 Cj，Cj 对于 Aj 的补集 Bj = Aj – Cj (如果是白名单，白名单包含的所有维度值构成的集合记作 Cj，记 Bj = Cj)。

笛卡尔乘积 B1 × B2 × B3 × ⋯ × Bj × ⋯ × Bn 中的元素用 d_k 表示(1 <= k <= B1 × B2 × B3 × … × Bj × … × Bn 中的元素的个数)，d_k 是一个 n 元组，表示成<v_1, v_2, …, v_j, …, v_n>。

任务 i 的数据源接入，就是从其他系统拉取待统计分析的日志数据。

任务 i 的多维数据的提取就是对于每一个到来的日志数据，从里面提取任务 i 配置的 n 个<维度名，维度的值数据提取方式，黑/白名单>三元组对应的 n 个维度的值数据。如果这些值数据都通过各自黑/白名单的筛选，这些值数据按照任务配置信息里维度的先后顺序构成的 n 元组就对应一个 d_k。

对应一个 d_k 的日志数据需要聚合到一起按照任务 i 配置信息里的度量类型和度量周期做周期性的度量计算，这是任务 i 的一个子任务。

任务 i 的多维聚合计算就是计算任务 i 的所有子任务。

任务 i 的多维聚合结果持久化比较简单，就是将多维聚合计算的结果持久化存储到 HBase 中。

这四个过程，通过编写一个 Storm topology 来具体实现。图 2 是 topology 的设计。

1) 数据源接入。这部分在 KafkaSpout 中实现。Topology 订阅哪个 Kafka topic、topology 的并行度等信息是在 topology 的配置文件里设置的，这样可以使 topology 适配各种 Kafka topic。KafkaSpout 从 Kafka partition 拉取订阅的数据，然后发到下游的 ExtractBolt，Tuple(依然是原始日志数据)从 kafkaSpout 到 ExtractBolt 之间是通过 ShuffleGrouping 策略发送的，即一个 Tuple 会被发送到一个随机选择的 ExtractBolt task。

2) 多维数据的提取。这部分在 ExtractBolt 中实现。ExtractBolt 从 MySQL 读取与该 topology 订阅的 topic 对应的任务配置信息。在 ExtractBolt 里，KafkaSpout 每传递过来一个 Tuple，配置的所有任务就会对该日志数据轮流解析，如图 3 的 a 部分。以一个任务 i 为例，如图 3 的 b 部分。

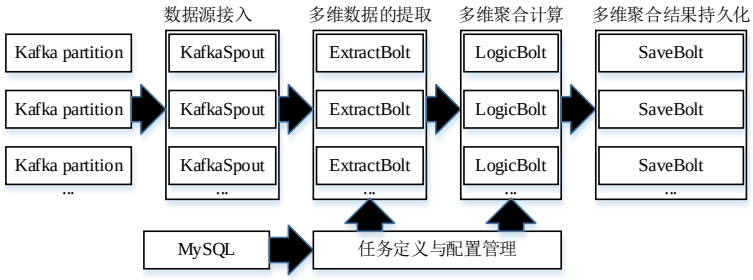
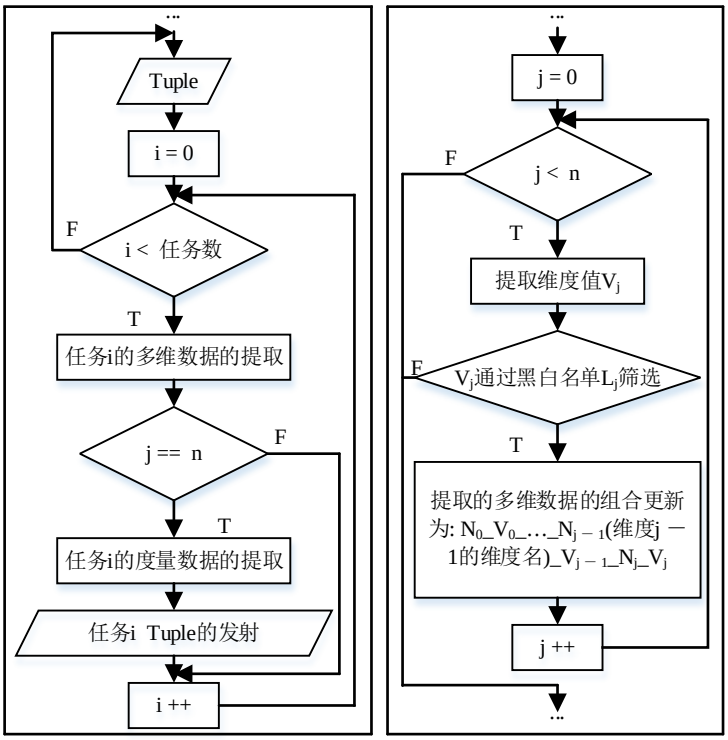


Figure 2. A storm topology
图 2. 一个 Storm topology



a. ExtractBolt的多任务　　b. 任务i的多维数据的提取

Figure 3. Extracting of multidimensional data
图 3. 多维数据的提取

对于第 j 个<维度名，维度的值数据提取方式，黑/白名单>三元组，按照维度的值数据提取方式从日志数据里提取维度的值数据，提取方式有分隔符、正则、先分隔符后正则三种。接着用黑/白名单筛选该值数据，黑/白名单可以用定值、正则或集合实现。若筛选通过，接着进行第 j + 1 个三元组对应的工作。如此，提取完任务 i 配置的 n 个维度的值数据，并且所有维度的值数据都通过黑/白名单的筛选后(这些值数据构成的重组就对应一个 d_k)，还需要根据任务 i 配置信息里的度量信息里的度量数据提取方式，从日志数据里提取度量数据，如图 3 中的 a 部分。接着构造 Tuple(任务 id_维度 1 名_维度 1 值数据_..._维度 n 名_维度 n 值数据，度量数据，度量类型，度量周期)，发射到下游 LogicBolt。Tuple 从 ExtractBolt 到 LogicBolt 是通过 FieldsGrouping 策略发送的，设置的 field 是 Tuple 中第一个字段，则任务 id 相同且维度的值数据相同的 Tuple 会发送到相同的一个 LogicBolt task 中，这样确保对应一个 dk 的所有日志数据能够聚合到一起进行接下来的多维聚合计算，即任务 i 的一个子任务的聚合计算。任务 i 的多维数据的提取完成后，进行第 i + 1 个任务的多维数据的提取。ExtractBolt 在多个 Executor 线程并行执行，每个线程处理不同的数据。

3) 多维聚合计算。这部分工作在 LogicBolt 中实现。LogicBolt 有个哈希表结构的属性，记录被该 LogicBolt 实例对应 Executor 线程处理的每个子任务的当前度量周期的度量结果和开始时间等信息。LogicBolt 收到一个 Tuple，首先解析 Tuple 里的度量类型，根据不同度量类型进入到不同算法中。每个算法的大致结构为：查看度量结果哈希表里面有没有该子任务的项，如果没有则表示这是用户提交的新任务或更新的旧任务的子任务，对 Tuple 里的度量数据做度量计算，最后用该计算结果、0 值分别作为当前度量周期的度量结果和开始时间构造表项并 put 到结果哈希表中。如果有，则读取度量结果，并结合 Tuple 里的度量数据计算本度量周期到目前为止的度量结果并 put 到度量结果表中。度量结果的表示根据度量类型的不同而不同，比如取平均度量类型，需要保存当前度量周期内的所有 Tuple 的度量数据的总和以及 Tuple 的总数。退出算法，接着处理下一个到来的 Tuple。LogicBolt 的 prepare 方法启动一个计时线程，计时线程周期性(周期的值由用户在 topology 的配置文件设置，该值就是度量周期的分度值)扫描度量结果哈希表每个表项的开始时间字段，并和当前时间比较，如发现有子任务的度量周期到点，则计时线程读取该子任务的度量结果并构造新的 Tuple(任务 id_维度 1 名_维度 1 值数据_..._维度 n 名_维度 n 值数据，度量类型，时间戳，度量结果)发送到 SaveBolt，最后将度量结果哈希表中该子任务当前度量周期的度量结果和开始时间清零，表示下一个度量周期的开始。Tuple 从 LogicBolt 到 SaveBolt 同样是通过 FieldsGrouping 策略发送的，设置的 field 是 Tuple 中第一个字段。LogicBolt 在多个 Executor 线程并行执行,每个线程处理不同子任务。

4) 多维聚合结果持久化。这部分工作在 SaveBolt 中实现。HBase 表的 rowKey 设计成“任务 id_维度 1 名_维度 1 值数据_..._维度 n 名_维度 n 值数据_度量类型_时间戳”，value 就是度量结果。SaveBolt 使用 LogicBolt 传递下来的 Tuple 信息构造 rowkey 和 value，持久化存储到 HBase 中。SaveBolt 在多个 Executor 线程并行执行，每个线程处理不同子任务。

3.2.2. 通过配置提交、更新、删除任务

设计如图 4，分为任务配置层、配置信息持久化层和分布式计算层。

1) 任务配置层。由前端用户系统实现，飞流使用 Web UI 实现，一个任务的配置信息和控制信息通过表单实时提交给后台配置信息持久化层。热提交、热更新和热删除。

2) 配置信息持久化层。由 MySQL 实现，有一张任务配置信息表，表里的一个元组表示一个任务的配置信息，有多少个元组，就有多少个任务。控制信息结合配置信息被转化成数据库的相应 INSERT、UPDATE 和 DELETE 操作，以实现插入、更新、删除一个元组。

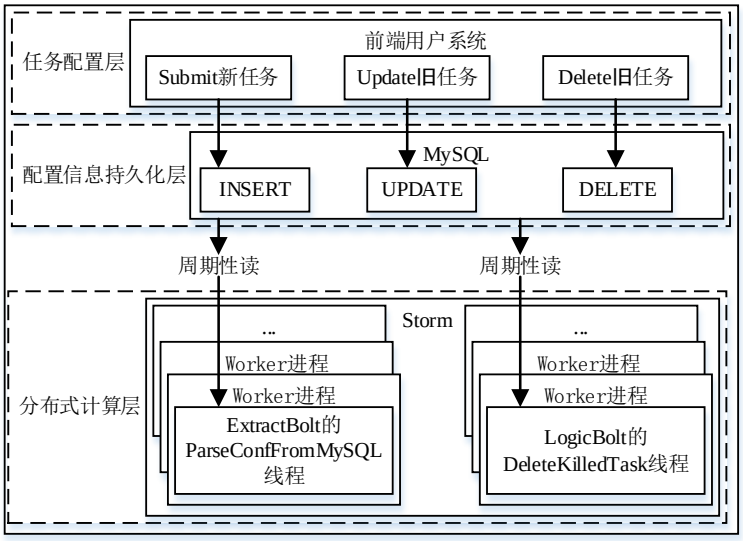


Figure 4. Hot submitting, hot updating and hot deleting
图 4. 热提交、热更新和热删除

3) 分布式计算层。该层是一个 Storm topology。topology 由多个 Worker 进程执行，每个进程可以有一个或多个 Executor 线程，一个 Executor 线程对应一个 Bolt 实例。飞流的设计里，ExtractBolt 实例对应的 Executor 线程执行数据处理之前，在 prepare 方法里启动一个 ParseConfFromMySQL 线程，ParseConfFromMySQL 线程周期性(周期的值，由用户在 topology 的配置文件里设置)读取 MySQL 的任务配置信息表，用最新的任务配置信息更新 ExtractBolt 实例中相应的任务配置信息属性，ExtractBolt 根据这些最新的配置信息进行多维数据的提取时就可以响应新任务的提交、旧任务的更新或删除。LogicBolt 对新任务的提交和旧任务的更新的响应在多维聚合计算部分已描述。LogicBolt 实例对应的 Executor 线程执行数据处理之前，在 prepare 方法里启动一个 DeleteKilledTask 线程，该线程周期性读取 MySQL 的任务配置信息表，并和度量结果哈希表比较，如果某个子任务存在于度量结果哈希表中，而在任务配置信息表中不存在该子任务对应的原任务的配置信息，则表示此原任务刚刚被用户删除，则从度量结果哈希表中删除此子任务对应的表项，这样就响应了用户旧任务的删除。这样的设计，做到了将任务的逻辑和程序本身解耦，程序不和具体的某个任务相关，而是灵活的适配各种任务。

4. 结论

本论文整合 Flume、Kafka、Storm、HBase 等开源系统设计了飞流大规模日志数据实时多维统计分析平台，解决了多种日志数据接入、实时多维度统计分析、用户简单配置即可提交、更新、删除统计分析任务等关键技术问题，提供了飞流平台上不同用户通过 web 配置界面同时提交、同时执行、同时删除不同的统计分析任务的功能，不需要用户进行大数据分析处理编程。飞流平台在互联网企业中实际应用效果较好，满足了业务部门和运维部门的大部分日志数据多维统计分析需求。未来，支持更多的度量类型和支持 SQL，还需要做一些工作。

参考文献 (References)

[1] Apache Flume (2017) Flume Homepage. http://flume.apache.org/

[2] Percy, M. (2017) Flume NG Performance Measurements. https://cwiki.apache.org/confluence/display/FLUME/Flume+NG+Performance+Measurements

[3] Apache Kafka (2017) Kafka Homepage. http://kafka.apache.org/

[4] Kreps, J. (2017) Benchmarking Apache Kafka: 2 Million Writes Per Second (On Three Cheap Machines). http://kafka.apache.org/performance

[5] Apache Spark Streaming (2017) Spark Streaming Homepage. http://spark.apache.org/streaming

[6] Apache Storm (2017) Storm Homepage. http://storm.apache.org/

[7] Naik, R. and Amin, S. (2017) Microbenchmarking Apache Storm 1.0 Performance. https://hortonworks.com/blog/microbenchmarking-storm-1-0-performance/

[8] 李川, 鄂海红, 宋美娜. 基于 Storm 的实时计算框架的研究与应用[J]. 软件, 2014, 35(10): 16-20.

[9] Apache HBase (2017) HBase Homepage. http://hbase.apache.org/

[10] Misty (2017) Testing HBase Performance and Scalability. https://wiki.apache.org/hadoop/Hbase/PerformanceEvaluation

[11] 张智, 龚宇. 分布式存储系统 HBase 关键技术研究[J]. 现代计算机, 2014(32): 33-37.

[12] 陈任飞, 吕玉琴, 侯宾. 基于 Flume/Kafka/Spark 的分布式日志流处理系统的设计与实现[EB/OL]. http://www.paper.edu.cn/html/releasepaper/2015/07/130/

[13] 薛瑞, 朱晓民. 基于 Spark Streaming 的实时日志处理平台设计与实现[J]. 电信工程技术与标准化, 2015(9): 55-58.

[14] 廖开际. 数据仓库与数据挖掘[M]. 北京: 北京大学出版社, 2008: 79-86.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 359-368
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74044

Hans汉斯

Image Analysis by Charlier Moments

Guanghui Shi, Xuan Wang

School of Physics and Information Technology, Shaanxi Normal University, Xi'an Shaanxi
Email: shiguanghui_0216@163.com

Received: Apr. 10th, 2017; accepted: Apr. 23rd, 2017; published: Apr. 27th, 2017

Abstract

The existing methods for extracting the translation and scale invariants from the discrete orthogonal moments are via a linear combination of the corresponding invariants of geometric moments or image normalization, which led to calculational errors. In this paper, a novel kind of discrete orthogonal moments named as Charlier moment is proposed based on the discrete Charlier polynomials, and then an approach to directly derive the translation and scale invariants from Charlier moments is also presented. Experimental results show the high classification and representation accuracy of these invariants as a result of direct calculation instead of the image normalization or a linear combination of the corresponding invariants of geometric moments. It is also shown that these invariants are relatively robust in the presence of image noise and are potentially useful as a kind of invariant descriptors in some image analysis and pattern recognition.

Keywords

Discrete Orthogonal Moments, Charlier Polynomials, Translation, Scale Invariants, Pattern Recognition

基于Charlier矩的图像分析

时光慧，王　晅

陕西师范大学物理学与信息技术学院，陕西　西安
Email: shiguanghui_0216@163.com

收稿日期：2017年4月10日；录用日期：2017年4月23日；发布日期：2017年4月27日

摘　要

针对离散正交矩的平移和尺度不变量只能通过图像归一化或者借助几何矩不变量的线性组合间接获取，会带来较大的表示误差，本文基于离散Charlier多项式，提出了一种新的离散正交矩——Charlier矩，并给出了Charlier矩平移与尺度不变量的直接计算方法。实验表明，由Charlier矩直接计算的平移和尺

文章引用: 时光慧, 王晅. 基于 Charlier 矩的图像分析[J]. 计算机科学与应用, 2017, 7(4): 359-368.
https://doi.org/10.12677/csa.2017.74044

度不变量与现有方法相比，具有较高的表示精度与分类准确率，而且对图像噪声有较强的稳定性，可以应用于图像不变分析与目标识别等应用领域。

关键词

离散正交矩，Charlier多项式，平移和尺度不变量，模式识别

1. 引言

1962 年 Hu [1]提出的图像几何矩理论受到广泛关注。图像几何矩不变描述量具有平移、旋转和尺度不变的特性，而且计算效率较高，可以应用于图像不变分析与机器视觉的各个领域。然而，几何矩的基函数是非正交的，导致了较大的信息冗余，所以基于几何矩重构图像是非常困难的，而且该类矩对图像噪声非常敏感，带来一定的表示误差与分类错误率。针对以上问题，一些研究工作者基于连续正交基函数提出了 Zernike [2] [3] [4]和 Legendre [5]等连续正交矩。由于连续正交矩的基函数是正交的连续函数，所以可以具有较小的信息冗余，基于连续正交矩可以对图像精确重构，此外，与几何矩相比，对噪声有更好的稳定性。但是，由于连续正交矩的定义涉及到二维连续积分形式，而数字图像是定义在离散域的数字矩阵，所以计算过程中需要进行坐标转换和积分近似，所以存在数值积分近似误差与几何误差[4]。这些误差会严重影响图像的重构精度与分类精度。

Mukundan [6]在 2001 年提出一种基于离散 Tchebichef 多项式的离散正交矩。随后，Yap [7]提出另一种基于 Krawtchouk 多项式的离散正交矩。离散正交矩的基函数与数字图像域完全匹配，定义与计算过程消除了由坐标转换和近似误差引起的数值积分近似误差与几何误差，具有良好的图像表征能力与分类精度，而且与连续正交矩相比，抵抗噪声的干扰能力也明显增强，然而，离散正交矩缺乏本质上的平移与尺度变换不变性，平移和尺度不变量是通过图像平移与尺度归一化方法或者借助几何矩不变量的线性组合间接获取通，这一过程会引入新的计算误差，从而影响离散正交矩的优良性能。本文基于 Charlier 离散正交多项式[8]，提出了一种新的离散正交矩，并基于该离散正交多项式的特殊性质，提出一种直接计算该矩的平移和尺度不变量的方法。实验结果验证了该方法所获得的 Charlier 离散正交矩尺度和平移不变量具有很好的表示与分类性能。

2. Charlier 离散正交矩

大小为 $N\times N$ 的图像函数为 $f(x,y)$，二维 Charlier 矩[8]定义如下：

$$A_{nm}=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\tilde{t}_n^{(a1)}(x)\tilde{t}_m^{(a1)}(y)\,f(x,y) \tag{1}$$

其中，$\tilde{t}_n^{(a1)}(x)$ 和 $\tilde{t}_m^{(a1)}(y)$ 分别表示 n 阶和 m 阶标准 Charlier 离散正交多项式。n 阶 Charlier 多项式可用超几何函数定义如下：

$$t_n^{(a1)}(x)={}_2F_0(-n,-x;;-1/a1) \tag{2}$$

其中，

$$ {}_2F_0\left(a,b;;x\right)=\sum_{k=0}^{\infty}\left(a\right)_k\left(b\right)_k\frac{x^k}{k!} \tag{3} $$

$$ \left(a\right)_k=a\left(a+1\right)\left(a+2\right)\cdots\left(a+k-1\right) \tag{4} $$

有

$$ \left\langle a\right\rangle_k=\left(-1\right)^k\left(-a\right)_k \tag{5} $$

因此，n 阶 Charlier 多项式可以写为：

$$ \begin{aligned} t_n^{(a1)}\left(x\right)&=\sum_{k=0}^{n}\frac{\left\langle n\right\rangle_k}{\left(-a1\right)^k k!}\left\langle x\right\rangle_k \\ &=\sum_{k=0}^{n}B_{n,k}\left\langle x\right\rangle_k \end{aligned} \tag{6} $$

其中，

$$ B_{n,k}=\frac{\left\langle n\right\rangle_k}{\left(-a1\right)^k k!} \tag{7} $$

$$ \left\langle x\right\rangle_k=\sum_{i=0}^{k}s\left(k,i\right)x^i \tag{8} $$

其中，$s\left(k,i\right)$ 表示第一类斯特林数，有以下递归关系：

$$ \begin{aligned} &s\left(k,i\right)=s\left(k-1,i-1\right)-\left(k-1\right)s\left(k-1,i\right),\ k\geq 1,i\geq 1 \\ &s\left(k,0\right)=0,\ s\left(0,0\right)=1. \end{aligned} \tag{9} $$

离散 Charlier 多项式 $t_n^{(a1)}\left(x\right)$ 和 $t_m^{(a1)}\left(x\right)$，$0\leq m,n\leq\infty$，$a1>0$，满足以下正交关系：

$$ \sum_{x=0}^{\infty}w^{(a1)}\left(x\right)t_n^{(a1)}\left(x\right)t_m^{(a1)}\left(x\right)=\rho_n^{(a1)}\delta_{nm} \tag{10} $$

