

硕士学位论文

基于卷积神经网络的手势识别研究

Gesture Recognition Based on Convolutional  
Neural Network

学科专业      控制工程

专业领域      信息科学

作者姓名      喻仲斌

指导教师      谢斌 副教授

中 南 大 学

2017 年 5 月

中图分类号 TP391

学校代码 10533

UDC 620

学位类别 专业学位

## 硕士学位论文

### 基于卷积神经网络的手势识别研究

Gesture Recognition Based on Convolutional  
Neural Network

作者姓名：喻仲斌  
学科专业：控制工程  
专业领域：信息科学  
研究方向：模式识别与图像处理  
二级培养单位：信息科学与工程学院  
指导教师：谢斌 副教授  
副指导教师：

论文答辩日期\_\_\_\_\_ 答辩委员会主席\_\_\_\_\_

中 南 大 学  
2017 年 5 月

## 学位论文原创性声明

本人郑重声明，所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了论文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中南大学或其他教育机构的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

作者签名：\_\_\_\_\_ 日期：\_\_\_\_\_年\_\_\_月\_\_\_日

## 学位论文版权使用授权书

本学位论文作者和指导教师完全了解中南大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交学位论文的复印件和电子版；本人允许本学位论文被查阅和借阅；学校可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用复印、缩印或其它手段保存和汇编本学位论文。

保密论文待解密后适应本声明。

作者签名：\_\_\_\_\_

导师签名\_\_\_\_\_

日期：\_\_\_\_\_年\_\_\_月\_\_\_日

日期：\_\_\_\_\_年\_\_\_月\_\_\_日

# 基于卷积神经网络的手势识别研究

**摘 要：**手势作为一种自然友好的人机交互方式，受到了越来越多人的青睐。但是手势含义丰富，跟踪识别困难，人工提取特征的手势识别方法，容易受到外界环境的影响，这使得手势识别依旧是一个具有挑战的课题。近年来深度学习的兴起，给手势识别带来了一种新思路和方法。本文通过对深度学习的研究，提出了基于卷积神经网络的手势识别方法，并设计了基于车载手势识别的原型系统。具体工作如下：

1.本文提出了基于多尺度卷积神经网络的静态手势识别方法。多尺度卷积神经网络利用不同的尺度特征图能够提取出比单尺度卷积神经网络更加精细的特征，提高手势识别的精度。与传统静态手势识别相比，基于多尺度卷积神经网络的静态手势识别方法对复杂环境适应性更好。

2.本文提出了一种基于 3D 卷积神经网络的动态手势识别方法。与传统动态手势识别方法相比，该方法通过 3D 卷积神经网络自动提取手势特征，无需人工提取手势特征。与 2D 卷积神经网络相比，3D 卷积神经网络采用 3D 卷积和 3D 池化，能够同时提取时间和空间特征。

3.本文设计并实现了车载手势识别原型系统。车载手势识别原型系统结合静态手势识别技术和动态手势识别技术，实现了对汽车内常用的四个应用（音乐播放，电话，导航和收音机）的手势操控。

图 33 幅，表 18 个，参考文献 78 篇

**关键词：**卷积神经网络；动态手势识别；静态手势识别；车载手势识别

**分类号：**TP391.4

# **Gesture Recognition Based on Convolutional Neural Network**

**Abstract:** As a natural and friendly way of human-computer interaction, gestures attract more and more people of all ages. But the gesture is rich in meaning, difficult in tracking and recognition. Artificial extraction of the characteristics of the gesture recognition method is vulnerable to the impact of the external environment, which makes gesture recognition still a challenging subject. With the rise of the depth learning this years, it brings gesture recognition a new way of thinking and created methods. Based on the study of depth learning, this paper proposes a gesture recognition method based on convolutional neural network, and designs a prototype system based on vehicle gesture recognition. Specific work is as follows:

1.This paper presents a static gesture recognition method based on multi-scale convolutional neural network. Multi-scale convolutional neural networks can extract more fine features than single-scale convolutional neural networks which use different scale feature maps to improve the accuracy of gesture recognition. Compared with traditional static gesture recognition, static gesture recognition method based on multi-scale convolutional neural network is more adaptable to complex environment.

2.This paper presents an algorithm based on 3D convolutional neural network for dynamic gesture recognition. Compared with the traditional dynamic gesture recognition method, this method can automatically extract the gesture feature through 3D convolutional neural network, and it is not necessary to extract the gesture feature manually. Compared with the 2D convolutional neural network , 3D convolutional neural network adopts 3D convolution kernel and 3D pooling window, and 3D convolutional neural network can extract time and space characteristics at the same time.

3.This paper designs and realizes the prototype system of vehicle gesture recognition. Vehicle gesture recognition prototype system combined with static gesture recognition technology and dynamic gesture

recognition technology, achieving the gesture control of four commonly-used applications (music player, telephone, navigation and radio).

This paper includes 33 images, 18 tables and 78 reference documents.

**Keywords:** Convolutional neural network; Dynamic gesture recognition; Static gesture recognition; Vehicle gesture recognition

**Classification:** TP391.4

# 目 录

摘 要 .....	I
Abstract.....	II
目 录 .....	IV
<b>1 绪论</b> .....	<b>1</b>
1.1 研究背景和意义.....	1
1.2 国内外研究现状.....	2
1.2.1 静态手势识别研究现状.....	2
1.2.2 动态手势识别研究现状.....	3
1.2.3 手势识别技术的商业应用.....	4
1.3 主要研究内容与论文结构.....	5
1.3.1 主要研究内容.....	5
1.3.2 论文结构.....	5
<b>2 静态手势识别</b> .....	<b>7</b>
2.1 静态手势样本集与预处理.....	7
2.2 卷积神经网络架构介绍.....	8
2.3 卷积神经网络的相关运算.....	9
2.3.1 前向传播.....	10
2.3.2 反向传播.....	11
2.4 多尺度卷积神经网络设计.....	12
2.4.1 多尺度卷积神经网络结构.....	13
2.4.2 基于多尺度卷积神经网络的静态手势识别网络设计.....	14
2.5 实验仿真与结果分析.....	16
2.5.1 实验环境介绍.....	16
2.5.2 单尺度特征与多尺度特征神经网络实验结果对比.....	17
2.5.3 与传统静态手势识别方法实验结果对比分析.....	18
2.6 本章小结.....	20
<b>3 动态手势识别</b> .....	<b>21</b>
3.1 传统动态手势识别.....	21
3.2 基于卷积神经网络的视频分类算法.....	22
3.3 3D 卷积神经网络结构设计.....	23

3.3.1	CNN 网络中的时空信息融合方式 .....	23
3.3.2	3D 卷积神经整体网络结构 .....	24
3.3.3	过拟合 .....	26
3.3.4	3D 卷积神经网络的训练 .....	27
3.4	数据集 .....	28
3.4.1	数据扩充 .....	30
3.4.2	数据预处理 .....	31
3.5	实验结果分析 .....	31
3.6	本章小结 .....	33
<b>4</b>	<b>车载手势识别原型系统设计 .....</b>	<b>34</b>
4.1	车载手势识别系统意义 .....	34
4.2	SR300 介绍 .....	34
4.3	软硬件平台框架设计 .....	35
4.4	车载手势识别系统实现 .....	37
4.4.1	功能模块设计与手势定义 .....	37
4.4.2	软件功能实现 .....	42
4.5	实验与分析 .....	44
4.6	本章小结 .....	47
<b>5</b>	<b>总结与展望 .....</b>	<b>48</b>
5.1	总结 .....	48
5.2	展望 .....	49
	参考文献 .....	50
	攻读硕士学位期间主要研究成果 .....	57
	致谢 .....	58



# 1 绪论

## 1.1 研究背景和意义

随着计算机技术、通讯技术、硬件设备等的飞速发展，人机交互已经在我们生活中越来越频繁，人们对用户体验要求越来越高。为了在竞争中占领制高点，各大科技公司对人机交互研究的投入越来越大，人机交互技术也得到了长足的发展。人类非语言沟通（手势，身体姿势和面部表情的沟通）占人类所有沟通的三分之二<sup>[1]</sup>。虽然人类的身体姿势和面部表情的沟通也是非语言沟通的一部分，但是更多表达的是情绪状态，但手势可以具有特定的语言内容，这使得手势成为最适用于通信和交流的最常见的身体语言类别之一<sup>[2]</sup>。并且手势具有自然、直观、易学等优点，这使得手势成为了一种新兴的人机交互方式。手势识别技术作为手势人机交互的技术支撑，自然成为了研究的热点。

手势识别从使用的硬件设备上进行分类，主要可以分为基于数据手套的手势识别技术和基于摄像头的手势识别技术<sup>[3]</sup>。基于数据手套的手势识别技术具有数据精确，识别率高，抗干扰性能好等优点，但是基于数据手套的识别技术存在设备昂贵，识别过程中必须戴上数据手套等缺点，人们逐渐放弃了这种手势识别技术的研究。基于摄像头的手势识别，运用计算机视觉技术，不需要佩戴昂贵的设备，人机交互过程更加自然，基于摄像头的手势识别技术已经成为手势识别的主流。但是基于摄像头的手势识别技术，识别率较低，受外界环境影响大，所以还需要不断地研究和改善算法，以克服这些缺点<sup>[4]</sup>。

手势识别从应用的领域进行分类，主要可以分为手语识别和人机交互。在人际交流中，手势作为一种肢体语言，也经常作为辅助交流的一种手段，尤其对于聋哑人，更是成为其主要的与人沟通的手段<sup>[5]</sup>。但是普通人通常很难理解手语，手语识别系统的开发对于聋哑人正常的融入社会具有重要意义。人机交互从最初以机器为中心转变成以人为中心，手势识别技术作为友好的交互方式受到很多关注<sup>[6][7]</sup>。通过手势进行人机交互，人们可以直接通过手掌和手指的变化直接操作，而不需要借助鼠标键盘等，这样的人机交互更加的自然，符合当前人机交互发展的趋势。尤其在新兴领域如虚拟现实、可穿戴设备等，传统的键盘鼠标这种交互方式已经完全不能适应，手势交互方式成为了很多科技企业的选择。在很多非接触式交互领域，手势交互也有其独特的优势，比如智能手术室、无人机控制、远程教学等<sup>[8]</sup>。

在最近几年，深度学习得到了快速发展，并且在机器视觉，自然语言理解等上取得了很好的成效<sup>[9]</sup>。Google、IBM、Facebook、百度等科技巨头都成立了自己的深度

学习实验室。深度学习是通过多层的网络,模拟大脑的工作模式,实现对数据的深层理解<sup>[10]</sup>。现有的 AlexNet、GoogLeNet 等算法已经在图像分类上取得了很好的成果<sup>[11][12][13]</sup>,这说明深度卷积神经网络在机器视觉领域具有良好的应用前景。但是采用深度学习算法的手势识别技术研究相对比较少,还有很多的难点需要去克服。

## 1.2 国内外研究现状

由于基于视觉的手势识别是现在手势识别的主流方向,所以本节将重点回顾基于视觉的手势识别最新的成果<sup>[14][15]</sup>。基于视觉的手势识别通常采用彩色(RGB)相机,近些年随着深度相机的发展,深度相机逐渐在手势识别领域得到应用<sup>[16]</sup>。虽然深度相机已经在计算机视觉中使用了多年,然而深度相机由于其高价格质量差使得使用受到限制。2010年由微软推出的低成本彩色深度(RGB-D)摄像机 Kinect, Kinect 能够提供高质量的深度图像,解决了复杂背景和照明变化等问题,使得深度相机在手势识别领域得到了广泛应用<sup>[17]</sup>。本节我们将从静态手势识别、动态手势识别和商业应用三个方面分别介绍国内外最新的手势识别技术研究与应用情况。

### 1.2.1 静态手势识别研究现状

静态手势识别方法可以分为 4 类: (a)基于无监督学习, (b)基于有监督学习, (c)基于 3D 模型, (d)基于动态时间规整的方法。在本小节将分别介绍基于这四种方法的静态手势识别方法现状。

无监督学习。Yang 等人<sup>[18]</sup>提出了一种分布式局部线性嵌入(DLLE)算法用于静态手势识别。局部线性嵌入(LLE)算法<sup>[19]</sup>是一种无监督学习算法,该算法将高维数据映射到低维空间,同时保持邻域关系。DLLE 能够发现输入数据的固有属性,提取数据的内在结构,如邻域关系。低维空间中的投影数据点之间的距离取决于输入图像的相似度,概率神经网络(PNN)根据低维空间中的距离,对不同手势进行分类识别。

有监督学习。上海交通大学 Xiaolong Teng 提出了一种有监督的 LLE(local linear embedding)算法用于中国手语识别<sup>[20]</sup>。该法通过肤色进行人手检测,利用人手的内在几何特征进行手势识别。电子科技大学吴杰在 LeNet-5 卷积神经网络基础上设计了基于深度卷积神经网络的静态手势识别方法<sup>[21]</sup>,采用卷积神经网络进行静态手势识别最大优势在于不用人工提取特征,网络通过训练自动学习特征。但是基于训练的方法的常见问题是他们对训练数据的依赖,为了提高一般性和用户独立性, Licsar 和 Sziranyi 提出了一种用户自适应手势识别系统,能够实现对新用户的快速适应<sup>[22][23]</sup>。Pisharady 提出了针对复杂背景手势检测和识别问题的解决方案<sup>[24]</sup>。该系统利用贝叶斯模型生成显着图,并从复杂背景中检测,识别和分割人手区域。Pisharady 提出了使用

形状, 纹理和颜色的组合特征, 通过支持向量机 (SVM) 分类器, 实现对手势的分割和识别。Huang 等人提出了一种在不同照明条件下进行静态手势识别的算法<sup>[25]</sup>。该算法使用自适应肤色模型方法实现照明条件的不变性。

3D 模型拟合用于静态手势识别。该方法估计所有关节角度, 将手形重建为体素模型, 然后在 3D 空间中完成 3D 模型和体素模型之间拟合<sup>[26]</sup>。该方法仅使用手形的几何信息和模型拟合的体素模型, 并且不需要任何启发式或先验信息。Yin 和 Xie 提出基于 3D 模型的手势识别方案<sup>[27]</sup>。手指的边缘点被提取为感兴趣的点, 他们利用手的拓扑特征来进行 3D 手势识别。

### 1.2.2 动态手势识别研究现状

动态手势识别的技术可以分为 4 类: (a) 基于统计学的动态手势识别方法, (b) 基于学习算法的动态手势识别方法, (c) 基于曲线拟合的动态手势识别方法, (d) 基于动态时间规整 (DTW) 的动态手势识别方法。本节我们将分别介绍这些动态手势识别方法, 并单独介绍近些年应用深度相机的手势识别方法。

基于统计学方法的动态手势识别方法。基于统计学方法的动态手势识别方法中应用最广的是基于隐马尔科夫模型 (HMM) 的动态手势识别方法。HMM 是统计学模型, 其被建模的系统被假定为具有未知参数的马尔可夫过程。HMM 使用网络的隐藏状态的状态转移和输出概率表示可观察符号序列的统计行为。基于 HMM 的动态手势识别方法主要利用输入图像序列的时间和空间特征进行手势识别。Chen 等人<sup>[28]</sup>利用傅里叶描述符和基于光流法的运动分析来分别表征空间和时间特征。每一种手势对应一个隐马尔科夫模型, 识别器识别输出概率最大的手势模型。HMM 的一种改进是识别过程中滤掉小概率模型。Lee 和 Kim<sup>[29]</sup>提出了一个基于 HMM 的阈值模型概念, 以过滤掉可能性较小的模型。通常 HMM 是基于均匀马尔科夫链, Marcel 等人<sup>[30]</sup>提出了一种 HMM 的扩展方法, 输入/输出隐马尔科夫模型 (IOHMM)。IOHMM 基于非均匀马尔可夫链, 其中输出和状态转移概率取决于输入。Yoon 等人<sup>[31]</sup>提出了一种动态手势识别方法, 将手位置, 角度和速度特征组合用于 HMM 以实现用于动态手势识别。手通过肤色分析进行定位, 并通过跟踪移动手区域的质心获取手的运动轨迹。Yoon 比较了三个特征: 位置, 角度和速度各自效用, 并得出结论, 角度特征是最有效的, 具有更好的辨别能力。Ramamoorthy 等人<sup>[32]</sup>实现了基于时间表征与静态手形识别系统相结合的 HMM 动态手势识别系统。Ramamoorthy 使用基于卡尔曼滤波器的手轮廓跟踪器提供手势的时间特征。使用基于轮廓判别分类器来识别形状。

基于学习算法的动态手势识别方法。基于学习算法的动态手势识别方法最主要的是使用人工神经网络的动态手势识别。Yang<sup>[33]</sup>利用时间延迟神经网络 (TDNN) 来学习 2D 手势运动轨迹实现动态手势识别。TDNN 是多层前馈网络, 其利用所有层之间

