中国科学院大学

University of Chinese Academy of Sciences

学士学位论文

多任务学习框架下注意力驱动的孤立手语识别

作者姓名：谈清扬

指导教师：陈熙霖 研究员 中国科学院计算技术研究所

柴秀娟 副研究员 中国科学院计算技术研究所

学位类别：工学学士

专　　业：计算机科学与技术

学院（系）：计算机与控制学院

2018 年 6 月

Attention-based Isolated Gesture Recognition with Multi-task Learning

A thesis submitted to the
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Bachelor of Engineering
in Computer Science and Technology

By

Tan Qingyang

Supervisors: Prof. Chen Xilin

Prof. Chai Xiujuan

School of Computer and Control Engineering, University of Chinese Academy of Sciences

June, 2018

中国科学院大学
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的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

手语是聋哑人群体内以及聋哑人与外界交流的主要途径，自动手语识别系统能够搭建聋哑人与健听人沟通的桥梁，改善聋哑人的生活质量，具有重要的科学研究价值和社会意义。随着深度学习技术的广泛研究，卷积神经网络等被引入这一领域并使得手语识别的性能有了显著的提高。已有的深度学习手语识别框架，普遍需要先行检测定位出手、脸等关键区域，减少背景和身体其他区域对识别的影响。然而，此类检测-识别分段进行的方法，需要进行独立的调参，和更多的预处理步骤，相对困难，且无法将区域识别和手语分类的特征信息进行有效共享。因此，本文提出在同一个深度学习网络架构内，将手部等关键时空区域的判别和手语分类相结合，利用注意力机制对有效特征进行关注，实现端到端的自动手语识别，有效提高手语识别的效率，并保持一定的识别精度。

关键词：手语识别，多任务学习，注意力机制，伪三维残差卷积网络，ChaLearn数据集

Abstract

For deaf-mute people, sign language is the main way to communicate inside their group and with the outside world. Automatic sign language recognition system can bridge the communication between deaf-mute people and hearing people, improve the quality of life of deaf-mute people, and has important scientific research value and social impact. With the extensive study of deep learning technology, convolutional neural networks were introduced into this field and the performance of sign language recognition has been significantly improved. Existing deep learning sign language recognition frameworks generally need to detect and locate key areas such as hand and face, and reduce the influence of background and other body regions on recognition. However, this kind of methods which break down detection and recognition into several modules, requires independent training for different modules and more preprocessing steps, is relatively difficult, and cannot effectively share the feature information of region recognition and sign language classification. Therefore, we propose to combine the identification of key temporal and spatial regions and sign language classification into one deep learning network architecture, using attention mechanisms to focus on effective features, and realizing end-to-end automatic sign language recognition. Our proposed architecture can enhance the efficiency of sign language recognition and maintain a comparable recognition accuracy with state-of-art work.

Keywords: Sign Language, Multi-task Learning, Attetion Mechanism, P3D, ChaLearn Dataset

目 录

图形列表

表格列表

第 1 章　引言

1.1　研究背景及意义

手势识别由于其在手语识别翻译、人机交互、机器人控制与虚拟现实中的广泛应用，近年来在机器视觉领域得到了广泛的关注。其中，手语是聋哑人群体内及聋哑人与外界交流的主要途径；而研究开发自动手语识别系统，能够搭建聋哑人与健听人沟通的桥梁，改善聋哑人的生活质量，具有重要的科学研究价值和社会意义。

手势与手语的识别涉及对人体动作的检测和理解，在实际应用中由于数据和任务的不同，面临着种种挑战。例如，手势手语数据会受用户肤色、服装、环境的背景、光照等因素的影响，造成识别成功率下降；同时，识别数据常为动态视频，用户不同的运动速度、手型、运动轨迹等会影响识别系统的泛化能力。除此之外，部分手势，在不同的情境下会有截然不同的意思，例如 “V” 字手势可能表达 “胜利” 之义，在有些情况之下也会表达数字 “2”。本文主要关注如何设计端到端的深度网络识别系统，实现对孤立手语词汇的有效识别。

手势识别系统的设计也伴随着数据采集设备的发展而进步。早期，研究者利用数据手套采集手形和位置的相关数据，并设计相关识别系统 (Wang et al., 2002; Kong et al., 2008)。然而，此类设备十分昂贵，且不便于使用，限制了手语识别系统在日常生活场景中的应用。所以，有些研究者利用普通摄像设备，简化手语数据的采集流程 (Wang et al., 2010)。但 RGB 视频数据极易受到光照和背景的影响，使得手的检测和连续手语视频的分割非常困难。随着微软 Kinect(Zhang, 2012) 等设备的引入，深度信息可以和颜色信息同时进行采集，给识别系统提供了颜色之外的距离信息，使设计便捷高效的手势识别系统成为可能。所以，更多的研究者也把关注点转向了如何提升在 RGB-D 手语数据上的识别性能，许多数据集也公开发布，包括本文实验采用的 ChaLearn LAP Large-scale Isolated Gesture Dataset（IsoGD）(Wan et al., 2016)。

1.2　国内外研究现状

手势与手语识别相关研究已经在学界开展数年。早期，研究者利用手工设计的特征来进行识别算法的设计，包括帧级特征和时空性特征。利用帧级特征的识别算法需要引入描述序列数据的模型，例如隐马尔科夫模型（HMM）(Malgireddy

et al., 2013; Yamato et al., 1992)，条件随机场（CRF）(Wang et al., 2006) 和动态时间规整算法（DTW）(Corradini, 2001) 等。与帧级特征不同，时空性特征能够描述手势的动态变化，能够更有效地处理视频数据。此类研究大部分从时间维度拓展了原有的二维特征，包括 3D Harris corner detector (Laptev, 2005)，3D scale-invariant feature transform（SIFT）(Scovanner et al., 2007)，speeded-up robust features（SURF）(Willems et al., 2008)，3D HOG (Klaser et al., 2008) 等。

ChaLearn LAP IsoGD (Wan et al., 2016) 等手势手语大型数据集的出现，为深度学习技术在手势手语识别中的应用提供了可能。卷积神经网络（CNN）能够自主地从图片或视频中提取高层次语义特征，吸引了大量研究者的关注。而此类算法需要大量数据的支撑，足够的数据才能训练得到有一定泛化能力和鲁棒性的模型。ChaLearn LAP 数据集从原有 ChaLearn Gesture Dataset（CGD）数据集 (Guyon et al., 2013) 整理得到，合并统一了一些具有很强相似性的手势类别，并手工标注划分了原有数据，得到了具有 249 类，总数接近五万样例的 IsoGD 和 ConGD 数据集，为手势手语算法的设计和评价提供了资源。

近两年，基于深度学习的手势手语识别算法随着数据的发展也有了一定的进展。静态数据上，Nagi et al. (2011) 将最大池化训练用于人机交互的手势识别系统；动态数据上，从Karpathy et al. (2014) 将深度学习引入视频分类开始，逐渐产生了基于 RGB 信息和光流的双流识别网络 (Simonyan et al., 2014)，以及引入时间维度进行三维卷积的 C3D (Tran et al., 2015) 等，研究者逐渐有了能够提取视频数据深度特征的工具，并在 ChaLearn LAP IsoGD 和 ConGD 数据上进行了有效的尝试 (Liu et al., 2017; Miao et al., 2017)。