其中，$\rho_n^{(a1)}$ 表示平方模，δ_{nm} 表示狄拉克函数，$w^{(a1)}\left(x\right)$ 是加权函数，有

$$ w^{(a1)}\left(x\right)=\frac{e^{-a1}a1^x}{x!} \tag{11} $$

$$ \rho_n^{(a1)}=\frac{n!}{a1^n} \tag{12} $$

由此，n 阶标准 Charlier 离散正交多项式 $\tilde{t}_n^{(a1)}\left(x\right)$ 为：

$$ \tilde{t}_n^{(a1)}\left(x\right)=t_n^{(a1)}\left(x\right)\sqrt{\frac{w^{(a1)}\left(x\right)}{\rho_n^{(a1)}}} \tag{13} $$

根据公式(6)，(7)，(8)和(13)，$\tilde{t}_n^{(a1)}\left(x\right)$ 又可以写成：

$$ \tilde{t}_n^{(a1)}\left(x\right)=\sqrt{\frac{w^{(a1)}\left(x\right)}{\rho_n^{(a1)}}}\sum_{k=0}^{n}B,k\sum_{i=0}^{k}s\left(k,i\right)x^i \tag{14} $$

可以证明，标准 Charlier 离散正交多项式存在以下递推关系，

$$A\tilde{t}_n^{(a1)}(x)=B\tilde{t}_{n-1}^{(a1)}(x)+C\tilde{t}_{n-2}^{(a1)}(x) \tag{15}$$

其中 $A=-a1$，

$$\begin{aligned} B&=(x-n+1-a1)\sqrt{\frac{a1}{n}},\\ C&=(n-1)\sqrt{\frac{a1^2}{n(n-1)}}. \end{aligned} \tag{16}$$

零阶和一阶 Charlier 多项式为：

$$\begin{aligned} \tilde{t}_0^{a1}(x)&=\sqrt{\frac{e^{-a1}a1^x}{x!}},\\ \tilde{t}_1^{(a1)}(x)&=\frac{a1-x}{a1}\sqrt{\frac{e^{-a1}a1^{x+1}}{x!}}. \end{aligned} \tag{17}$$

根据以上的递推关系，可以实现 Charlier 基函数的快速计算，从而提高 Charlier 矩的计算速度。

3. Charlier 矩不变量的快速计算

3.1. Charlier 矩的平移不变量

给定图像 $f(x,y)$，为了计算其平移与尺度变换不变量，进行以下的预处理，

$$\tilde{f}(x,y)=\left[w^{(a1)}(x)w^{(a1)}(y)\right]^{(-1/2)}f(x,y) \tag{18}$$

计算预处理后的图像对应用的质心坐标，有

$$\begin{aligned} x_0&=\frac{\sum\sum xf(x,y)}{\sum\sum f(x,y)},\\ y_0&=\frac{\sum\sum yf(x,y)}{\sum\sum f(x,y)}. \end{aligned} \tag{19}$$

对应的 Charlier 中心矩为

$$\begin{aligned} A'_{nm}&=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\tilde{t}_n^{(a1)}(x-x_0)\tilde{t}_m^{(a1)}(y-y_0)\tilde{f}(x,y)\\ &=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\bar{t}_n^{(a1)}(x-x_0)\bar{t}_m^{(a1)}(y-y_0)f(x,y) \end{aligned} \tag{20}$$

$\bar{t}_n^{(a1)}(x)$ 表示归一化后的 Charlier 多项式：

$$\bar{t}_n^{(a1)}(x)=\frac{1}{\sqrt{\rho_n^{(a1)}}}t_n^{(a1)}(x) \tag{21}$$

运用式子(8)和(14)，又可写为：

$$\bar{t}_n^{(a1)}(x)=\sum_{k=0}^{n}\bar{B}_{n,n-k}\langle x\rangle_k \tag{22}$$

其中有，

$$\bar{B}_{n,n-k}=\frac{B_{n,n-k}}{\sqrt{\rho_n^{(a1)}}} \tag{23}$$

平移 Charlier 多项式 $\bar{t}_n^{(a1)}(x-x_0)$ 可以表示为下面可分离形式：

$$\bar{t}_n^{(a1)}(x-x_0)=\sum_{k=0}^{n}\bar{v}_{n,n-k}(-x_0)\bar{t}_n^{(a1)}(x) \tag{24}$$

有

$$\begin{aligned}&\bar{v}_{n,n-k}(-x_0)\\&=\frac{1}{\bar{B}_{n-k,n-k}}\left[\sum_{l=0}^{k}\binom{n-k+l}{l}\bar{B}_{n,n-k+l}\langle -x_0\rangle_l-\sum_{i=0}^{k-1}\bar{B}_{n-i,n-k}\bar{v}_{n,n-i}(-x_0)\right]\end{aligned} \tag{25}$$

类似地，y 轴方向上的平移 Charlier 多项式可以写成：

$$\bar{t}_m^{(a1)}(y-y_0)=\sum_{s=0}^{m}\bar{v}_{m,m-s}(-x_0)\bar{t}_m^{(a1)}(y) \tag{26}$$

和

$$\begin{aligned}&\bar{v}_{m,m-s}(-y_0)\\&=\frac{1}{\bar{B}_{m-s,m-s}}\left[\sum_{r=0}^{s}\binom{m-s+r}{r}\bar{B}_{m,m-s+r}\langle -y_0\rangle_r-\sum_{j=0}^{s-1}\bar{B}_{m-j,m-s}\bar{v}_{m,m-j}(-y_0)\right]\end{aligned} \tag{27}$$

公式(20)表示的 Charlier 中心距又可以写成：

$$A'_{nm}=\sum_{k=0}^{n}\sum_{s=0}^{m}\bar{v}_{n,n-k}(-x_0)\bar{v}_{m,m-s}(-y_0)A_{n-k,m-s} \tag{28}$$

由上可知，Charlier 中心距可以用 Charlier 矩线性表示。根据上述的递推关系，Charlier 矩的平移不变量可以快速的计算出来。

3.2. 尺度不变量

假设原始图像在横轴和纵轴上的尺度因子分别为 a 和 b，则预处理后的图像函数 $\tilde{f}(x,y)=\left[w^{(a1)}(x)w^{(a1)}(y)\right]^{(-1/2)}f(x,y)$ 尺度化后的 Charlier 矩为：

$$\begin{aligned}A''_{nm}&=ab\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\tilde{t}_n^{(a1)}(ax)\tilde{t}_m^{(a1)}(by)\tilde{f}(x,y)\\&=ab\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\bar{t}_n^{(a1)}(ax)\bar{t}_m^{(a1)}(by)f(x,y)\end{aligned} \tag{30}$$

由公式(10)和(23)，我们可以把归一化的多项式写成：

$$\begin{aligned}\bar{t}_n^{(a1)}(x)&=\sum_{k=0}^{n}\bar{B}_{n,n-k}\langle x\rangle_k\\&=\sum_{k=0}^{n}\sum_{i=0}^{k}\bar{B}_{n,n-k}s(k,i)x^i\\&=\sum_{i=0}^{n}\sum_{k=0}^{n-i}\bar{B}_{n,n-k}s(n-k,i)x^i=\sum_{i=0}^{b}C(n,i)x^i\end{aligned} \tag{31}$$

其中，

$$C(n,i)=\sum_{k=0}^{n-i}\bar{B}_{n,n-k}s(n-k,i) \tag{32}$$

因此，横轴方向的尺度化 Charlier 多项式可以表示为：

$$\overline{t}_n^{(a1)}(ax)=\sum_{i=0}^{n}C(n,i)a^i x^i \tag{33}$$

未经过尺度化的多项式和尺度化后的多项式之间的数学关系为：

$$\sum_{k=0}^{n}\gamma_{n,k}\overline{t}_n^{(a1)}(ax)=a^n\sum_{k=0}^{n}\gamma_{n,k}\overline{t}_n^{(a1)}(x), \tag{34}$$

其中，

$$\gamma_{n,n}=1,$$

$$\gamma_{n,k}=\sum_{r=0}^{n-k-1}\frac{-C(n-r,k)\gamma_{n,n-r}}{C(k,k)},\quad 0\le k<n. \tag{35}$$

类似的，纵轴方向的 Charlier 多项式可以表示为：

$$\sum_{s=0}^{m}\gamma_{m,s}\overline{t}_m^{(a1)}(by)=b^m\sum_{s=0}^{m}\gamma_{m,s}\overline{t}_m^{(a1)}(y) \tag{36}$$

因此，有如下：

$$\begin{aligned}\eta_{nm}&=\sum_{k=0}^{n}\sum_{s=0}^{m}\gamma_{n,k}\gamma_{m,s}A''_{nm}\\&=a^{n+1}b^{m+1}\sum_{k=0}^{n}\sum_{s=0}^{m}\gamma_{n,k}\gamma_{m,s}A_{nm}\end{aligned} \tag{37}$$

通过消除因子 a 和 b，我们可以推导出 Charlier 矩的尺度不变量：

$$\phi_{nm}=\frac{\eta_{nm}\eta_{00}^{r+1}}{\eta_{n+r,0}\eta_{0,m+r}},\quad n,m=0,1,2,\cdots,\quad r=1,2,3,\cdots \tag{38}$$

根据(32)、(33)、(35)表示的递推公式，我们可以快速计算出来尺度不变量。由上式可知，尺度不变量来也可以表示平移不变量，即尺度不变量也可以表示为尺度平移不变量。

4. 仿真实验

为了验证本文方法的性能，选择大小为英文字母、汉字“幕”二值图像与“招财猫”及蝴蝶灰度图像进行了实验仿真，并于最新的图像平移与尺度不变量计算方法-Tchebichef 不变量[9]进行了性能比较。仿真实验的主要目的是首先测试本文方法计算的平移不变量与尺度不变量的稳定性，其次测试基于上述不变量的分类性能及对图像噪声干扰的鲁棒性。

4.1. 平移不变量的稳定性

实验采用大小为 60×60 的汉字“幕”和 64×64 的英文字符“Z”，在 x 与 y 方向进行了平移量为−2到 2，增量为 1 的平移操作，通过本文方法计算平移不变量，相应结果见表 1 和表 2，可以看出，由本文方法计算的平移不变量具有很好的稳定性。

4.2. 尺度不变量

实验采用大小为128×128 灰度图像“招财猫”，分别进行了尺度因子为 0.6 到 1.4，增量为 0.2 的尺度变换，对变换结果利用本文方法计算尺度变换不变量与文献[9]方法计算 Tchebichef 不变量，相应结果见表 3 和表 4，可以看出，由本文方法计算的尺度变换平移不变量具有很好的稳定性，而文献[9]方法计算的 Tchebichef 不变量变化范围较大，说明本文方法具有更好的计算精度与不变表示能力。

Table 1. Selected orders of Charlier central moments for an English letter “Z”
表 1. 英文字符“Z”部分 Charlier 中心矩

	$\Delta i=-2$ $\Delta j=-2$	$\Delta i=-1$ $\Delta j=-2$	$\Delta i=-1$ $\Delta j=-1$	$\Delta i=0$ $\Delta j=0$	$\Delta i=1$ $\Delta j=1$	$\Delta i=2$ $\Delta j=2$
ψ_{20}	401.07082	402.89158	401.07082	401.07082	401.07082	401.07082
ψ_{02}	480.82693	479.89774	480.82693	480.82693	480.82693	480.82693
ψ_{21}	666.89656	669.68833	666.89656	666.89656	666.89656	666.89656
ψ_{12}	722.30390	723.41046	722.30390	722.30390	722.30390	722.30390
ψ_{30}	476.06276	478.78428	476.06276	476.06276	476.06276	476.06276
ψ_{03}	604.67024	603.29863	604.67024	604.67024	604.67024	604.67024

Table 2. Selected orders of Charlier central moments for a Chinese letter “幕”
表 2. 汉字“幕”部分 Charlier 中心矩

	$\Delta i=-2$ $\Delta j=-2$	$\Delta i=-2$ $\Delta j=-1$	$\Delta i=-1$ $\Delta j=-1$	$\Delta i=-1$ $\Delta j=0$	$\Delta i=0$ $\Delta j=1$	$\Delta i=2$ $\Delta j=2$
ψ_{20}	979.68870	974.13980	982.89549	977.39687	977.75855	979.34989
ψ_{02}	1033.1864	1036.7779	1035.8423	1039.5638	1037.7183	1027.2964
ψ_{21}	1557.9146	1552.4218	1561.6946	1556.3197	1554.9388	1553.4436
ψ_{12}	1594.1952	1594.7600	1597.4277	1598.1656	1595.0396	1585.5860
ψ_{30}	1164.8134	1156.7273	1168.4153	1160.3794	1160.7488	1164.5483
ψ_{03}	1245.6679	1251.5537	1248.1471	1254.2098	1250.9622	1236.5141

Table 3. The scale invariant descriptors of Charlier moments for “Fortune Cat”
表 3. 招财猫 Charlier 矩尺度不变量

	Scale				
	0.6	0.8	1.0	1.2	1.4
ψ_{20}	9495.04371	9371.42680	9365.37329	9335.56486	9310.38017
ψ_{02}	12034.0387	11876.4830	11860.8917	11806.7750	11788.7129
ψ_{21}	12757.5156	12611.0302	12607.0146	12569.6818	12541.9991
ψ_{12}	13616.5974	13459.0287	13459.9855	13414.0276	13392.8197
ψ_{30}	10926.0734	10786.7232	10776.0769	10743.3130	10712.2394
ψ_{03}	13946.7244	13763.5516	13750.4789	13689.3063	13668.3996

Table 4. The scale invariant descriptors of Tchebichef moments for “Fortune Cat”
表 4. 招财猫 Tchebichef 矩尺度不变量

	Scale				
	0.6	0.8	1.0	1.2	1.4
ψ_{20}	15.1957322	26.8063918	41.3923290	59.7420159	81.4111538
ψ_{02}	19.2891761	33.9719517	52.4218221	75.5562791	103.082012
ψ_{21}	10.3460797	18.2854908	28.2480322	40.7828456	55.6053873
ψ_{12}	11.0494524	19.5150548	30.1592498	43.5223601	59.3775296
ψ_{30}	13.40296	23.6665504	36.5412983	52.7559806	71.8836758
ψ_{03}	17.1388877	30.1978445	46.6273912	67.2225386	91.7207660

4.3. 分类性能

实验采用大小为 64×64 的 26 个英文字符二值图像数据集与蝴蝶灰度图像数据集，进行了平移与尺度变换不变分类，选用以下的 6 个不变量作为分类特征，

$$V=\left[\psi_{20};\ \psi_{02};\ \psi_{21};\ \psi_{12};\ \psi_{30};\ \psi_{03}\right] \tag{39}$$

选择欧氏距离作为衡量特征向量的相似性度量，欧氏距离定义为

$$d\left(z_s,z_t^{(k)}\right)=\sum_{j=1}^{T}\left(z_{sj}-z_{tj}^{(k)}\right)^2 \tag{40}$$

其中 z_s 是未知样本的 T 维特征向量，$z_t^{(k)}$ 是 k 类的训练向量，选取最近邻分类方法进行分类，性能测评指标采用分类正确率 η，定义如下

$$\eta=\frac{\text{正确识别的图片}}{\text{全部的图片}}\times100\% \tag{41}$$

实验 1：采用 26 个大写字母图像作为训练集，对于训练集中的每幅图像进行平移变换和尺度变换，尺度因子 $a=b=\{0.8,0.9,1.0,1.1,1.2\}$，平移量为 $\Delta i,\Delta j=-2,-1,0,1,2$，并对每幅变换后的图像感染密度为 0%~4%，增量为 1%的椒盐噪声得到测试集，测试集中的部分图像见图 1，应用本文方法计算出的 Charlier(C)与文献[9]方法计算的 Tchebichef(T)不变量进行分类，分类结果见表 5。可以看出，在未感染噪声的情况下，本文方法可以对所有全部测试图像进行正确分类，而文献[9]的方法正确分类率仅有 37%。随着噪声感染的密度增加，正确识别率也明显降低，但均显著高于文献[9]的方法，说明本文方法在噪声干扰的情况下，也能获得一定的识别精度。

实验 2：采用七张蝴蝶灰度图片作为训练集，对于训练集中的每幅图像进行平移变换和尺度变换，尺度因子 $a=b=\{0.8,0.9,1.0,1.1,1.2\}$，平移量为 $\Delta i,\Delta j=-2,-1,0,1,2$，并对每幅变换后的图像感染密度为 0%~4%，增量为 1%的椒盐噪声得到测试集，测试集中的部分图像见图 2，应用本文方法计算出的 Charlier(C)与文献[9]方法计算的 Tchebichef(T)不变量进行分类，分类结果见表 6。可以看出，对于灰度图像，在未感染噪声的情况下，本文方法正确识别率显著高于文献[9]的方法。对噪声干扰的鲁棒性也明显好于文献[9]的方法。

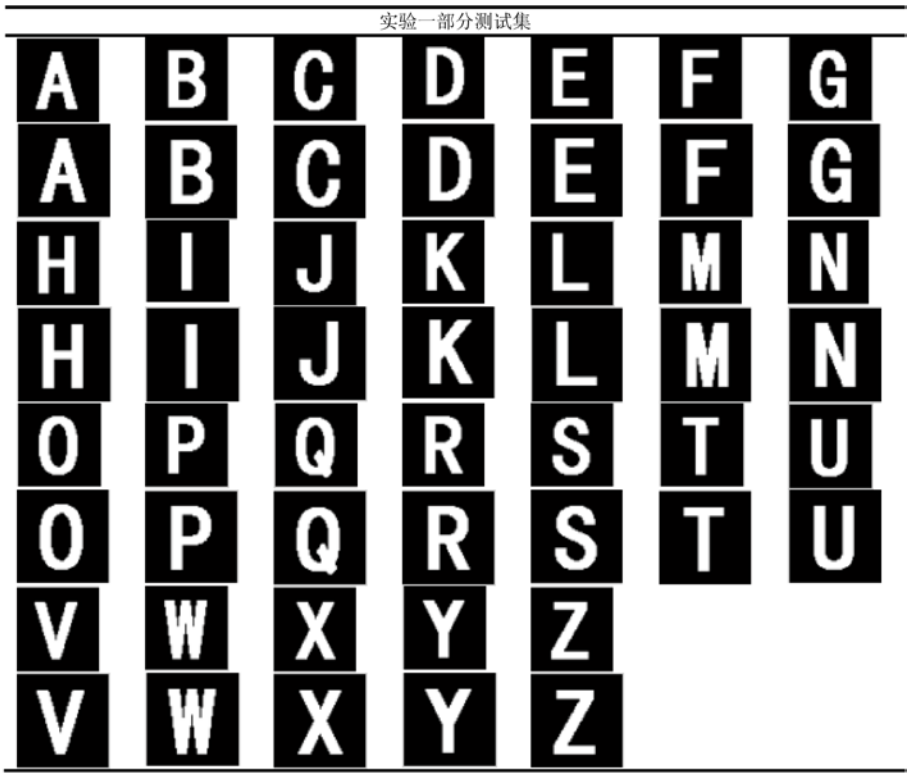


Figure 1. Part of the first testing set
图 1. 实验一部分测试集

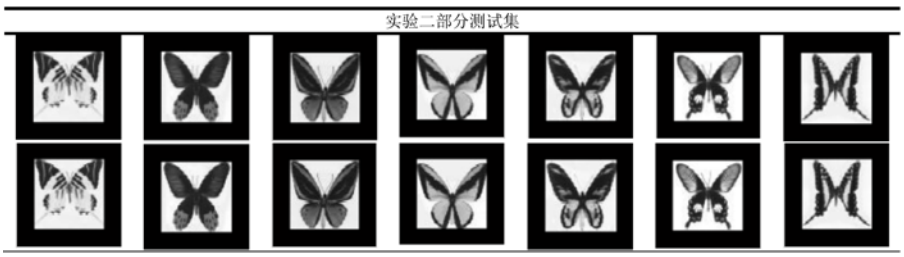


Figure 2. Part of the second testing set
图 2. 实验二部分测试集

Table 5. Classification results of uppercase letters
表 5. 大写字母的分类结果

噪声密度	0.00	0.01	0.02	0.03	0.04
T	0.3718	0.3333	0.3077	0.2949	0.2821
C	1.0	0.9615	0.8077	0.7308	0.6795

Table 6. Classification results of butterfly images
表 6. 蝴蝶图片的分类结果

噪声密度	0.00	0.01	0.02	0.03	0.04
T	0.4286	0.4286	0.4386	0.3810	0.3810
C	0.9524	0.9048	0.9048	0.8571	0.8571

5. 结论

本文基于 Charlier 多项式提出了一种新的离散正交矩——Charlier 矩，并给出了 Charlier 矩平移与尺度不变量的直接计算方法。与现有离散矩的平移与尺度不变量的计算方法相比，具有更好的计算精度。实验验证了本文方法较现有最新方法在二值图像与灰度图像都具有非常好的稳定性与分类精度，并对噪声干扰也具有较高的鲁棒性，作为平移与尺度不变特征描述可以应用于数字图像分析与机器视觉的相关领域。

参考文献 (References)

[1] Hu, M.K. (1962) Visual Pattern Recognition by Moment Invariants. *IEEE Transactions on Information Theory*, **2**, 179-187.

[2] Kamila, N.K. and Mahapatra, S.N. (2005) Invariance Image Analysis Using Modified Zernike Moments. *Pattern Recognition*, **28**, 747-753.

[3] Singh, C. and Walia, E. (2010) Fast and Numerically Stable Methods for the Computation of Zernike Moments. *Pattern Recognition*, **43**, 2497-2506.

[4] Chong, C.-W., Raveendran, P. and Mukundan, R. (2003) Translation Invariants of Zernike Moments. *Pattern Recognition*, **36**, 1765-1773.

[5] Sun, Z.-H. (2014) Generalized Legendre Polynomials and Related Super Congruences. *Journal of Number Theory*, **143**, 293-319.