的滑动窗口来表示事件之间的时间关系。Chan 等人<sup>[34]</sup>提出了 HMM 和循环神经网络 (RNN) 的组合模型, 实现了比单独使用的 HMM 或 RNN 更好的性能。使用基于傅立叶描述子的形状特征, 并作为初始姿态分类的网络的径向基函数 (RBF) 输入。RBF 网络的手势似然向量连同运动信息作为两个独立分类器 HMM 和 RNN 的输入。近年来深度学习的兴起, 也出现了一些基于深度卷积神经网络的动态手势识别方法。Pavlo Molchanov 等<sup>[35]</sup>提出了一种基于双卷积的神经网络用于动态手势识别。Pavlo Molchanov 直接将视频输入到卷积神经网络中, 通过大量的训练, 让神经网络自动提取手势特征, 而不采用人为指定的特征进行手势识别。

基于曲线拟合的动态手势识别方法。Shin 等人<sup>[36]</sup>提出了一种使用 Bezier 曲线进行轨迹分析和动态手势分类的几何方法。通过将曲线拟合 3D 手势运动轨迹来识别手势。该方法加入了手势速度特征, 能够从具有速度变化的轨迹中进行准确识别手势。

基于动态时间规整的手势识别。动态时间规整算法是典型的模板匹配算法, 它是一种基于动态规划思想对非线性时间进行归一化再模式识别的算法。J. Alon<sup>[37]</sup>等提出了一种基于动态时间规整思想的改进算法: 动态时间空间规整算法 (DSTW)。该算法对测试样本和模板在空间和时间上都进行了规整, 以此来识别动态手势。华南理工大学的邹洪<sup>[38]</sup>提出了一种基于光流特征的 DTW 动态手势识别算法, 该算法通过采用光流和高斯金字塔的方法提取光流直方图特征, 然后将测试视频和模板视频直方图转换成一组序列号, 然后通过 DTW 算法计算序列号相似度实现手势识别。

基于深度相机的动态手势识别方法。Gallo 等人<sup>[39]</sup>设计了一种基于 Kinect 的手势识别系统。通过手区域的拓扑分析, 可以识别各种手势, 例如缩放, 点击和旋转等功能。Giulio Marin 等人<sup>[40]</sup>使用 Leap Motion 和 Kinect 获取指尖的位置和方向特征, 通过 SVM 分类器进行手势识别。

### 1.2.3 手势识别技术的商业应用

在商业应用领域, 手势识别技术得到了很多科技巨头的青睐, 并展开了相应的研究<sup>[41]</sup>。谷歌将推出手势识别芯片 Soli, 该芯片能够识别“细微手势”, 可以用于移动设备, 电脑和虚拟现实等需要手势识别的电子设备上。韩国三星公司已经将手势识别的技术应用到手机上, 能够通过手机前置摄像头实现拍照, 屏幕滑动等功能。美国手势识别传感器公司 Leap motion 已经成功推出了二代的手势识别传感器<sup>[42][43]</sup>, 并在虚拟现实领域和游戏上得到了广泛的应用。芯片巨头 Intel 也推出了相应的 3D 实感摄像头 Realsense<sup>[44][45]</sup>, 并在联想 ThinkPad S5 Yoga、宏基 Aspire V Nitro 笔记本上得到应用。OPPO 公司在 2016 年推出的旗舰机 OPPO R9s 中使用了基于静态手势识别的拍照技术, 通过张开手掌, 5 秒倒计时后自动拍照。2016 年北京暴风科技有限公司推出了基于手势识别 VR 设备魔镜 Matrix。智能电视企业乐视也推出了“超级手势体感摄像

头”，用于对电视的操控。宝马公司推出的 IDriver 系统，通过 3D 传感器能够识别手势<sup>[46]</sup>。IDriver 系统能够识别轻拍、向左、向右移动等手势。

## 1.3 主要研究内容与论文结构

### 1.3.1 主要研究内容

传统手势识别通过人工提取手势特征用于手势识别，但是人工提取特征的手势识别方法容易受到环境影响，特征的设计和选择对实验结果影响非常大，所以手势识别模型设计难度非常大。本文提出了使用卷积神经网络的手势识别方法。基于卷积神经网络的手势识别方法能够利用卷积层自动学习手势特征，克服了人工提取特征的弊端。在本论文中，我们从下面三个方面进行手势识别的研究。

1) 设计多尺度卷积神经网络。针对静态手势特点，本文设计了多尺度卷积神经网络。多尺度卷积神经网络设计最大的难点在基础网络结构的设计和不同尺度特征的选择。本文测试了 CaffeNet、VGG\_CNN\_F、VGG\_CNN\_M、VGG\_CNN\_S 四种当前性能优越的深度卷积神经网络，通过实验结果对比，本文选取 CaffeNet 作为多尺度卷积神经网络作为基础网络结构。本文利用贪心算法思想选取了第 2、4、5 个卷积层输出的特征图作为本文设计的多尺度卷积神经网络的特征，实验表明设计的网络具有良好的性能。

2) 设计 3D 卷积神经网络。针对动态手势识别特点，本文设计了 3D 卷积神经网络。3D 卷积神经网络设计的难点主要是时空信息融合方式的选择、过拟合和训练速度的提升。在 CNN 网络中时空信息融合方式有 3 种：早期融合、晚期融合、缓慢融合。本文通过实验对照，选择了缓慢融合方式作为 3D 卷积神经网络的时空信息融合方式。针对过拟合问题，我们在全连接层引入 Dropout 结构，Dropout 结构能够在每次训练过程以一定概率使一部分神经元不激活。为了加速 3D 卷积神经网络的训练速度，本文引入 Nesterov 加速梯度法加快 3D 卷积神经网络的训练。

3) 设计并实现了车载手势识别原型系统。本文基于卷积神经网络的手势识别方法，利用英特尔推出的实感摄像头 SR300 作为传感器，设计并实现了车载手势识别系统。车载手势识别系统实现了汽车内常用的音乐播放、导航、电话、收音机四个应用的手势操控。

### 1.3.2 论文结构

本论文一共分为五个章节，各章介绍如下：

第一章介绍手势识别研究的意义和现状，并介绍了本文的主要研究内容和论文结构。

第二章提出了基于多尺度的卷积神经网络的静态手势识别方法，详细介绍了卷积神经网络的前向传播和参数更新过程，介绍了基于多尺度的卷积神经网络的设计过程，并通过与单尺度卷积神经网络和传统手势识别方法进行实验对比，发现了基于多尺度卷积神经网络的静态手势识别的优缺点。

第三章介绍了传统动态手势识别技术的优缺点，分析了基于深度卷积神经网络的视频分类方法以及 3 种时空信息融合方式，并设计了基于 3D 卷积神经网络的动态手势识别方法。

第四章介绍手势识别在汽车领域的应用的意义，设计了一套车载手势原型系统，并对设计过程和原理进行了详细的叙述和分析。

第五章总结了本文的研究内容，指出在实验和方法上的优势和不足。

## 2 静态手势识别

通过手势进行人机交互，我们的目的是希望机器理解手势的含义。而静态手势只涉及到单幅图片，其本质就是对单幅图片进行分类识别。而卷积神经网络在当前是图像分类理解领域最好的算法，所以本章将采用卷积神经网络对静态手势进行识别。本章将介绍手势样本集，深度网络架构，卷积神经网络的相关运算以及基于多尺度特征的卷积神经网络设计。最后通过多组对比实验，分析总结算法的优势和不足。

### 2.1 静态手势样本集与预处理

在本章中使用的手势样本集是采用 Sebastien Marcel 手势数据集<sup>[47]</sup>。该数据集包含三种手势数据，Jochen Triesch 静态手势数据集、Sebastien Marcel 静态手势数据集和 Sebastien Marcel 动态手势数据集。Sebastien-Marcel 手势数据集有复杂背景和简单背景两种，动态手势包含有点击、旋转、停止等多种控制手势。本章采用了 Sebastien Marcel 手势数据集的静态手势数据集，并进行了扩充，使手势集的训练样本数量达到了 10000 个，每个手势分别包含 1667 个样本，测试样本集达到 1500 个样本，每个手势包含 125 个复杂背景样本和 125 个简单背景样本。该静态手势集定义了 6 个手势，具体如图 2-1、图 2-2 所示：



图 2-1 简单背景样本集



图 2-2 复杂背景样本集

为了更好使用卷积神经网络进行手势识别，还需将每一种图片的大小统一。在本文中，静态手势识别的图片统一使用  $66 \times 76$  的图片。为了减少计算量我们将图片进行灰度化处理。





图 2-3 数据集部分数据截图

## 2.2 卷积神经网络架构介绍

深度学习的先驱者 LeCun Y 等人设计多种深度学习网络，已在机器视觉模式识别等领域展现出优秀的性能<sup>[48][49]</sup>。常见的卷积神经网络的结构如下图：



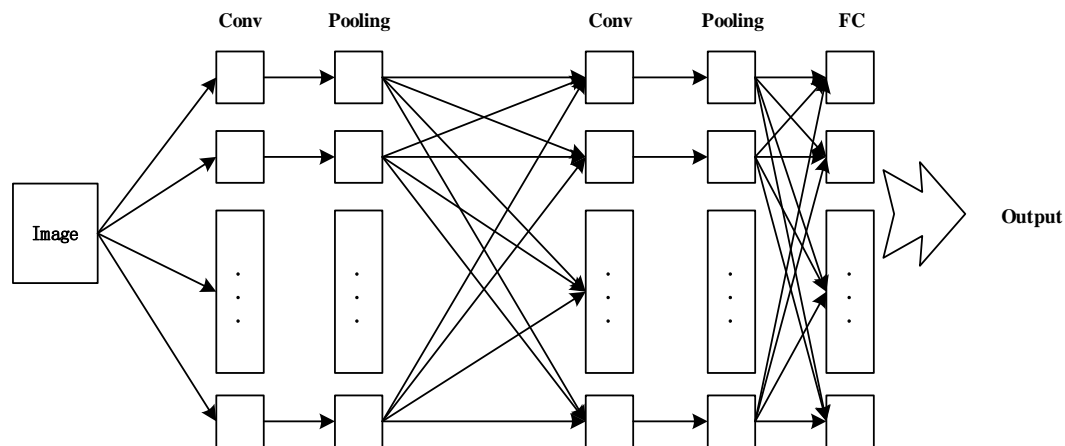


图 2-4 卷积神经网络结构

(Conv 表示卷积层, Pooling 表示池化层, FC 表示全连接层)

卷积神经网络主要包括卷积层、池化层、全连接层、输出层。

**卷积层:**卷积神经网络与普通神经网络最大的不同在于卷积神经网络拥有卷积层,能够直接对输入图片进行图像特征提取。在一个卷积层中通常有多个卷积核,每一个卷积核对应输出一张特征图。

**池化层:**池化层是对卷积层输出的特征图进行下采样计算,但仍然保留图中最重要的信息。池化的方法有:最大值池化,平均值池化。池化层主要有以下几个作用:

- 1.减少特征图的维度,减少网络对能存的消耗。
- 2.减少神经网络中的参数数量,减少计算量。
- 3.减少图像中平移,失真等的影响。

**全连层:**全连接层相当于传统的多层感知器。在全连接层中,每个神经元都与前一层的每个神经元相连。输入图像经过多层的卷积、池化等操作,再与全连接层相连时已经呈现出高层特征,全连接层使用这些高层特征进行图像分类。

## 2.3 卷积神经网络的相关运算

卷积神经网络的训练过程主要分为两个部分:第一部分是前向传播,第二部分是反向传播。前向传播过程中主要涉及的是离散卷积运算和池化;反向传播过程是利用实际输出与期望输出的“误差”更新神经网络中的参数,实现对神经网络的训练。

### 2.3.1 前向传播

离散卷积在图像处理中占有很重的地位，离散卷积公式如下：

$$y(n) = \sum_{i=-\infty}^{\infty} x(i)h(n-i) = x(n)h(n) \quad (2-1)$$

在卷积神经网络中前向传播中的卷积操作公式为：

$$x_j^l = f \left( \sum_{i \in M_j} x_i^{l-1} k_{i,j}^l + b_j^l \right) \quad (2-2)$$

其中 $l$ 表示第 $l$ 层， $j$ 表示卷积层的第 $j$ 个核， $M$ 表示卷积核所在区域 $k$ 表示卷积核 $b$ 表示偏置， $x$ 表示特征图对应位置的值， $f$ 表示激活函数。具体卷积操作如下图所示：

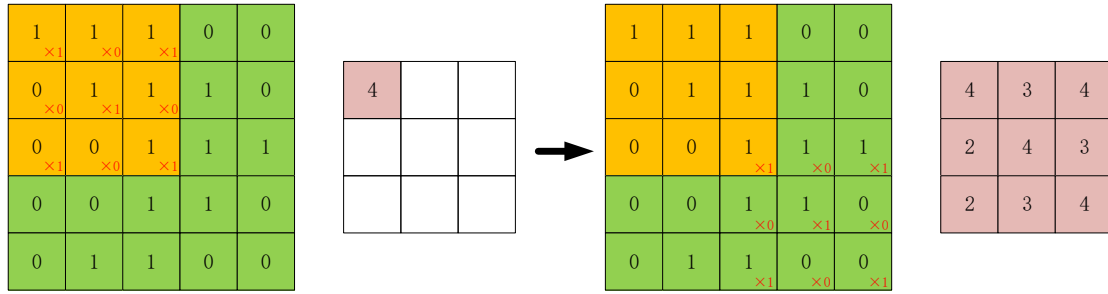


图 2-5 卷积计算示意图

在卷积神经网络中常用的池化（下采样）运算有：最大池化、均值池化和高斯池化。在卷积神经网络设计过程中，池化层的设计只需要定义池化窗口的大小、池化方法和步长。最大池化过程如下图所示：

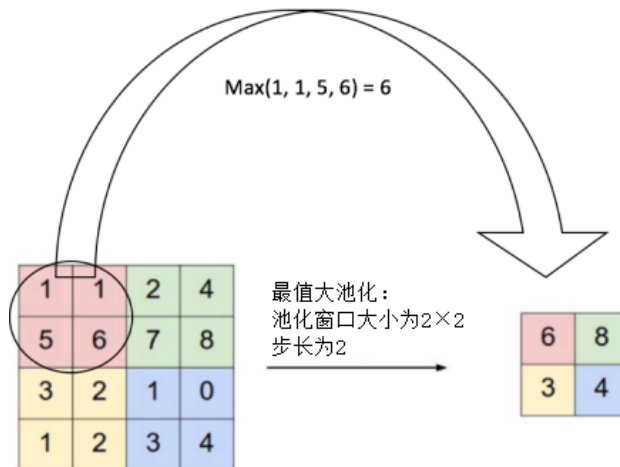


图 2-6 最大池化计算示意图

### 2.3.2 反向传播

卷积神经网络中的反向传播主要作用是对卷积神经网络中的参数进行更新，从而实现对神经网络的训练<sup>[50]</sup>。卷积神经网络的反向传播算法中可以分为三种情况：第一种情况全连接网络部分的参数更新；第二种情况是第  $l+1$  层是池化层  $l$  层是卷积层；第三种情况是第  $l+1$  层是卷积层第  $l$  层是池化层。

反向传播算法是“误差反向传播”的简称，通常与梯度下降法结合用来训练人工神经网络。该方法通过计算神经网络中代价函数对所有参数的梯度，用来更新参数值，使得代价函数不断减小，实现对神经网络的训练。

第一种情况：全连接网络部分的参数更新。全连接层的反向传播计算过程需要利用残差这一中间变量。残差的计算分为两种情况，一种是输出层的残差计算；一种是隐含层的残差计算

$$\delta_j = (d_{q,h} - x_{out,j}) g(x_j) \quad (2-3)$$

$$\delta_j^l = \left( \sum_{h=1}^{n^{l+1}} \delta_h^{l+1} w_{h,j}^{l+1} \right) g(x_j^l) \quad (2-4)$$

公式（2-3）为输出层残差计算公式，公式（2-4）为隐含层残差计算公式。其中  $d_{qh}$  表示期望输出。 $x_{out,j}$  表示实际输出， $g(x_j)$  表示激活函数的导数， $x_j$  表示上一个层的输出， $h$  表示第  $h$  个神经元， $j$  表示第  $j$  个输入。

根据反向传播算法公式，关于全连接网络层的权重和偏置更新公式如下：

$$\Delta \vec{W}^l = -\eta \cdot \vec{x}^{l-1} \cdot (\vec{\delta}^l)^T \quad (2-5)$$

$$\Delta \vec{b}^l = \vec{\delta}^l \quad (2-6)$$

$\Delta \vec{W}^l$  表示第  $l$  层的权值， $\eta$  表示学习率， $\vec{\delta}^l$  表示  $l$  层的残差， $\vec{x}^{l-1}$  表示  $l-1$  层的输出， $\Delta \vec{b}^l$  表示第  $l$  层的偏置。

第二种情况：第  $l+1$  层是池化层  $l$  层是卷积层。由于经过第  $l+1$  层的池化操作，卷积层输出的特征图的维度和池化层的输出的特征图维度存在着很大的差别，这使得  $l+1$  层的残差维度也和卷积层的维度也将不一致，所以在进行残差的计算时，需要对残差进行上采样（ $\text{up}(\delta_j^{l+1})$ ）运算。 $\text{up}(\delta_j^{l+1})$  使用 Kronecker 乘积恢复与卷积层的矩阵的大小。下面是残差计算公式为：

$$\delta_j^l = \beta_j^{l+1} \left( f'(u_j^l) \circ \text{up}(\delta_j^{l+1}) \right) \quad (2-7)$$