1.3 本文研究内容及主要贡献

本文设计了多任务学习的深度学习网络架构，将手部等关键时空区域的判别和手语分类相结合，通过注意力机制使手脸检测的学习结果辅助提升手语识别的精度。通过此架构，可以减少对输入数据的预处理步骤，提升识别效率。也能够将不同任务的特征信息有效共享，使学习到的特征更紧致且更有益于目标任务的学习，最终提升手语识别分类性能。

本文探索了如何在视频分类卷积神经网络有效引入注意力机制。现有的卷积网络注意力工作多集中于二维数据 (Wang et al., 2017; Fu et al., 2017)，即图片分类，仅关注空间维度。本文利用伪 3D 残差卷积网络空间 (Qiu et al., 2017) 维度和时间维度分离的特点，同时进行特征注意力遮罩分析，利用了视频输入的时序

性特点，提升了性能。

在多任务学习方面，本文亦探索了在视频识别三维卷积网络中引入检测位置的模式，包括先学习位置后学习手语分类的串行模式，和同时学习的并行模式。对这两种模式进行了性能评估，分析了精度差异问题的可能成因。

结合多任务学习框架和注意力机制，本文所提模型在 ChaLearn IsoGD (Wan et al., 2016) 上达到了与当前最高水平可比的性能。

1.4 本文组织结构

为使读者对研究背景和设计思路有充分了解，引言后本文将按照相关研究工作，框架整体及分模块介绍，实验分析结果及结束语依次展开说明。

第 2 章　相关工作

除手势手语识别相关研究以外，本课题也涉及了一般类三维卷积视频识别网络，物体检测和语义分割，基于卷积网络的注意力机制。本章将对这些领域的相关工作进行介绍。

2.1　三维卷积视频识别网络

视频数据可视为多帧连续变化的二维图像数据的组合，设计识别视频数据的卷积神经网络，即需要设计恰当的能处理时间轴连续信息的模型。三维卷积神经网络 (C3D) 形式上与二维卷积相仿，但加上了时间维度，使其卷积核天然的具有分析时间-空间信息的能力，已被证明对于视频动作类输入具有很好的识别能力 (Tran et al., 2015)。然而，由于其利用三维卷积，相比于二维卷积参数更多，计算量更大，利用现有视频数据集难以训练。同时早期方法直接从随机初始化参数开始训练，无法利用在更大型的二维图像数据集上的预训练结果，使得整体网络层数较浅，具性能瓶颈。

近年来，研究者针对这些弱点，提出了各类方法来提高性能。Qiu et al. (2017) 提出伪三维残差卷积网络 (P3D)，将三维卷积拆分成在时域和空域的二维卷积和一维卷积，有效降低网络整体浮点运算量，同时可利用图片识别领域现有网络进行精细优化，使识别性能有了巨大提升。P3D 主要构建模块如图2.1所示。

同时期，Carreira et al. (2017) 提出双流扩大三维卷积网络 (i3D)，逆向思维，将图像看做连续多帧的“假视频”，将 $N \times N$ 二维卷积向时间轴扩大 N 帧，形成 $N \times N \times N$ 三维卷积，并将参数核整体除以 n，得以使三维卷积网络利用二维卷积预训练参数结果，还优化调整了时间轴池化设置，提高了性能。P3D 网络性能稍次于 i3D 网络，但由于 P3D 网络更加轻便易训练，本文选取了 P3D 网络作为基础识别骨架网络。

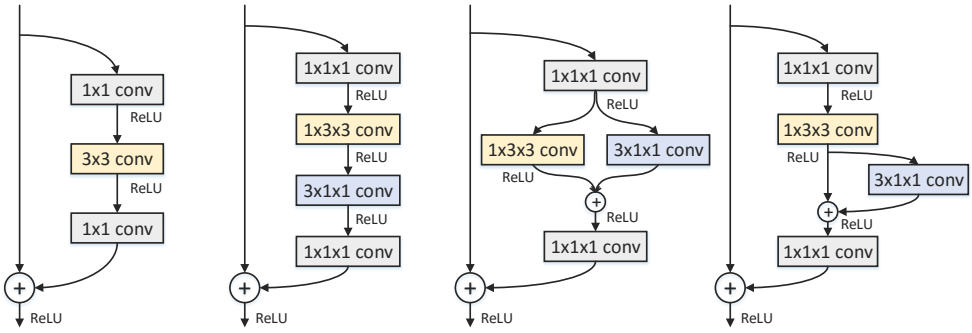


图 2.1 残差网络单元和伪三维卷积的瓶颈构建模块示意（自 (Qiu et al., 2017)）(a) 残差单元，(b) P3D-A，(c) P3D-B，(d) P3D-C。

Figure 2.1 Bottleneck building blocks of Residual Unit and Pseudo-3D. (From (Qiu et al., 2017)) (a) Residual Unit, (b) P3D-A, (c) P3D-B, (d) P3D-C.

2.2 基于卷积网络的注意力机制

深度神经网络已在各大图像数据集上证明了其提取高层语义、高判别力的优势，研究者现正更关注于如何更有效利用卷积特征，在细粒度划分的数据集上提高性能。有两类主要方法，如残差注意力模块 (Wang et al., 2017)，通过让网络自学习掩模，加上利用残差网络设计的经验，引入原特征跳跃链接，使识别性能有了很大的提升。该模块学习的掩模可视化效果如图片2.2所示。

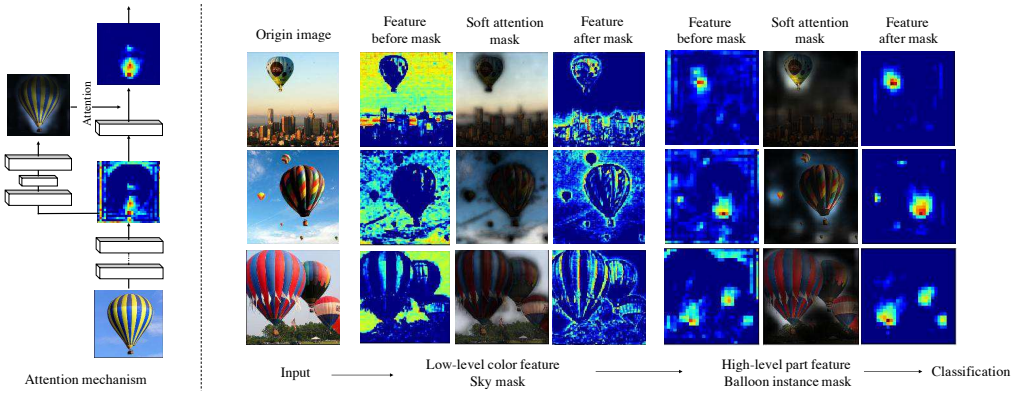


图 2.2 残差注意力模块掩模可视化效果（自 (Wang et al., 2017)）

Figure 2.2 Illustration of Mask Generated from Attention Residual Learning (From (Wang et al., 2017))

又如利用注意力提取网络 (APN) 的循环注意力卷积神经网络 (RA-CNN)(Fu et al., 2017)，主动识别有效区域，并利用双线性插值进行尺度调整，继而在放大

图片上再次识别，具体流程见图2.3。通过一般的分类损失和多尺度概率比较损失（Rank Loss) 迭代优化，该框架能够同时训练卷积神经网络参数和注意力提取网络参数，且对该方法优化梯度的分析证明了比较损失对于注意力参数优化的有效性。这一方法被证明在鸟类分类等关注细节的分类任务上具有很强大的性能。