[6] Mukundan, R., Ong, S.H. and Lee, P.A. (2001) Image Analysis by Tchebichef Moments. *IEEE Transactions on Image Processing*, **10**, 1357-1364. https://doi.org/10.1109/83.941859

[7] Yap, P.T., Paramesran, R. and Ong, S.H. (2003) Image Analysis by Krawtchouk Moments. *IEEE Transactions on Image Processing*, **12**, 1367-1377. https://doi.org/10.1109/TIP.2003.818019

[8] Zhu, H., Liu, M., Shu, H., Zhang, H. and Luo, L. (2010) General Form for Obtaining Discrete Orthogonal Moments. *IEEE Transactions on Image Processing*, **4**, 335-352. https://doi.org/10.1049/iet-ipr.2009.0195

[9] Zhu, H., Shu, H., Xia, T., Luo, L. and Coatrieux, J. (2007) Translation and Scale Invariants of Tchebichef of Moments. *Pattern Recognition*, **37**, 2530-2542.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 369-376
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74045

Hans汉斯

Aurora Image Classification Based on Global Features

Zhenting Cao, Xuan Wang

School of Physics and Information Technology, Shanxi Normal University, Xi'an Shaanxi
Email: 18392518428@163.com

Received: Apr. 10th, 2017; accepted: Apr. 23rd, 2017; published: Apr. 30th, 2017

Abstract

Aurora classification is of great significance to the research on the way and degree of the influence of solar activity on the earth. The existing methods for aurora image classification are mainly based on the local features extracted from the aurora images. These local features are sensitive to noise and variances of the position and orientation, so their classification accuracies and robustness are insufficient for complicated applications. This paper proposed a novel aurora image classification method based on the global feature descriptors. In the proposed method, the aurora image is projected to Radon domain via Radon transform, and then, the variances of columns are determined, the rotation invariant features are obtained via circular shift operation on the variance sequence to let the maximum value in the first place, which completes the rotation normalization of the variance sequence. A nearest neighbor classifier based on Euclidean distance is used for classification. Experimental results show that the proposed approach yields a better performance in terms of the correct classification percentages compared with the aurora image classification method based on representative local feature. It is also shown that the proposed approach yields observably low computational cost and relatively high robustness to noise, the variations of orientation and position of aurora images.

Keywords

Aurora Classification, Global Feature, Radon Transform

基于全局特征的极光图像分类

曹振婷，王　晅

陕西师范大学物理学与信息技术学院，陕西　西安
Email: 18392518428@163.com

收稿日期：2017年4月10日；录用日期：2017年4月23日；发布日期：2017年4月30日

文章引用：曹振婷，王晅. 基于全局特征的极光图像分类[J]. 计算机科学与应用, 2017, 7(4): 369-376.
https://doi.org/10.12677/csa.2017.74045

摘　要

极光分类对于太阳活动对地球的影响方式的研究具有非常重要的意义，现有的极光分类方法主要是基于极光图像的局部特征，这些局部特征对噪声干扰及极光图像的位置、方向等变化较为敏感，很难满足实际应用的要求。本文提出了一种新的基于全局特征的极光图像分类方法，在该方法中，极光图像通过Radon变换投影到Radon域，然后计算投影矩阵中每列的方差作为特征，为了实现方向变化不变性，对该方差序列进行循环移位使得该序列方差最大的值居于首位，进行旋转归一化处理，然后，应用基于欧氏距离的最近邻分类器实现极光图像分类。实验表明，该方法分类精度明显高于现有的基于局部特征的分类方法，对噪声干扰及极光图像位置、方向的鲁棒性显著高于现有方法，而且计算效率也高于基于局部特征的分类方法。

关键词

极光分类，全局特征，Radon变换

1. 引言

极光是太阳风所携带的高能带电粒子与地球高空大气层中的中性气体相撞而产生的唯一能够通过肉眼观测到的具有极区特征的地球物理现象。极光的合理分类对研究空间天气和太阳风——磁层之间的耦合作用具有极其重要的意义。

2004 年，Syrjäsuo 等人[1]对加拿大 CANOPUS (Canadian Auroral Network for the Open Program Unified Study)项目[2]的数据进行分析后，将全天空极光图像分为三种典型的极光形态：弧型(auroral arcs)、斑块型(patchy auroras)和欧米伽型(Omega-bands)。在他们的工作中，首先对极光图像进行分割和极光区域检测，然后使用傅里叶描述子来提取极光图像的形状特征。该类方法由于仅关注极光图像的轮廓结构，适用于形状单一、轮廓明确的极光弧，很难应用于具有复杂结构极光图像的分类问题中。

针对上述问题，随后出现了基于子空间分解的极光图像分类方法[3]，该类方法对极光图像应用主分量分析(PCA)、线性判别分析(LDA)进行分解，其分解系数作为极光特征实现分类。与基于轮廓结构的方法相比，子空间分解适合于具有复杂结构极光图像的分类，但是，其特征的表示能力与分类精度有限，达不到实际应用的要求。

一些研究者将图像分析中的纹理识别方法应用到极光分类中，主要关注极光图像中的纹理特征，高凌君等[4]首先提取极光样本图像的 Gabor 特征，利用 K-均值聚类算法进行训练样本选择，然后引入 AdaBoost 算法进行分类器的构建实现极光图像的检测。韩慎苗等[5]采用多阶统计特征结合小波分解，通过提取图像的灰度分布、灰度共生矩阵和行程长度矩阵等特征来表征纹理信息，并结合 KNN 和后向神经网络(BPNN)进行自动分类。该类方法分类精度普遍高于基于子空间分解的极光图像分类方法，但是，对噪声干扰与极光图像的位置、角度及尺度变化较为敏感，分类精度还不能完全满足应用的要求。

在数字图像分析领域，局部不变描述符，如尺度变换不变特征(SIFT)、局部二值模式(LBP)在机器视

觉的相关应用领域，如图像区域匹配、目标检测、基于内容的图像检索等表现出很好的性能，为了进一步提升极光分类的分类精度及对噪声干扰的鲁棒性，Wang 等[6]把局部二值模式(LBP)和分块策略相结合来获得极光图像的特征描述，采用 K-近邻分类器(KNN)将极光分为弧状、帷幔状、热斑点状和射线状四种类型，随后，Yang [7]等对上述 LBP 极光特征进行了改进，提出一种 ULBP 算子提取极光图像的局部不变特征空间，并应用隐马尔可夫模型(HMM) 实现极光分类。该类方法获得了较高的分类精度，但对极光位置、方向变化和噪声干扰的鲁棒性仍然达不到应用的要求。而在实际应用中，同类极光在天空中出现的位置会发生变化，其相对于采集设备的角度也会发生变化。另外，由于灯光、星光等或采集设备自身的因素，生成的极光图像会带有一定的噪声。因此，如何提取与表示分类能力强、对极光位置、方向变化和噪声干扰的稳定性好的极光特征仍然是该领域极具挑战性的问题。

本文针对上述问题，提出一种极光全局特征的提取方法，该特征对极光位置、方向的变化具有不变性，而且对噪声干扰具有较高的稳定性。仿真实验表明，本文算法分类精度明显高于现有的基于局部特征的分类方法，对噪声干扰及极光图像位置、方向的鲁棒性显著高于现有方法，而且计算效率较高，可以应用于极光图像的自动分类中。

2. 本文算法

给定极光图像 $f(x,y)$，首先通过下式对其进行 Radon 变换，Radon 变换可以理解为图像函数 $f(x,y)$ 沿包含该函数的平面内的一组直线的线积分有[8]

$$P(t,\theta)=R(t,\theta)\{f(x,y)\}=\iint f(x,y)\delta(t-x\cos\theta-y\sin\theta)\mathrm{d}x\mathrm{d}y \tag{1}$$

其中，$\delta(t)$ 是单位脉冲函数，t 代表坐标原点 O 到直线的距离，$\theta\in[0,\pi)$ 代表直线与 x 轴之间的夹角(或直线的法线与轴的夹角)(如图 1 所示)。

然后，对极光图像 $f(x,y)$ 的 Radon 变换结果 $P(t,\theta)$，定义角 θ 对应的每个方向灰度积分的相对变换强度 η_θ 作为极光特征

$$\eta_\theta=\frac{\sum_{t=1}^{N}\left|P(t,\theta)-\mu_\theta\right|}{N} \tag{2}$$

其中，N 为 $P(t,\theta)$ 的行数，μ_θ 为每列的均值，有

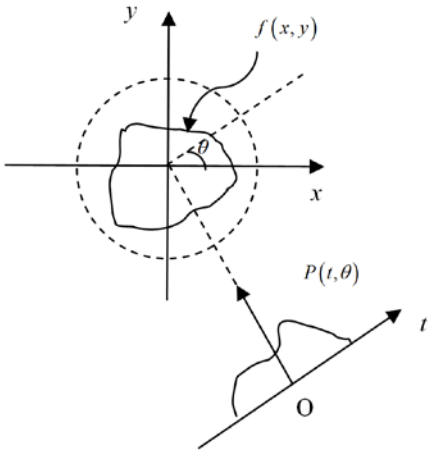


Figure 1. Radon transform of a two-dimensional function *f*(*x*, *y*)
图 1. 二维函数 $f(x,y)$ 的 Radon 变换

$$\mu_\theta = \frac{\sum_{t=1}^{N} P(t,\theta)}{N} \tag{3}$$

对计算所得的相对变换强度η_θ序列进行循环移位，使得其最大值位于首位，将该相对变换强度序列作为极光特征，根据 Radon 变换定义，该特征序列反映了极光图像各方向的积分强度变化情况，极光图像的分类标准也是依据于极光的形状，而这些形状变化会反映到各方向的积分强度变化上，因此，该极光特征可以很好的体现分类标准。另外，由于该特征来源于对极光图像全部象素的积分变换，应属于图像的全局特征。

可以证明，该特征具有如下性质：

1) 具有对极光出现位置变化的平移不变性

如果$f(x-x_0,y-y_0)$表示的极光图像为原极光图像$f(x,y)$平移量为(x_0,y_0)的平移结果，根据(1)式，$f(x-x_0,y-y_0)$的 Radon 变换$p'(t,\theta)$为

$$p'(t,\theta) = \iint f(x-x_0,y-y_0)\delta(t-x\cos\theta-y\sin\theta)\mathrm{d}x\mathrm{d}y \tag{4}$$

令$x_1 = x-x_0$，$y_1 = y-y_0$，有

$$p'(t,\theta) = \iint f(x_1,y_1)\delta(t-x_1\cos\theta-y_1\sin\theta-x_0\cos\theta-y_0\sin\theta)\mathrm{d}x_1\mathrm{d}y_1 \tag{5}$$

令$t_0 = x_0\cos\theta+y_0\sin\theta$，有

$$p'(t,\theta) = \iint f(x_1,y_1)\delta(t-x_1\cos\theta-y_1\sin\theta-t_0)\mathrm{d}x_1\mathrm{d}y_1 = p(t-t_0,\theta) \tag{6}$$

其中，$p(t,\theta)$为原图像$f(x,y)$对应的 Radon 变换，可以看出，极光出现位置变化会导致 Radon 变换结果在t方向的平移，如果极光图像全部在观察域内，$p(t,\theta)$的所有数据都会出现在$p'(t,\theta)$，只是位置不同，所以有

$$\mu'_\theta = \frac{\sum_{t=1}^{N} P(t-t_0,\theta)}{N} = \frac{\sum_{t=1}^{N} P(t,\theta)}{N} = \mu_\theta \tag{7}$$

$$\eta'_\theta = \frac{\sum_{t=1}^{N} \left|P(t-t_0,\theta)-\mu'_\theta\right|}{N} = \frac{\sum_{t=1}^{N} \left|P(t,\theta)-\mu_\theta\right|}{N} = \eta_\theta \tag{8}$$

2) 具有对极光出现方向变化的旋转不变性

如果$f_1(x,y)$表示的极光图像为原极光图像$f(x,y)$以中心为轴旋转α度的旋转结果，有

$$f_1(x,y) = f(x\cos\alpha-y\sin\alpha, x\sin\alpha+y\cos\alpha) \tag{9}$$

根据(1)式，$f_1(x,y)$的 Radon 变换$p_1(t,\theta)$为，

$$\begin{aligned} p_1(t,\theta) &= \iint f_1(x,y)\delta(t-x\cos\theta-y\sin\theta)\mathrm{d}x\mathrm{d}y \\ &= \iint f(x\cos\alpha-y\sin\alpha, x\sin\alpha+y\cos\alpha)\delta(t-x\cos\theta-y\sin\theta)\mathrm{d}x\mathrm{d}y \end{aligned} \tag{10}$$

令$x_1 = x\cos\alpha-y\sin\alpha$，$y_1 = x\sin\alpha+y\cos\alpha$，则$x = x_1\cos\alpha+y_1\sin\alpha$，$y = y_1\cos\alpha-x_1\sin\alpha$，有

$$p_1(t,\theta) = \iint f(x_1,y_1)\delta(t-x_1\cos(\alpha+\theta)-y_1\sin(\alpha+\theta))\mathrm{d}x_1\mathrm{d}y_1 = p(t,\theta+\alpha) \tag{11}$$

可以看出，图像的旋转只会导致变换结果在θ方向的平移，也就是特征值不变，只是序列顺序发生了变化，对η_θ序列进行循环移位，使得其最大值位于首位，可以对其方向进行归一化，实现方向不变。

当然，上述结论是基于连续积分推导的，在实际计算中，由于旋转本身会带来图像的重新采样，所以会有一定误差存在。

3) 对高期噪声的鲁棒性较高

假设图像 $f(x,y)$ 被均值为 0 和方差为 σ^2 的加性噪声 $\eta(x,y)$ 所污染，则有如下关系：

$$\hat{f}(x,y)=f(x,y)+\eta(x,y) \tag{12}$$

那么

$$R(t,\theta)\{\hat{f}(x,y)\}=R(t,\theta)\{f(x,y)\}+R(t,\theta)\{\eta(x,y)\} \tag{13}$$

根据 Radon 变换是对图像的线性积分，在连续的情况下，在各点和各个方向上噪声的 Radon 变换是一个常量，并且该常量等于噪声的均值，也就是 0，所以就有：

$$R(t,\theta)\{\hat{f}(x,y)\}=R(t,\theta)\{f(x,y)\} \tag{14}$$

这意味着零均值的加性噪声在图像 Radon 变换以后没有什么影响。考虑到实际情况中，图像是由有限个点组成的，于是有如下关系：

$$SNR_P \approx 10\log_{10}(1.7N_R)+SNR_i \tag{15}$$

其中 N_R 是图 2 所示的内截圆半径，SNR_P 是投影信号的信噪比，SNR_i 是原始信号的信噪比。从上式可以看出，相对于原图像来说信噪比增加了 $10\log_{10}(1.7N_R)$，在实际当中 $10\log_{10}(1.7N_R)$ 是一个很大的值，如果 $N_R=64$ 的话，则 $10\log_{10}(1.7N_R)\approx 20(\text{dB})$。由此可见，该算法对加性噪声的鲁棒性较强。

3. 实验结果与分析

本文所用的实验数据是中国北极黄河站极光观测系统采集的全天空极光数据，原始图像大小为 512×512 的灰度图像，经过一系列图像预处理(包括暗电流、灰度增强、旋转和噪声)后，用于实验仿真的极光图像为 440×440 的灰度图像。本数据库拥有四种类型的极光图像，分别为弧状、帷幔状、射线状和热斑点状。如图 3 所示。

实验所用的数据集包括 2000 幅手动标记过的极光图像，每种极光类型有 500 幅。实验所用计算机配置为：Intel(R) Core(TM)2 Duo CPU，1.99GB 内存。实验所运行的软件环境为 MatlabR2010a，操作系统为 Windows xp。设计了五个实验验证所提出的极光图像的特征提取方法的有效性。实验中每一类极光图

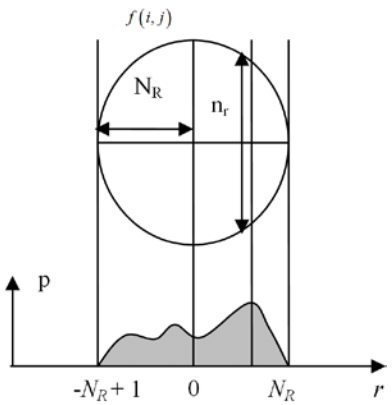
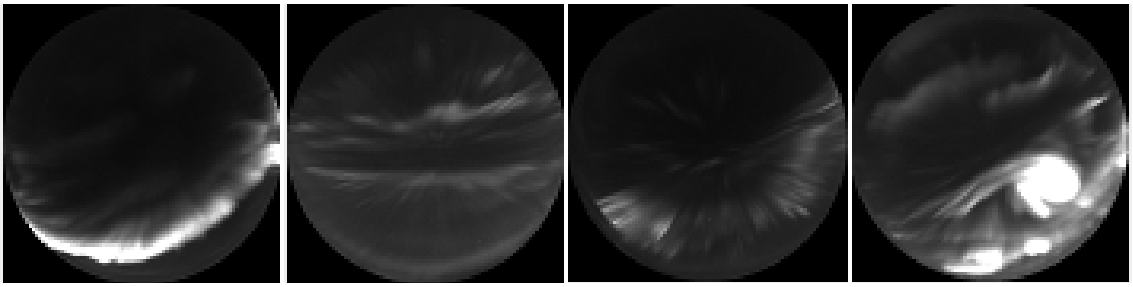


Figure 2. Diagram for calculations in the Radon domain
图 2. Radon 变换的示意图



(a) (b) (c) (d)

Figure 3. Typical categories of aurora: (a) arc aurora; (b) drapery aurora; (c) radial aurora; (d) hot-spot aurora
图 3. 极光典型类别；(a) 弧状极光；(b) 帷幔状极光；(c) 射线状极光；(d) 热斑点状极光

像取出 100 张作为训练集，其余 400 张作为测试集。

3.1. 本文所提方法与传统方法的对比实验

为了有效的避免过学习以及欠学习状态的发生，本实验利用基于 m 重交叉验证(m-fold Cross Validation)下的最近邻(NN)分类器来评价其分类精度。m 重交叉验证具体操作是将原始极光数据均分成 m 组，将每个子集数据分别做一次验证集，其余的 m−1 组子集数据作为训练集，这样会得到 m 个模型，用这 m 个模型最终的验证集的分类准确率的平均数作为此 m 重交叉验证下分类器的性能指标。为了探讨训练样本数目对实验结果的影响，实验中 m 的值从 5 递增取到 15。对于一个给定的 m，实验独立重复 200/m 倍，共进行 200 次实验以得到一个可信的结果。

本实验用来测试本文所提特征提取方法的分类精度，并且将目前最具代表性的特征提取方法均匀局部二值模式(ULBP)作为对比。实验中 ULBP 的特征向量长度为 59(P = 8，R = 2)。实验结果如图 4 所示，可以看出 Radon 变换的基本分类正确率高于 ULBP。

3.2. 对极光图像进行平移的对比实验

鉴于极光图像的特殊性，本实验需要对极光图像先进行加黑框操作再做平移，平移后两种方法的正确率如表 1 所示，从而可得 Radon 变换的平移不变性更佳。

3.3. 对极光图像进行旋转的对比实验

为了验证本文所提方法的旋转不变性，将数据集中的每一张极光图像都进行 1 到 180 度中一个随机角度的旋转，然后对比基于两种特征提取方法下的正确分类率。实验结果如表 2 所示，从表中可以看出 Radon 变换的旋转不变性略优于传统方法 ULBP。

3.4. 对极光图像进行加噪的对比实验

本实验的设计是为了验证所提方法的抗噪性，用均值为 0，方差 v 从 0 以 0.2 的步长增加到 1 的高斯噪声去感染数据集中的每一幅极光图像，之后进行特征提取及分类实验对比。从图 5 的实验结果可以得出 ULBP 在加噪情况下分类正确率急剧下降，而相同情况下 Radon 变换表现出较好的抗噪性。

3.5. 特征提取的时间复杂度对比

评价一个算法的有效性，除了正确率，效率也是至关重要的一个因素，故而设计本实验来对比两种方法的时间复杂度，表 3 的实验结果表明本文的方法在时间复杂度上远远低于 ULBP 方法，几乎减少了近 6 倍。综合以上实验可以看出，本文的方法在分类精度和分类效率上都有极大的改善。

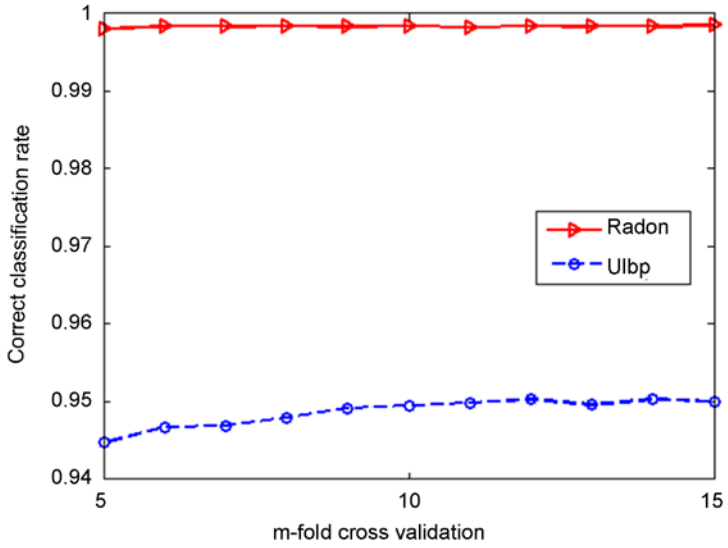


Figure 4. Basic classification accuracy
图 4. 基本分类正确率

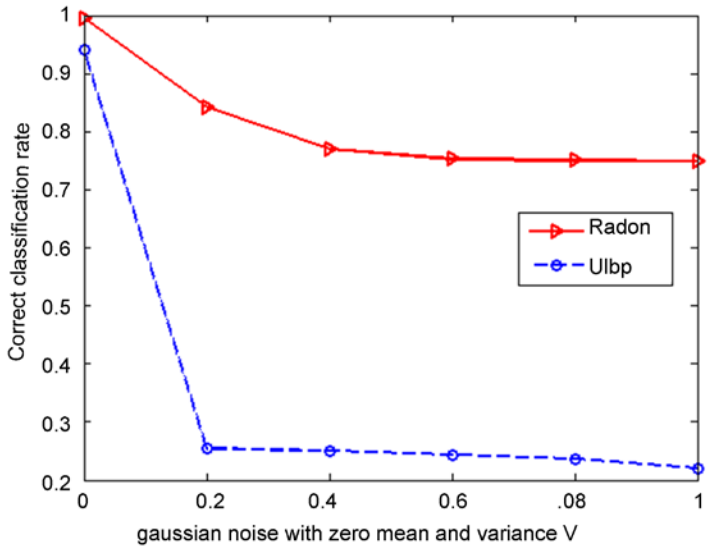


Figure 5. Anti noise ability comparison
图 5. 抗噪性比较

Table 1. Classification accuracy of translation
表 1. 平移分类正确率

	Radon	Ulbp
分类正确率	0.9925	0.8088

Table 2. Classification accuracy of rotation
表 2. 旋转分类正确率

	Radon	Ulbp
分类正确率	0.8119	0.8075

Table 3. Feature extraction time of the two algorithms
表 3. 两种算法的特征提取时间

	Radon	Ulbp
消耗时间(秒/幅)	0.28	1.91

4. 总结

本文提出一种新的、全局的特征提取方法用于极光图像分类，为极光研究领域提供了新的有效工具。为了验证所提方法的可用性，文中设计了五个实验与传统的用于极光图像特征提取的方法均匀局部二值模式进行比较。实验结果表明，本文所提方法在平移、旋转和加噪情况下的分类精度均高于传统方法，同时大大减小了特征提取的时间复杂度。考虑到极光图像的拍摄环境和复杂结构，这种鲁棒性较好的特征提取方法更有利于对极光的进一步研究。

目前为止，极光的分类研究所依据的极光分类机制完全是依靠人的主观判断，通过手工标记极光图像来评估分类精度，这样的分类机制并不一定是对极光最真实的分类。而聚类分析是探索数据的本质特征，从而避免了这一问题，所以在接下来的工作中可以考虑将聚类分析应用到海量极光数据的分析中来。

参考文献 (References)

[1] Syrjäsuo, M.T., Donovan, E.F., Qin, X. and Yang, Y.-H. (2006) Automatic Classification of Auroral Images in Substorm Studies. *Proceedings of the 8th International Conference on Substorms*, Banff, 27-31 March 2006, 309-313.