$f'(u_j^l)$  表示卷积层（ $l$  层）的激活函数的一阶导数， $\beta$  表示池化层的权值， $j$  表示卷积层的第  $j$  个卷积核。 $f'(u_j^l) \circ \text{up}(\delta_j^{l+1})$  表示矩阵的相乘，即对应位置元素

相乘。

偏置更新

$$\Delta b_j^l = \sum_{u,v} (\delta_j^l)_{u,v} \quad (2-8)$$

$u, v$  表示残差矩阵的位置。

$$\Delta k_{ij} = \sum_{u,v} \eta (\delta_j^l)_{uv} (p_i^{l-1})_{uv} \quad (2-9)$$

$p_i^{l-1}$  表示  $x_j^{l-1}$  在卷积过程中与  $k_{i,j}^l$  相乘的结果  $x$  表示的第  $l-1$  层输出的值,  $\eta$  表示学习率。

第三种情况: 第  $l+1$  层是卷积层第  $l$  层是池化层。这种情况池化层和卷积层也存在维度差。所以在计算池化层的残差时, 也需要对残差矩阵的周边补零, 进行扩充, 使得计算出的残差维度和池化层的维度一样。

$$\delta_i^l = f'(u_i^l) \odot \left( \sum_{j \in M} \delta_j^{l+1} \odot K_j \right) \quad (2-10)$$

$f'(u_i^l)$  表示激活函数的一阶导数,  $i$  表示第  $l$  层 (采样层) 的第  $i$  张图,  $M$  表示用到了  $l$  层的第  $i$  张图的  $l+1$  层的图集合,  $K$  表示卷积核, 表示  $\delta_j^{l+1} \odot K_j$  一种矩阵运算。矩阵运算过程首先将矩阵  $K_j$  以左边为轴进行 180 度翻转, 然后以上边为轴进行 180 度翻转, 再对  $\delta_j^{l+1}$  进行四周进行填充 0, 使得  $\delta_j^{l+1}$  进行卷积运算后维度在与  $K_j$  卷积后维度与池化层的一致, 然后进行卷积运算。

偏置更新

$$\Delta b_i = \sum_{uv} (\delta_i^l)_{uv} \quad (2-11)$$

池化权值更新

$$\Delta \beta_i = \eta \sum_{uv} (\delta_i^l * d_i^l)_{uv} \quad (2-12)$$

$d_i^l$  表示  $l$  层的下采样值输出值,  $u, v$  表示坐标位置,  $\eta$  表示学习率。

## 2.4 多尺度卷积神经网络设计

常见的深度学习网络如 AlexNet, CaffeNet 等网络结构, 只利用了单个输出层提取的特征, 然后将这些特征用于图像的分类识别<sup>[51]</sup>。这些网络只利用了最后提取的高层特征进行图像的分类识别, 这导致这些网络能够很好的区分人、狗、和汽车等这些区别很明显的物体。但是这种网络结构往往很难区分需要精细特征

才能判断的对象，比如区分汽车的型号，动物的物种等。而静态手势识别正是要精细特征才能解决的问题，所以在静态手势识别中采用常见的单一特征卷积神经网络很难取得很好的识别效果。Songfan Yang 等人<sup>[51]</sup>在 2015 年提出了一种基于多尺度特征的卷积神经网络用于图像分类识别技术。根据生理学家的对哺乳动物视觉系统的研究，对于图像表示应该从不同分辨率来进行描述。这就是 Songfan Yang 等人提出的基于多尺度特征的卷积神经网络的核心思想。通过提取不同尺度下的特征，更加准确的表示了图像，这就能够使得卷积神经网络的识别率得到提升。

#### 2.4.1 多尺度卷积神经网络结构

多尺度在机器视觉中是一个比较经典的概念。但是多尺度的概念在卷积神经网络中还很少运用。下面就是 Songfan Yang 提出的多尺度神经网络结构示意图。

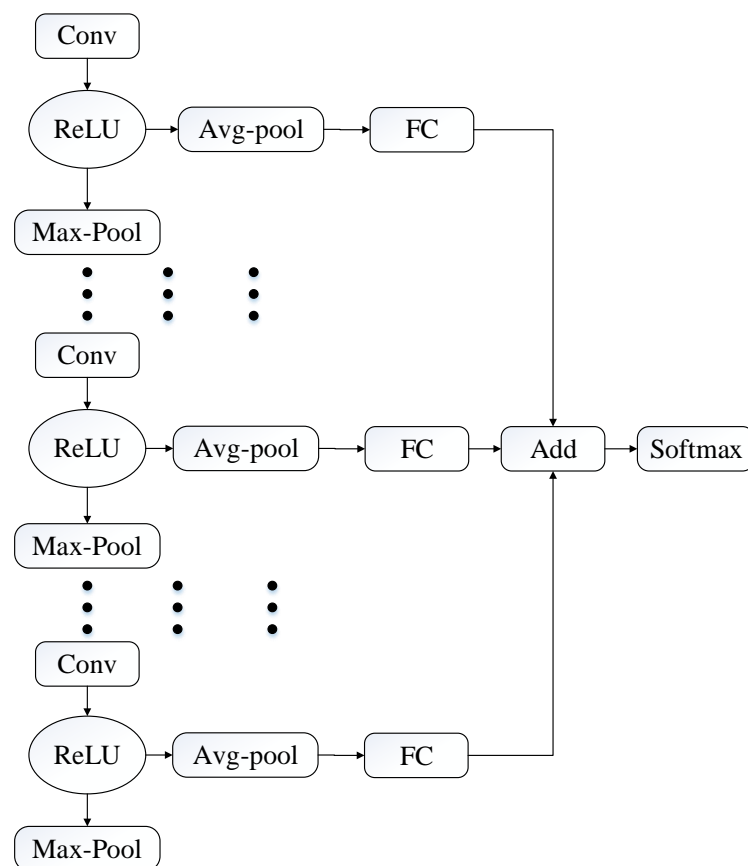


图 2-7 多尺度卷积神经网络示意图

(Conv 表示卷积层, ReLU 表示激活层, Max-Pool 表示最大池化层, Avg-Pool 表示平均池化层, FC 表示全连接层, Softmax 表示 Softmax 层)

Songfan Yang 提出的多尺度深度卷积神经网络结构是将每一个激活层 (ReLU 层) 后输出的特征图分两路输出，一路输出沿着正常的传播方向输出；

一路直接输出，经过均值池化后接入全连接层。最后将各个全连接层输出的特征向量进行特征融合，输入 Softmax 层进行分类识别。

通过多层的特征提取，多尺度卷积神经网络能够利用低层、中层和高层图像特征进行图像分类识别，使得图像的分类识别能够更加精细化。而且多尺度卷积神经网络利用已经计算出来的特征图，并没有给神经网络增加很大的计算量。

#### 2.4.2 基于多尺度卷积神经网络的静态手势识别网络设计

基于多尺度卷积神经网络的静态手势识别网络设计主要有三个难点，第一、神经网络层数的确定；第二、尺度特征的选择；第三、如何避免过拟合。

在神经网络设计中，为了达到良好的实验效果，我们需要确定合适的神经网络层数。如果设计的神经网络层数太少，会导致神经网络性能不能满足要求，识别效果不好，如果选择的层数过多，很容易出现过拟合现象，而且训练时间很长，对实验设备的要求也跟高，所以有必要选择合适的网络层数。本文参考现有的神经网络模型，以此为基础来进行微调，通过实验对比，选取最适合手势识别的神经网络模型。本文对 CaffeNet、VGG\_CNN\_F、VGG\_CNN\_M、VGG\_CNN\_S<sup>[53][54]</sup> 深度卷积神经网络进行了测试，并最终以 CaffeNet 网络模型为基础设计了多尺度网络模型。

尺度特征的选取对于实验结果影响很大，如果每一个激活层输出的特征都叠加到一起很容易出现过拟合现象，并且由于层数的增加，会占用很大的运行内存。如果选的尺度特征过少，并不能达到预期的实验效果，所以选取合适的尺度特征非常的必要。在本文中对尺度特征的选择是基于贪心算法的思想，通过实验对比得出。

本文设计的多尺度卷积神经网络为了避免过拟合，将代价函数正则化。代价函数正则化是在代价函数中加入一个额外的正则化项。加入正则化项的代价函数变为：

$$C=C_0+\frac{\lambda}{2}\sum_{\omega}\omega^2 \quad (2-13)$$

C 代表新代价函数， $C_0$  代表原代价函数， $\lambda$  为参数， $\omega$  权重。

新的代价函数对权值求偏导可以知：

$$\frac{\partial C}{\partial \omega}=\frac{\partial C_0}{\partial \omega}+\lambda\omega \quad (2-14)$$

对于权值的学习变为：

$$\omega' = \omega - \eta \frac{\partial C_0}{\partial \omega} - \eta \lambda \omega = (1 - \eta \lambda) \omega - \eta \frac{\partial C_0}{\partial \omega} \quad (2-15)$$

新的权值更新规则出现了  $1 - \eta \lambda$ ，其中  $\eta$  是学习率， $\eta \lambda$  称为权值衰减率，通过调节  $\lambda$  的大小，改变整体的权值大小。当  $\lambda$  比较大时，训练好的模型权值比较小，比较小的权值对训练数据中的噪声不敏感，从而能够减少过拟合现象的出现。

最终我们设计的卷积神经网络结构如下图：

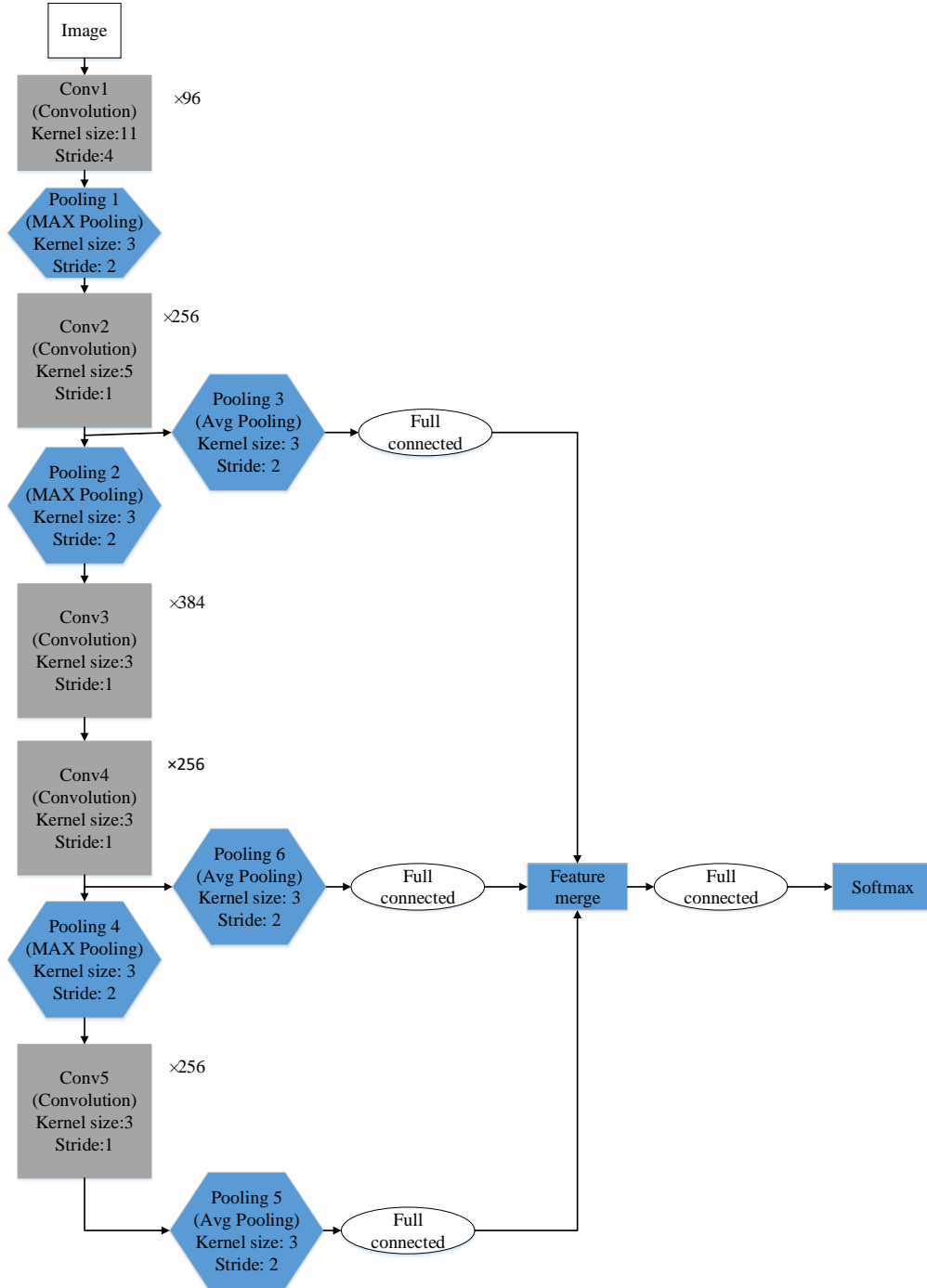


图 2-8 本文设计的多尺度卷积神经网络

在本网络包含 5 个卷积层和 6 个池化层和 3 个全连接层。第 1 个卷积层卷积核大小为  $11 \times 11$  步长为 4 包含 96 个卷积核；第 2 个卷积层卷积核大小为  $5 \times 5$  步长为 1 包含 256 个卷积核；第 3 个卷积层卷积核大小为  $3 \times 3$  步长为 1 包含 384 个卷积核；第 4 个卷积层卷积核大小为  $3 \times 3$  步长为 1 包含 256 个卷积核；第 5 个卷积层卷积核大小为  $3 \times 3$  步长为 1 包含 256 个卷积核；第 1 个池化层池化窗口大小  $3 \times 3$ ，步长为 2，采用最大池化方式；第 2 个池化层池化窗口大小  $3 \times 3$ ，步长为 2，采用最大池化方式；第 3 个池化层池化窗口大小  $3 \times 3$ ，步长为 2，采用最大池化方式；第 4 个池化层池化窗口大小  $3 \times 3$ ，步长为 1，采用最大池化方式；第 5 个池化层池化窗口大小  $3 \times 3$ ，步长为 2，采用平均池化方式；第 6 个池化层池化窗口大小  $3 \times 3$ ，步长为 2，采用平均池化方式。根据 Songfan Yang 等人设计多尺度卷积神经网络的原则，选取经过激活函数 ReLU 输出的特征图作为多尺度特征，本网络中我们选取的是第 2 个卷积层、第 4 个卷积层和第 5 个卷积层输出的特征图分别进行池化处理，通过一个全连接层后进行特征融合，最后输入 Softmax 层进行分类识别。

## 2.5 实验仿真与结果分析

为了对本章采用的基于多尺度卷积神经网络的静态手势识别方法进行评价，我们将从两个方面进行：1. 采用单尺度特征的卷积神经网络和采用多尺度特征的卷积神经网络识别进行对比；2. 传统静态手势识别方法和基于多尺度卷积神经网络的静态手势识别方法的结果进行比较。

### 2.5.1 实验环境介绍

本文的实验硬件环境是采用 Nvidia 的 GTX1060 显卡和 Intel 的 6 代 i7 处理器，GTX1060 显卡具有 6.1 的算力，拥有 6G 显存容量，能够为深度卷积神经网络提供强大的计算能力。

本实验的软件环境采用的 Ubuntu16.04 系统和伯克利视觉和学习中心（BVLC）开发的 Caffe 框架。Ubuntu16.04 是基于 Debian 发行版和 GNOME 桌面环境开发的一款 Linux 操作系统，具有友好的交互界面和良好的稳定性。Caffe 框架是目前最流行的深度学习框架<sup>[52]</sup>，Caffe 框架具有灵活的框架结构，通过简单的一条命令实现 CPU 和 GPU 之间的切换来训练；Caffe 框架是一个开源框架，在 Caffe 发布的第一年就有 1000 个开发者参与框架的完善，具有分布开发者和使用者；Caffe 具有良好的性能，单个 NVIDIA K40 GPU 一天能够处理 6 千万张图片。



### 2.5.2 单尺度特征与多尺度特征神经网络实验结果对比

关于采用单尺度特征的卷积神经网络和采用多尺度特征的卷积神经网络识别结果的对比,在本章中,我们实验了当前比较流行的几种单尺度特征的深度卷积神经网络 CaffeNet、VGG\_CNN\_F、VGG\_CNN\_M、VGG\_CNN\_S,并且我们参考这些网络结构,同时设计实现了这些网络结构的多尺度深度卷积神经网络模型,然后经行了实验对比。本文实验采用的训练样本为 10000,测试样本为 1500,这些样本既包含简单背景也包含复杂背景,实验结果如下表:

表 2-1 各网络识别精度、训练所需时间以及所需内存

网络结构	测试精度	迭代 1 次时间	网络所需内存
CaffeNet	83.7%	0.43s	2.5G
Multi_Scale-CaffeNet	90.3%	0.51s	3.4G
VGG_CNN_F	81.9%	0.40s	2.2G
Multi_Scale-VGG_CNN_F	87.1%	0.55s	3.33G
VGG_CNN_M	79.7%	0.56s	2.3G
Multi_Scale-VGG_CNN_M	86.8%	0.64s	3.6G
VGG_CNN_S	74.3%	0.65s	2.6G
Multi_Scale-VGG_CNN_S	85.3%	0.68s	3.7G

在单尺度的卷积神经网络结构中,通常是将全连接网络的最后一层的输出作为特征,CaffeNet 特征向量维度为 4096,VGG\_CNN\_F 特征向量维度为 1000,VGG\_CNN\_M 特征向量维度为 1000,VGG\_CNN\_S 特征维度 1000。多尺度卷积神经网络的特征维度主要取决与两个方面的选择:第一是特征图的选择;第二是特征图池化窗口大小的选择。Multi\_Scale-CaffeNet 本文选择第 2, 4, 5 个卷积层输出的特征图加入,特征维度变为 9216;Multi\_Scale-VGG\_CNN\_F 选择第 1, 3, 5 个卷积层输出的特征图加入,特征维度变为 2000;Multi\_Scale-VGG\_CNN\_M 选择第 1, 3, 5 卷积层输出的特征图加入,特征维度变为 2000;Multi\_Scale-VGG\_CNN\_S 在选择第 1, 4, 5 层输出的特征加入,特征维度变为 2000。这些卷积神经网络引入多尺度特征进行实验,特征维度大致增加了 2 倍。从表 2-1 可以看出,多尺度卷积神经网络特征维度增加,识别率也得到了很大的提升,说明引入多尺度特征能够提高卷积神经网络静态手势的识别率。但是网络训练的时间并没有出现大幅度增加,这是因为训练卷积神经网络卷积计算是耗时最大的操作,而在本文设计的多尺度卷积神经网络中并没有进行比原网络更多的卷积计算,所以网络训练所需的时间并没有大幅度增加。从内存使用的情况看,

由于我们增加了网络的层数，所以网络需要保存的中间变量增加，这就使得训练网络所需的内存增加比较大。

本文的设计多尺度卷积神经网络是在 CaffeNet 基础上，加入多尺度特征实现的。下面是 Multi\_Scale-CaffeNet 网络迭代次数和识别精度变化曲线。

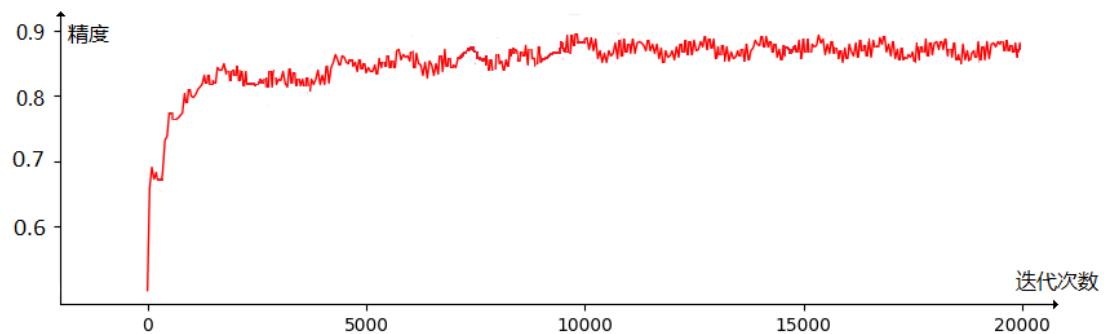


图 2-9 手势识别精度与迭代次数变化图

迭代次数和精度变化图可以看出网络在前 1300 次训练迭代，识别精度迅速提升，随着迭代次数的增加精度提升缓慢，在后 10000 次到 20000 次迭代过程中出现震荡，识别精度几乎没有得到提升。迭代前期精度提高迅速，是因为神经网络迅速向极小值点靠近，所以精度提升快；当迭代到一定次数，离极小值点越来越近，这就出现了震荡。

### 2.5.3 与传统静态手势识别方法实验结果对比分析

传统静态手势识别方法比较多，本文选取了最具有代表性的静态手势识别方法动态时间规整算法进行比较。

基于动态时间规整算法的静态手势识别首先通过皮肤检测和边缘检测，将手势的轮廓检测出来，然后将手势轮廓用时间序列化曲线描述，然后计算样本与测试集欧式距离，选取距离最小的作为识别的结果。



图 2-10 简单背景手势原始图像



图 2-11 简单背景手势轮廓检测结果



图 2-12 复杂背景手势原始图像

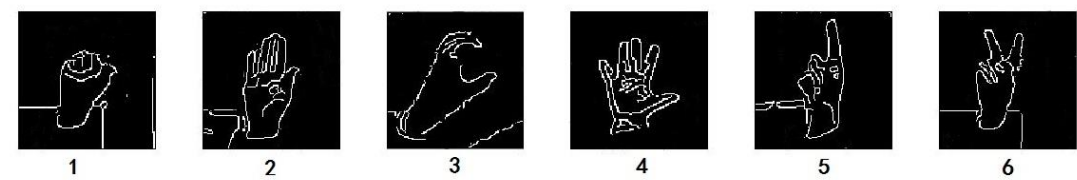


图 2-13 复杂背景手势轮廓检测结果

下面是基于动态时间规整算法的静态手势识别与基于多尺度卷积神经网络的静态手势识别结果如下：

表 2-2 简单背景静态手势识别率

手势分类	1	2	3	4	5	6
动态时间规整	92.1%	91.5%	93.0%	91.5%	92.1%	93.3%
多尺度卷积神经网络	92.1%	92.1%	92.7%	91.6%	92.1%	92.9%

表 2-3 复杂背景静态手势识别率

手势分类	1	2	3	4	5	6
动态时间规整	79.7%	79.1%	78.2%	81.4%	82.1%	85.6%
多尺度卷积神经网络	85.1%	85.7%	86.4%	87.0%	85.2%	87.7%

从复杂背景和简单背景的实验结果可以看出基于动态时间规整的静态手势识别在简单背景下的手势识别率和多尺度卷积神经网络的手势识别率差别并不是很大。这是因为在简单背景下对于手势的分割手型的轮廓提取都能获得较好的结果，这就使得识别结果相对比较好。当背景变复杂，手势分割与轮廓提取难度增大，所以实验结果和卷积神经网络相比就存在较大的差距。这说明基于多尺度卷积神经网络的静态手势识别对于复杂环境的适应能力更好。

从单个手势的识别率角度分析可以看出，多尺度卷积神经网络各手势之间的识别率差异相比基于动态时间规整算法的手势识别率差异更小，这说明基于多尺度卷积神经网络的手势识别方法，通过训练提取了手势更加本质的特征，而基于动态时间规整的静态手势方法利用手型作为手势的特征，并不适合所有的静态手

势。

## 2.6 本章小结

本章首先介绍了静态手势样本库；接着介绍了卷积神经网络基本结构和卷积神经网络的前向传播和反向传播过程；然后分析了单尺度卷积神经网络，提出了基于多尺度卷积神经网络的静态手势识别方法；最后通过与多种单尺度卷积神经网络的对比实验，证明了基于多尺度卷积神经网络的静态手势识别方法优于单尺度卷积神经网络；通过与传统静态手势识别方法进行对比发现，基于卷积神经网络的静态手势识别方法对于复杂环境适应性明显优于传统静态手势识别方法。

### 3 动态手势识别

将手势识别技术应用于人机交互领域如：智能电视的操控，车载多媒体交互，都需要用到动态手势识别。动态手势能够为人机交互提供良好的用户体验，能够使得手势的交互更加丰富。传统的动态手势识别都是人工提取手势特征进行识别，往往手势识别特征的选取对实验结果影响非常大，使得模型的设计变得非常困难，而且设计的模型很难适应复杂多变的环境，手势识别效果并不理想。卷积神经网络不需要人工提取手势特征，卷积神经网络能够从训练样本中自主学习样本特征。随着卷积神经网络的发展，卷积神经网络从最初只能对图片进行分类识别，发展出了对视频的处理的能力。本章将从卷积神经网络对视频分类处理上着手，并结合动态手势识别的特点，提出一种基于 3D 卷积神经网络的动态手势识别方法。

#### 3.1 传统动态手势识别

与静态手势相比，动态手势的识别需要同时结合时间信息和空间信息，这使得动态手势识别难度更大。传统动态手势识别方法依赖于手势特征的选取，大部分动态手势识别是利用手在空间中的运动轨迹作为特征来实现动态手势的识别。下面本文将分析常用的几种传统动态手势识别方法的优劣<sup>[55]</sup>。

隐马尔科夫模型是一种统计模型，创立于 20 世纪，并在语音识别领域、模式识别领域取得了许多重要的成果<sup>[56]</sup>。基于隐马尔科夫模型的动态手势识别方法是把手形及运动轨迹作为手势识别的特征进行识别。通过对人手轮廓的跟踪，获得手部形状和多自由度（手心空间坐标  $(x, y, z)$  以及欧式位姿  $(\alpha, \beta, \theta)$ ）的手势的运动轨迹，这样就实现了对视频中手势特征的提取。将提取的动态手势特征作为隐马尔科夫模型的观察值序列，并作为训练好的隐马尔科夫模型的输入值，通过前向算法求出最大似然值，然后输出识别的手势。基于隐马尔科夫模型的动态手势识别将每一种手势对应一个隐马尔科夫模型，其优点是对复杂的手势动作识别精度较高，新增手势时容易添加或修改手势库。但是随着手势数量的增多，隐马尔科夫模型越来越复杂，训练难度也会越来越大。

基于压缩时间轴的动态手势识别方法，首先将手从各帧图像中分割出来，然后将这一系列的图像进行归一化处理，形成一张静态的图片<sup>[57]</sup>。然后通过静态手势识别的方法进行动态手势的识别。这种手势识别的好处是能够很好的识别手型不同的手势，但是由于丢失了很多手势的空间特征，这导致这种手势识别的方法对空间变化复杂的手势识别率很低。

动态时间规整算法本质上是一种模板匹配算法<sup>[55]</sup>。由于在动态手势识别的过程中每一个手势的时间序列长度都不一样,所以在计算测试模板和参考模板的相似性的时候采用动态时间规整算法应对手势时间序列长度不一致的情况。动态时间规整算法将手势运动轨迹作为动态手势特征,通过计算测试样本与参考样本之间的欧式距离实现对动态手势的识别。与基于隐马尔科夫模型的动态手势识别相比动态时间规整算法实现简单,计算量较小,但是基于动态时间规整的手势识别算法对复杂手势识别率较低。

### 3.2 基于卷积神经网络的视频分类算法

卷积神经网络对于图片的分类理解取得了很好的结果,随着研究的深入,逐渐发展出了针对视频分类的卷积神经网络。在 2015 年 Andrej Karpathy 总结了卷积神经网络在视频处理上的难点<sup>[58]</sup>。第一、卷积神经网络对视频的处理不能只简单的考虑对单帧静态图像的理解,还需要考虑到整个运动过程。第二、卷积神经网络对视频的训练比对图片的训练涉及的参数更多,模型训练时间更长,所以如何优化神经网络结构也是卷积神经网络重要考虑的因素。

目前基于卷积神经网络的视频分类的算法主要有结合长短时记忆 (LSTM) 网络的视频分类算法、排名池化网络结构的视频分类算法、3D 卷积神经网络的视频分类算法。

长短时记忆网络是循环神经网络 (RNN) 中的一种<sup>[59][60]</sup>,最初在自然语言处理领域中使用。长短时记忆网络最大的优势在于能够利用历史数据,通过“串行”的方式实现对整体的理解。在视频分类处理中,往往需要对视频中整体特征进行综合才能实现对视频的正确理解,而 LSTM 网络非常适合处理这种问题,使得 LSTM 在视频分类中得到了很多的应用。但是 LSTM 结构复杂,这样使得在视频分类处理中计算量会非常大,训练的参数非常多,这就对硬件和数据样本提出了非常高的要求。

基于排名池化 (Rank-Pooling) 的视频分类算法<sup>[61]</sup>,是对视频中的单帧图像用卷积神经网络进行特征提取,然后再结合所有的卷积神经网络提取的特征,在 Temporal Pool 层进行 Rank-Pooling 计算,最后将特征输入 Softmax 层进行分类识别。

基于 3D 卷积神经网络的视频分类算法将视频看成 3 维矩阵输入卷积神经网络,通过 3D 卷积和 3D 池化提取时空特征。与 2D 卷积神经网络的最大不同在于 3D 卷积神经网络采用 3D 卷积核和 3D 池化,不仅能够提取空间特征,还能提取时间特征。

### 3.3 3D 卷积神经网络结构设计

在本小节将介绍本文设计的基于 3D 卷积神经网络的动态手势识别方法。本文将从以下几个方面介绍：CNN 网络中时空信息的融合方式，3D 卷积神经网络整体结构，3D 卷积神经网络的训练。

#### 3.3.1 CNN 网络中的时空信息融合方式

相比图像分类识别，视频分类识别不仅要考虑空间特征提取，而且需要考虑时间特征的提取。视频的分类需要融合时间信息和空间信息，才能提取出有效的特征。卷积神经网络处理视频的时空信息融合网络结构有三种，早期融合网络结构，晚期融合网络结构，缓慢融合网络结构<sup>[58]</sup>。

**早期融合网络结构：**早期融合网络结构是将所有视频帧图像以 3 维矩阵的形式输入到第一个卷积层，通过一个与视频长度相同的卷积核进行卷积操作，实现时空信息的融合。具体结构如下图所示：

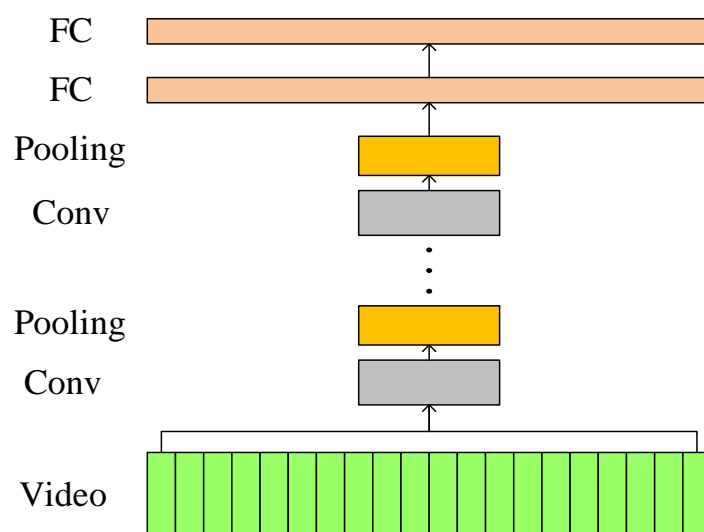


图 3-1 早期融合示意图

(FC 表示全连接层，Pooling 表示池化层，Conv 表示卷积层)

**晚期融合网络结构：**视频的单帧图像分别输入到独立的卷积神经网络中，通过卷积池化操作，提取每一帧图像的特征，然后在最后的全连接网络中将所有帧的特征的合并，实现时空信息的融合。具体网络结构如下图：

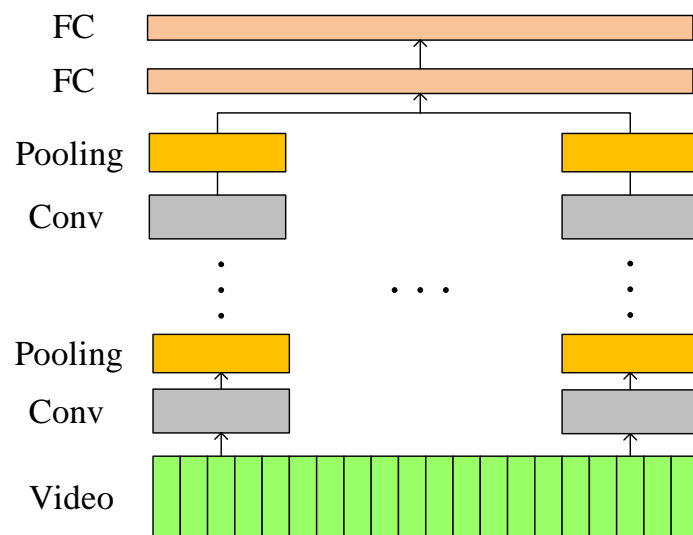


图 3-2 晚期融合示意图

(FC 表示全连接层, Pooling 表示池化层, Conv 表示卷积层)

缓慢融合网络结构: 缓慢融合结构是将视频分成多个 3 维矩阵, 分别输入到多个 3D 卷积层, 经过几次卷积, 池化操作后, 将各个卷积层逐步合并, 最后合并到一个卷积层。结构如下图所示:

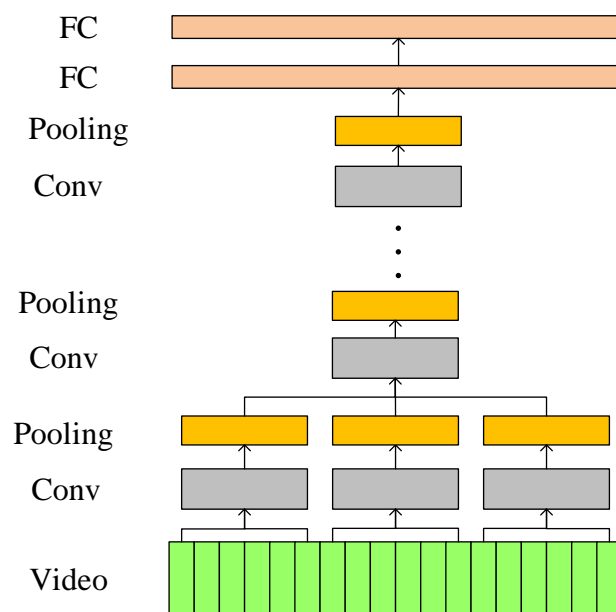


图 3-3 缓慢融合示意图

(FC 表示全连接层, Pooling 表示池化层, Conv 表示卷积层)

### 3.3.2 3D 卷积神经网络整体网络结构

Du Tran 等在 2015 年提出了一种 3D 卷积神经网络来学习视频中的时空特征, 并将学习的特征与简单的线性分类器结合, 在视频分类任务中产生了良好的效果。



与 2D 卷积神经网络相比, 拥有 3D 卷积和 3D 池化的卷积神经网络能够更好的对时空信息进行建模。3D 卷积神经网络的卷积和池化都是同时作用在时间和空间域上, 而 2D 的卷积神经网络仅仅作用在空间域<sup>[65]</sup>。下图是 2D 卷积神经对单帧图像、2D 卷积神经网络对视频和 3D 卷积神经网络对视频进行卷积运算示意图。

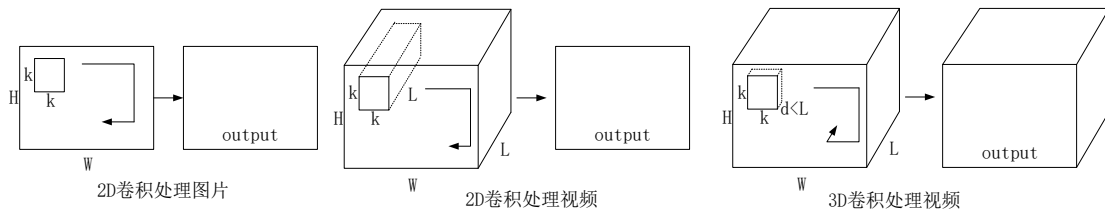


图 3-4 2D 卷积操作和 3D 卷积操作示意图

2D 卷积神经网络对单帧图像进行卷积运算上输出单张图, 因此 2D 卷积神经网络进行卷积运算之后就丢失输入信号的时间信息。同样, 2D 池化也和 2D 卷积一样, 也会丢失时间信息。2D 卷积神经网络对多帧图像进行卷积运算也是输出单张图片, 因此在进行第一次卷积之后时间信息也会丢失。只有 3D 卷积神经网络保留了输入视频中的时间信息。在本文的设计的卷积神经网络中, 采用的缓慢融合网络结构, 将视频帧分成三份, 分别转换成 3 维矩阵输入到 3D 卷积神经网络中。与早期融合网络结构和晚期融合结构相比, 采用缓慢融合网络结构能够充分的提取时空特征, 对时空信息的利用率大大提升, 所以本文采用缓慢融合网络结构。下面是本文设计的 3D 卷积神经网络结构:

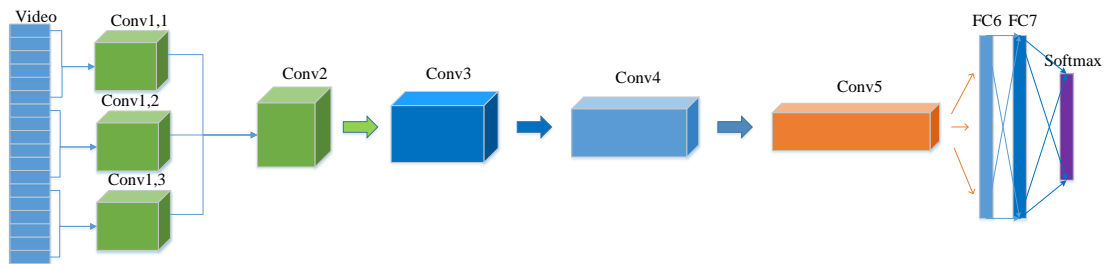


图 3-5 3D 卷积神经网络结构图

从上图可以看出, 本文设计的 3D 卷积神经网络时空信息融合方式采用缓慢融合方式, 先用三个卷积层对视频进行一次卷积操作, 然后合并输入下一个卷积层, 合并之后再经过四个卷积层和两个全连接层, 最后输入 Softmax 层进行分类识别。在本图中的卷积层包含卷积和池化两个操作。第一个卷积层的卷积核大小都为  $3 \times 3 \times 3$ , 都只有一个卷积核, 采用均值池化方法, 池化窗口大小为  $2 \times 2 \times 2$ ; 第二个卷积层的卷积核大小是  $3 \times 3 \times 2$ , 一共有 4 个核, 池化窗口大小是  $2 \times 2 \times 2$ , 采用均值池化方法; 第三个卷积层的卷积核大小是  $5 \times 5 \times 3$ , 一共有 8 个核, 池化窗口大小是  $2 \times 2 \times 2$ , 采用均值池化方法; 第四个卷积层的卷积核大

小是  $5 \times 5 \times 3$ ，一共有 32 个核，池化窗口大小是  $1 \times 2 \times 1$ ，采用均值池化方法；第五个卷积层的卷积核大小是  $3 \times 5 \times 3$ ，一共有 64 个核和池化窗口大小是  $2 \times 2 \times 1$ ，采用均值池化方法。经过两个全连接网络，再将结果输入到 Softmax 层，Softmax 层的输出结果表示为一个一维向量 P，P 计算公式如下：

$$P_j = \frac{e^{z_j}}{\sum_q e^{z_q}} \quad (3-1)$$

j 代表第 j 个输出神经元；z 代表 Softmax 层的加权输入；q 代表神经元数量。

### 3.3.3 过拟合

深度卷积神经网络包含多个非线性隐含层，具有很强的拟合能力，可以在输入和输出之间产生非常复杂的对应关系。本文为了避免过拟合现象，主要做了两方面工作：第一对数据集进行扩充；第二对网络结构进行改进。本文将在 3.4.1 节讲解如何对 VIVA 动态手势数据集进行扩充。在本节将分析在网络结构方面的改进来避免过拟合。本文在设计网络中引入 Dropout 层来避免过拟合<sup>[66]</sup>。Dropout 层是让神经网络中的神经元以一定的概率不工作。Dropout 层的操作具体如下图：

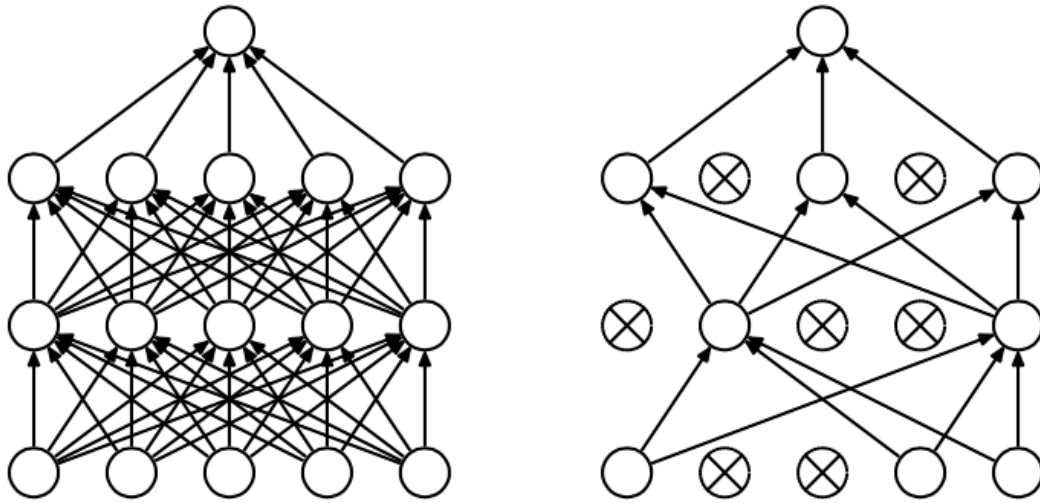


图 3-6 左图正常的神经网络，右图引入 Dropout 层的神经网络

本文在全连接层后加入 Dropout 层，Dropout 层具体实现公式如下：

没有 Dropout 层的神经网络传播过程

$$z_i^{l+1} = w_i^{l+1} y^l + b_i^{l+1} \quad (3-2)$$

$$y_i^{l+1} = f(z_i^{l+1}) \quad (3-3)$$

有 Dropout 层的神经网络传播过程

$$r_j^l \sim \text{Bernoulli}(p) \quad (3-4)$$

$$\tilde{y}^l = r^l * y^l \quad (3-5)$$

$$z_i^{l+1} = w_i^{l+1} \tilde{y}^l + b_i^{l+1} \quad (3-6)$$

$$y_i^{l+1} = f(z_i^{l+1}) \quad (3-7)$$

$z$  代表加权输入,  $l$  代表第  $l$  层  $f$  为激活函数,  $w$  表示权重,  $b$  表示偏置,  $\text{Bernoulli}(p)$  函数以概率  $p$  生成 1。

Dropout 层之所以能够降低过拟合, 有以下几个原因<sup>[66]</sup>:

- (1) 取平均的作用。通常我们用相同的训练集去训练不同的神经网络 (没有 Dropout 层), 一般情况下会得到不同的结果。如果这些神经网络都出现了过拟合现象, 那么通常我们会综合这些神经网络的结果, 进行取“平均”得到最好的结果。如果这些神经网络出现相反方向的过拟合, 那么取平均就可以实现相互抵消。神经网络中加入 Dropout 层实际上相当于一个样本集训练了多个神经网络, 能够将不同方向的过拟合实现抵消, 从而避免了过拟合。
- (2) 减少神经元之间的共适应。Dropout 层的加入使得训练的神经网络并不是每次都一样, 这样就能够减少某些神经元对特定神经元的依赖, 这样就迫使神经元去学习更加具有鲁棒性的特征, 而不会对一些特定特征过于敏感。这样就减少了过拟合的出现。
- (3) 在自然界中, 物种的进化都是自然选择的结果<sup>[67]</sup>, 更适应环境的会留下。在神经网络中加入 Dropout 层实际上相当于生物进化中的突变。经过不断迭代, 能够筛选出更适应这个模型的神经元, 这些神经元权值较大, 能够对整个神经网络的结果产生更大的影响, 不适合的神经元权值将很小, 对整个神经网络的影响较小。

### 3.3.4 3D 卷积神经网络的训练

卷积神经网络的训练其实是通过优化卷积神经网络的参数使得代价函数值越来越小。在本文设计 3D 卷积神经网络中, 代价函数采用 log-likelihood 函数:

$$L(W, n) = -\frac{1}{n} \sum_{i=0}^n \ln(P_i) \quad (3-8)$$

$n$  代表数据集大小,  $P_i$  表示分类器的输出值。本文通过随机梯度下降方法对参数进行更新, 并通过 Nesterov 加速梯度法对随机梯度进行优化<sup>[68][69]</sup>。随机梯度下降法在深度学习中得到广泛的运用, 与批量梯度下降法相比, 随机梯度下降

法收敛速度更快，而且不容易陷入局部最小值。但是随机梯度也存在不足，就是学习率的设定。当学习率设置太小，会导致收敛速度太慢；当学习率设置过大容易出现代价函数震荡，甚至发散的情况。本文通过引入 Nesterov 加速梯度法对随机梯度下降法进行优化。Nesterov 加速梯度法引入一个动量参数  $v_i$  来使随机梯度加速前进，减少在极小值附近的震荡。参数更新具体实现如下：

$$\nabla w_i = \frac{1}{n} \sum_j \frac{\partial L}{\partial w_{i-1}} \quad (3-9)$$

$$v_i = \gamma v_{i-1} - \lambda \nabla w_i \quad (3-10)$$

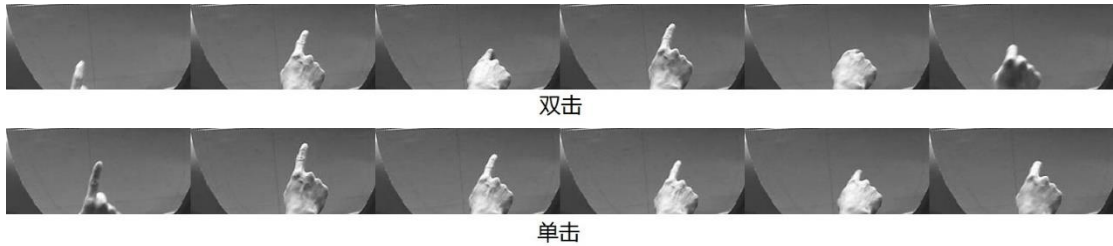
$$w_i = w_{i-1} + \gamma v_i - \lambda \nabla w_i \quad (3-11)$$

$w$  表示参数； $v_i$  表示动量参； $\gamma$  表示常数，取 0.9； $L$  是代价函数； $\lambda$  表示学习率； $i$  表示第  $i$  次迭代； $n$  表示随机梯度法中 mini-batches 样本数量，本文设置的 mini-batches 大小为 50。

Dropout 层中 Bernoulli( $p$ )函数中参数  $p$  本文设定为 0.5。Dropout 层参数在反向传播中没有激活的神经元不参与参数的更新。

### 3.4 数据集

本文采用的数据集是 VIVA (Vision for Intelligent Vehicles and Applications) 数据集<sup>[70]</sup>。VIVA 数据集考虑了与驾驶员，乘客，车辆动力学和车辆周围环境以及交通基础设施相关的参数的感测，分析，建模和预测中的问题。VIVA 数据集分为人手检测数据集，人手跟踪数据集，手势识别数据集，人脸数据集，交通灯数据集，标志牌数据集。本文采用的是手势识别数据集。手势识别数据集旨在研究复杂背景环境下的动态手势识别。该数据集定义了 20 个动态手势，2920 个动态手势视频片段，包含彩色视频信息和深度信息。VIVA 手势识别数据集是在变化的照明条件下收集的数据。手势由驾驶员座椅中的对象的右手或前排乘客座椅中的对象的左手执行，手势涉及手和手指运动。VIVA 动态手势集手势示意图如下所示：





画圈



画半圈



画叉



画勾



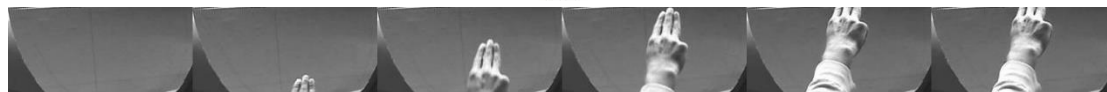
画加号



上挥



下挥



前推



下拉



左横移



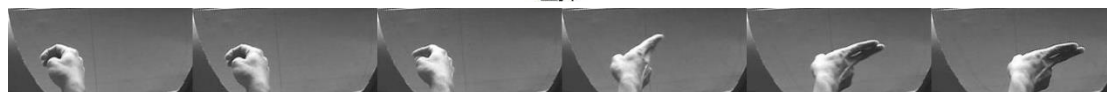
右横移



握拳



左挥



右挥

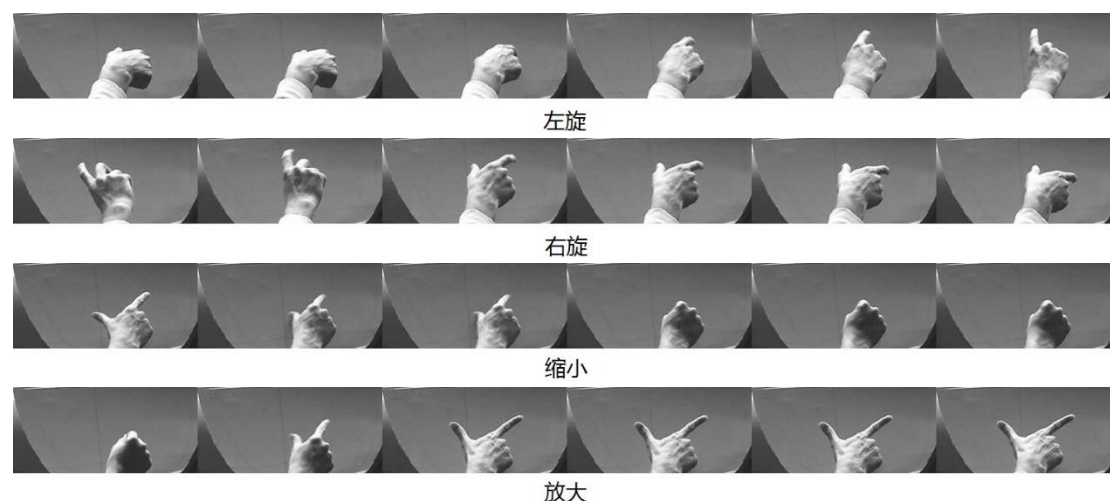


图 3-7 动态手势示意图

### 3.4.1 数据扩充

在卷积神经网络的训练中，由于 VIVA 数据集包含的数据集只有 2920 个动态手势视频，这样很容易出现过拟合现象，所以本文需要对数据集进行补充。

数据的扩充本文采用虚拟样本<sup>[71][72]</sup>对数据进行扩充。虚拟样本概念在 1992 年由 Poggio 和 Vetter 提出，虚拟样本是指通过利用待研究领域的先验知识，结合已知的样本产生辅助样本。本文采用虚拟样本生成方法是对已有样本集进行几何变换对数据集进行扩充。具体扩充方法有：1) 利用水平镜像将视频帧的每一幅图像做镜像处理，形成新的视频序列；2) 将视频帧的顺序颠倒，形成倒序的视频；3) 镜像和倒序同时作用，形成新的视频；4) VIVA 数据集的视频帧大小为  $115 \times 250$ ，本文随机裁剪出一个大小为  $100 \times 226$  的视频。具体操作结果如图 3-8 所示：