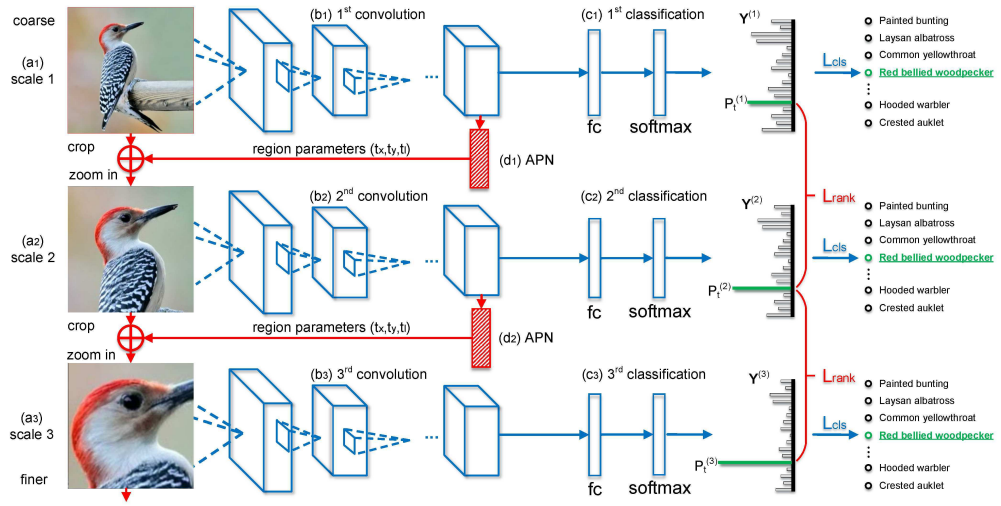


图 2.3 循环注意力卷积神经网络框架（自 (Fu et al., 2017))

Figure 2.3 Framework of RA-CNN (From (Fu et al., 2017))

手语识别之于一般的视频识别，相当于利用局部（手，脸，肘等）的特征，和关键帧（挥打重要手势）的信息，进行细粒度的分类。因此，本文提出要在手语识别网络中引入注意力机制，并同时利用多任务学习对注意力相关的参数进行优化。前述卷积网络上的注意力工作主要集中于二维图像识别领域，本文将注意力机制引入到视频分析领域，对时间维度的注意力提取进行了有效的尝试。

2.3 深度学习下的物体检测和语义分割

近年的手语识别工作，主要利用物体检测类手段对手部进行位置确定，并进行预处理 (Liu et al., 2017)。Girshick et al. (2014) 的工作区域卷积神经网络 (R-CNN) 可视为深度学习时代物体检测的开山之作。R-CNN 接受输入图像后，使用区域生成 (Region Proposal) 算法提取约 2000 个可行区域，变形到统一尺寸后，提取卷积特征，进行目标分类和包围框的回归。后续工作 Fast R-CNN(Girshick, 2015)，Faster R-CNN(Ren et al., 2015) 又提出区域池化 (RoI Pooling) 和区域生成网络 (Region Proposal Network) 加快了物体检测的速度，并提高了精度。但现阶段精度较高的深度学习物体检测方法仍需要区域生成和特征提取两步骤，复杂

且速度较慢，不适合与手语任务在同一框架下相结合进行多任务学习。本文仅利用物体检测方法生成手语视频手脸位置的训练数据。

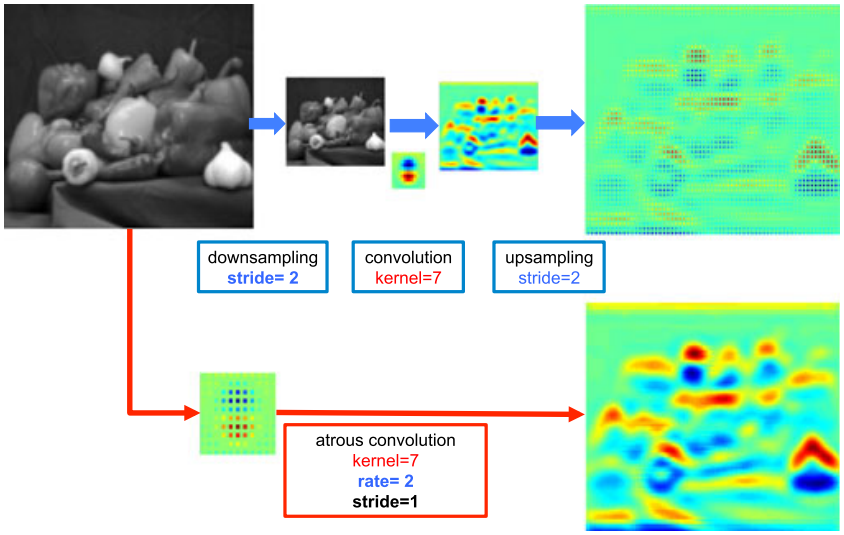


图 2.4 2 维空洞卷积示意（自 (Chen et al., 2018)）

Figure 2.4 Illustration of atrous convolution in 2-D. (From (Chen et al., 2018))

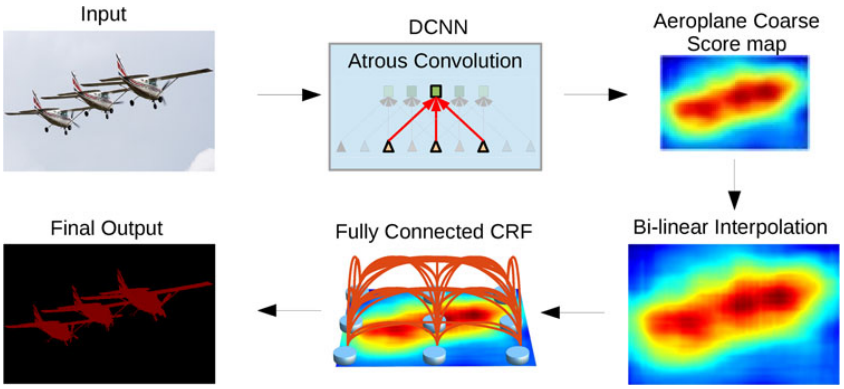


图 2.5 DeepLab 语义分割框架流程图（自 (Chen et al., 2018)）

Figure 2.5 Framework of DeepLab Semantic Segmentation Model (From (Chen et al., 2018))

在计算视觉领域，语义分割类任务与物体检测类任务具有一定相似性，都是对图像中各部分进行理解。但物体检测类任务主要用于检测并确定图像中主要物体位置，而语义分割强调对图片中逐像素的类别判断。从方法上来说，即是对各个像素进行物体分类。Chen et al. (2018) 的 DeepLab 系列工作在这一任务上处于领先地位。此系列工作主要利用空洞卷积（见图2.4）提取深层的语义信息，再利用反卷积或双线性插值扩大到原图尺寸，并利用条件随机场进行修正，具体流程可见图片2.5。空洞卷积和深层次信息反卷积均有助于扩大卷积感受野，使得

计算分割的特征能够了解全图的信息。本文在进行多任务学习时，借鉴了语义分割的形式，降采样提取特征，并升采样扩大得到较大尺寸的物体分布概率图，以此学习手脸等重要部位的位置信息。