[2] Donovan, E.F., Trondsen, T.S., Cogger, L.L. and Jackel, B.J. (2003) Auroral Imaging in Canadian CANOPUS and NORSTAR Programs. *Proceedings of Atmospheric Studies by Optical Methods*, Longyearbyen, 13-17 August 2003, 109-112.

[3] 王倩, 梁继民, 高新波, 等. 基于表象特征的极光图像分类方法研究[C]//全国日地空间物理学术讨论会. 第十二届全国日地空间物理学术讨论会论文摘要集. 北京: 中国学术期刊电子出版社, 2007: 71.

[4] 高凌君, 高新波, 梁继民. 结合样本选择和 AdaBoost 的日侧冕状极光检测算法[J]. 中国图象图形学报, 2010, 15(1): 116-121.

[5] Han, S.M., Wu, Z.S., Wu, G.L., *et al*. (2011) Automatic Classification of Dayside Aurora in All-Sky Images Using a Multi-Level Texture Feature Representation. *Advanced Materials Research*, **341-342**, 158-162. https://doi.org/10.4028/www.scientific.net/AMR.341-342.158

[6] Wang, Q., Liang, J., Hu, Z.J., *et al*. (2010) Spatial Texture Based Automatic Classification of Dayside Aurora in All-Sky Images. *Journal of Atmospheric and Solar-Terrestrial Physics*, **72**, 498-508.

[7] Yang, Q., Liang, J., Hu, Z., *et al*. (2012) Auroral Sequence Representation and Classification Using Hidden Markov Models. *IEEE Transactions on Geoscience & Remote Sensing*, **50**, 5049-5060. https://doi.org/10.1109/TGRS.2012.2195667

[8] Wang, X., Guo, F.X., Xiao, B. and Ma, J.F. (2010) Rotation Invariant Analysis and Orientation Estimation Method for Texture Classification Based on Radon Transform and Correlation Analysis. *Journal of Visual Communication and Image Representation*, **21**, 29-32.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 377-387
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74046

Hans汉斯

Structure Optimization Design of Spot Car's Body Based on HyperStudy

Shuai Wang, Chang Sun, Yaqin Feng, Jian Wang

Traffic & Transportation School, Dalian Jiaotong University, Dalian Liaoning
Email: 236155867@qq.com

Received: Apr. 16th, 2017; accepted: Apr. 27th, 2017; published: Apr. 30th, 2017

Abstract

Hypermesh and Ansys are combined by HyperStudy perfectly, which are most frequently used CAE simulation software in the engineering field so far. HyperStudy can quickly optimize and improve new rail cars, and greatly shorten the time to produce new rail cars. The internal optimization method can meet the needs of the current optimization analysis. The optimized results can be used to find the optimal solution quickly, and use the best way to complete the production of the vehicle.

Keywords

HyperStudy, ANSYS, Spot Car, Optimization Design

基于HyperStudy的点焊车车体结构优化设计

王　帅，孙　畅，奉雅琴，王　剑

大连交通大学交通运输工程学院，辽宁 大连
Email: 236155867@qq.com

收稿日期：2017年4月16日；录用日期：2017年4月27日；发布日期：2017年4月30日

摘　要

HyperStudy把目前使用最多的CAE仿真软件，HyperMesh及ANSYS完美的结合到了一起。运用HyperStudy能够快速的对出产的新车局部问题进行优化改进，大大缩短了出产新车的时间。内部自带的优化方法能满足目前各种优化分析的需要，用其优化的结果能快速的找到优化最优解，并以最优的方式

文章引用: 王帅, 孙畅, 奉雅琴, 王剑. 基于 HyperStudy 的点焊车车体结构优化设计[J]. 计算机科学与应用, 2017, 7(4): 377-387. https://doi.org/10.12677/csa.2017.74046

完成整车的制造。

关键词

HyperStudy，ANSYS，点焊车，优化设计

1. 引言

现代轨道车辆中应用最多的焊接方式有对接焊、搭接焊、点焊等。点焊结构具有质量轻、静强度高、可靠性好和易于实现自动化等优点[1]，被中车各生产厂所接受。对点焊车车体的安全校核，基于有限元方法进行仿真分析，是广泛采用的一种方法。张军等[2]构建出较好的车辆优化数学模型并验证了其可行性；陈勇敢等[3]对车体焊点布置方法及优化软件进行了对比，指出了 OptiStruct 和 ANSYS 对焊点布局优化的利弊，并运用这两款软件对应用模态分析对焊点进行了布局优化。谢素明等[4]解释了结构稳定性的算法原理，并运用 OptiStruct 软件对不锈钢点焊车车体局部运用子结构技术进行了稳定性分析，并通过优化手段解决了某点焊车车体的失稳问题，同时提出可用变密度法来进行焊点的布局优化。潘迪夫等[5]在论文中提出可运用多目标遗传算法对机车部件进行优化。

目前的优化软件较多，HyperStudy 可调用各种分析软件进行形状和尺寸优化，吴洪亮[6]运用，HyperStudy 与 OptiStruct 结合对汽车车体进行了尺寸优化，刘磊[7]对航天器结构做了优化，而焦柯[8]的论文中，表明修改某单元内部参数就可以完成尺寸优化，厂内目前的分析软件主要以 ANSYS 为主，若直接用 HyperStudy 与 ANSYS 连接，能大大提高车体研发效率。通过以上论文的启发，焊点在 ANSYS 中也是以网格形式模拟，若直接修改网格参数，运用 HyperStudy 的内部优化方法，不断修改网格参数，得到满足优化中目标函数的数据，从而完成对焊点布局的优化操作。本文工作基于 HyperStudy 优化平台，结合 HyperMesh 建模和 ANSYS 分析功能，对新型不锈钢点焊车车体进行分析与优化，包括补强板设计位置、厚度等，局部结构的优化设计能在节省材料的同时大大降低应力值。

2. 车体有限元模型

新型不锈钢点焊车车体钢结构整体由底架、侧墙、车顶、端墙等组成。车体整体外裹不同厚度蒙皮，内为不同尺寸 Z 型钢焊接而成的骨架。内外部由焊点连接。车体受力后，力通过各焊点遍布整个车体。车体端部底架及枕梁大部分用 09CuPCrNi 不锈钢，其余部分采用 SUS301L 系列不锈钢，其力学性能如表 1 所示。

HyperMesh 是一个高质量高效率的有限元前处理软件，它提供了高度交互的可视化环境帮助用户建立产品的有限元模型。HyperMesh 高质量高效率的网络划分技术可以完成全面的杆梁、板壳、四面体和六面体网格的自动和半自动划分，大大简化对复杂几何进行仿真建模的过程[9]。基于 HyperMesh 软件对上述车体进行有限元建模，有限元模型主要以四节点壳单元(ANSYS SHELL181)为主，车体焊点由两点梁单元(ANSYS BEAM188)模拟，设备质量由质量单元(ANSYS MASS21)模拟，连接关系由柔性元(RBE3)和刚性元(CERIG)模拟。车体单元总数为 1,745,762 个，其中梁单元总数为 37,612 个，节点总数为 1710453

个。车体有限元模型分别如图 1 所示，梁单元分布情况如图 2 所示。

3. 计算结果及分析

3.1. 整车应力分析

车体静强度分析包括垂向、纵向、扭转等 14 个工况，其中最为恶劣的是整备状态车钩 1500 KN 压缩

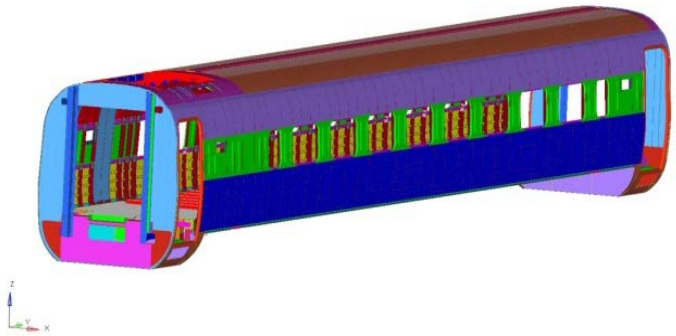
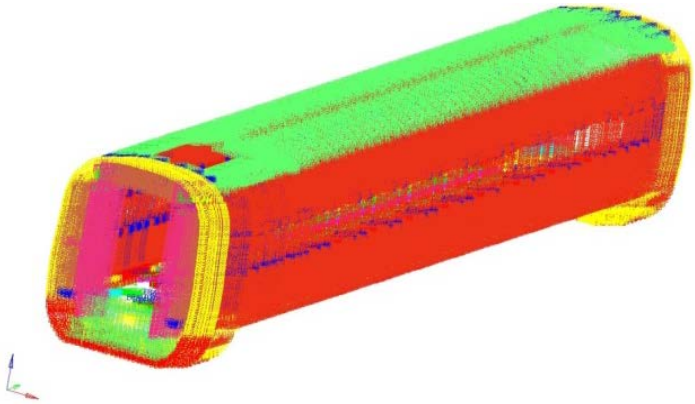
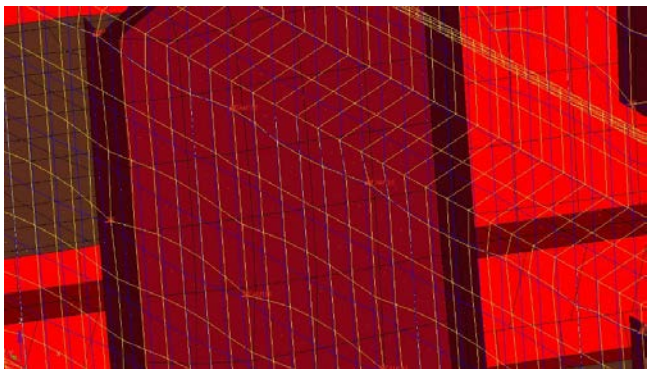


Figure 1. Geometric model of vehicle
图 1. 车体几何模型



(a)



(b)

Figure 2. Distribution of solder joints. (a) Overall distribution of solder joints; (b) Local distribution map of solder joints
图 2. 焊点分布图。 (a) 焊点整体分布图；(b) 焊点局部分布图

Table 1. Mechanical Properties of stainless steel high speed EMUs
表 1. 不锈钢高速动车组车体材料力学性能表

材料名称	密度 kg/m^3	弹性模量 MPa	泊松比	屈服极限 MPa
09CuPCrNi-A	7.297	210,000	0.3	345
SUS301L-DLT	7.297	183,000	0.3	345
SUS301L-ST	7.297	183,000	0.3	410
SUS301L-MT	7.297	183,000	0.3	480
SUS301L-HT	7.297	183,000	0.3	685

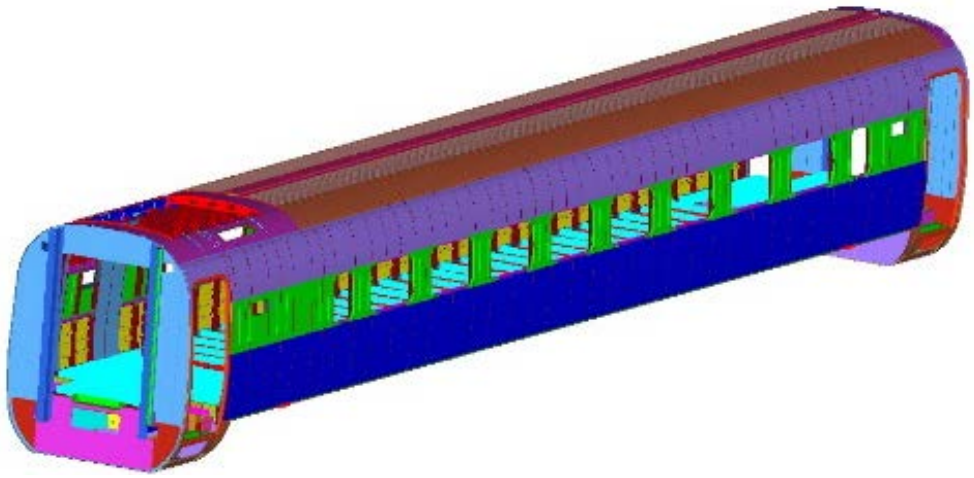


Figure 3. The preparation conditions of 1500 KN get off the hook compression working load diagram
图 3. 整备状态下车钩处 1500 KN 压缩工况加载示意图

工况。对纵截面进行横向位移约束及纵向垂向扭转约束，为模拟真是实验状态，左右车钩处各施加 750 KN 的纵向压力，车体地板梁处施加 151 KN 的垂向力，施加状况如图 3 所示。

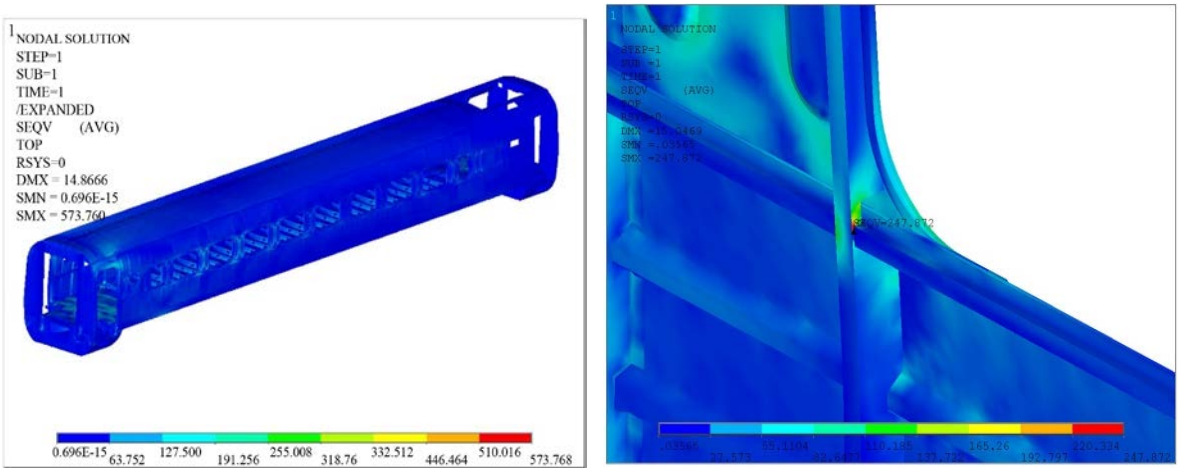
提取压缩工况计算结果，车体各窗角附近应力值均在 200 MPa 以上，最大窗角处应力值到达 247 MPa，具体位置如图 4 所示。

3.2. 修改窗角设计

窗角处计算应力值达到较大值 247 MPa，超出屈服强度，需要对窗角结构进行修改设计，以优化结构降低应力。为计算方便计算，修改设计过程中采用子模型技术，把窗角处分离出来，引入总体结构中边界条件，对子模型单独进行优化计算。修改设计选定在窗角四周均加入 2 mm 厚 L 型补强板，同时在横梁和立柱交叉处焊接 2 mm 厚十字立板，如图 5 所示，此时应力值下降到 202 MPa。

4. 基于 HyperStudy 的结构优化

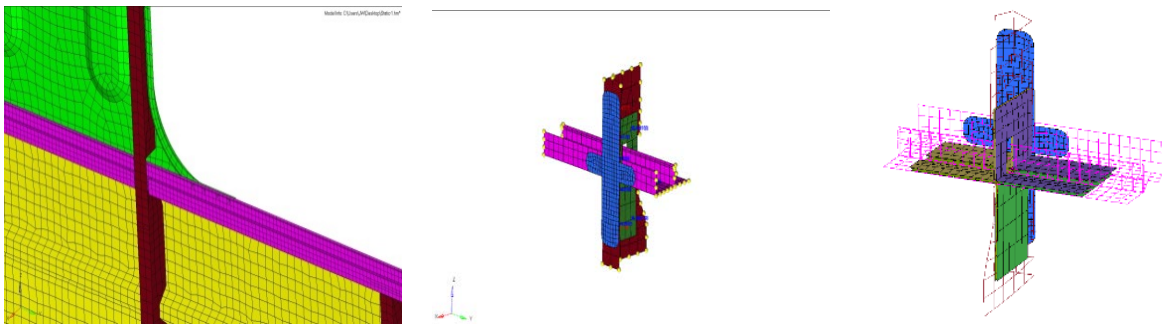
图 6 是两图对比，说明这块区域应力在整车上，和分离出来应力最大值差不多，所以可用子模型结构来替换整车结构，进行计算分析。虽然增加补强板后，应力值下降到了厂里要求的水平，但从板厚的选择，焊点的布置并不是最优解，在此处可利用 HyperStudy 平台，对子模型中的 5 块补强板及与之关联的 34 个焊点进行优化，优化方法选用序列二次规划法(SQP 法)及遗传算法(GA 法)，对编程后的 cdb 文件，运用 ANSYS11.0 版本求解器优化后使分布达到最优解。补强板结构优化设计过程如图 7 所示。



(a) (b)

Figure 4. The preparation conditions of 1500 KN coupler compression condition stress nephogram. (a) Overall stress nephogram of car body; (b) Local stress nephogram of car body

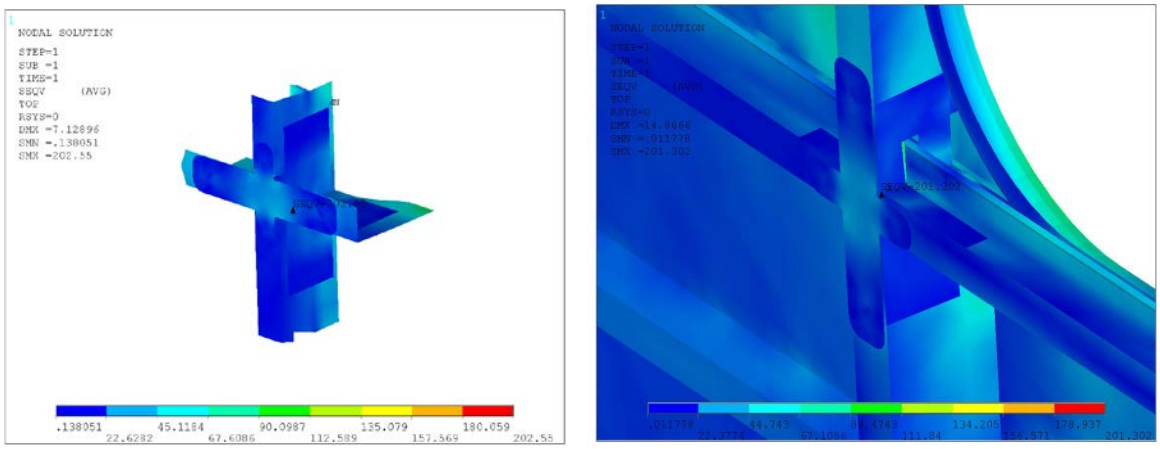
图 4. 整备状态车钩 1500 KN 压缩工况应力云图。(a) 车体整体应力云图；(b) 车体局部应力云图



(a) (b) (c)

Figure 5. Model of window angle reinforcement. (a) Original structure of window corner; (b) Plus cross plate structure; (c) Reinforced plate structure

图 5. 窗角补强方案模型。(a) 窗角处原结构；(b) 加十字板结构；(c) 加强板结构



(a) (b)

Figure 6. Modified stress nephogram. (a) Sub model stress nephogram; (b) The whole vehicle local stress nephogram

图 6. 修改结构后应力云图。(a) 子模型应力云图；(b) 整车局部应力云图

4.1. 优化方法介绍[10]