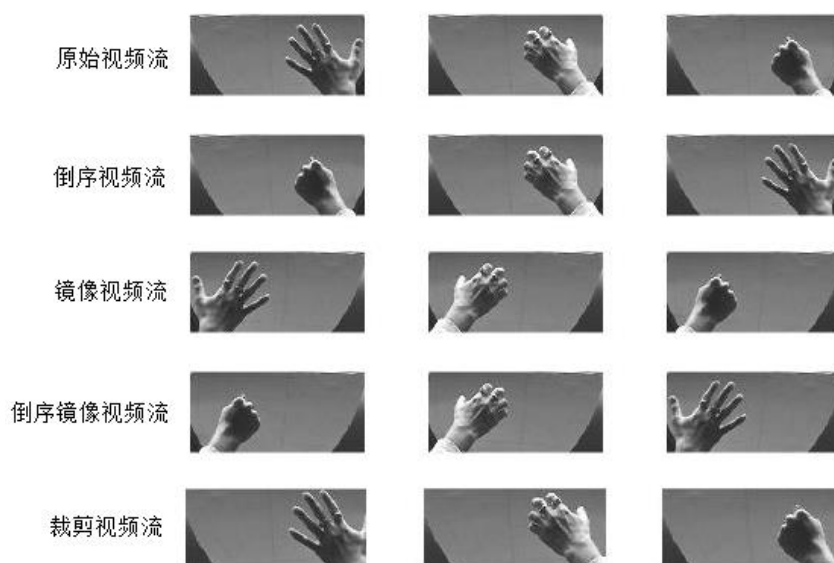


图 3-8 数据扩充示意图



### 3.4.2 数据预处理

在 VIVA 手势数据集中,不同的手势视频有不同的长度,在本章设计的 3D 卷积神经网络中,采用 30 帧长度的视频作为神经网络的输入。所以需要对数据进行预处理,使得数据长度符合神经网络的输入要求。通过使用最近邻域法丢弃或复制帧,使得每个手势的视频长度都为 30 帧。

## 3.5 实验结果分析

为了测试本文设计的 3D 卷积神经网络的性能,本文采用 VIVA 数据进行测试和训练,我们从数据集中随机抽取十分之一作为测试样本,十分之九作为训练样本。本章我们将比较三种不同时间空信息融合结构的卷积神经网络的手势识别结果。同时本节还比较了 Ohn-Bar 和 Pavlo Molchanov 等人在 VIVA 数据集上进行手势识别的结果。

本章采用的实验环境是 Nvidia GTX1060 GPU,软件环境采用的是 Facebook C3D 开源 Caffe 框架,该框架是对 BVLC Caffe 框架的改进版本,能够进行 3D 卷积和池化。

下面是早期融合结构、晚期融合结构和缓慢融合结构的卷积神经网络在随着迭代次数的增加,测试精度的变化曲线。

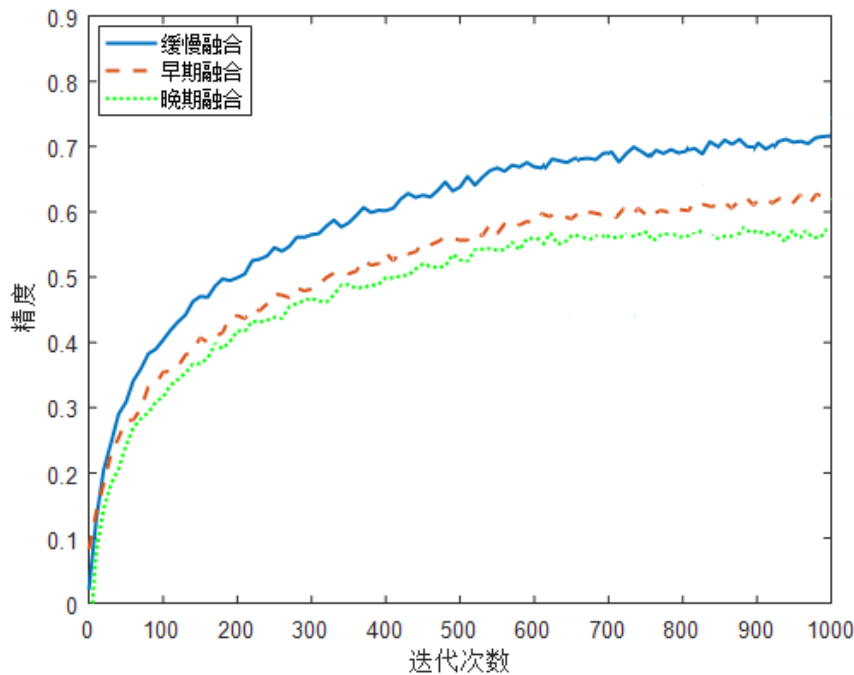


图 3-9 不同时空信息融合方式随迭代次数精度变化图

从上迭代次数和精度变化情况来看，缓慢融合识别率最高，早期融合结构识别率次之，最差的是晚期融合结构。从融合方式来看，早期融合入直接在第一层就进行了卷积操作之后，就只剩下一张特征图，这就没有充分利用到视频帧中的时间信息，只是提取了视频帧中的空间特征；而手势动作和时间变化速度等都相关，在损失了时间信息之后空间特征的提取也将不充分，这就是早期融合结构结果较差原因。晚期融合虽然没有在网络的第一层就全部损失掉时间信息，但是并没有将每一帧之间的动作关联起来，使用卷积运算提取特征阶段，没有融合时间信息和空间信息，时间特征没有得到充分利用，这就使得他的精度低于缓慢融合的 3D 卷积神经网络。

2014 年 Ohn-Bar<sup>[73]</sup>基于 VIVA 数据集提出基于 HOG 特征的动态手势识别；2015 年 Pavlo Molchanov<sup>[35]</sup>等人基于 VIVA 数据集提出了基于双卷积(LRN+HRN)的动态手势识别方法。下表是本文提出的方法与 Ohn-Bar 和 Pavlo Molchanov 提出的方法手势识别率的比较。

表 3-1 不同方法识别率

方法	HOG	LRN+HRN	3DCNN
识别率	64.5%	77.5%	71.2%
单个手势识别时间	220ms	112ms	45ms

从上表可以看出，本文提出的基于 3D 卷积神经网络的手势识别方法的识别率比 Pavlo Molchanov 等人提出的双卷积神经网络的手势识别方法的识别率略低，但高于基于 HOG 特征的手势识别方法的识别率。基于双卷积神经网络的手势识别方法需要对两组视频流进行处理，这使得基于双卷积神经网络的计算量比本文提出的 3DCNN 卷积神经网络更加大，这也将导致基于双卷积神经网络的手势识别时间更长，实时性能更差。

本文定义了 20 个手势，各手势识别情况如下表所示：

表 3-2 手势识别率

手势	双击	单击	画圈	画半圈	画勾	画叉	画加号	上挥	下挥	握拳
识别率	64.2%	64.4%	63.1%	64.6%	61.5%	60.1%	58.9%	75.2%	75.5%	76.8%



表 3-3 手势识别率

手势	前推	下拉	左横 移	右横 移	左挥	右挥	左旋	右旋	缩小	放大
识别 率	74.5%	74.1%	79.1%	79.9%	70.2%	69.8%	69.9%	70.7%	66.9%	67.2%

从上表结合手势动作视频可以看出动态手势识别率高的手势存在以下几个特点：1、手运动轨迹简单；2、与其他手势区分度高；3、完成手势时间较长。画加，画叉，画勾三个手势空间运动轨迹复杂，而且三个手势之间区分度低，这导致这些手势识别率较低。单击，双击两个手势主要的区分点在于手指点击的次数，而完成点击动作时间非常短，可能仅有一帧图像捕获了点击动作，甚至没有捕获到点击动作，这就使得这两个手势相互干扰很大，导致单击和双击手势识别率较低。

### 3.6 本章小结

本章首先分析了传统动态手势识别方法的优缺点。接着介绍了基于卷积神经网络的视频分类算法；然后详细阐述了 CNN 网络中的时空信息融合方式，分析 2D 卷积神经网络与 3D 卷积神经网络的区别，设计了基于缓慢融合的 3D 卷积神经网络的动态手势识别方法。为了避免过拟合，本章引入 Dropout 层，详细分析了 Dropout 层的原理和降低过拟合的原因。本章还对动态手势样本集进行了扩充。最后本章通过实验对比，证明了采用缓慢融合结构的 3D 卷积神经网络方法的更适合动态手势识别，通过与基于 HOG 特征和基于双卷积神经网络(LRN+HRN)的动态手势识别方法进行实验对比发现，基于 3D 卷积神经网络的动态手势识别方法识别率高于基于 HOG 特征的传统手势识别，但是略低于基于双卷积神经网络的手势识别方法，但是基于 3D 卷积神经网络的动态手势识别方法计算量小于基于双卷积神经网络的动态手势识别方法，在相同硬件条件下实时性更好。

## 4 车载手势识别原型系统设计

手势识别技术可以应用到很多场合,结合第二章和第三章的静态和动态手势识别技术,利用英特尔实感摄像头 SR300 和 Surface Pro 4 实现了车载手势识别原型系统。通过 SR300 SDK 提供的人手检测函数检测感应区域是否存在手,如果有手存在,则从 SR300 中读取 RGB 图像信息输入手势识别系统中进行手势识别。该原型系统实现了对汽车内最常用的四个应用(收音机、电话、音乐播放、导航)的手势操控,并在真实汽车内部进行了实验测试,驾驶体验得到显著改善,有效的提高了人机交互效率。

### 4.1 车载手势识别系统意义

汽车是我们生活中的重要工具,但是汽车现有的车载交互方式大多是通过按钮来进行人机交互的。随着现代科技的发展,汽车也由原先简单的交通工具变成了一个集娱乐、信息交流的智能移动平台。这使得汽车的车载交互方式的成为现在汽车创新的一个热点。奔驰推出了 COMAND 系统,通过触摸屏、按钮、语言识别来实现车载电子设备的交互<sup>[74]</sup>;宝马公司推出的 IDriver 系统通过手势识别技术和触摸屏实现对车载电子设备的交互。

手势识别技术相较于传统的触摸屏技术和按钮具有更好的用户体验,而且驾驶者不需要将眼睛的注意力从路面转向车内就可以实现对车载电子设备的控制,这大大降低了交通事故发生的概率<sup>[75]</sup>。如果利用语音识别进行交互<sup>[76][77][78]</sup>,将受到很大的外界环境因素的干扰,比如汽车内人的谈话声音,如果隔音效果不好的汽车很容易受到汽车外部声音的干扰。对于不同地区的人,存在不同的方言,这使得语音识别的难度大大增加,很难研究出具有通用意义的识别系统。所以语音识别并不能很好的满足汽车内人机交互的要求。手势识别技术则不同,受到外界干扰较小,而且设计出几个通用的手势并不困难,而且运用手势识别技术进行人机交互,具有良好的用户体验,所以研发车载手势识别系统具有重要的实际意义。

### 4.2 SR300 介绍

SR300 传感器是由英特尔公司推出的实感摄像头,能够提供每秒高达 60fps 的高清视频图像信息。SR300 实感摄像头实物图如下:



图 4-1 SR300 实物图

SR300 包含一个红外摄像头、一个高清彩色摄像头和一个红外发射器。SR300 具体参数如下表：

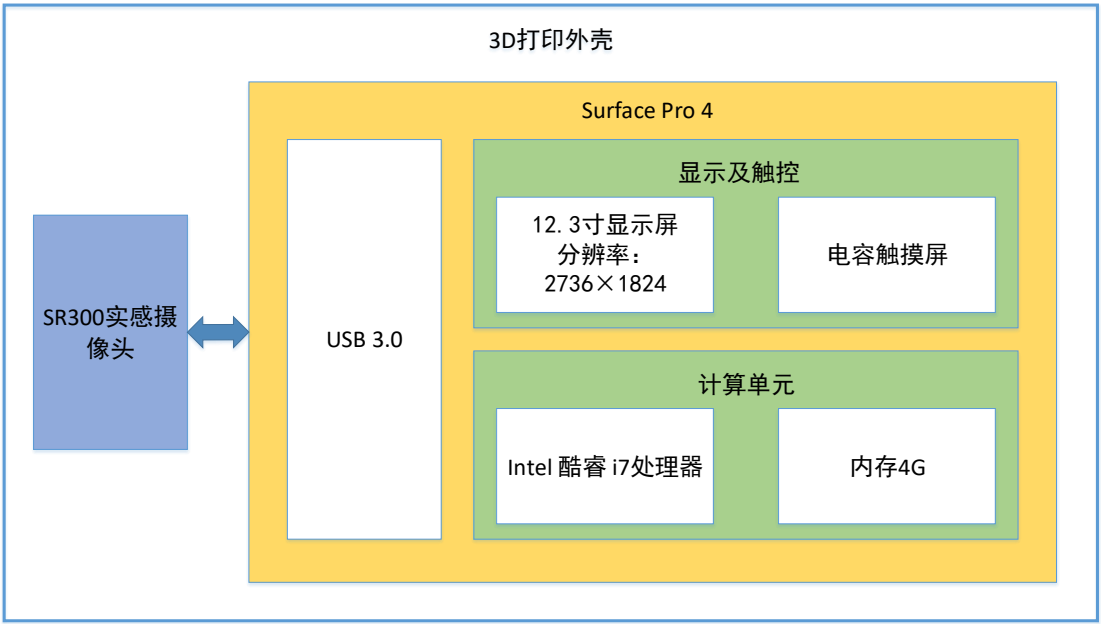
表 4-1 SR300 传感器参数

SR 300	
有效探测范围	0.2 - 1.2 米
彩色摄像头	1080p (30 FPS),720p (60 FPS)
IR 摄像头	640x480 200 FPS
主板接口	USB 3.0, 5V
所需操作系统	Microsoft Windows 10 64 位
开发语言	C++,C#, Visual Basic, Java, JavaScript

与 Kinect 相比 SR300 体积更小，通过 USB 接口就可以实现供电，不需要额外的电源，非常适合汽车这种空间狭窄的环境，所以选择 SR300 作为车载手势识别系统的传感器。

4.3 软硬件平台框架设计

本系统原型样机由一个 3D 打印的外壳进行整体装配，装配内容包括 Pad（Surface Pro 4）和 SR300 实感摄像头。Surface Pro 4 是微软公司推出的性能优越的平板电脑，采用第六代酷睿™ i7 处理器，具有强大的计算能力，Surface Pro 4 厚度只有 8.4 毫米，所以本文采用 Surface Pro 4 作为车载手势识别的核心平台。原型样机硬件系统框架图如下：



本系统软件是以 Caffe、SR300 软件开发工具包(SDK)、OpenCV、Windows10 为核心构建的手势识别操控系统。通过 SR300 SDK 提供人手检测函数检测感应区域是否有手,如果有手则获取图像信息,结合 OpenCV 对图像信息进行预处理,再将图像信息输入 Caffe 搭建的神经网络,实现动态手势和静态手势相结合的手势识别系统。

软件框架结构如下图:

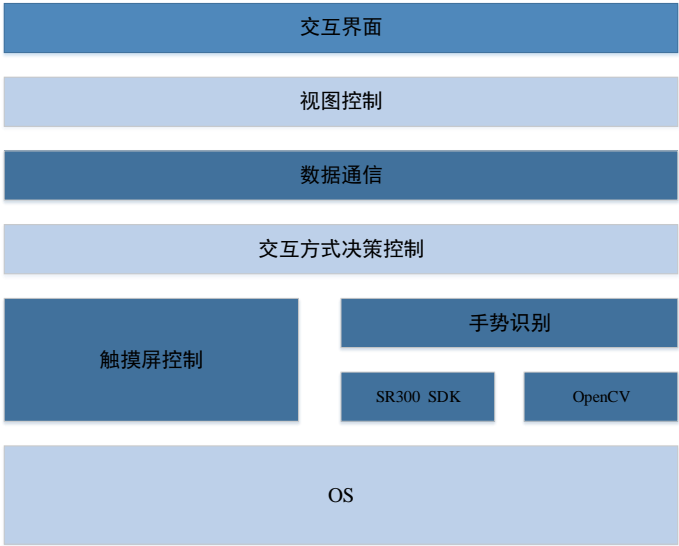


图 4-3 软件结构图

Caffe 是最流行的深度学习框架之一，在 2.5.1 节中有详细介绍。SR300 软件开发工具包是英特尔公司免费提供给开发者工具包（SDK），通过 SR300 SDK 可