第 3 章　多任务学习框架下注意力驱动的孤立手语词识别框架

本章节将对手语识别框架的具体设计进行描述。图3.1展示了该框架的整体流程。本识别框架主要包括三个主要部分。首先，我们对数据进行必要但简单的预处理步骤，即将 RGB 通道和深度通道的视频进行对齐。然后，我们利用结合注意力机制与多任务学习的深度神经网络分别对 RGB 视频和深度视频进行特征提取。最后，将两个通道的视频特征进行融合并得到最后的手语识别结果。

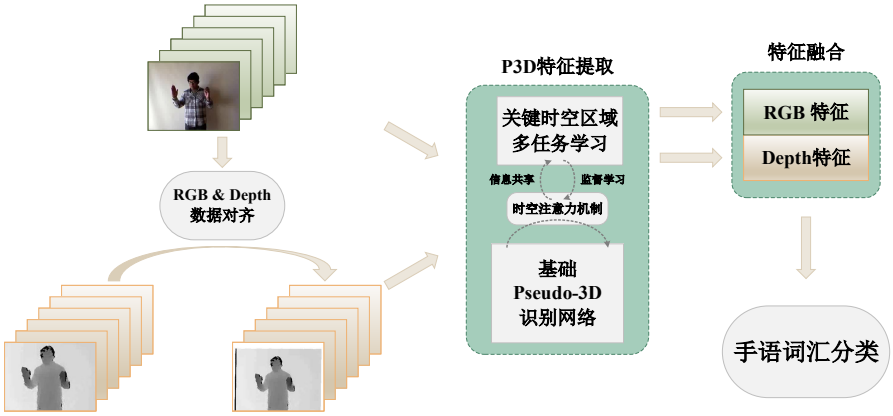


图 3.1 手语识别框架流程图

Figure 3.1 The pipeline of the proposed sign language recognition framework

在训练阶段，本框架仍需提取手和脸的位置信息作为训练数据。但测试时，由于位置的粗提取内涵在了多任务学习了之中，无需单独划分出一个步骤。章节4的实验数据证明了本方法在保证一定性能的情况下，有效减少了数据预处理步骤，减少了手语识别的整体耗时。

3.1　数据预处理

Liu et al. (2017) 指出，虽然 ChaLearn 数据集中手语视频的 RGB 和深度通道是同时被采集的，但由于相机参数等的细微不同，两个通道可能并没有完全对齐。所以，我们需要利用 RGB 通道和深度通道之间的对应关系，将深度通道的数据对齐到相应的 RGB 数据上。我们使用Bradski et al. (2008) 的相机校准技术来实现此步骤。图3.2给出了数据对齐的例子。

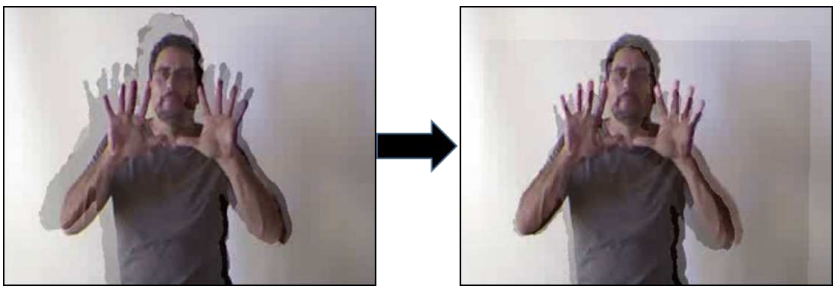


图 3.2 手语视频 RGB 通道和深度通道对齐实例（自 (Liu et al., 2017)）

Figure 3.2 An example of coordinate alignment between RGB and depth channels (From (Liu et al., 2017))

3.2 基础识别网络

章节2.1已提到，本文设计框架使用 P3D 网络 (Qiu et al., 2017) 作为基础识别网络。此节将对 P3D 网络的具体实现细节进行介绍。

(Qiu et al., 2017) 受到在图像识别领域领先的残差卷积网络 (ResNet)(He et al., 2016) 启发，通过具有残差单元的伪三维卷积瓶颈 (Bottleneck) 构建模块，来识别空间-时间特征。P3D 网络将一般同时作用在空间-时间维度的三维卷积，拆分成了空间维度的二维卷积，和时间维度的一维卷积，有效降低了计算复杂度和参数量，易于训练。通过叠加瓶颈模块，和池化操作，最终构建成了完整的 P3D 网络。

一般的残差单元包括了一个短路连接（输入直接做相等映射）和一个待学习的非线性函数 F（一般由卷积操作和非线性激活函数组成），可用下式进行描述：

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{F}(\mathbf{x}_t), \tag{3.1}$$

其中 $\mathbf{x}_t$ 和 $\mathbf{x}_{t+1}$ 分别代表 t 层残差单元的输入输出。如图2.1a所示，为了使残差单元提取重要信息，残差单元（黄色底色部分）内嵌在了瓶颈模块之中。瓶颈模块利用首尾 1×1 卷积操作进行降低以及回升维度的操作，这样残差单元的 3×3 卷积操作将作用在降维后的关键信息上，防止过拟合。

伪三维卷积模块主要针对视频输入，对残差单元进行修改。由于将统一同时进行的三维卷积拆分为空间和时间上两次卷积，需要考虑这两次卷积是不是要在结构上相互影响，以及如果有影响时两者间影响程度，所以 (Qiu et al., 2017) 提出了三种结构，以满足不同需求。

（1）P3D-A（图2.1b）：这种结构将时间卷积 (**T**) 接在了空间卷积 (**S**) 之后，认为时间卷积会直接受到空间卷积结果的影响，公式为：

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{T}(\mathbf{S}(\mathbf{x}_t)). \tag{3.2}$$

（2）P3D-B（图2.1c）：这种结构时间卷积和空间卷积平行放置，认为两者之间没有直接的互相影响，相加之后共同通向下一层，公式为：

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{T}(\mathbf{x}_t) + \mathbf{S}(\mathbf{x}_t). \tag{3.3}$$

（3）P3D-C（图2.1d）：这种结构是上述两者结构的结合，考虑空间卷积对时间卷积的影响，但同时保留了空间卷积信息流向下一层的直接通路，公式为：

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{T}(\mathbf{S}(\mathbf{x}_t)) + \mathbf{S}(\mathbf{x}_t). \tag{3.4}$$

三种结构在后文将统一描述为

$$\mathbf{x}_{t+1} = \mathbf{P3D}_{Bottleneck}(\mathbf{T}(\mathbf{x}_t), \mathbf{S}(\mathbf{x}_t)), \tag{3.5}$$

其中 $\mathbf{P3D}_{Bottleneck}$ 为三种瓶颈结构之一。(Qiu et al., 2017) 实验证明了，三种结构性能相差不多，但如果在同一个网络中混合三种瓶颈构建模块，总正确率会有提升。本文按照这一思路构建了基础的手语词汇识别网络，具体层数等参数设置见章4。

3.3 时空注意力机制

手语识别主要利用手部的细节完成任务，例如手的运动轨迹，手形，手和脸的相对位置等。Liu et al. (2017) 显式地在预处理步骤利用双流 Faster R-CNN 提取了手部的位置，并通过对视频的预处理，保留视频中手部区域，置黑其它无关区域，减少了无关因素如背景，服装，身体躯干等对识别的影响。但物体检测方法需要较长的运行时间，这一预处理步骤会使分类时间增长。本文提出利用注意力机制，隐式地完成对输入视频中手脸等重要区域的关注，有助于网络选取重点关注区域进行识别。同时，由于注意力机制内嵌在了识别网络中，与卷积核参数同时训练，能够有效增强正确特征的优化，减少错误梯度的影响。