4.1.1. 序列二次规划法(SQP 法)介绍

序列二次规划法(SQP 法)是一种用于求解有约束问题的优化方法。其基本思想是在原有模型的基础上构建一个拉格朗日二阶近似模型，并通过求解该近似模型已确定搜索方向:

$$x^{(k)} = x^{(k-1)} + as$$

在搜索过程中，设计约束被线性化，得：

$$\min \tilde{f}(s) = f\left(x^{(k)} + \nabla_x^T fs + \frac{1}{2}s^T C_{(k)}s\right)$$

$$\tilde{g}_i(s) = g_i\left(x^{(k)}\right) + \nabla_X^T g_j s \le 0$$

求解此类二次问题即得优化目标。搜索方向的确定服从以下基本原则：从当前设计点出发，通过搜索得到的新的设计点需要使目标目标函数的取值更优，且满足约束条件。

4.1.2. 遗传算法(GA 法)介绍

遗传算法是以进化理论为依托的一类机器学习技术。其实一种群体进化技术，个体通常用二进制来进行编码。在进行适应度(Fitness)计算时，需对个体进行解码。适应度是评价某个解优劣性的指标。按照达尔文的适者生存进化理论，具有高实用度的个体更有可能被选择(Selection)来繁殖下一代。

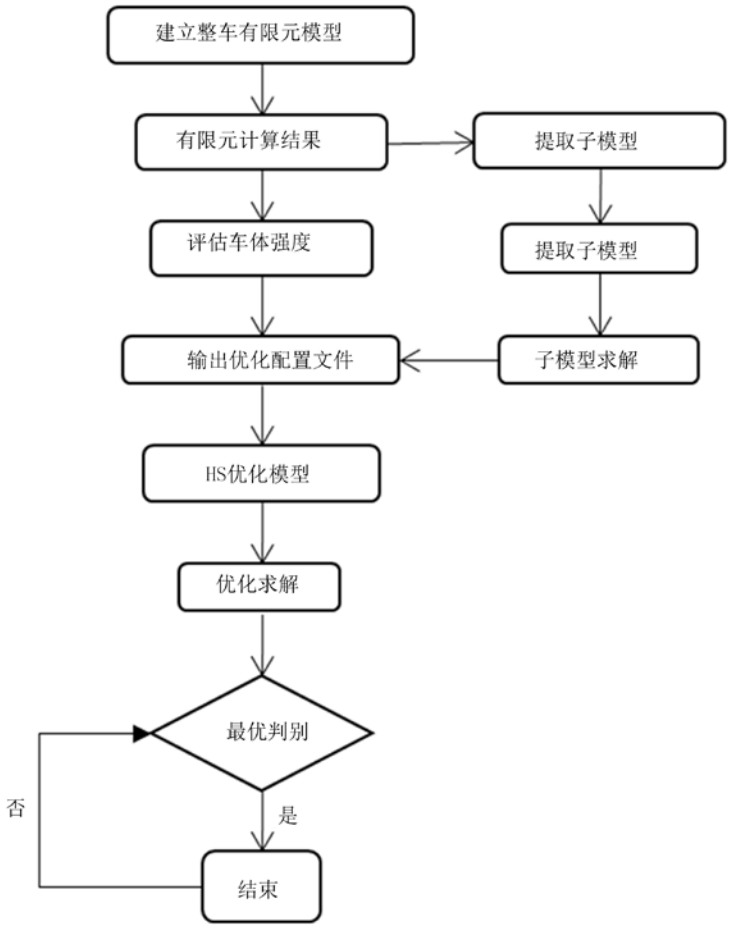


Figure 7. The flow chart of the optimization of the body reinforcing plate
图 7. 车体补强板处优化的流程图

基于所选个体，通过遗传操作算子将繁殖(Reproduction)下一代。一个个体与所选的另一个个体按照一定概率进行交叉(Cross over)和变异(Mutation)。所产生的个体成为下一代的候选解。这一过程将被重复许多代以使种群不断进化，从而获得优化问题的解。

4.2. 焊点布局优化(SQP 法)

4.2.1. 仅对 beam 梁拓扑优化

在优化求解器中选择 SQP 方法，设计变量：单独设置 34 个 beam 梁材料的弹性模量值为 183,000 MPa，下限为 1 MPa，上限为 183,000 MPa；设置约束条件：子模型所有节点最大应力值小于等于 180 MPa；设置目标函数：弹性模量总和最小。设置好后执行优化，优化结果如图 8 所示。根据要求，去掉小于 1000 MPa 弹性模量的 beam 梁 4 个，beam 梁优化后形状改变图如图 9，应力云图如图 10，在 bottom 面，最大值为 193.107 MPa。

4.2.2. 对结构尺寸优化

再对 5 个补强板进行优化，同理在优化求解器中选择 SQP 方法，设计变量：单独设置 5 块板厚度值为 4mm，下限为 1.5 mm，上限为 4 mm；设置约束条件：子模型所有节点最大应力值小于等于 180 MPa；设置目标函数：应力值最低。设置好后执行优化，优化结果如图 11 所示，最大应力值降低到 184.6 MPa，

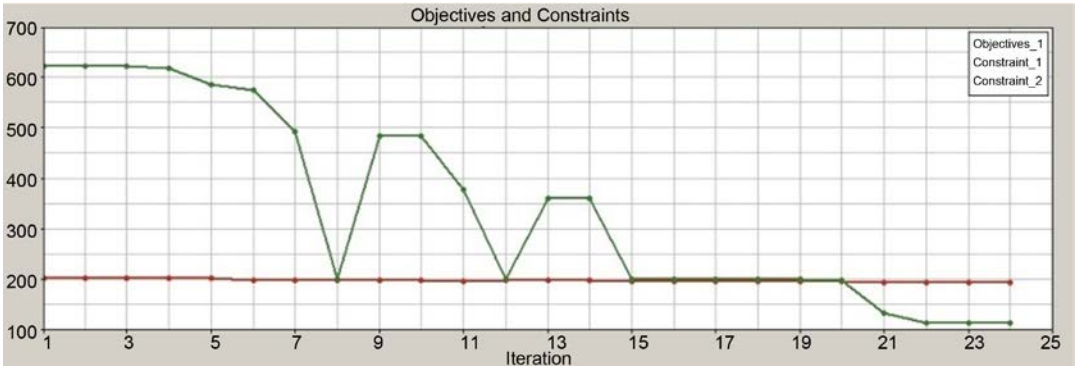
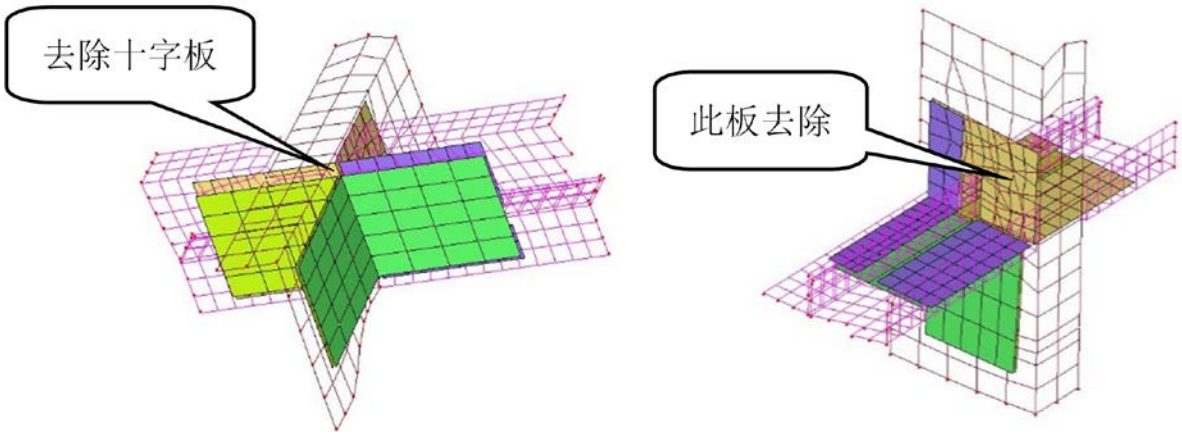


Figure 8. Solder joint layout optimization history
图 8. 焊点布局优化历史



(a) (b)

Figure 9. Shape optimization of beam after optimization
图 9. beam 梁优化后形状改变图

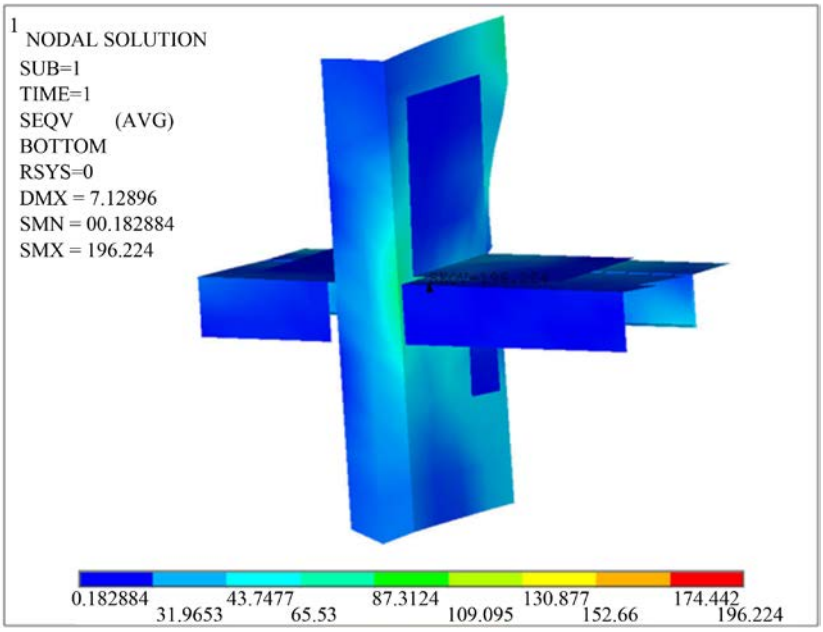


Figure 10. Stress distribution of weld layout optimization
图 10. 焊点布局优化后应力分布图

结果表明若想继续降低应力需增加全部板厚。

4.3. 焊点布局优化(GA 法)

4.3.1. 仅对 beam 梁拓扑优化

在优化求解器中选择 GA 方法，设计变量：单独设置 34 个 beam 梁材料的弹性模量值为 183,000 MPa，下限为 1 MPa，上限为 183,000 MPa；设置约束条件：子模型所有节点最大应力值小于等于 180 MPa；设置目标函数：弹性模量总和最小。设置好后执行优化，优化结果如图 12 所示。根据要求，去掉小于 1000 MPa 弹性模量的 beam 梁 6 个。形状改变图如图 13 所示；得应力云图如图 14 所示，在 bottom 面，最大值为 193 MPa。

4.3.2. 对结构尺寸优化

再对 5 个补强板进行优化，同理在优化求解器中选择 GA 方法，设计变量：单独设置 5 块板厚度值为 4 mm，下限为 1.5 mm，上限为 4 mm；设置约束条件：子模型所有节点最大应力值小于等于 180 MPa；设置目标函数：应力值最低。设置好后执行优化，优化结果如图 15。根据要求，分别改变厚度以降低应力值，最大应力值降低到 181.9 MPa，优化结果与用 SQP 优化结果基本一致。

4.4. 计算结果对比

优化方法	设计变量	目标函数	改前应力值	改后应力值
序列规划法 SQP	34 个 beam 梁	弹性模量总和最小	202MPa	193MPa
序列规划法 SQP	5 块板厚	节点应力值最低	193MPa	185MPa
遗传算法 GA	34 个 beam 梁	弹性模量总和最小	202Mpa	193MPa
遗传算法 GA	5 块板厚	节点应力值最低	193MPa	182MPa

5. 结论

本次对车体补强结构整体进行了优化，解决了直接用 HyperMesh 导出 CDB 文件进行优化的难题。使用序列规划法优化，结果显示有 4 个 beam 梁的弹性模量值在 1000 以下，得到优化结果后取出此 4 个 beam 梁并去除与之相连的局部补强板；使用遗传算法优化，结果显示有 6 个 beam 梁的弹性模量值在 1000 以下，得到优化结果后取出此 6 个 beam 梁并去除与之相连的局部补强板。结构改变后放入 ANSYS

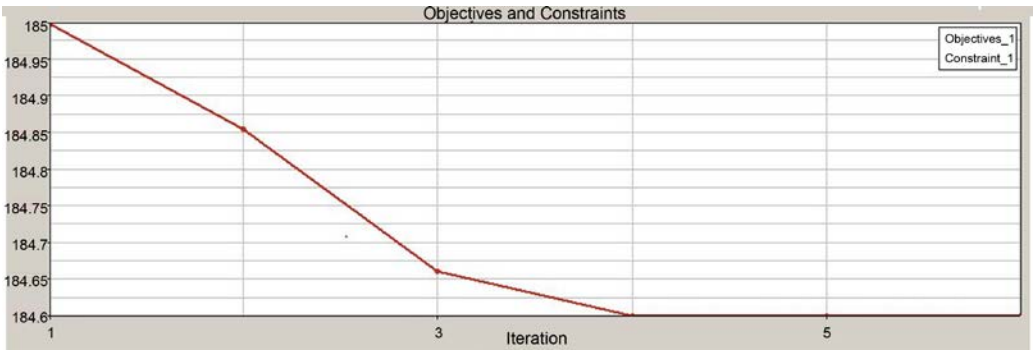


Figure 11. Numerical simulation of the target constraint function and variable iteration step
图 11. 目标约束函数及变量迭代步骤数值变换图

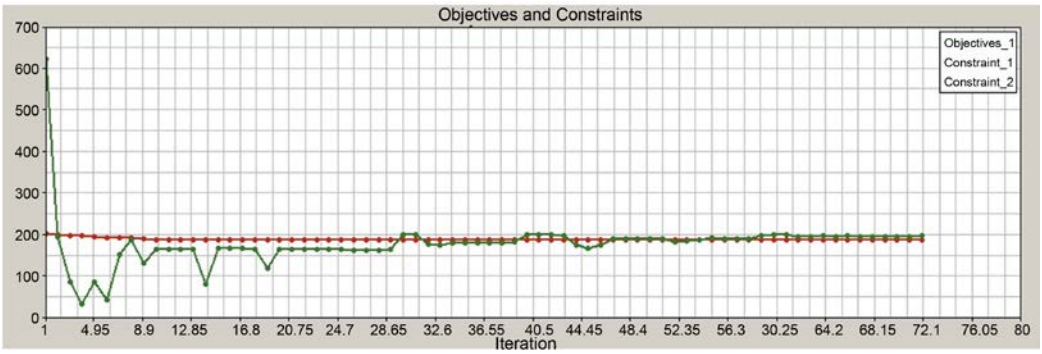


Figure 12. Iterative history of solder joint layout optimization
图 12. 焊点布局优化迭代历史

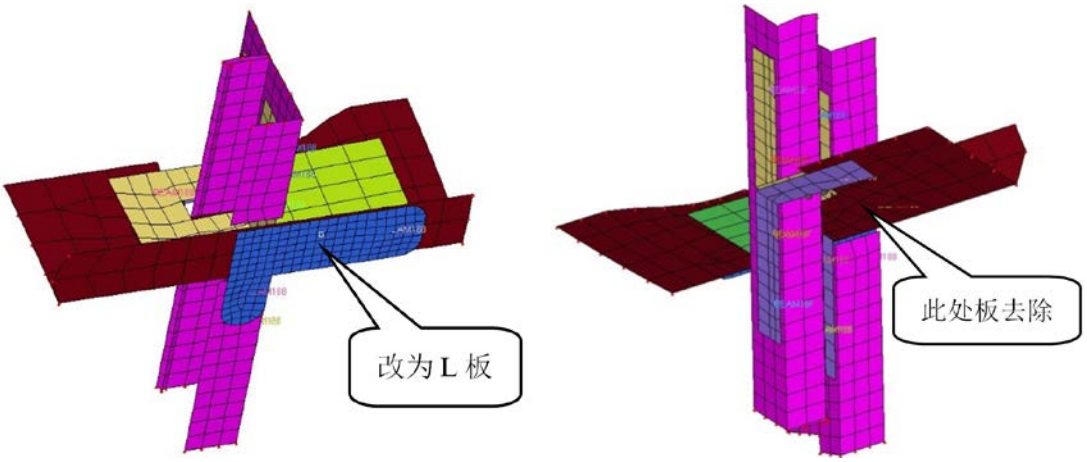


Figure 13. Shape change of beam after optimization
图 13. Beam 梁优化后形状改变图

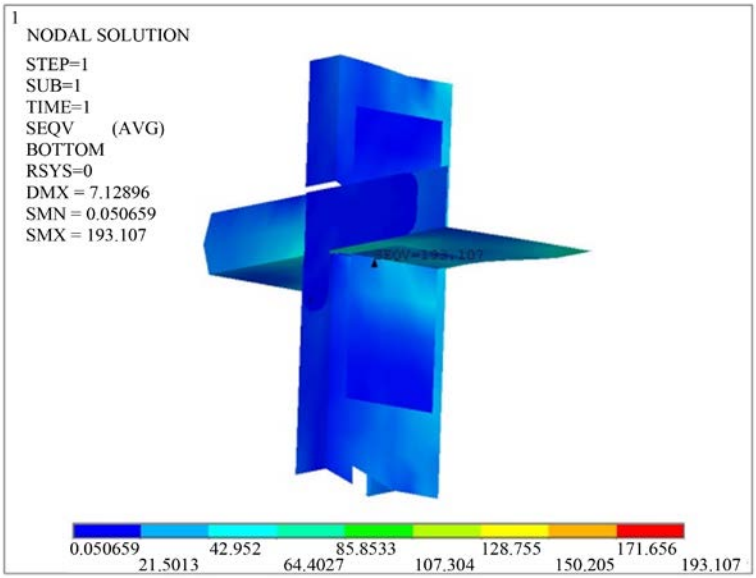


Figure 14. Stress distribution map of solder joint optimization
图 14. 焊点优化应力分布图

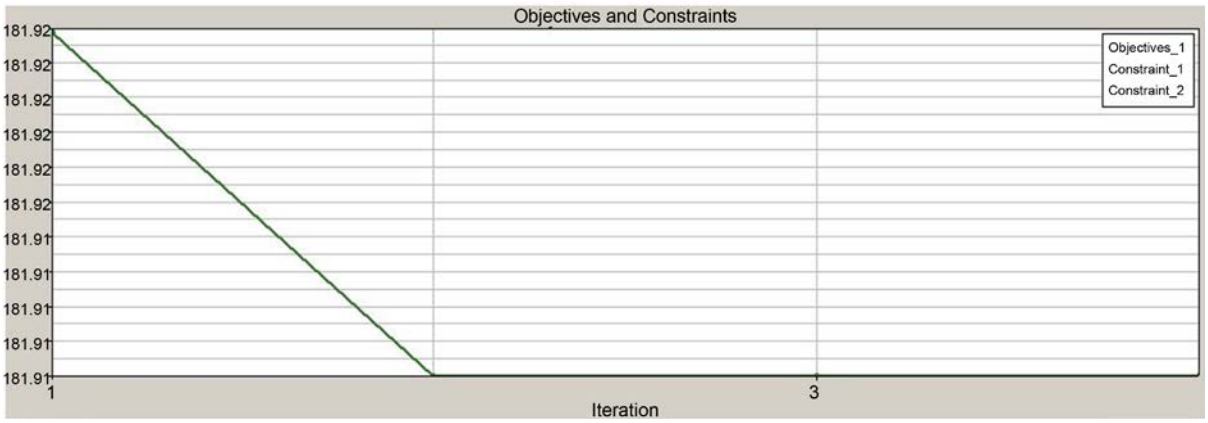


Figure 15. Numerical simulation of the target constraint function and variable iteration step
图 15. 目标约束函数及变量迭代步骤数值变换图

重新计算，得到结果与直接优化产生一致，至此表明优化结果能合理的应用到实际中，并能通过优化达到减材减重的效果。

参考文献 (References)

[1] 徐杰. 轿车车身焊点布置优化研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2010.

[2] 张军, 段丽芳, 李向伟, 等. 基于有限元分析的铁路货车车体优化设计[J]. 大连交通大学学报, 2011, 32(2): 1-4.

[3] 陈勇敢, 毕传兴, 张永斌, 等. 拓扑优化在某 SRV 白车身焊点缩减中的应用[J]. 汽车工程, 2011, 33(8): 733-737.

[4] 谢素明, 穆伟, 高阳. 不锈钢点焊车体结构稳定性分析及局部焊点布局优化[J]. 大连交通大学学报, 2013, 34(4): 12-16.

[5] 潘迪夫, 朱亚男. 基于多目标遗传算法的机车二系支承载荷调整优化方法[J]. 铁道科学与工程学报, 2011, 8(2): 76-80.

[6] 吴洪亮, 陈文琳. 基于 HyperStudy 的多工况铝合金车门尺寸优化[J]. 汽车工程师, 2014(4): 34-36.

[7] 刘磊, 马爱军, 刘洪英, 等. 基于 HyperStudy 的航天器结构多目标优化设计[J]. 智能制造, 2016(6): 34-36.
[8] 茹卫东, 欧旻韬, 焦柯. HyperStudy 在框架结构抗震优化设计中的应用[J]. 广东土木与建筑, 2011(8): 3-5.
[9] 王钰栋, 金磊, 洪清泉. HyperMesh & HyperView 应用技巧与高级实例[M]. 北京: 机械工业出版社, 2012.
[10] 洪清泉. OptiStruct & HyperStudy 理论基础与工程应用[M]. 北京: 机械工业出版社, 2013.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 388-397
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74047

Hans汉斯

Uncertainty Measures for Continuous-Valued Information Systems

Xin Xu

School of Computer and Control Engineering, Yantai University, Yantai Shandong
Email: 1845691121@qq.com

Received: Apr. 16th, 2017; accepted: Apr. 26th, 2017; published: Apr. 30th, 2017

Abstract

Approach to uncertainty measures is one of the hottest topics in area of artificial intelligence, which attracts attention from many researchers. Relevant research results have been applied in data mining, decision analysis, pattern recognition and artificial intelligence. In this paper, the uncertainty measures of continuous-valued information systems have been investigated systematically by using binary relation and entropy. Based on the approximation accuracy, knowledge granulation and information entropy in Pawlak rough set theory, we propose the rough accuracy, knowledge granulation and knowledge entropy in continuous-valued information systems. We also have the comparative study about three measures in the paper. The proposed uncertainty measures for continuous-valued information systems could provide the theoretical foundation for knowledge reduction and representation in continuous information systems.