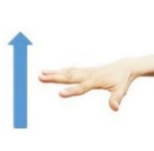
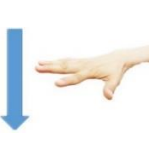
以获取彩色图像信息和深度信息。OpenCV 是英特尔发起并参与开发的开源计算机视觉库，OpenCV 库提供了丰富的图像处理函数，能够大大加快开发进度。

4.4 车载手势识别系统实现

4.4.1 功能模块设计与手势定义

本文设计实现了针对汽车内最常用的四个应用（收音机、电话、音乐播放、导航）的手势操控。当驾驶员进入汽车内，打开该系统，系统进入主界面，如果在 SR300 感应区域做出相应的动态或静态手势操作就可以实现相应功能模块的操作。本文设计了 5 个模块来实现对收音机、电话、音乐播放、导航这四个应用的操作和不同应用的切换。为了简化操作，获得更好的用户体验，我们将复用手势功能，最终选取了识别率最高且符合操作习惯的 6 个手势作为本系统的操控手势，具体如下表：

表 4-2 手势示意图





编号	1	2	3	4	5	6
示意图						

1、主界面模块：主界面模块是启动的首个模块，拥有收音机、电话、音乐播放、导航这四个应用的入口。主界面主要功能是实现不同应用的入口选择和不同应用切换中的一个过渡界面。主界面模块实现应用选择操作，本文设计了手横移动作来实现，为了避免干扰，本文选择三个手指并拢的操作方式。手向左横移界面向左切换，手向右横移界面向右切换。本文定义了一个最常用的“V”字形静态手势作为确定功能手势。当选择中了某个应用，在感应区域做出一个“V”字形静态手势，进入相应的应用功能界面。当处在其他应用模块，需要返回主界面，本文定义了一个返回主界面的动作，在感应区域做出“C 型”静态手势，实现返回主界面的功能。



图 4-4 主界面

表 4-3 主界面模块手势功能表

功能模块	手势示意图	静态/动态	说明
主界面		动态	将左边应用入口移动到中间
		动态	将右边应用入口移动到中间
		静态	进入选定的应用
		静态	从其他模块返回主界面

2、电话模块：电话功能要求实现电话的接听与呼叫，电话接听分为两个操作：接听和挂断。电话的呼叫功能涉及到三个操作：联系人选择、呼叫操作和挂断操作。实际使用电话过程中可以分为两种场景，一种正处在通话过程中，一种选择联系人准备呼叫。由于场景不一样，本文可以对同一个手势赋予不同的功能，简化手势操作。本文综合电话接听和呼叫这两个功能的操作，本文定义了三个手势是实现电话接听与呼叫功能。在本原型系统中定义手向右横移为接听，向左横移为挂断，“V”字型静态手势为呼叫选中的联系人，手向左向右横移进行联系人的选择。



图 4-5 电话模块界面

表 4-4 电话模块手势功能表





功能模块	手势示意图	静态/动态	说明
电话模块		动态	向左移动通讯录
		动态	向右移动通讯录
		静态	呼叫选中的联系人
		动态	接听来电
		动态	拒接来电或结束通话

3、收音机模块：收音机只有两个操作：电台选择，音量调节。电台的选择本原型系统定义手的横移来选择电台，对于音量的调节，为了更加符合用户的操作习惯和干扰的去除，首先识别一个五指张开掌心朝下的静态手势，激活音量调节功能。激活音量调节功能之后根据手上下移动的距离实现音量大小的调控



图 4-6 收音机模块界面

表 4-5 收音机模块手势功能表

功能模块	手势示意图	静态/动态	说明
收音机模块		动态	向左切换收听的电台
		动态	向右切换收听的电台
		动态	手放入感应区 激活音量调整， 增大音量
		动态	手放入感应区 激活音量调整， 减小音量







4、音乐播放模块：音乐播放的操控由切换歌曲，音量调节，开始播放，暂停播放这四个操作组成，所以需要定义四种手势，由于音乐播放器的操作和收音机模块的操作类似，从简化操作的角度考虑，我们共用收音机的几个手势。手的横移操作定义为切换歌曲功能，音量的增加和收音机的一样。本文定义“V”字静态手势实现开始播放和暂停操作。“V”字静态手势播放或者暂停功能是根据当前播放器的状态来确定，当音乐播放器处于播放状态，检测到“V”字静态手势就是暂停操作，如果音乐播放器处在暂停状态，则为播放操作。





图 4-7 音乐播放模块界面

表 4-6 音乐播放模块手势功能表

功能模块	手势示意图	静态/动态	说明
音乐播放模块		动态	向左移动播放列表
		动态	向右移动播放列表
		静态	暂停正在播放的歌曲
		静态	开始播放当前选中的歌曲
		动态	手放入感应区 激活音量调整， 增大音量
		动态	手放入感应区 激活音量调整， 减小音量

5、导航模块：导航功能由于涉及到目的地设定，本文只设定了一个确定功能的手势在设定目的地后启动导航。本文定义了一个“V”字形的静态手势来启动导航。



图 4-8 导航模块界面

表 4-7 导航模块手势功能表

功能模块		手势示意图	静态/动态	说明
导航模块	开始导航		静态	进入导航状态

4. 4. 2 软件功能实现

因为存在多个功能之间的切换，所以需要为各个功能模块定义优先级来避免功能切换出现混乱。各模块优先级定义如下：电话模块优先级为 1 级，导航模块第 2 级，收音机模块、音乐播放器模块为第 3 级。之所以设置优先级是从用户习惯和用户体验的角度考虑设置的优先级，1 为最高级，3 为最低级。当低优先级的模块在运行时，高优先级的如果发生请求，则停止低优先级的模块，跳转到高优先级的模块运行，不能同时运行两个优先级一样的模块。主界面模块是这些模块的入口和功能切换的过渡界面，所以主界面不受优先级的限制。下面是软件设计流程图：

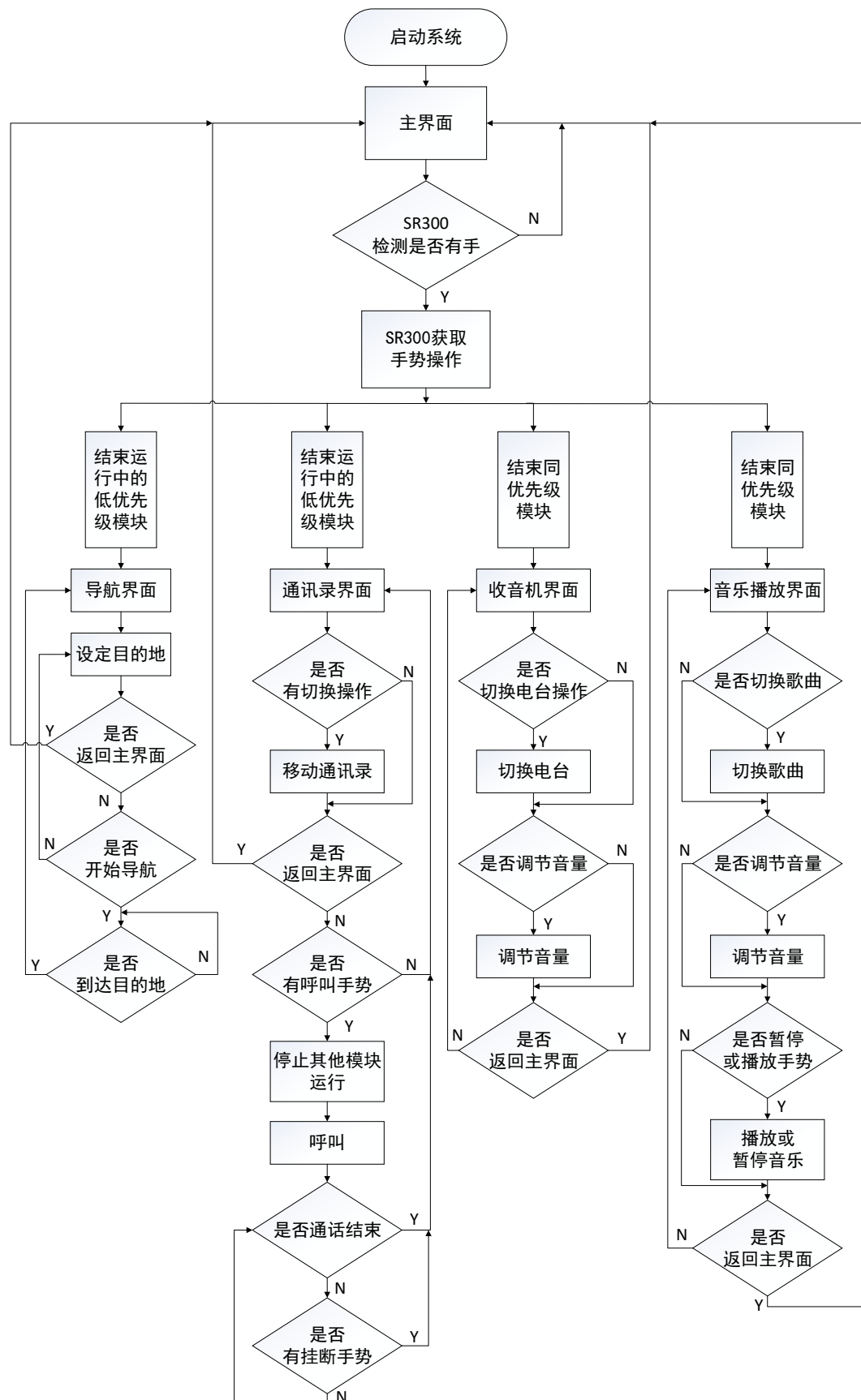


图 4-9 原型系统软件流程图

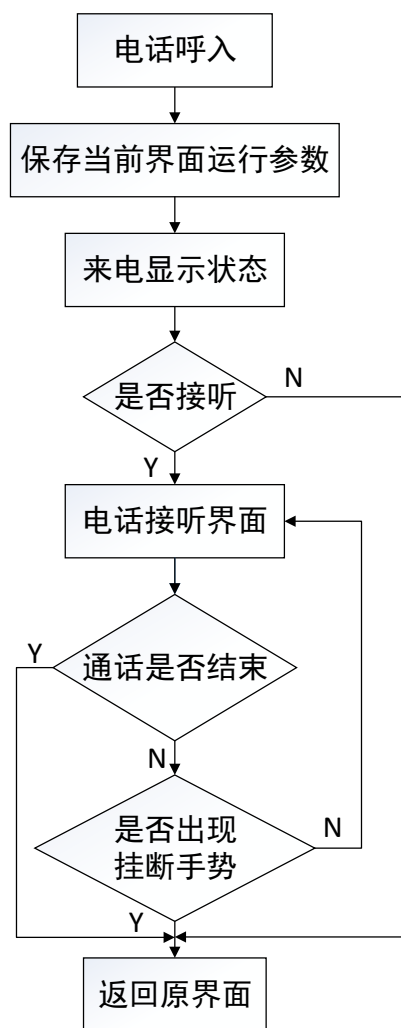


图 4-10 电话呼入软件流程图

在本系统软件流程图如上图示。在本系统中的手势识别部分将定义好的每一个手势封装成了函数，当判断是否进行操作时，首先检测感应区域是否有手，如果有手则通过 SR300 彩色摄像头获取图像信息进行手势识别。在实际开发过程中，通过创建子线程进行手势识别的运算，这样能够加快 UI 线程的刷新速度，使界面保持流畅。电话的呼入是个随机事件，我们通过设定一个子线程对电话状态进行实时的监控。当电话呼入时需要先将当前用户界面状态保存，当通话结束时，恢复原界面。在通话过程中一直保持在通话界面，而不能切换到其他界面，由于存在手势复用，这样能保证通话不受干扰。

## 4.5 实验与分析

为了验证本章设计的原型系统的性能，本文在真实的环境下进行了实验验证。本章设计原型系统是建立在前两章的基础上，本章原型系统的神经网络采用的是第二章和第三章训练好的网络。本节对车载手势识别原型系统的 5 个模块（主界

面模块、电话模块、导航模块、音乐播放模块、收音机模块) 的操控分别进行了测试。我们选取了 32 个人, 分别在驾驶员位置和副驾驶位置做了 50 次实验。



图 4-11 现场实验图

本系统中的手势选择的是前两章中实验中识别率最好的且符合人机交互习惯的几个手势作为车载手势交互系统的操作手势。我们首相通过 SR300 提供的 SDK, 检测在 SR300 感应区是否有手的存在; 如果有人手存在, 通过运动检测, 判断手是否处在运动中, 如果手正在运动, 通过 SR300 彩色摄像头记录手的运动情况。正常人完成一个手势时间大约在 0.5s, 本文原型系统采用 60fps 的帧率大约捕获 30 帧图像。根据第三章设计动态神经网络结构, 本文对于动态手势记录 30 帧图像视为一个动态手势, 在完成一个动态手势的记录后 0.5s 后再进行手势检测。获取手运动视频后输入到动态手势 3D 卷积神经网络中, 进行识别。当手没有运动时, 我们将手的图像信息输入到静态手势识别网络中。与其他手势不同的是音量的调节手势, 为了防止干扰, 在进行音量调节时, 需要先张开手掌, 识别五指张开的静态手势, 激活音量调节, 根据手上下移动距离调节音量大小。

表 4-8 主模块手势识别率

功能模块		识别率
主界面模块	切换到上一个应用	83.2%
	切换到下一个应用	83.4%
	进入应用	96.4%
	返回主界面	97.8%

表 4-9 电话模块手势识别率

功能模块		识别率
电话模块	切换到上一条通讯录	83.6%
	切换到下一条通讯录	83.2%
	呼叫	96.3%
	接听	84.2%
	挂断	83.8%

表 4-10 收音机模块手势识别率

功能模块		识别率
收音机模块	向左切换电台	83.2%
	向右切换电台	82.8%
	音量增加	98.1%
	音量减小	98.2%

表 4-11 导航模块手势识别率

功能模块		识别率
导航模块	开始导航	96.7%

表 4-12 音乐播放模块手势识别率

功能模块		识别率
音乐播放模块	上一曲	83.2%
	下一曲	83.4%
	暂停	96.5%
	开始	96.4%
	音量增加	97.1%
	音量增加	97.1%

从表 4-8、表 4-9、表 4-10、表 4-11 和表 4-12 各模块的手势识别率可以看出，静态手势识别率相对比动态手势识别率高。动态手势低于静态手势识别率原因有 2 点：第一、一次静态手势可以捕获多帧关于手的图像信息，进行多次识别，结合多次识别结果做出判断，从而使得静态手势识别率较高，而动态手势则一次只能检测一个。本文设计的原型系统对于静态手势的识别通过获取 3 张手彩色图像，把识别结果相同的 2 张图片结果的作为最后的输出；如果 3 张图片识别结果都不一样则重新获取 3 张手彩色图像进行识别。第二、由于不同人做动作的时间存在差异，动态手势的视频可能不完整，而静态手势不存在手势图像信息不完整的情况，这也使得动态手势的识别率要低于静态手势。音量调节手势识别率教高，这是因为音量调节手势实质上是获取手的图像，然后通过静态手势识别系统判断手是否是五指张开的手势，然后通过 SR300 获取手的上下位移距离，然后将位移量转化为音量调节的量，所以音量调节的手势识别率较高。

## 4.6 本章小结

本章运用第二章和第三章的静态和动态手势识别技术实现了车载手势识别原型系统的开发。首先阐述了车载手势识别的意义；接着介绍了 SR300 传感器，并详细阐述了车载手势识别系统的硬件框架和软件框架设计；然后详细介绍了车载手势识别各模块设计过程和软件流程图；最后选取了 32 个人，对车载手势识别系统各模块进行了测试。

## 5 总结与展望

### 5.1 总结

深度学习近年在诸多领域都取得了很多成果，尤其在计算机视觉方面，深度学习可以说开启了一个全新的时代。本文通过对深度卷积神经网络的研究，实现了基于视觉的静态手势识别和动态手势识别，并完成了车载手势识别原型系统的设计与实现。具体工作总结如下：

1. 针对静态手势识别，本文采用 **Sebastien Marcel** 手势数据集，并对手势集进行了扩充，使得样本达到 10000。根据静态手势识别需要精细特征识别的特点，结合 **Songfan Yang** 等人提出的多尺度特征思想，在 **CaffeNet** 网络基础上设计了适合静态手势识别的多尺度卷积神经网络，通过与单尺度卷积网络实验对比发现，基于多尺度卷积神经网络的静态手势识别方法识别率明显高于基于单尺度卷积神经网络方法。通过与传统方法的对比，分析了多尺度卷积神经网络用于静态手势识别的优缺点。