手语视频不同帧所包含的信息量不同。对于 ChaLearn 手势识别集合来说，大部分手语视频在开始或结束时，由于人还没有将手抬起或已经放下，有大量的冗余帧，对于手语识别没有价值。Liu et al. (2017) 的后续工作提出，利用手部检

测结果，定义手语动作开始结束手部高度阈值，剪裁前后冗余帧后，将视频送入网络进行识别，使输入更加紧致。但这一方法仍需要额外的数据预处理步骤，本文通过利用时间维度引入注意力机制解决冗余帧的问题。同时，部分手语视频分类关键为手形有动作变化的少数帧，此类模式不易用手部位置等信息形式化定义。本文利用注意力机制关注时间维度有效帧，希望通过自学习达到对这类关键帧的定位。

本文生成注意力信息主要采用语义分割和骨架关键点识别常见的漏斗结构(Hourglass Structure)。漏斗结构前部的向下网络结构能够生成低分辨率但具有强大语义信息的特征图，向上逆卷积结构能够恢复出密集的特征，使得能够进行逐像素的推测。本文在空间维度和时间维度均设置注意力分支生成，利用了 P3D 构建模块空间卷积 **S** 和时间卷积 **T** 分离的特点，将空间和时间注意力分别结合到 **S** 和 **T** 之上。空间和时间注意力分别在空间维度和时间维度进行降、升采样，以形成漏斗结构。Wang et al. (2017) 指出，如果将注意力分支与特征点乘，会导致性能下降，所以提出了残差注意力学习：

$$\mathbf{x}_{t+1} = (1 + \mathbf{M}(\mathbf{x}_t)) * \mathbf{F}'(\mathbf{x}_t), \tag{3.6}$$

其中 $\mathbf{F}'$ 为利用残差单元生成的特征，$\mathbf{M} \in [0, 1]$ 为学习到的注意力分支。残差注意力学习借鉴了残差单元相等映射连接的形式，利于在深层次网络中进行注意力的学习，性能理论上不会弱于没有注意力的网络（$\mathbf{M} = 0$）。本文亦采取了残差注意力的连接形式。

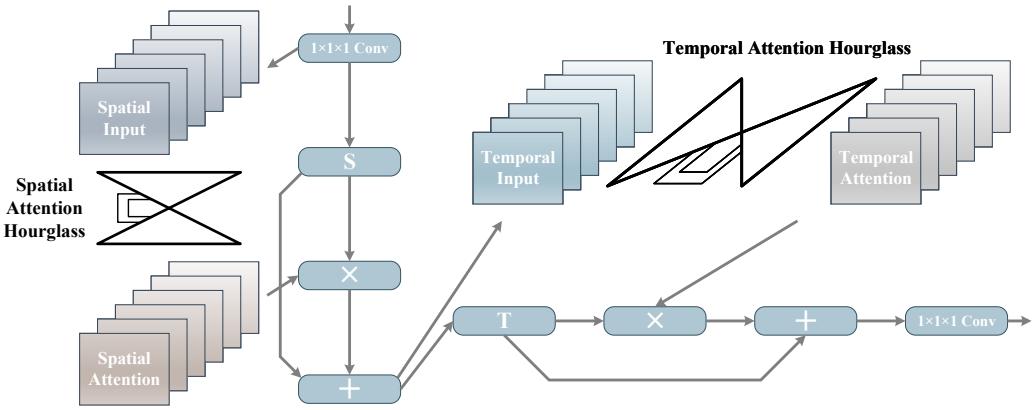


图 3.3 空间时间注意力机制下的 P3D 瓶颈结构（以 P3D-A 为例）

Figure 3.3 P3D Bottleneck building blocks with Spatial-Temporal Attention (Taking P3D-A as an example)

结合残差注意力后的 P3D 瓶颈结构形式化描述为

$$\mathbf{x}_{t+1} = \mathbf{P3D}_{Bottleneck}((1+\mathbf{M_T}(\mathbf{x}_t)) * \mathbf{T}(\mathbf{x}_t), (1+\mathbf{M_S}(\mathbf{x}_t)) * \mathbf{S}(\mathbf{x}_t)), \tag{3.7}$$

其中 $\mathbf{M_T}$ 和 $\mathbf{M_S}$ 为时间和空间维度上生成的注意力信息。图3.3以 P3D-A 为例展示了本文设计的空间时间注意力瓶颈结构，具体参数设置见章节4.2。

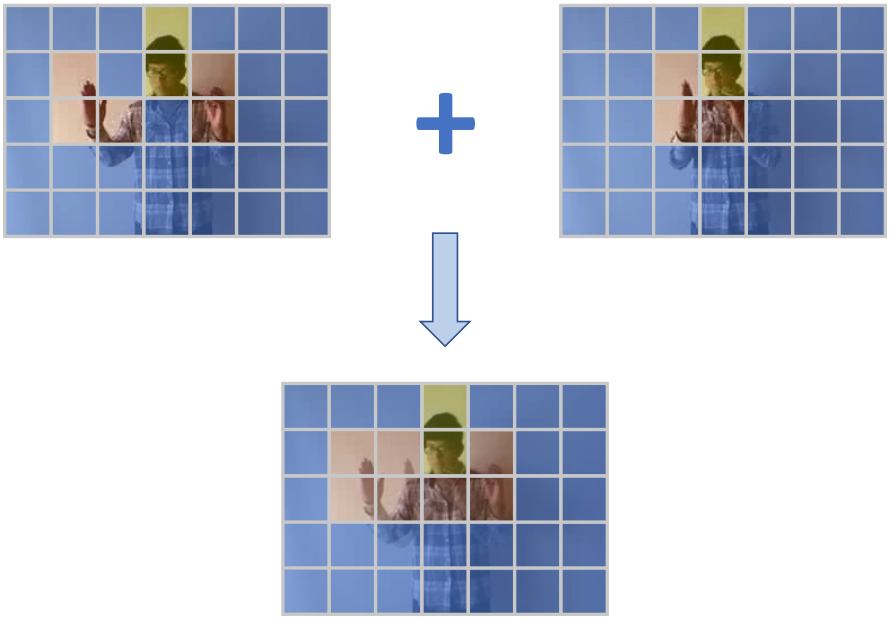


图 3.4 相邻帧位置掩模信息合并

Figure 3.4 Combination of contiguous frames' position-information mask

3.4 多任务学习

完全从零学起注意力信息，是具有一定难度的。而我们所希望学到的注意力信息，主要是关注到手脸等有效区域。因此，本文通过多任务学习的模式，将Liu et al. (2017) 所采集训练的手脸位置信息，作为训练数据，对注意力信息进行一定程度上的有监督学习。利用 Faster R-CNN 等物体检测方法提取的位置信息为包围框数据，本文将其视作一个粗糙的重点区域矩形掩模 (Mask)。在 P3D 网络各层，特征组织形式为时长 $T\times$ 高 $H\times$ 宽 W 的三维张量。我们将原始覆盖在每一帧上的掩模，进行相邻帧合并，并依据特定大小进行放缩裁剪，得到可以进行训练的手脸位置信息：

$$Position\_Info(i,j,k) \in \{0,1,2\} \quad (i,j,k) \in T \times H \times W, \tag{3.8}$$