Keywords

Continuous-Valued Information Systems, Rough Sets, Uncertainty Measure, Entropy

连续值信息系统的不确定性度量

许　鑫

烟台大学计算机与控制工程学院，山东　烟台
Email: 1845691121@qq.com

收稿日期：2017年4月16日；录用日期：2017年4月26日；发布日期：2017年4月30日

摘　要

不确定性的度量方法是人工智能研究的重要课题之一，受到国内外专家学者的广泛关注，相关研究成果已经成功的应用于数据挖掘，决策分析，模式识别与人工智能领域中。通过二元关系与熵，对连续值信

文章引用: 许鑫. 连续值信息系统的不确定性度量[J]. 计算机科学与应用, 2017, 7(4): 388-397.
https://doi.org/10.12677/csa.2017.74047

息系统中的不确定性度量进行了系统研究。基于经典Pawlak粗糙集理论中的近似精度、知识粒度与信息熵，提出了连续值信息系统的粗糙度、知识粒度与知识熵，并对三种度量方式进行了比较分析。三种不确定性度量方式的提出，为连续值信息系统知识约简与表示的研究提供了理论基础。

关键词

连续值信息系统，粗糙集，不确定性度量，熵

1. 引言

由于客观世界中广泛存在不确定性，不确定性问题的研究工作一直备受关注[1]-[8]。某些应用背景中，知识库中的数据以连续值形式呈现，例如某季节的气温范围、某公司中员工的年龄区间、以及工业批量产品制造的误差范围。连续值决策系统的知识库中描述的知识的属性并非同等重要，其中有些属性是冗余的。冗余的属性不利于对知识的泛化与规则的提取，同时也加大了系统存储的负担，造成了资源的浪费。

度量不确定性的程度称为不确定性度量，最早的度量方法是由Kolmogorov于1933年提出的概率论。随着通信技术的发展，Shannon于1948年提出信息熵的概念，解决了对信息量的度量问题。近些年来，随着粗糙集理论研究的不断深入，粗糙集理论中的不确定性研究成为热点问题。国际上，Düntsch等[1]利用Shannon熵提出了三个模型选择准则，这些标准用于指导人们如何挑选最优条件属性集合来描述一个决策值，并用于刻画粗糙集预测的质量。Wierman [2]从公理化角度出发，给出了一种不确定性度量，称为粒度度量，在五条公理约束下，可以证明其提出的粒度度量与Shannon熵具有相同的形式。Beaubouef [3]应用Shannon熵分别研究了粗糙集中概念的粗糙度和关系数据库中的粗糙度。国内，1997年，苗夺谦[9]将信息熵的概念引入粗糙集理论研究中。文献[10]提出了知识的粗糙性、信息熵与互信息等概念，并讨论了知识粗糙性与信息熵之间的关系。2002年，苗夺谦在文献[11]中通过等价关系的基数定义了知识粒度与分辨度的概念。作为两个概念的应用，分别介绍了重要度在求最小约简、协调度在构造决策树方面的应用。王国胤[12]提出了代数观点与信息观点下的粗糙约简。黄兵等人于2004年在文献[13]中基于一般二元关系提出了信息系统的广义粗糙熵概念。2006年，Liang等在文献[14]中将完备信息系统中的知识粒度、分辨度与信息熵等概念扩展到不完备信息系统中，分别提出了知识粒度、信息熵与粗糙熵等概念，用来度量不完备信息系统中的不确定性。冯琴荣在文献[15] [16]中提出了基于数学期望的知识粒度定义，将每个知识粒看作是一个一维对象，粒的基数看作是其长度，从而针对信息系统定义了知识粒度的测度，该测度定义为划分中粒长度的数学期望。类似于不完备信息系统中的不确定性度量[17]，Xu等人于2009年在文献[18]中提出了序信息系统中的知识粒度、知识熵与知识的不确定性度量。2011年，王国胤等在文献[19]中综述了知识不确定性问题的粒计算模型，从粒计算模型的角度分析了模糊集、粗糙集以及商空间理论模型中的不确定性问题，并对知识不确定性问题的研究工作进行了讨论和总结，对有待研究的重要问题进行了展望。基于扩展的条件信息熵，Dai等[20]在2013年对区间值信息系统中的不确定性度量的进行了相关研究。

经典粗糙集中完备信息系统的不确定性度量有三类：1)信息系统的粗糙度；2)知识粒度(或分辨度)；3)信息熵。通过推广完备信息系统中的不确定性度量，本文在第 3 小节提出了三种连续值信息系统中的不确定性度量：1)连续值信息系统中的粗糙度；2)连续值信息系统中的知识粒度(分辨度)；3)连续值信息系统中的知识熵。在第 4 小节对三种度量方法进行了比较。

2. 连续值信息系统中的粒表示

Guan 等在文献[21]中提出了连续值信息系统，采用相似度量方法 $r\left(x_i,x_j\right)=1-\max\left\{\left|c_{ik}\left(x_i\right)-c_{ik}\left(x_i\right)\right|\,|\,c_{ik}\in B\right\}$ 来度量对象 x_i 与 x_j 的相似程度。由相似性度量方法，可给出连续值信息系统中的相似关系为：$R_\beta^B=\left\{\left(x_i,x_j\right)\in U\times U,r_{i,j}^B\left(B\right)=1\right\}$。通过相似关系 R_β^B，给出连续值信息系统中的两种粒化方式：1)相似类(与对象 x_i 满足相似关系 R_β^B 的对象集合)：$T_\beta^B\left(x_i\right)=\left\{x_j\in U\,|\,r\left(x_i,x_j\right)\in R_\beta^B\right\}$；2)极大相容类(满足类中任意两个对象均满足相似关系 R_β^B 的最大集合)：$K_\beta^B\left(x_i\right)$，且 $T_\beta^B\left(x_i\right)=\bigcup\left\{K_\beta^B\left(x_i\right)\in CCR_\beta^B\left(x_i\right)\right\}$。将极大相容类作为连续值信息系统中的基本信息粒，给出粗糙上下近似分别为：

$$\overline{ER}_\beta^B\left(X\right)=\left\{x\,|\,\exists K\in CCR_\beta^B\left(x\right):K\cap X\neq\varnothing\right\};\quad \underline{AR}_\beta^B\left(X\right)=\left\{x\,|\,\forall K\in CCR_\beta^B\left(x\right):K\subseteq X\right\}。$$

对连续值信息系统更加详细的粒化描述与性质，请参见文献[21]。限于本文篇幅，这里不再一一累述。

3. 连续值信息系统的不确定性度量

3.1. 连续值信息系统的粗糙度

Pawlak 教授研究近似精度、近似质量时提出了粗糙度的概念。粗糙集的粗糙度通过集合的上、下近似来定义，它充分反映了由于集合边界域的存在所引起的不确定性。类似地，连续值信息系统中的粗糙度用来度量连续值信息系统中的不确定性。

定义 1 给定的连续值信息系统 $CIS=\left(U,C\cup\left\{d\right\},F,f_d\right)$，对任意的 $X\subseteq U$，$B\subseteq C$，$\beta\in\left[0.5,1\right]$，集合 X 的粗糙近似精度为：

$$\lambda_\beta^B\left(X\right)=\frac{\left|\underline{AR}_\beta^B\left(X\right)\right|}{\left|\overline{ER}_\beta^B\left(X\right)\right|}=\frac{\left|\underline{AR}_\beta^B\left(X\right)\right|}{\left|U\right|-\left|\underline{AR}_\beta^B\left(X\right)\right|}。$$

其中：

$\overline{ER}_\beta^B\left(X\right)=\left\{x\,|\,\exists K\in CCR_\beta^B\left(x\right):K\cap X\neq\varnothing\right\}$ 是集合 X 的上近似算子，

$\underline{AR}_\beta^B\left(X\right)=\left\{x\,|\,\forall K\in CCR_\beta^B\left(x\right):K\subseteq X\right\}$ 是集合 X 的下近似算子，

$\left|S\right|$ 表示集合 S 的基数。

定义连续值信息系统中的粗糙度为：

$$\theta_\beta^B\left(X\right)=1-\lambda_\beta^B\left(X\right)=1-\frac{\left|\underline{AR}_\beta^B\left(X\right)\right|}{\left|\overline{ER}_\beta^B\left(X\right)\right|}=\frac{\left|BR_\beta^B\left(X\right)\right|}{\left|\overline{ER}_\beta^B\left(X\right)\right|}。$$

其中，$BR_\beta^B\left(X\right)=\overline{ER}_\beta^B\left(X\right)-\underline{AR}_\beta^B\left(X\right)$。

粗糙度 $\theta_\beta^B\left(X\right)$ 有如下性质：

1) $\theta_\beta^B\left(X\right)$ 值与连续值信息系统的不确定性成正比，即，$\theta_\beta^B\left(X\right)$ 越大，不确定性越大；反之，$\theta_\beta^B\left(X\right)$ 越小，不确定性越小；

2) $\theta_\beta^B\left(X\right)=0$ 时，$\overline{ER}_\beta^B\left(X\right)=\underline{AR}_\beta^B\left(X\right)$，集合 X 是精确集，不确定度为 0；$0<\theta_\beta^B\left(X\right)<1$ 时，集合

X 是不确定集，不确定度的值 $\theta_\beta^B(X)\in(0,1)$； $\theta_\beta^B(X)=1$ 时， $\underline{AR}_\beta^B(X)=0$，集合 X 是完全不确定集，不确定性为最大值 1。

3.2. 连续值信息系统中的知识粒度

为方便引入连续值信息系统中的知识粒度，先给出一些基本定义。

设连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， $A,B\subseteq C$，有

$$T_\beta(A)=\left\{T_\beta^A(x_1),T_\beta^A(x_2),\cdots,T_\beta^A\left(x_{|U|}\right)\right\}，\ T_\beta(B)=\left\{T_\beta^B(x_1),T_\beta^B(x_2),\cdots,T_\beta^B\left(x_{|U|}\right)\right\}。$$

定义连续值信息系统中的二元关系“$\preceq$”，“$\approx$”与“$\prec$”如下：

$T_\beta(A)\preceq T_\beta(B)\Leftrightarrow$ 若对于任意的 $i\in\{1,2,\cdots,|U|\}$，有 $T_\beta^A(x_i)\subseteq T_\beta^B(x_i)$，其中 $T_\beta^A(x_i)\in T_\beta(A)$，$T_\beta^B(x_i)\in T_\beta(B)$。简记为 $A\preceq B$。

$T_\beta(A)\approx T_\beta(B)\Leftrightarrow$ 若对于任意的 $i\in\{1,2,\cdots,|U|\}$，有 $T_\beta^A(x_i)=T_\beta^B(x_i)$，其中 $T_\beta^A(x_i)\in T_\beta(A)$，$T_\beta^B(x_i)\in T_\beta(B)$。简记为 $A\approx B$；

$T_\beta(A)\prec T_\beta(B)\Leftrightarrow T_\beta(A)\preceq T_\beta(B)$ 且 $T_\beta(A)\neq T_\beta(B)$，简记为 $A\prec B$。

若 $A\prec B$，称属性集 B 对应的论域分类比属性集 A 对应的分类粗，或称属性集 A 对应的论域分类比属性集 B 对应的分类细；若 $A\approx B$，称属性集 B 对应的论域分类与属性集 A 对应的分类相等。

定理 1 连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$，记 $T=\{T_\beta(A)\mid A\subseteq C\}$，则 $(T,\preceq)$ 是一个偏序集。

证明

令 $L,M,N\subseteq C$，有

$$T_\beta(L)=\left\{T_\beta^L(x_1),T_\beta^L(x_2),\cdots,T_\beta^L\left(x_{|U|}\right)\right\},$$

$$T_\beta(M)=\left\{T_\beta^M(x_1),T_\beta^M(x_2),\cdots,T_\beta^M\left(x_{|U|}\right)\right\},$$

$$T_\beta(N)=\left\{T_\beta^N(x_1),T_\beta^N(x_2),\cdots,T_\beta^N\left(x_{|U|}\right)\right\}。$$

1) 对任意的 $x_i\in U$，有 $T_\beta^L(x_i)=T_\beta^L(x_i)$ 成立，因此 $L\preceq L$。

2) 假设 $M\preceq N$ 和 $N\preceq M$。由上面的定义，可得：

$M\preceq N\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使得 $T_\beta^M(x_i)\subseteq T_\beta^N(x_i)$，其中，$T_\beta^M(x_i)\in T_\beta(M)$，$T_\beta^N(x_i)\in T_\beta(N)$；

$N\preceq M\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使得 $T_\beta^N(x_i)\subseteq T_\beta^M(x_i)$，其中，$T_\beta^N(x_i)\in T_\beta(N)$，$T_\beta^M(x_i)\in T_\beta(M)$。

因此，有 $T_\beta^N(x_i)\subseteq T_\beta^M(x_i)\subseteq T_\beta^N(x_i)$，即 $T_\beta^M(x_i)=T_\beta^N(x_i)$。所以，对于任意的 x_i，都有 $T_\beta^M(x_i)=T_\beta^N(x_i)$，即 $M\approx N$。

3) 假设 $L\preceq M$ 和 $M\preceq N$。由上面的定义，可得：

$L\preceq M\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使 $T_\beta^L(x_i)\subseteq T_\beta^M(x_i)$，其中 $T_\beta^L(x_i)\in T_\beta(L)$，$T_\beta^M(x_i)\in T_\beta(M)$；

$M\preceq N\Leftrightarrow$ 对于任意的 $i\in\{1,2,\cdots,|U|\}$，使 $T_\beta^M(x_i)\subseteq T_\beta^N(x_i)$，其中 $T_\beta^M(x_i)\in T_\beta(M)$，$T_\beta^N(x_i)\in T_\beta(N)$。

因此，对于任意的 $i\in\{1,2,\cdots,|U|\}$，有 $T_\beta^L(x_i)\subseteq T_\beta^N(x_i)$，即 $T_\beta^L(x_i)\subseteq T_\beta^N(x_i)$，所以 $L\preceq N$。

考虑到上述三点，$(T,\preceq)$ 是偏序集。

在本章中，将运用这种偏序关系对连续值信息系统中的不确定性进行研究。

Yao 等在文献[22]中给出了信息系统中粒度的一般性定义，为构建和比较知识粒度提供了有利条件。

定义 2 [22]设信息系统为 $IS=(U,AT,V,f)$，对任意的 $A\subseteq AT$，有 $GD(A)$ 满足：

1) **非负性**：$GD(A)\geq 0$；

2) **恒等性**：对于 $\forall A,B\subseteq AT$，若 $A\approx B$ 时，有

$$GD(A)=GD(B)；$$

3) **单调性**：对于 $\forall A,B\subseteq AT$，若 $A\prec B$ 时，有

$$GD(A)<GD(B)。$$

则称 $GD(A)$ 为信息系统 $IS=(U,AT,V,f)$ 上关于属性集 A 的知识粒度。算子“$\approx$”与“$\prec$”是信息系统 IS 中的粒度偏序关系。在粗糙集理论中，不同的知识粒度实质上是对信息细化的不同层次的平均度量。

为度量完备信息系统中的知识不确定性，苗夺谦等在文献[11]中首先给出了完备信息系统中知识粒度的定义：

定义 3 设 $IS=(U,AT,V,f)$ 是一个完备信息系统，$U/IND(A)=\{X_1,X_2,\cdots,X_m\}$，则 IS 关于属性 A 的知识粒度(Knowledge Granularity)定义为

$$GD(A)=\frac{1}{|U|^2}\sum_{i=1}^{m}|X_i|^2。$$

其中，$\sum_{i=1}^{m}|X_i|^2$ 是由 $\bigcup_{i=1}^{m}(X_i\times X_i)$ 决定的等价关系中元素数目。

定理 2 [23]完备信息系统中的知识粒度 $GD(A)$ 是定义 2 意义下的一个知识粒度。

考虑定义 3，设 $X(x_i)$ 是论域 U 中对象 x_i 的等价类，给出完备信息系统中知识粒度的另外一种表示为：

$$GD(A)=\frac{1}{|U|^2}\sum_{i=1}^{m}|X_i|^2=\frac{1}{|U|^2}\sum_{i=1}^{|U|}|X(x_i)|=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}。$$

$X(x_i)$ 是关于对象 $x_i\in U$ 的等价类，$X(x_i)$ 的基是论域 U 中与对象 x_i 满足等价关系的对象数目。$|U|\times|U|$ 是论域中对象间全部的关系数，且 $GD(A)\in[1/|U|,1]$。

1) 当论域的等价类划分最细，即每个划分仅包含单个元素时，论域中对象最易分辨：

$$GD(A)=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}=\sum_{i=1}^{|U|}\frac{1}{|U|\times|U|}=\frac{1}{|U|}。$$

2) 当论域的等价类划分最粗，即论域的划分是整个论域，论域中对象完全不可分辨：

$$GD(A)=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}=\sum_{i=1}^{|U|}\frac{|U|}{|U|\times|U|}=1。$$

$GD(A)=\sum_{i=1}^{|U|}\frac{|X(x_i)|}{|U|\times|U|}$ 可看作论域划分产生的对象关系和与论域中所有对象总和 $|U|\times|U|$ 的比，所以，易将完备信息系统中基于等价类的知识粒度扩展到连续值信息系统中的知识粒度为：

定义 4 连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$，对任意的 $B\subseteq C$，CIS 中属性集 B 的知识粒度定义为：

$$CKG_\beta(B)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}|\cup K_\beta^B|=\sum_{i=1}^{|U|}\frac{|\cup K_\beta^B|}{|U|\times|U|}=\sum_{i=1}^{|U|}\frac{|T_\beta^B(x_i)|}{|U|\times|U|}。$$

其中， $K_\beta^B \in T_\beta^B(x_i)$， $\left|\cup K_\beta^B\right|$ 是 $T_\beta^B(x_i)$ 中所有元素并的基，且 $\left|\cup K_\beta^B\right| = \left|T_\beta^B(x_i)\right|$。

考虑 $KG_\beta(B)$ 的取值情况，给出如下三个定理：

定理 3 (极小值)连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$， R_β^B 是相似关系。连续值信息系统 CIS 中相对于属性集 B 的粒度最小值是 $1/|U|$，当且仅当 $R_\beta^B = \breve{R}_\beta^B$，其中 $\breve{R}_\beta^B$ 为恒等相似关系，有

$$U/\breve{R}_\beta^B = \left\{T_\beta^B(x_i) = (x_i) : x_i \in U\right\} = \left\{\{x_1\}, \{x_2\}, \cdots, \left\{x_{|U|}\right\}\right\}。$$

证略。

定理 4 (极大值)连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$， R_β^B 是相似关系。连续值信息系统 CIS 中相对于属性集 B 的粒度最大值是 1，当且仅当 $R_\beta^B = \hat{R}_\beta^B$，其中 $\hat{R}_\beta^B$ 为全域相似关系，有

$$U/\hat{R}_\beta^B = \left\{T_\beta^B(x_i) = U : x_i \in U\right\} = \{U, U, \cdots, U\}。$$

证略。

定理 5 (边界性)连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$， R_β^B 是相似关系。连续值信息系统 CIS 中相对于属性 B 的粒度存在边界为：

$$1/|U| \le CKG_\beta(B) \le 1，$$

其中： $KG_\beta(B) = 1/|U|$，当且仅当 $R_\beta^B = \breve{R}_\beta^B$；

$KG_\beta(B) = 1$，当且仅当 $R_\beta^B = \hat{R}_\beta^B$。

证略。

定理 6 连续值信息系统 CIS 中的知识粒度 $CKG_\beta(B)$ 是定义 2 意义下的信息粒度。

证明：

1) 显然， $KG_\beta(B) \ge 0$。

2) 令 $A, B \subseteq C$，则连续值信息系统 CIS 下的论域分类可表示为

$$T_\beta(A) = \left\{T_\beta^A(x_1), T_\beta^A(x_2), \cdots, T_\beta^A\left(x_{|U|}\right)\right\}，\ T_\beta(B) = \left\{T_\beta^B(x_1), T_\beta^B(x_2), \cdots, T_\beta^B\left(x_{|U|}\right)\right\}。$$

若 $A \approx B$，则对于任意的 $i \in \{1, 2, \cdots, |U|\}$，都有 $T_\beta^A(x_i) = T_\beta^B(x_i)$，即 $\left|T_\beta^A(x_i)\right| = \left|T_\beta^B(x_i)\right|$。因此，有

$$CKG_\beta(A) = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^A(x_i)\right| = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^B(x_i)\right| = CKG_\beta(B)。$$

3) 令 $A, B \subseteq C$，且 $A \prec B$，则对于任意的对象 $u_i \in U$，有 $T_\beta^A(x_i) \subseteq T_\beta^B(x_i)$。所以

$$CKG_\beta(A) = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^A(x_i)\right| < \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|T_\beta^B(x_i)\right| = CKG_\beta(B)。$$

由此可得， $CKG_\beta(B)$ 是定义 2 意义下的一个知识粒度。

连续值信息系统中的知识粒度可以表示知识分辨能力， $CKG_\beta(B)$ 越小，分辨能力越强； $CKG_\beta(B)$ 越大，分辨能力越小。为更符合认知与方便计算，提出分辨度的定义为：

定义 5 连续值信息系统 $CIS = (U, C \cup \{d\}, F, f_d)$，对任意的 $B \subseteq C$，连续值信息系统 CIS 的分辨度(Discernibility)定义为：

$$CDS_\beta(B) = 1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|\cup K_\beta^B\right| = \sum_{i=1}^{|U|}\frac{1}{|U|}\left(1 - \frac{\left|T_\beta^B(x_i)\right|}{|U|}\right)。$$