2. 针对动态手势识别，本文分析了基于动态时间规整，基于隐马尔科夫模型，基于压缩时间轴的动态手势识别方法优缺点。接着分析了 CNN 网络 3 种不同时空信息的融合方式，并从实验证明了 3 种时空信息融合方式在动态手势识别方面性能，分析了 2D 卷积神经网络与 3D 卷积神经网络的区别，最后设计出基于缓慢融合的 3D 卷积神经网络进行动态手势识别。本文的动态手势识别的样本采用的是 **VIVA** 动态手势数据集，由于 **VIVA** 动态手势集中样本数量不足，只有 2920 个，采用卷积神经网络容易出现过拟合现象，所以本文从数据量和网络结构优化两个方面来避免过拟合。本文通过水平镜像，倒序，裁剪等操作生成虚拟样本，对数据集进行扩充。通过在网络中加入 **Dropout** 层，减少过拟合的出现。卷积神经网络对于视频的处理计算量非常大，为了加快代价函数的收敛，本文引入 **Nesterov** 加速梯度法使神经网络更快的找到极小值点。通过与基于 **HOG** 特征和基于双卷积神经网络的动态手势识别方法实验对比发现，基于 3D 卷积神经网络的动态手势识别方法速度最快，实时性能好，识别率较好。

3. 设计实现了车载手势识别原型系统。本文结合第二章和第三章的手势识别技术，设计并实现了车载手势识别系统。在该系统中设计了 5 个模块，用于实现汽车内部常用的 4 个应用（收音机、电话、音乐播放、导航）。本文详细阐述了系统软硬件架构设计，各个模块的功能定义以及操作各模块的手势设计，并分析了软件实现流程图。最后在真实环境下测试了各功能模块的手势操控识别率，



并分析了动态手势与静态手势识别率的差异。

## 5.2 展望

虽然手势识别研究已经进行了几十年，并且取得了不小的进展，出现了大量的研究成果，但是手势识别技术并不成熟。从公布的公共手势数据库来看，与人脸识别数据库相比，手势库整体数据量偏小，背景环境单一。从目前公布的手势研究成果来看，手势识别率都比较高，但是往往是在光照变化小，背景简单的环境下进行测试得出的结论。由于手势本身存在着很大的个体差异性，如老年人做挥手的动作速度往往比年轻人要慢，所以如何克服个体差异也是手势识别的一个难点。本文中提出的静态手势识别和动态手势识别方法都通过实验验证取得了较好的结果，并实现了车载手势识别系统，但是还存在很大的改进空间。

1. 本文用于卷积神经网络的训练数据集都采用的是公开数据集，虽然对于 Sebastien Marcel 手势数据集进行了扩充，但是这样多的样本集数量还是较少。当设计层数更多的更复杂的深度学习网络，这样量级的数据集还不能满足要求，所以对手势数据库的扩充对于手势识别的研究非常有意义。

2. 随着 AI 技术的飞速发展，汽车领域也即将掀起一轮变革。当前各大汽车厂商如宝马，奔驰等都在进行汽车内新的交互方式的研究；美国科技巨头 Google 推出的无人驾驶汽车，甚至连方向盘都去掉了。汽车在未来将作为一种移动平台而存在，而不是单纯的交通工具。随着汽车功能定位的改变，汽车内的人机交互方式必将改变，而手势作为一种非常友好的人机交互方式很有可能运用到未来汽车上。本文设计了一种车载手势识别系统，但是功能相对还比较单一，手势识别的稳定性不够，所以还需丰富该系统的功能以及提高手势识别率。

## 参考文献

- [1] Pisharady P K, Saerbeck M. Recent methods and databases in vision-based hand gesture recognition: A review[J]. Computer Vision and Image Understanding, 2015, 141(2015): 152-165.
- [2] Kelly S D, Manning S M, Rodak S. Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education[J]. Language and Linguistics Compass, 2008, 2(4): 569-588.
- [3] Berman S, Stern H. Sensors for gesture recognition systems[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, 42(3): 277-290.
- [4] Rautaray S S, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey[J]. Artificial Intelligence Review, 2015, 43(1): 1-54.
- [5] Mohandes M, Liu J, Deriche M. A survey of image-based arabic sign language recognition[C]. Systems, Signals & Devices (SSD), 2014 11th International Multi-Conference on. IEEE, 2014: 1-4.
- [6] Hasan H, Abdul-Kareem S. Fingerprint image enhancement and recognition algorithms: a survey[J]. Neural Computing and Applications, 2013, 23(6): 1605-1610.
- [7] Karam M. A framework for research and design of gesture-based human computer interactions[D]. University of Southampton, 2006.
- [8] Hasan H, Abdul-Kareem S. Human-computer interaction using vision-based hand gesture recognition systems: a survey[J]. Neural Computing and Applications, 2014, 25(2): 251-261.
- [9] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
- [10] Schmidhuber J. Deep learning in neural networks: An overview[J]. Neural networks, 2015, 61(78): 85-117.
- [11] Deng L, Yu D. Deep learning: methods and applications[J]. Foundations and Trends® in Signal Processing, 2014, 7(3-4): 197-387.
- [12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing

- systems. 2012: 1097-1105.
- [13] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
- [14] Mitra S, Acharya T. Gesture recognition: A survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007, 37(3): 311-324.
- [15] Suarez J, Murphy R R. Hand gesture recognition with depth images: A review[C]. RO-MAN, 2012 IEEE, 2012: 411-417.
- [16] Khan R Z, Ibraheem N A. Hand gesture recognition: a literature review[J]. International journal of artificial Intelligence & Applications, 2012, 3(4): 161.
- [17] Zhang Z. Microsoft kinect sensor and its effect[J]. IEEE multimedia, 2012, 19(2): 4-10.
- [18] Ge S S, Yang Y, Lee T H. Hand gesture recognition and tracking based on distributed locally linear embedding[J]. Image and Vision Computing, 2008, 26(12): 1607-1620.
- [19] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.
- [20] Teng X, Wu B, Yu W, et al. A hand gesture recognition system based on local linear embedding[J]. Journal of Visual Languages & Computing, 2005, 16(5): 442-454.
- [21] 吴杰. 基于深度学习的手势识别研究[D]. 电子科技大学, 2015.
- [22] Licsár A, Szirányi T. Dynamic training of hand gesture recognition system[C]. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004, 4: 971-974.
- [23] Licsár A, Szirányi T. User-adaptive hand gesture recognition system with interactive training[J]. Image and Vision Computing, 2005, 23(12): 1102-1114.
- [24] Pisharady P K, Vadakkepat P, Loh A P. Attention based detection and recognition of hand postures against complex backgrounds[J]. International Journal of Computer Vision, 2013, 101(3): 403-419.
- [25] Huang D Y, Hu W C, Chang S H. Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination[J]. Expert Systems with Applications, 2011, 38(5): 6031-6042.
- [26] Ueda E, Matsumoto Y, Imai M, et al. A hand-pose estimation for vision-based human interfaces[J]. IEEE Transactions on Industrial Electronics, 2003, 50(4):

- 676-684.
- [27] Yin X, Xie M. Estimation of the fundamental matrix from uncalibrated stereo hand images for 3D hand gesture recognition[J]. Pattern Recognition, 2013, 36(3): 567-584.
- [28] Chen F S, Fu C M, Huang C L. Hand gesture recognition using a real-time tracking method and hidden Markov models[J]. Image and vision computing, 2003, 21(8): 745-758.
- [29] Lee H K, Kim J H. An HMM-based threshold model approach for gesture recognition[J]. IEEE Transactions on pattern analysis and machine intelligence, 1999, 21(10): 961-973.
- [30] Marcel S, Bernier O, Viallet J E, et al. Hand gesture recognition using input-output hidden markov models[C]. Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000: 456-461.
- [31] Yoon H S, Soh J, Bae Y J, et al. Hand gesture recognition using combined features of location, angle and velocity[J]. Pattern recognition, 2001, 34(7): 1491-1501.
- [32] Ramamoorthy A, Vaswani N, Chaudhury S, et al. Recognition of dynamic hand gestures[J]. Pattern Recognition, 2003, 36(9): 2069-2081.
- [33] Yang M H, Ahuja N, Tabb M. Extraction of 2d motion trajectories and its application to hand gesture recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(8): 1061-1074.
- [34] Shen X, Hua G, Williams L, et al. Dynamic hand gesture recognition: An exemplar-based approach from motion divergence fields[J]. Image and Vision Computing, 2012, 30(3): 227-235.
- [35] Molchanov P, Gupta S, Kim K, et al. Hand gesture recognition with 3D convolutional neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 1-7.
- [36] Shin M C, Tsap L V, Goldgof D M B. Gesture recognition using Bezier curves for visualization navigation from registered 3-D data[J]. Pattern Recognition, 2004, 37(5): 1011-1024.
- [37] Alon J, Athitsos V, Yuan Q, et al. A unified framework for gesture recognition and spatiotemporal gesture segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(9): 1685-1699.

- [38] 邹洪. 实时动态手势识别关键技术研究[D].华南理工大学,2011.
- [39] Gallo L, Placitelli A P, Ciampi M. Controller-free exploration of medical image data: Experiencing the Kinect[C]. Computer-based medical systems (CBMS), 2011 24th international symposium on. IEEE, 2011: 1-6.
- [40] Marin G, Dominio F, Zanuttigh P. Hand gesture recognition with leap motion and kinect devices[C]. Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014: 1565-1569.
- [41] Hasan H, Abdul-Kareem S. Human-computer interaction using vision-based hand gesture recognition systems: a survey[J]. Neural Computing and Applications, 2014, 25(2): 251-261.
- [42] Weichert F, Bachmann D, Rudak B, et al. Analysis of the accuracy and robustness of the leap motion controller[J]. Sensors, 2013, 13(5): 6380-6393.
- [43] Khademi M, Mousavi Hondori H, McKenzie A, et al. Free-hand interaction with leap motion controller for stroke rehabilitation[C]. Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems. ACM, 2014: 1663-1668.
- [44] Draelos M, Qiu Q, Bronstein A, et al. Intel realsense= real low cost gaze[C]. Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015: 2520-2524.
- [45] House R, Lasso A, Harish V, et al. Evaluation of the Intel RealSense SR300 camera for image-guided interventions and application in vertebral level localization[C]. SPIE Medical Imaging. International Society for Optics and Photonics, 2017: 101352Z-101352Z-7.
- [46] Reuschenbach A, Wang M, Ganjineh T, et al. iDriver-Human Machine Interface for Autonomous Cars[C]. Information Technology: New Generations (ITNG), 2011 Eighth International Conference on. IEEE, 2011: 435-440.
- [47] Sébastien Marcel. Hand Posture and Gesture Datasets[DB/OL]. [2001-10-16]. <http://www-prima.inrialpes.fr/FGnet/data/10-Gesture/gestures/main.html>.
- [48] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1717-1724.
- [49] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.
- [50] Fredric M Ham,Ivica Kostanic.神经计算原理[M].叶世伟,王娟.译.北京:机械工

业出版社,2007.

- [51] Yang S, Ramanan D. Multi-scale recognition with DAG-CNNs[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 1215-1223.
- [52] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]. Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.
- [53] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014:79-82.
- [54] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[J]. arXiv preprint arXiv:1405.3531, 2014: 215-223.
- [55] 易靖国, 程江华, 库锡树. 视觉手势识别综述 [J]. 计算机科学, 2016, (S1): 103-108.
- [56] 王西颖, 戴国忠, 张习文, 等. 基于 HMM-FNN 模型的复杂动态手势识别 [J]. 软件学报, 2008, 19(9): 2302-2312.
- [57] 任海兵, 祝远新, 徐光, 等. 基于视觉手势识别的研究—综述[J]. 电子学报, 2000, 28(2): 118-121.
- [58] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.
- [59] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.
- [60] Wu Z, Wang X, Jiang Y G, et al. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification[C]. Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 461-470.
- [61] Fernando B, Gould S. Learning end-to-end video classification with rank-pooling[C]. Proc. of the International Conference on Machine Learning (ICML). 2016.
- [62] Ye H, Wu Z, Zhao R W, et al. Evaluating two-stream CNN for video classification[C]. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 2015: 435-442.
- [63] Chéron G, Laptev I, Schmid C. P-cnn: Pose-based cnn features for action recognition[C]. Proceedings of the IEEE International Conference on Computer

- Vision. 2015: 3218-3226.
- [64] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.
- [65] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]. Advances in neural information processing systems. 2014: 568-576.
- [66] Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
- [67] Livnat A, Papadimitriou C, Pippenger N, et al. Sex, mixability, and modularity[J]. Proceedings of the National Academy of Sciences, 2010, 107(4): 1452-1457.
- [68] Su W, Boyd S, Candes E. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights[C]. Advances in Neural Information Processing Systems. 2014: 2510-2518.
- [69] Giselsson P, Doan M D, Keviczky T, et al. Accelerated gradient methods and dual decomposition in distributed model predictive control[J]. Automatica, 2013, 49(3): 829-833.
- [70] Vision for Intelligent Vehicles and Applications (VIVA) Challenge [DB/OL]. [2012-02-01]. <http://cvrr.ucsd.edu/vivachallenge/>
- [71] 于旭,杨静,谢志强. 虚拟样本生成技术研究[J]. 计算机科学, 2011, (3): 16-19.
- [72] Niyogi P, Girosi F, Poggio T. Incorporating prior information in machine learning by creating virtual examples[J]. Proceedings of the IEEE, 1998, 86(11): 2196-2209.
- [73] Ohn-Bar E, Trivedi M M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations[J]. IEEE transactions on intelligent transportation systems, 2014, 15(6): 2368-2377.
- [74] Molchanov P, Gupta S, Kim K, et al. Multi-sensor system for driver's hand-gesture recognition[C]. Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. IEEE, 2015, 1: 1-8.
- [75] Riener A. Gestural interaction in vehicular applications[J]. Computer, 2012, 45(4): 42-47.
- [76] Saini P, Kaur P. Automatic speech recognition: A review[J]. International journal of Engineering Trends & Technology, 2013: 132-136.

- 
- [77] Laroche J, Murgia C. Noise suppression assisted automatic speech recognition: U.S. Patent 9,558,755[P]. 2017-1-31.
- [78] Oh S Y, Chung K Y. Target speech feature extraction using non-parametric correlation coefficient[J]. Cluster Computing, 2014, 17(3): 893-899.



## 攻读硕士学位期间主要研究成果

### 参与科研项目情况

- [1] 校企合作项目：华为车联网，2016.06-2016.12，主要研究人员
- [2] 校企合作项目：华为触觉互联网，2016.03-2017.03，主要研究人员
- [3] 校企合作项目：多模脑部图像的肿瘤特征分析，2017.03-至今，主要研究人员

### 已发表论文

- [1] 一款非接触式手势操控的车载音乐播放器[A]. UXPA 中国.User Friendly 2014 暨 UXPA 中国第十一届用户体验行业年会论文集[C].UXPA 中国:,2014:4.

### 硕士期间参加学术交流活动

- [1] 2016 年 12 月 中国云计算大会(CCCC2016)(湖南长沙)

## 致谢

时光飞逝，研究生生活即将结束，很快将告别学生身份，走上工作岗位。回头看这些年的学习生活，有太多的不舍。

首先最想感谢的是我的导师谢斌老师。记得刚刚考完研找谢老师的时候，给谢老师发了一条短信，没想到的是很快就得到了回复，很快约好见面时间。很幸运能够成为谢老师的学生，每当自己在学习生活中遇到困难，迷茫的时候，总能得到谢老师的指点和帮助。以前觉得谢老师对学生要求太过严格，但是当过完这三年才会慢慢理解老师严格要求的意义，才能理解老师的良苦用心。

其次能够在智能所学习，我深感荣幸，同时要特别向蔡自兴教授致谢，有幸接受过蔡老师的教导，他对学术严谨的态度以及对自己的严格要求让我时刻不能放松对自己的要求，他的言传身教让我铭记于心。衷心祝愿蔡老师健康长寿。

感谢尤作师兄，作哥是来中南大学认识的第一个师兄，刚来中南大学读研究生的时候自己编程能力等各方面还很差，也没有很好的适应研究生的生活，作哥一步一步帮我克服了这些困难。从作哥身上不仅仅学习到了知识，更重要的感受到了作哥做事的专注，作哥的勤奋。

感谢马超民博士对于论文写作的帮助和指导，感谢魏楠、李沁、宋迪和彭盛楠帮我修改论文。

此外还要感谢唐璘老师、肖晓明老师、余伶俐老师、王勇老师、李仪老师、陈白帆老师、刘丽珏老师、高琰老师、郭璠老师、谭平老师、邹逸群老师在我研究生期间给我的指导和帮助，谢谢各位老师。

很幸运能够生活在智能所这样一个大家庭，同届的赵庆会、林修明、杨贵徽、黄佳伟、尹大庆、魏文燕、雷晓亮、周博茂，和你们一起度过的时光将是我最好的记忆，希望大家将来越来越好，毕业后能够常联系。还要感谢我的同门小师弟小师妹，李沁、梁照栋、宋迪、何小宇，你们这些新鲜血液的加入让“谢门”生气勃勃。还要感谢李凡、刘方、梁春芳、王琦、刘春发、龙子威、李路、罗舒宁、莫斯尧、顾磊等师弟师妹，认识你们很开心，祝你们学业进步，生活开心。

最后还要特别感谢的父母和家人，谢谢你们这么多年的养育。没有你们的无私奉献，就没有今天的我。不论遇到什么困难，我总相信只要有你们在，我不会孤单，无论什么困难我们都能一起克服。

2017年4月12日于  
中南大学升华楼