其中 $\{0,1,2\}$ 分别代表 “其它类别”（背景躯干等），“脸”，“手”。图3.4展示了相邻帧掩模信息合并的效果。

同时需要注意到，引入的注意力学习模块的参数与本身卷积层的参数相当。在 ChaLearn 仅有五万样例的情况下，很容易发生参数过拟合，导致测试和验证集性能下降，我们也在具体实验中注意到了这一现象（见章4.4）。引入多任务学习，对注意力模块的参数加以限制，也能防止产生过拟合，对性能的提升起到一定的理论保证。

多任务学习主要有串行和并行两种模式。若有 *A*、*B*、*C* 三个任务，如果任务学习有一定先后，后一个任务的学习在一定程度上依赖于前一个任务的结果，形成如图3.5a的结构，即串行模式。如果任务学习可以并行进行，仅依赖一些基础信息，形成如图3.5b的结构，那么则为并行学习模式。

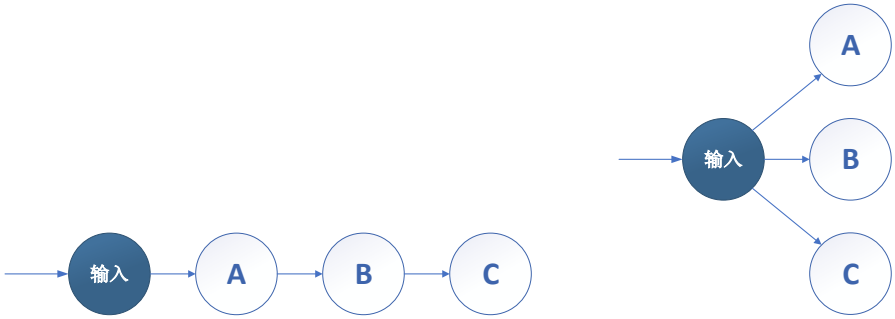


图 3.5 多任务学习模式 (a) 串行模式，(b) 并行模式。

Figure 3.5 Mutli-task learning mode (a) serial mode, (b) parallel mode.

在本框架下，如果采用串行模式，则认为手语分类依赖于手脸位置的学习。将学习到的注意力信息转化为手脸位置信息后，直接用手脸位置信息对卷积产生的特征进行修正，对式3.7修改后形式为：

$$\mathbf{P}_{position}(\mathbf{x}_t) = \mathbf{Softmax}(\mathbf{L}(\mathbf{M}(\mathbf{x}_t)))$$

$$\mathbf{x}_{t+1}(i,j,k) = \mathbf{F}'(\mathbf{x}_t)(i,j,k) * (1 + \mathbf{P}_{position}(\mathbf{x}_t)(i,j,k|i,j,k\ area\ contains\ hand\ or\ face)), \tag{3.9}$$

其中 **L** 为将注意力信息从原通道维数降到三维（对应“其它类别”，“脸”，“手”三个类别）的 1×1 卷积，**Softmax** 为归一化到 $[0,1]$ 形成概率形式的函数，$\mathbf{P}_{position}$ 为学习到的手脸位置概率。在训练时，将 $Position\_Info$ 作为训练数据，利用交叉熵对 $\mathbf{P}_{position}$ 进行学习。但在实验阶段，发现此模式会对手语识别性能产生干扰，相关数据和分析见章4。

如果采用并行模式，则认为手语分类和手脸位置的学习之间没有直接的影响。注意力信息转化为手脸位置信息 $\mathbf{P}_{position}$ 后，利用相同的方法对手脸位置进

行学习，但仍按照式3.7利用原注意力信息对特征进行修正。实验证明此模式对性能有显著提升，相关实验结果及分析见章4。

3.5 特征融合

由于 RGB 数据和深度数据本身没有对齐，预处理的对齐操作也可能不精确，Liu et al. (2017) 指出不宜将两种数据混合到一个网络中进行训练和测试。且混合后数据通道数过多，在现有深度学习框架下内存消耗过大，不易训练。所以，本文也采取了两种数据分开训练，然后再进行特征融合的模式。融合的方式借鉴了Narayana et al. (2018)，形式化为：

$$\begin{aligned} f_1 &= P3D(v_{RGB}) \\ f_2 &= P3D(v_{Depth}) \\ F_c &= \sum_{i=1}^{2} w_{ic} f_{ic} \quad c \in [1, Total\ Classes], \end{aligned} \tag{3.10}$$

其中 v_{RGB} 和 v_{Depth} 为 RGB 和深度输入视频，$P3D$ 代表了结合注意力机制和多任务学习的 P3D 识别网络，f_1 和 f_2 分别为在 RGB 和深度输入上学习到的特征，F 为连接后得到的总特征，下标 c 代表在每一类手势上的值，w_{ic} 为逐通道逐类权重。

分别在两类输入上预训练得到初始模型后，固定卷积参数，训练融合权重，以希望更好地使两类输入相互配合，提升识别性能。

第 4 章　实验与分析

本章节包括数据集介绍，框架的具体实现与实验分析的相关内容。

4.1　数据集

本课题利用 ChaLearn LAP IsoGD 手势识别孤立词数据集 (Wan et al., 2016) 进行主要实验，此数据集共包含 249 类手势，划分为训练集（35878 个视频序列），验证集（5784 个视频序列），测试集（6271 个视频序列）。所有手势均包括同时由 Kinect 设备 (Zhang, 2012) 采集的 RGB 和 Depth 通道数据。