其中， $K_\beta^B \in T_\beta^B(x_i)$， $\left|\cup K_\beta^B\right|$ 是 $T_\beta^B(x_i)$ 中所有元素并的基，且 $\left|\cup K_\beta^B\right| = \left|T_\beta^B(x_i)\right|$。

定理 7 (极小值)连续值信息系统 $CIS=\left(U,C\cup\{d\},F,f_d\right)$，$R_\beta^B$ 是相似关系。连续值信息系统 CIS 中相对于属性集 B 的分辨度最小值是 0，当且仅当 $R_\beta^B=\widehat{R}_\beta^B$，其中 $\widehat{R}_\beta^B$ 为全域相似关系，有

$$U/\widehat{R}_\beta^B=\left\{T_\beta^B\left(x_i\right)=U:x_i\in U\right\}=\{U,U,\cdots,U\}。$$

定理 8 (极大值)连续值信息系统 $CIS=\left(U,C\cup\{d\},F,f_d\right)$，$R_\beta^B$ 是相似关系。连续值信息系统 CIS 中相对于属性集 B 的分辨度最大值是 $1-1/|U|$，当且仅当 $R_\beta^B=\breve{R}_\beta^B$，其中 $\breve{R}_\beta^B$ 为恒等相似关系，有

$$U/\breve{R}_\beta^B=\left\{T_\beta^B\left(x_i\right)=\{x_i\}:x_i\in U\right\}=\left\{\{x_1\},\{x_2\},\cdots,\left\{x_{|U|}\right\}\right\}。$$

定理 9 (边界性)连续值信息系统 $CIS=\left(U,C\cup\{d\},F,f_d\right)$，$R_\beta^B$ 是相似关系。连续值信息系统 CIS 中相对于属性 B 的分辨度存在边界为：

$$0\le CDS_\beta\left(B\right)\le 1-1/|U|，$$

其中，$CDS_\beta\left(B\right)=0$，当且仅当 $R_\beta^B=\widehat{R}_\beta^B$；

$CDS_\beta\left(B\right)=1-1/|U|$，当且仅当 $R_\beta^B=\breve{R}_\beta^B$。

连续值信息系统中的分辨度可以更加直观的表示知识的分辨能力：$CDS_\beta\left(B\right)$ 越大，分辨能力越强；$CDS_\beta\left(B\right)$ 越小，分辨能力越小。这符合人们的直观理解也便于计算。

3.3. 连续值信息系统的知识熵

粗糙集理论研究中，论域 U 上的一个等价关系(即划分)可以看作是定义在 U 的子集组成的 σ-代数上的一个随机变量。其概率分布可通过如下方法来确定。

定义 6 [9]设 U 为论域，P 为 U 上的等价关系，P 在 U 上导出的划分为 π，其中 $\pi=\{X_1,X_2,\cdots,X_n\}$，则 P 在 U 的子集组成的 σ-代数上定义的概率分布为

$$[\pi;p]=\begin{bmatrix}X_1 & X_2 & \cdots & X_n\\ p\left(X_1\right) & p\left(X_2\right) & \cdots & p\left(X_n\right)\end{bmatrix}$$

其中，$p\left(X_i\right)=\dfrac{|X_i|}{|U|}$，$i=1,2,\cdots,n$。

有了知识概率分布的定义后，根据信息论就可以定义知识熵的概念[9]。

定义 7 [9]设 P 是知识库 $K=(U,\mathbf{R})$ 中的知识，$U/P=\{X_1,X_2,\cdots,X_n\}$，$P\in\mathbf{R}$，定义知识 P 的熵 $H\left(P\right)$ 为

$$H\left(P\right)=-\sum_{i=1}^{n}p\left(X_i\right)\log p\left(X_i\right)。$$

对于 $H\left(P\right)$ 取值范围，有

1) 当等价关系是恒等关系时，

$$H\left(P\right)=-\sum_{i=1}^{n}p\left(X_i\right)\log p\left(X_i\right)=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}=-\sum_{i=1}^{|U|}\frac{1}{|U|}\log\frac{1}{|U|}=\log|U|。$$

2) 当等价关系是全域关系时，

$$H\left(P\right)=-\sum_{i=1}^{n}p\left(X_i\right)\log p\left(X_i\right)=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}=-\frac{|U|}{|U|}\log\frac{|U|}{|U|}=0。$$

故，$0\le H\left(P\right)\le\log|U|$。

等价关系下，若将每一个对象 x_i 单独看待，它所在的等价类 $X(x_i)$ 看作邻域，那么对象 x_i 所在等价类的基即为论域中与对象 x_i 满足等价关系的对象数目。等价类 $X(x_i)$ 的基越大，粒度越大， $X(x_i)$ 中与对象 x_i 满足等价关系的对象数目就越多，分辨能力就越弱；反之亦然。通过将邻域从等价类扩展到相似类，将相似类 $T_\beta^B(x_i)$ 中的元素个数看作与对象 x_i 满足相似关系的对象集合。$T_\beta^B(x_i)$ 的基越大，与对象 u_i 满足相似相容关系的对象数目越多，分辨能力越弱[24]。基于这种考虑，同时保持熵的良好性质(易计算性，单调性)，给出一种连续值信息系统的不确定性度量如下：

定义 8 连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， R_β^B 是相似关系。$U/R_\beta^B=\{T_\beta^B(x_i):x_i\in U\}$，给出连续值信息系统 CIS 的知识熵 $CKE_\beta(B)$ 为：

$$CKE_\beta(B)=\sum_{i=1}^{|U|}\frac{1}{|U|}\cdot\log_2\left|T_\beta^B(x_i)\right|。$$

定理 10 (**极小值**)连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， R_β^B 是相似关系。连续值信息系统 CIS 中 $CKE_\beta(B)$ 的最小值是 0，当且仅当 $R_\beta^B=\breve{R}_\beta^B$，其中 $\breve{R}_\beta^B$ 为恒等相似关系，有

$$U/\breve{R}_\beta^B=\{T_\beta^B(x_i)=\{x_i\}:x_i\in U\}=\{\{x_1\},\{x_2\},\cdots,\{x_{|U|}\}\}。$$

定理 11 (**极大值**)连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， R_β^B 是相似关系。连续值信息系统 CIS 中 $CKE_\beta(B)$ 的最大值是 $\log_2|U|$，当且仅当 $R_\beta^B=\hat{R}_\beta^B$，其中 $\hat{R}_\beta^B$ 为全域相似关系，有

$$U/\hat{R}_\beta^B=\{T_\beta^B(x_i)=U:x_i\in U\}=\{U,U,\cdots,U\}。$$

定理 12 (**边界性**)连续值信息系统 $CIS=(U,C\cup\{d\},F,f_d)$， R_β^B 是相似关系。连续值信息系统 CIS 中 $CKE_\beta(B)$ 的边界为：

$$0\le CKE_\beta(B)\le\log_2|U|,$$

其中， $CKE_\beta(B)=0$，当且仅当 $R_\beta^B=\breve{R}_\beta^B$；

$CKE_\beta(B)=\log_2|U|$，当且仅当 $R_\beta^B=\hat{R}_\beta^B$。

定义 8 给出的 $CKE_\beta(B)$ 保持了熵良好的单调性与计算的方便性。所以，可把 $CKE_\beta(B)$ 作为一种连续值信息系统中的不确定性度量。

4. 三种不确定性度量的比较

下面讨论连续值信息系统中 $CKG_\beta(B)$， $CDS_\beta(B)$ 及 $CKE_\beta(B)$ 之间的关系，如表 1 所示。当 R_β^B 由最粗的全域关系变为最细的恒等关系时，粒度 $KG_\beta(B)$ 由 1 减小到 $1/|U|$，分辨度 $CDS_\beta(B)$ 由 0 增大到 $1-1/|U|$， $CKE_\beta(B)$ 由 $\log_2|U|$ 减少到 0。故从全域关系变为恒等关系，粒度会越来越细，分辨度越来越高，知识熵越来越低。

Table 1. The comparison of three kinds of measure
表 1. 三种度量方式的比较

T_β^B	$\breve{T}_\beta^B$	$\hat{T}_\beta^B$	取值范围
$CKG_\beta(B)$	$1/\|U\|$	1	$1/\|U\|\le CKG_\beta(B)\le 1$
$CDS_\beta(B)$	$1-1/\|U\|$	0	$0\le CDS_\beta(B)\le 1-1/\|U\|$
$CKE_\beta(B)$	0	$\log_2\|U\|$	$0\le CKE_\beta(B)\le\log_2\|U\|$

5. 结束语

粗糙集理论中知识不确定性主要来源于两方面：不可分辨关系与粗糙上下近似。当边界域中的上、下近似不相等时，边界域不为空，即存在不确定性。本文基于经典粗糙集中的粗糙度、知识粒度与信息熵提出了连续值信息系统中的三种不确定性度量方法：粗糙度；知识粒度(分辨度)与知识熵。给出了三种度量的取值范围，并对三种度量间的关系进行了分析研究。基于本文提出的不确定性度量方法，接下来将在连续值信息系统中的属性约简方面开展进一步的研究工作。

参考文献 (References)

[1] Düntsch, I. and Gediga, G. (1998) Uncertainty Measures of Rough Set Prediction. *Artificial Intelligence*, **106**, 109-137. https://doi.org/10.1016/S0004-3702(98)00091-5

[2] Wierman, M. (1999) Measuring Uncertainty in Rough Set Theory. *International Journal of General Systems*, **28**, 283-297. https://doi.org/10.1080/03081079908935239

[3] Beaubouef, T., Petry, F.E. and Arora, G. (1998) Information-Theoretic Measures of Uncertainty for Rough Sets and Rough Relational Databases. *Information Sciences*, **109**, 185-195. https://doi.org/10.1016/S0020-0255(98)00019-X

[4] Qian, Y.H., Liang, J.Y. and Wang, F. (2009) A New Method for Measuring the Uncertainty in Incomplete Information Systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **17**, 855-880. https://doi.org/10.1142/S0218488509006303

[5] Liang, J.Y. and Shi, Z.Z. (2004) The Information Entropy, Rough Entropy and Knowledge Granulation in Rough Set Theory. I*nternational Journal of General Systems*, **12**, 37-46. https://doi.org/10.1142/s0218488504002631

[6] Yao, Y.Y. and Zhao, L.Q. (2012) A Measurement Theory View on the Granularity of Partitions. *Information Sciences*, **213**, 1-13. https://doi.org/10.1016/j.ins.2012.05.021

[7] 刘财辉, 苗夺谦, 岳晓冬, 赵才荣. 知识不确定性度量及其关系研究[J]. 计算机科学, 2015, 41(3): 66-70.

[8] 张楠, 苗夺谦, 岳晓冬. 区间值信息系统的知识约简[J]. 计算机研究与发展, 2010, 47(8): 1362-1371.

[9] 苗夺谦. Rough Set 理论及其在机器学习中的应用研究[D]: [博士学位论文]. 北京: 中国科学院自动化研究所, 1997.

[10] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.

[11] 苗夺谦, 范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践, 2002, 22(1): 48-56.

[12] Wang, G. (2003) Rough Reduction in Algebra View and Information View. *International Journal of Intelligent Systems*, **18**, 679-688. https://doi.org/10.1002/int.10109

[13] 黄兵, 周献中, 史迎春. 基于一般二元关系的知识粗糙熵与粗集粗糙熵[J]. 系统工程理论与实践, 2004, 24(1): 93-96.

[14] Liang, J., Shi, Z., Li, D. and Wierman, M. (2006) Information Entropy, Rough Entropy and Knowledge Granularity in Incomplete Information Systems. *International Journal of General Systems*, **35**, 641-654. https://doi.org/10.1080/03081070600687668

[15] Feng, Q.R., Miao, D.Q., Zhou, J. and Cheng, Y. (2010) A Novel Measure of Knowledge Granularity in Rough Sets. *International Journal of Granular Computing, Rough Sets and Intelligent Systems*, **1**, 233-251. https://doi.org/10.1504/IJGCRSIS.2010.029580

[16] 冯琴荣. 基于多维数据模型的粒计算方法研究[D]: [博士学位论文]. 上海: 同济大学, 2009.

[17] 王国胤, 苗夺谦, 吴伟志, 梁吉业. 不确定信息的粗糙集表示和处理[J]. 重庆邮电大学学报(自然科学版), 2010, 22(5): 541-544.

[18] Xu, W.H., Zhang, X.Y. and Zhang, W.X. (2009) Knowledge Granulation, Knowledge Entropy and Knowledge Uncertainty Measure in Ordered Information Systems. *Applied Soft Computing*, **9**, 1244-1251. https://doi.org/10.1016/j.asoc.2009.03.007

[19] 王国胤, 张清华, 马希骜, 杨青山. 知识不确定性问题的粒计算模型[J]. 软件学报, 2011, 22(4): 676-694.

[20] Dai, J.H., Wang, W.T., Xu, Q. and Tian, H.W. (2012) Uncertainty Measurement for Interval-Valued Decision Systems Based on Extended Conditional Entropy. *Knowledge-Based Systems*, **27**, 443-450. https://doi.org/10.1016/j.knosys.2011.10.013

[21] Guan, Y.Y., Wang, H.K., Wang, Y. and Yang, F. (2009) Attribute Reduction and Optimal Decision Rules Acquisition

for Continuous Valued Information Systems. *Information Sciences*, **179**, 2974-2984. https://doi.org/10.1016/j.ins.2009.04.017

[22] Yao, Y.Y. and Zhao, L.Q. (2012) A Measurement Theory View on the Granularity of Partitions. *Information Sciences*, **213**, 1-13. https://doi.org/10.1016/j.ins.2012.05.021

[23] 张楠. 区间值信息系统与知识空间的粒计算方法研究[D]: [博士学位论文]. 上海: 同济大学, 2012.

[24] 苗夺谦, 王珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论[J]. 模式识别与人工智能, 1998, 11(1): 34-40.

Computer Science and Application 计算机科学与应用, 2017, 7(4), 398-406
Published Online April 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.74048

Hans汉斯

Gait Recognition Algorithm Based on Sparse Representation of Joint Multi-Feature Dictionary

Xin Hu, Xiaohong Wu, Xiang Lei, Xiaohai He

School of Electronic Information, Sichuan University, Chengdu Sichuan
Email: 936690180@qq.com

Received: Apr. 13th, 2017; accepted: Apr. 25th, 2017; published: Apr. 30th, 2017

Abstract

Most of the existing gait recognition algorithms extract the single feature using model features or global features. However, these algorithms usually have a poor robustness and a low recognition rate in practical situations such as multi-angle. To solve this problem, a gait recognition algorithm based on joint sparse representation of multi-feature dictionaries is proposed in this paper. In this algorithm, three characteristics in different particle size are selected: Procrustes Mean Shape, Gait Energy Image and Region Area Sequence which is structured in this article. Feature training dictionaries are constructed and a multidisciplinary sparse representation to feature samples is done. Finally, the test sample category is obtained by calculating the minimum cumulative residual and achieves the integration of feature layer. Experimental results show that the multi-feature joint recognition method used in this paper has a higher recognition rate and a certain robustness at multiple angles compared to single feature extraction and recognition. This paper basically fulfills the complementary information between features.

Keywords

Gait Recognition, Joint Sparse, RAS

基于联合多特征字典稀疏表示的步态识别算法

胡　欣，吴晓红，雷　翔，何小海

四川大学电子信息学院，四川　成都
Email: 936690180@qq.com

收稿日期：2017年4月13日；录用日期：2017年4月25日；发布日期：2017年4月30日

文章引用: 胡欣，吴晓红，雷翔，何小海. 基于联合多特征字典稀疏表示的步态识别算法[J]. 计算机科学与应用, 2017, 7(4): 398-406. https://doi.org/10.12677/csa.2017.74048

摘 要

现有的步态识别算法多采用模型特征或整体特征进行单一特征提取，在多视角等实际情况中算法鲁棒性较差、识别率较低。针对这一问题，本文提出了一种基于联合多特征字典稀疏表示的步态识别算法。该算法选择三种不同粒度的特征：均值形状PMS、步态能量图GEI与自建特征-区域面积序列RAS，构建特征训练字典并对特征样本进行多任务联合稀疏表示，最后通过计算最小累计残差得到测试样本类别，实现特征层融合。实验结果表明，相比单一特征提取与识别，所采用的多特征联合识别方法识别率更高，且在多视角下具有一定鲁棒性，实现了特征之间的信息互补。

关键词

步态识别，联合稀疏，RAS

1. 引言

步态识别是运用步态的普遍唯一性[1]，通过提取人走路过程中步态特征，经过分类识别，最终实现辨别人物身份，达到分析行为特征的目的。一般包括轮廓检测、特征分析、分类识别三个主要研究步骤。其中特征提取是步态识别的关键环节，特征提取的好坏对最终识别率影响很大。步态特征的提取方法主要分为基于模型与基于整体两种方式。典型的基于模型的特征处理有钟摆模型[2]、棍状模型[3]和椭圆模型[4]。基于模型的特征提取对于遮挡和噪声有较强的鲁棒性，但计算量大，对序列变化敏感度高，难以实现实时处理。基于整体的特征提取也称为基于全局的特征提取，无需建立先验模型，而是通过分析目标动态变化与轮廓信息提取合适的特征，相对基于模型的特征提取实时性好，但对于视角和服饰变化等鲁棒性较差。N.V.Boulgouris 等人采用随机变换提取单帧特征，并使用 LDA 对周期累计特征矢量进行降维[5]。Dong Xu 等人提出一种新的局部分布式特征(PDF)，他们将每帧动态能量图(GEI)表示为局部增量 Gabor 特征组，将其从 5 个尺度和 8 个方向组合为新特征矢量[6]。Mohan Kumar HP 等人提出变式能量图(CEI)，将特征值重构为区间值，通过随机变换提取判决特征，实验表明该特征对于穿衣、背包等变化具有良好的识别鲁棒性[7]。

当前，步态识别的最新研究在国外各学术期刊和会议上频繁出现，主要聚焦在视角变化、衣帽穿戴、实时处理等问题上，步态识别的研究越来越接近实际，如何提高算法的鲁棒性和实际情况下的识别率成为未来的一个研究难点。现有的算法多采用整体特征或模型特征进行单一特征提取与识别，整体特征反映的是目标的轮廓特点，粒度较粗；而模型特征则反映了目标的局部动态特征，粒度又过细，在实际问题中具有一定局限性。针对这些问题，本文提出一种新的区域面积序列特征(RAS)，将人体通过采样划分为若干区域矢量，提取身体不同部位在运动过程中的面积变化特征，并将均值形状 PMS、动态能量图 GEI、区域面积序列 RAS 这三类不同粒度的步态特征构建成特征字典，最后采用基于联合多特征字典稀疏表示算法，去除不同特征间冗余信息的同时实现特征融合。在 CASIA B 标准库上进行的仿真实验表明，该算法在 90 度视角下能够达到较高识别率，在三种不同视角下鲁棒性均优于对应的单特征识别。

2. 多特征提取

步态是一种连续不断的运动，是一个过程，有别于传统的人脸、指纹等一代生物技术识别方法，而且步态序列中每一个运动目标的身体特征也都是不一样的，如身高体重、形态和服装等。因此我们可以用形状特征来作为它的静态特征，并根据目标的运动情况提取它的动态特征。

2.1. PMS 特征提取

Procrustes 均值形状(PMS)分析法是方向统计学中一种广泛使用的方法，通常适用于编码二维形状，并提供了一种好的工具来寻找一组图形的均值形状，因此它可以用来描述人体步态特征[8]。

采用 PMS 分析的特征提取方法如下：

Step 1. 采用 Canny 算子提取图像步态轮廓，将所有轮廓点表示为以人体质心为坐标原点的极坐标值，并以等间隔的角度对轮廓线进行采样，得到轮廓的一维极坐标向量表示。

Step 2. 通过极坐标到直角坐标的逆变换将每个形状表示为一个复数向量 $Z=[z_1,z_2,\cdots,z_k]^{\mathrm{T}}$。其中 $z_i=x_i+jy_i$，(x_i,y_i) 为采样后的轮廓坐标，k 为采样频率。为将形状置于坐标空间中心位置，定义中心向量 $U=[u_1,u_2,\cdots,u_k]^{\mathrm{T}}$，其中 $u_i=z_i-\overline{z}$，$\overline{z}=\sum_{i=1}^{k}z_i/k$；

Step 3. 当序列中含有 n 帧图像时，可得到 n 个中心配置向量，计算配置矩阵 $S=\sum_{i=1}^{n}\frac{U_iU_i^{-1}}{U_i^{-1}U_i}$，得到矩阵的特征值及对应特征向量；

Step 4. PMS 的均值形状 $\bar{U}$ 对应 S 的最优配置，即最大特征值对应的特征向量，将其作为步态序列的统计特征用于识别。

图 1 给出了三个图像序列的 PMS 均值形状，从图中可以看出不同个体在同一视角下的均值形状差别较大。

2.2. GEI 特征提取

用加权平均方法将一个周期内步态序列图像合成的一幅能量图像，称之为步态能量图(GEI)。GEI 既包含了原有周期图像序列中走路频率、相位变化等动态信息，同时有效的减少了统计步态图像序列包含