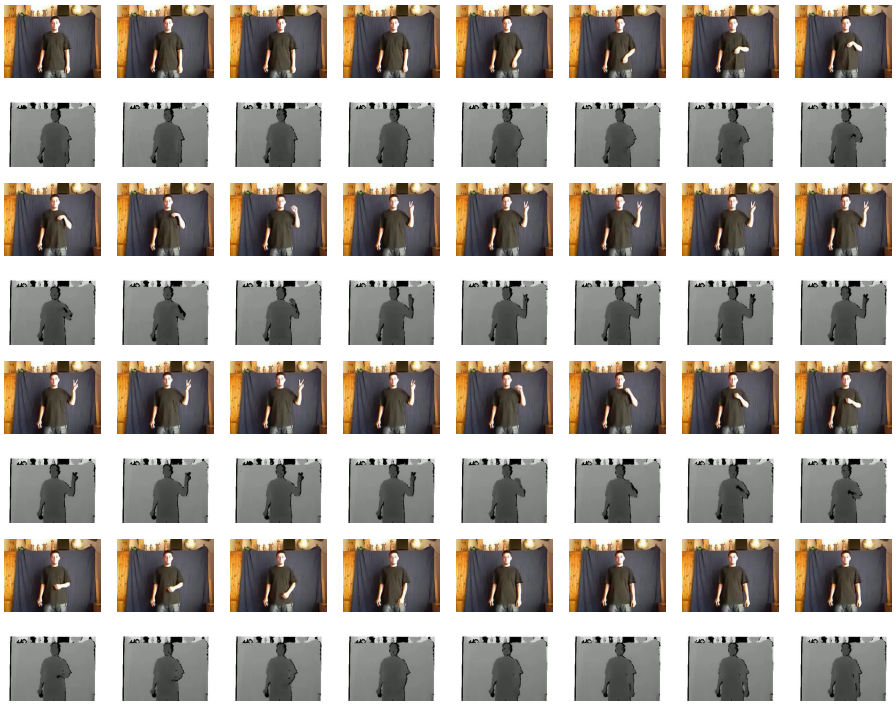
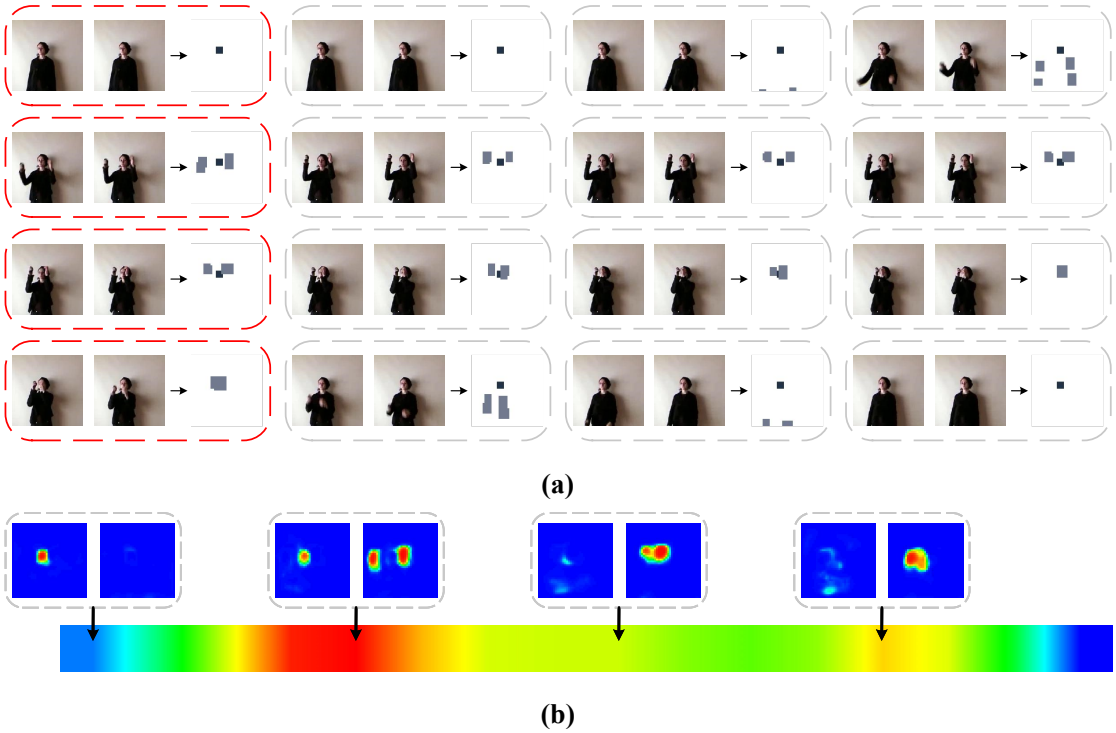


图 4.1 Chalearn 孤立词数据集样例

Figure 4.1 Sample of ChaLearn LAP IsoGD Dataset

数据集中视频每帧大小为 320 × 240，对齐后的输入视频将统一成 32 帧。图4.1展示了一个数据样例。训练时，为了增广数据，一半的数据将调整大小为 242×182，另一半的数据将随机调整大小为 $w \times h \quad w \in [160, 242]\ h \in [160, 182]$；一半的数据保持水平方向不变，另一半的数据水平翻转（数据集中大部分手势不受左右手影响）；最后所有视频会被随机裁剪为 160×160 输入给识别网络。测试时，为保持性能稳定，所有数据会调整大小为 242×182，不做水平翻转，选取画面中部进行裁剪。

除了输入视频之外，本文利用了Liu et al. (2017) 提供的双流 Faster R-CNN 提取手脸位置。相比于一般的 Faster R-CNN 模型，此模型将对齐后的 RGB 和深度图像同时提取特征并连接，之后再送入区域生成和区域池化，再进行分类和包围框回归，之后对每类结果进行非极大值抑制 (non-maximum suppression)，得到最后的检测结果。图4.2a展示了针对一个输入样例，依据节3.4的规则从手脸位置结果整理得到的 Res 2 层（定义见后一小节）手脸掩模。Res 2 层掩模为两帧合并的结果，图中虚线框包围的是相邻帧视频数据及其对应的掩模，浅灰色为手部位置，深灰色为脸部位置。在生成掩模时，手部位置优先。同时，由于掩模为两帧合并的结果，手部位置可能由于快速移动产生空隙。



(a)

(b)

图 4.2 多任务学习掩模及注意力机制可视化 (a) 双流 Faster R-CNN 提取位置产生掩模示例，(b) 手脸检测结果及时间注意力

Figure 4.2 Visualization of multi-task learning and attetion mechanism (a) Example of mask from 2-Stream Faster R-CNN detection results, (b) Hand & face detection and temporal attention

4.2 模型参数及实验设置

针对数据输入大小和数据集分类，本文设计了相应的网络架构，并利用 PyTorch 深度学习平台 (Paszke et al., 2017) 进行搭建。具体的网络参数见表4.1，网络结构示意见图4.3。图4.3中可以看出，为了防止过拟合，我们对网络进行了一

定的压缩：在 Res 3-5 层，每两个子层有一个注意力结构，同时 Res2-4 时空注意力数据将共享一个多任务学习模块。本文仿照Wang et al. (2017) 的工作设计了时空注意力的漏斗结构，包含了向下降采样，和向上升采样的过程，并在漏斗两侧进行跳跃 (Skip) 连接。表4.2展示了 Res 3 层的时空注意力结构的参数设置，其余层与之相仿。

表 4.1 特征提取网络参数细节

Table 4.1 Parameter details of feature extractor network

Layer	Ouput Size	Pseduo-3D	Attention	Multi-task
Res 1	$32\times80\times80$	$1\times7\times7,\ 64,\ stride\ 1\times2\times2$	-	-
Max pooling	$16\times40\times40$	$2\times3\times3,\ stride\ 2$	-	-
Res 2	$16\times40\times40$	$\begin{pmatrix} 1\times1\times1,\ 64 \\ S:\ 1\times3\times3,\ 64 \\ T:\ 3\times1\times1,\ 64 \\ 1\times1\times1,\ 256 \end{pmatrix}\times3$	$\begin{pmatrix} Spatial \\ Temporal \end{pmatrix}\times3$	$\begin{pmatrix} Spatial\ \&\ Temporal: \\ 1\times1\times1,3 \end{pmatrix}\times3$
Max pooling	$8\times40\times40$	$2\times1\times1,\ stride\ 2\times1\times1$	-	-
Res 3	$8\times20\times20$	$\begin{pmatrix} 1\times1\times1,\ 128 \\ first\ sub\ layer\ stride\ 1\times2\times2 \\ S:\ 1\times3\times3,\ 128 \\ T:\ 3\times1\times1,\ 128 \\ 1\times1\times1,\ 512 \end{pmatrix}\times8$	$\begin{pmatrix} Spatial \\ Temporal \end{pmatrix}\times4$	$\begin{pmatrix} Spatial\ \&\ Temporal: \\ 1\times1\times1,3 \end{pmatrix}\times4$
Max pooling	$4\times20\times20$	$2\times1\times1,\ stride\ 2\times1\times1$	-	-
Res 4	$4\times10\times10$	$\begin{pmatrix} 1\times1\times1,\ 256 \\ first\ sub\ layer\ stride\ 1\times2\times2 \\ S:\ 1\times3\times3,\ 256 \\ T:\ 3\times1\times1,\ 256 \\ 1\times1\times1,\ 1024 \end{pmatrix}\times36$	$\begin{pmatrix} Spatial \\ Temporal \end{pmatrix}\times18$	$\begin{pmatrix} Spatial\ \&\ Temporal: \\ 1\times1\times1,3 \end{pmatrix}\times18$
Max pooling	$2\times10\times10$	$2\times1\times1,\ stride\ 2\times1\times1$	-	-
Res 5	$2\times5\times5$	$\begin{pmatrix} 1\times1,\ 512 \\ first\ sub\ layer\ stride\ 2\times2 \\ S:\ 3\times3,\ 512 \\ 1\times1,\ 2048 \end{pmatrix}\times3$	$(Spatial)\times2$	$(Spatial:\ 1\times1\times1,3)\times2$
Max pooling	$1\times5\times5$	$2\times1\times1$	-	-
Average pooling	$1\times1\times1$	$1\times5\times5$	-	-
FC, Softmax	249			