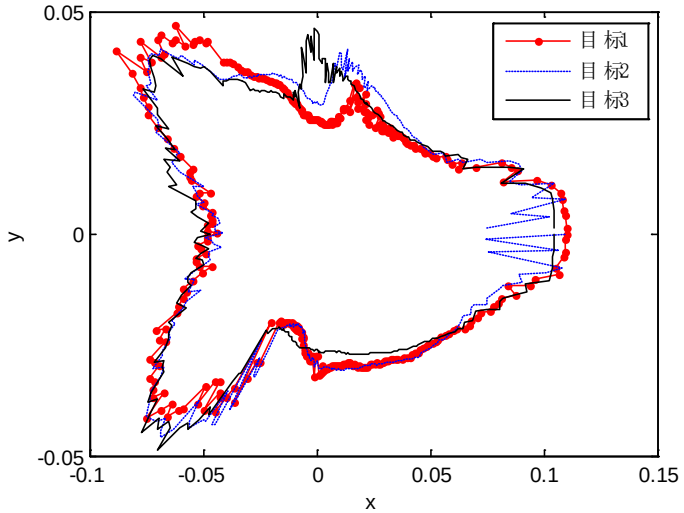


Figure 1. The mean shape of the three goals
图 1. 三个目标的均值形状

信息所需的数据量[9]。

对于一个周期内包含 n 帧图像的步态序列，其 GEI 的定义见式(1)：

$$G(x,y)=\frac{1}{n}\sum_{i=1}^{n}F_i(x,y) \tag{1}$$

其中 $F_i(x,y)$ 表示第 i 帧图像在 (x,y) 处的像素值。本文从测试库中选取的 2 个人 11 帧的步态序列及其能量图如图 2 所示。

在生成步态能量图后，采用主成分分析法 PCA 对 GEI 进行特征降维[10]。针对含有 56 个步态能量图的训练集，设定 PCA 降维数为 25，图 2 中两幅步态能量图在特征空间的表示如图 3 所示。

2.3. RAS 特征提取

为了描述运动人体各区域随时间变化状态，挖掘运动人体步态图像各区域所包含的深层信息，本文提出一种新的区域面积序列特征 RAS。该特征将二值化图像中人体各区域随时间的面积变化特征转化为特征向量来表示运动过程中的周期性状态变化。

以人体质心点与轮廓采样点连线，将每帧步态图像分为若干区域，各个区域的面积可采取等分角度采样方法计算，如图 4 所示。若将图像等分为 m 个区域，则这 m 个区域可用 $m/2$ 条经过人体质心的直线约束划分，直线的斜率 k_i 对应角度 θ_i，$\theta_i=2\pi i/m, i=0,1,2,\cdots,m$ 的正切值，以第一象限为例，$0<\theta<\pi/2$ 所划分的各区域可表示为：

$$\begin{cases}\Omega_i = f(x,y)\\ st.\begin{cases}y\le\tan(\theta_{i+1})x\\ y\ge\tan(\theta_i)x\end{cases}\end{cases} \tag{2}$$

式(2)中 $f(x,y)$ 为轮廓采样点坐标，Ω_i 为坐标值。进行等分角度提取后，设各区域面积为

Figure 2. Gait sequences and energy maps
图 2. 步态序列及其能量图

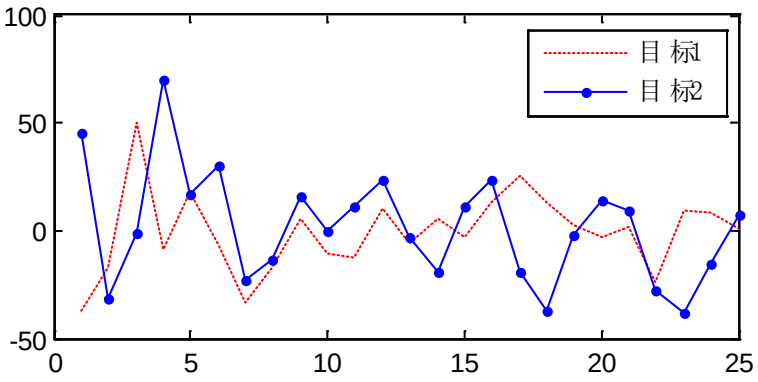


Figure 3. Dimensionality reduction by PCA
图 3. PCA 降维

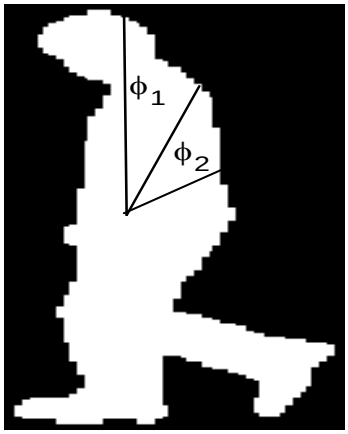


Figure 4. Feature extraction by RAS
图 4. RAS 特征提取

$\phi_{j,i}, i=1,2,\cdots,m$，则第 j 副步态能量图的区域面积特征向量 $\Phi_j=\left[\phi_{j,1},\phi_{j,2},\cdots,\phi_{j,m}\right]^{\mathrm{T}}$，对于包含 n 幅图像的步态序列，其 RAS 特征向量表示为 $\Phi=\left[\Phi_1;\Phi_2;\cdots;\Phi_n\right]$。

提取出 RAS 特征后，采用 PCA 主成分分析法对 RAS 特征进行降维处理。

图 5 给出了按角度 24 等分采样的一幅步态图像的 RAS 特征。

图 6 给出了图 2 中第一个步态序列对应的 RAS 特征。

3. 基于联合多字典稀疏表示的步态识别

基于联合多字典稀疏表示的步态识别算法框架如图 7 所示。算法首先通过计算步态轮廓均值形状得到 PMS 静态特征，同时分别对步态序列周期图像进行加权平均获得 GEI 动态特征，提取区域面积序列构成 RAS 动态特征，并对两种动态特征进行 PCA 降维以减少信息冗余。其次，基于 PMS、GEI、RAS 特征针对步态序列分别构建对应特征字典，并在构建的特征字典上进行多任务联合稀疏表示，约束条件为每种特征字典上的稀疏表示系数趋于一致(同样的稀疏表示图形)。最后，通过 Accelerated Proximal Gradient (APG)算法求得联合多任务稀疏表示系数的最优解，利用它计算三个特征字典累计残差最小即可得到测试样本类别。

式(3)为联合多字典稀疏表示计算公式：

$$\min_{w}\frac{1}{2}\sum_{k=1}^{K}\left\|Y^k-\sum_{j=1}^{J}X_j^k w_j^k\right\|_2^2+\lambda\sum_{j=1}^{J}\left\|w_j\right\|_2 \tag{3}$$

其中，Y 为测试步态序列，K 为特征字典数目，本文选用了 3 种特征，因此设定 $K=3$，X 为训练步态序列，w 为求解的联合多任务稀疏表示系数。这是一个基于 L_1/L_2 混合范数约束的多任务最小二乘回归问题，可以通过 APG 算法求解得到。

通过 APG 算法求得联合多任务稀疏表示系数的最优解 $\hat{w}$ 后，算法的决策条件由 3 个特征字典的累计重建残差最小得出：

$$j^*=\arg\min_{j}\frac{1}{2}\sum_{k=1}^{K}\theta^k\left\|Y^k-X_j^k\hat{w}_j^k\right\|_2^2 \tag{4}$$

累计重建残差 j 最小类则为识别结果。其中 θ 为每个特征字典赋予的权重，权重越大表示该特征对识别结果的影响越大。本文设置 θ 均为 1，表示赋予每个特征字典同样权值，对算法识别率的影响保持等量。

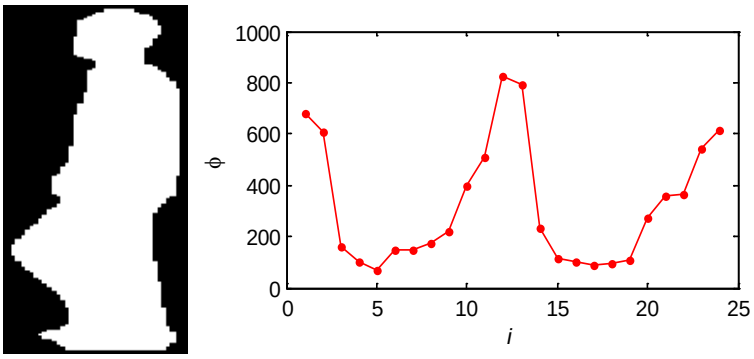


Figure 5. RAS features
图 5. RAS 特征

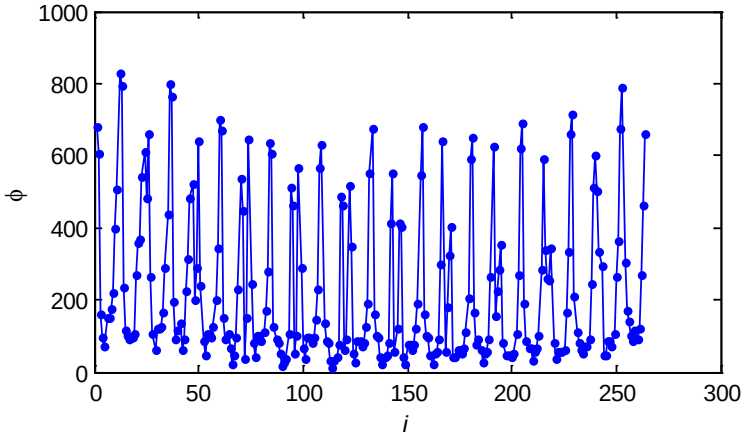


Figure 6. The RAS feature corresponding to gait sequence
图 6. 步态序列对应的 RAS 特征

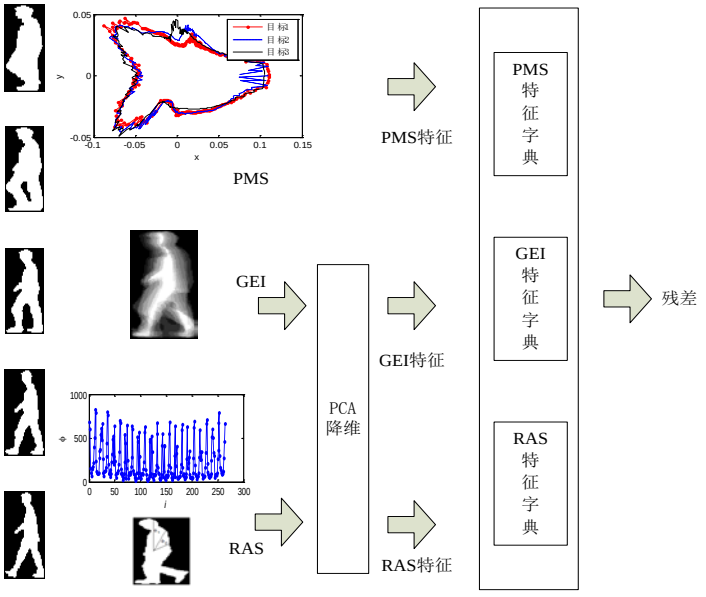


Figure 7. The framework of gait recognition algorithm based on joint multi-dictionary sparse representation
图 7. 基于联合多字典稀疏表示的步态识别算法框架

4. 实验步骤与结果分析

4.1. 实验步骤

当前各国为便于研究建立了多个步态数据库用于公开研究测试，但相对其它生物特征数据库数量偏少，样本规模和复杂度有待提升。CASIA (Chinese Academy of Sciences Institute of Automation)步态数据库由中科院自动化研究所建立，分为小规模库(Dataset A)、多视角库(Dataset B)、红外库(Dataset C)3 个子数据库[11]。其中 Dataset B 标准库于 2005 年建立，由 124 人从 0˚至 180˚以 11 种不同视角分别拍摄，每个视角包含 6 个正常行走序列、2 个穿大衣行走序列、2 个背包行走序列，总计 13640 个图像序列，序列规格统一为 25 fps，每帧 320 * 240 像素，使用较为广泛。

本文选用 CASIA B 库进行仿真实验。从 CASIA B 库的 124 人中随机选取 56 人作为训练样本，每个样本分别包含 180˚、90˚、54˚三种视角下的 1 个正常行走序列(11 帧步态图，归一化为 150 * 90 像素，周期由宽高比变化确定)，训练库共计 1848 帧步态图像。

测试库由标准库中余下的 68 个样本组成，测试时从对应视角下正常行走序列中随机选取步态图像，测试步骤如下：

Step 1. 将测试样本图像规整到 150 * 90 大小，提取 RMS、GEI 与 RAS 特征；

Step 2. 采用 PCA 方法对 GEI 与 RAS 特征进行降维处理，对每种特征采用最近邻分类器计算单一特征下的识别率；

Step 3. 采用联合多特征字典稀疏表示方法，计算三个特征融合下的识别率，与单一特征下的识别结果进行比较；

Step 4. 调整 PCA 降维数，测试维数对识别结果的影响；

Step 5. 改变稀疏度，测试稀疏度对识别结果的影响；

Step 6. 对比三个视角下的识别性能。

4.2. 实验结果与分析

表 1 给出了 90˚视角下采用步骤 2 中单一特征识别与本文多特征联合识别的结果对比。其中 PCA 降维数为 35，稀疏度为 9，PMS 特征提取采样间隔为 1˚，RAS 特征提取采样间隔为 15˚。

可以看出，在单一特征识别方面，GEI + PCA 特征识别率最高，PMS 特征次之，RAS + PCA 特征识别率最低。而本文所采取的静态特征 PMS 与动态特征 GEI、RAS 进行多特征融合方式能够得到最高识别率，实现了各个特征之间的信息互补。

表 2 给出了 PCA 降维数对识别率的影响，PCA 降维数分别取 40~15，间隔为 5，稀疏度为 9，PMS 特征提取采样间隔为 1˚，RAS 特征提取采样间隔为 15˚。从变化趋势来看，本文的多特征联合识别方法识别率随着 PCA 维数的降低而降低，降维数为 35 时可以在兼顾算法效率的同时取得最优识别率。

表 3 给出了不同稀疏度对识别率的影响，稀疏度分别取 1~11，PCA 降维数为 35，PMS 特征提取采样间隔为 1˚，RAS 特征提取采样间隔为 15˚。可见，识别率随着稀疏的增大趋势性增大，在稀疏度等于 9 时识别率达到最大，当稀疏度继续增大时，识别率有所波动。

表 4 给出了 180˚、90˚、54˚不同视角下几种特征提取的识别率。本文所采用的多特征联合识别方法在几种视角下均取得了较高的识别率。对比不同视角的识别率，90˚视角较 54˚视角与 180˚视角的识别率高，这是因为在视角为 90˚时，人体步态的大多数动态特征(如腿部、手臂的摆动等)均能从步态图像序列中有效提取，然而在 54˚、180˚视角下，腿部、手臂等部位的动态信息不够明显，不能被很好的提取，或者人体图像的一部分被分割成背景，导致其识别率较低。

Table 1. Comparison of single feature recognition and multi-feature joint recognition results at 90 degree perspective
表 1. 90˚视角单一特征识别与多特征联合识别结果比较

特征提取	算法识别率
PMS	78.6%
GEI + PCA	84.0%
RAS + PCA	66.1%
本文 (PMS + GEI + RAS**)**	**87.5%**

Table 2. The influence of PCA dimension reduction on recognition results at 90 degree single view
表 2. 90˚单一视角 PCA 降维数对识别结果的影响

降维数	40	35	30	25	20	15
本文(PMS + GEI + RAS**)**	**87.5%**	**87.5%**	**80.3%**	**82.4%**	**80.5%**	**78.5%**

Table 3. Influence of different sparse degree on multi-feature joint recognition
表 3. 不同稀疏度对多特征联合识别结果的影响

稀疏度	识别率
1	58.9%
2	69.6%
3	73.2%
4	80.4%
5	80.4%
6	83.9%
7	85.7%
8	75.0%
9	87.5%
10	85.7%
11	85.7%

Table 4. Recognition results at different perspectives
表 4. 不同视角下的识别结果

视角 / 特征提取	180˚	90˚	54˚
PMS	41.1%	78.6%	33.9%
GEI + PCA 降维	46.3%	84.0%	35.7%
RAS	23.2%	66.1%	14.3%
本文(PMS + GEI + RAS**)**	55.4%	**87.5%**	46.4%

5. 小结

针对步态识别中单一特征提取与分类在实际场景分析中识别率较低的问题，本文选择三种不同粒度的特征 PMS、GEI 与 RAS，通过联合多字典稀疏表达构建特征字典，实现特征层融合。在 CASIAB 标准

库上进行的仿真实验结果表明，本文所采用的多特征联合识别算法在不同视角下均优于单特征识别，降低冗余度的同时实现了特征之间的信息互补。

参考文献 (References)

[1] Murray, M.P. (1967) Gait as a Total Pattern of Movement. *American Journal of Physical Medicine*, **46**, 290-333.

[2] Chai, Y., Ren, J., Han, W., *et al*. (2011) Human Gait Recognition: Approaches, Datasets and Challenges. *Proceedings of* 2011 4*th IEEE International Conference on Imaging for Crime Detection and Prevention*, London, 3-4 November 2011, 1-6.

[3] Yoo, J. and Marker, N. (2011) Automated Markerless Analysis of Human Gait Motion for Recognition and Classification. *ETRI Journal*, **33**, 259-266. https://doi.org/10.4218/etrij.11.1510.0068

[4] Boulgouris, N.V., Hatzinakos, D. and Plataniotis, K.N. (2005) Gait Recognition: A Challenging Signal Processing Technology for Biometric Identification. *IEEE Signal Processing Magazine*, **22**, 78-90. https://doi.org/10.1109/MSP.2005.1550191

[5] Boulgouris, N.V. and Chi, Z.X. (2007) Gait Recognition Using Radon Transform and Linear Discriminant Analysis. *IEEE Transactions on Image Processing*, **16**, 857-860. https://doi.org/10.1109/TIP.2007.891157

[6] Xu, D., Huang, Y., Zeng, Z. and Xu, X. (2012) Human Gait Recognition Using Patch Distribution Feature and Locality-Constrained Group Sparse Representation. *IEEE Transactions on Image Processing*, **21**, 316-326.

[7] Nagendraswamy, H.S. (2014) Change Energy Image for Gait Recognition: An Approach Based on Symbolic Representation. *International Journal of Image, Graphics and Signal Processing*, **6**, 1. http://www.mecs-press.org/

[8] Wang, L., Tan, T., Hu, W., *et al*. (2003) Automatic Gait Recognition Based on Statistical Shape Analysis. *IEEE Transactions on Image Processing*, **12**, 1120-1131. https://doi.org/10.1109/TIP.2003.815251

[9] Han, J. and Bhanu, B. (2006) Individual Recognition Using Gait Energy Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 316-322. https://doi.org/10.1109/TPAMI.2006.38

[10] Liu, N., Lu, J. and Tan, Y. (2011) Joint Subspace Learning for View-Invariant Gait Recognition. *IEEE Signal Processing Letters*, **18**, 431-434. https://doi.org/10.1109/LSP.2011.2157143

[11] Online CASIA Database Information. http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp

Call For Papers

Computer Science and Application

计算机科学与应用

国际中文期刊征文启事

http://www.hanspub.org/journal/csa
ISSN: 2161-8801(Print) ISSN: 2161-881X(Online)

《计算机科学与应用》是一本关注计算机应用领域最新进展的国际中文期刊，主要刊登计算机基础学科、人工智能、计算机仿真及应用领域内最新技术及成果展示的相关论文。本刊支持思想创新、学术创新，倡导科学，繁荣学术，集学术性、思想性为一体，旨在为了给世界范围内的科学家、学者、科研人员提供一个传播、分享和讨论计算机科学领域内不同方向问题与发展的交流平台。该期刊由汉斯出版社出版，全球发行，中国教育图书进出口公司负责引进及在中国内地的销售。现诚邀相关领域的学者投稿。

投稿领域

计算机科学技术基础学科	Basic Disciplines of Computer Science and Technology
自动机理论	Automata Theory
可计算性理论	Computability Theory
人工智能	Artificial Intelligence
计算机系统结构	Computer Architecture
计算机软件	Computer Software
计算机工程	Computer Engineering
计算机元器件	Computer Components
计算机处理器技术	Computer Processor Technology
计算机存储技术	Computer Storage Technologies
计算机外围设备	Computer Peripheral
计算机应用	Computer Application
计算机仿真	Computer Simulation
计算机图形学	Computer Graphics
计算机图象处理	Computer Image Processing
计算机辅助设计	Computer Aided Design
计算机过程控制	Computer Process Control
计算机信息管理系统	Computer Information Management System
计算机科学技术其他学科	Other Related Areas

论文检索

本刊论文已被知网(CNKI)、国家哲学社会科学文献中心、读秀学术、开元知海·e读、全国期刊联合目录数据库(UNICAT)、中国科学院国家科学图书馆、Academic Journals Database、Academic Keys、Base Search、CALIS、citefactor、Cornell.edu、CNPIEC、Ebsco、EZB、Google Scholar、Gold Rush、Journalseek、NewJour、NYULibraries、Open Access Library、Open J-Gate、Open Science Directory (EBSCO)、Polish Scholarly Bibliography (PBN)、Research Bible、Scilit、SHERPA/ROMEO、SJSU、Technische Informationsbibliothek (TIB)、TOC Premier、Trueserials.com、The Open Access Digital Library、Ulrichsweb、Washington.edu、WorldCat、Worldwidescience、WRLC Catalog、WZB (Berlin Social Science Center)、Yale University Library、Zeitschriftendatenbank (ZDB)等数据库收录。

征文要求及注意事项

1. 稿件务求主题新颖、论点明确、论据可靠、数字准确、文字精炼、逻辑严谨、文字通顺，具有科学性、先进性和实用性；
2. 稿件必须为中文，且须加英文标题、作者信息、摘要、关键词和规范的参考文献列表；
3. 稿件请采用WORD排版，包括所有的文字、表格、图表、附注及参考文献；
4. 从稿件成功投递之日起，在2个月内请勿重复投递至其他刊物。本刊不发表已公开发表过的论文。文章严禁抄袭，否则后果自负；
5. 本刊采用同行评审的方式，审稿周期一般为5~7日。

欲了解更多信息请登录 http://www.hanspub.org/journal/csa 联系邮箱：csa@hanspub.org

Hans汉斯

计算机科学与应用

主编：戴元顺　电子科技大学教授，千人计划入选者
黄廷祝　电子科技大学教授
叶培新　南开大学教授

主办：汉斯出版社
编辑：《计算机科学与应用》编委会

网址：http://www.hanspub.org/journal/csa

出版：汉斯出版社