训练时，本文采用批随机梯度下降法，动量设置为 0.9，参数权值衰减为 0.5。初始训练率为 0.0001，每训练 5000 轮训练率下降 0.9。单通道模型训练在 80000 轮时结束，融合模型训练在 10000 轮时结束。单通道模型初始训练时，会使用Qiu et al. (2017) 在 Kinetics 视频分类数据集 (Carreira et al., 2017) 上的预训练结果。

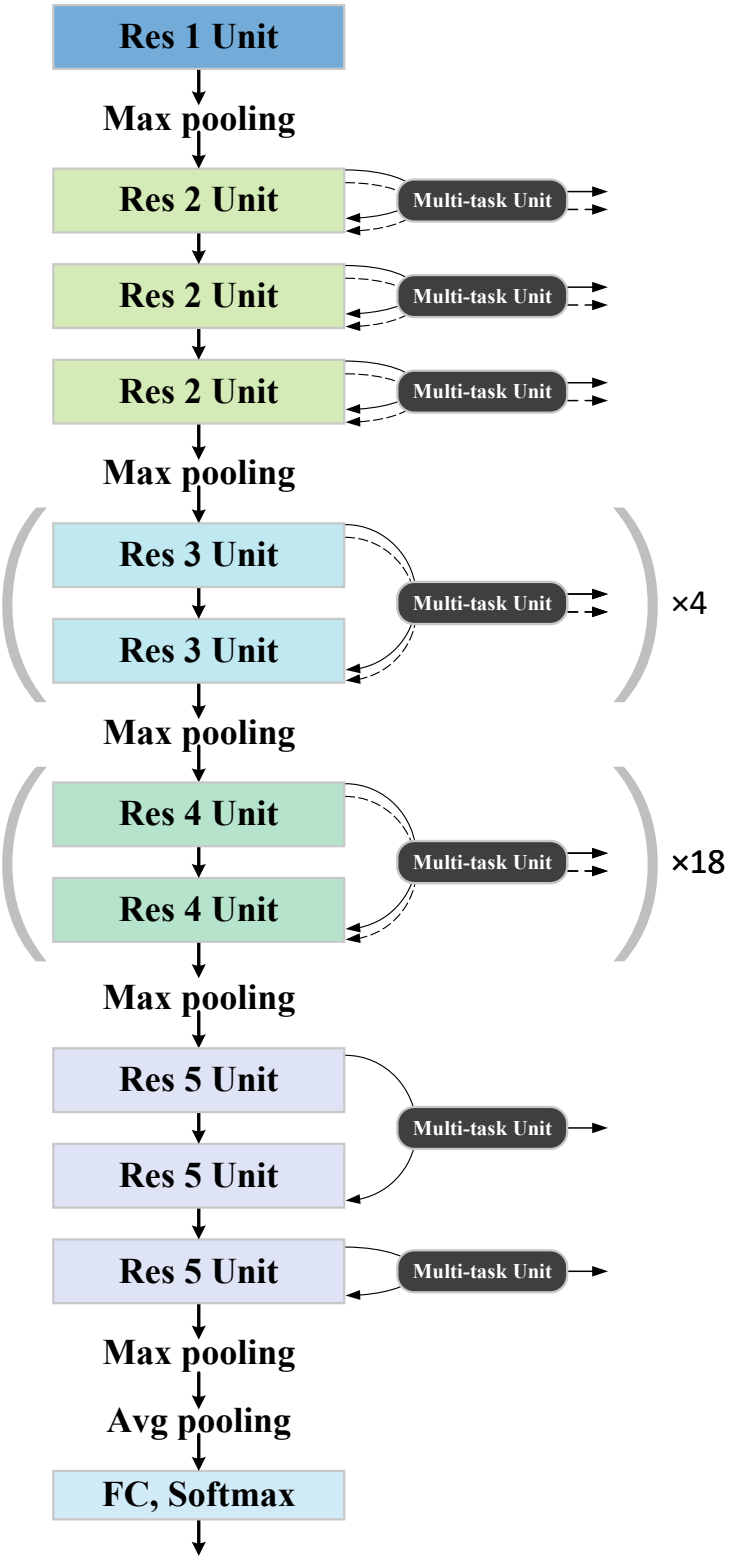


图 4.3 特征提取网络结构

Figure 4.3 Struture of feature extractor network

表 4.2 注意力结构参数示意（以 Res 3 层为例）

Table 4.2 Example of attention structure's parameters (Taking Res 3 as an example)

Layer	Spatial Attetion	Size	Temporal Attention	Size
1	$1\times1\times1, 32$	$8\times20\times20$	$1\times1\times1, 32$	$8\times20\times20$
2	Max pooling	$8\times10\times10$	Max pooling	$4\times20\times20$
3	$1\times3\times3, 32$	$8\times10\times10$	$3\times1\times1, 32$	$4\times20\times20$
4	Max pooling	$8\times5\times5$	Max pooling	$2\times20\times20$
5	$1\times3\times3, 32$	$8\times5\times5$	$3\times1\times1, 32$	$2\times20\times20$
6	Upsample	$8\times10\times10$	Upsample	$4\times20\times20$
7	$1\times3\times3, 32$	$8\times10\times10$	$3\times1\times1, 32$	$4\times20\times20$
8	Add with layer 3	$8\times10\times10$	Add with layer 3	$4\times20\times20$
9	Upsample	$8\times20\times20$	Upsample	$8\times20\times20$
10	$1\times3\times3, 32$	$8\times20\times20$	$3\times1\times1, 32$	$8\times20\times20$
11	$1\times1\times1, 128$	$8\times20\times20$	$1\times1\times1, 128$	$8\times20\times20$
12	$1\times1\times1, 128$	$8\times20\times20$	$1\times1\times1, 128$	$8\times20\times20$

4.3 多任务及注意力机制可视化

为了更加直观地说明多任务及注意力机制的作用，我们对 Res 2 层的手脸位置检测和时间维度注意力进行了可视化。图4.2b虚线框中热点图，展示了图4.2a红色虚线框里的相邻帧手脸检测结果。其中，左侧为脸部结果，右侧为手部结果。可见，本框架多任务学习能够学习到手和脸的大致位置信息。但由于手和脸具有一定相似性，可能会产生手脸间的误判。存在遮挡情况下，多任务机制的脸部识别结果不好。4.2b下部色带为时间注意力，取 16 个时间 “帧” 全图时间注意力平均数值后，再均一化，纯红色为强度最高，纯蓝色为强度最低，中间依次过渡。此示意图证明了注意力能够使网络关注到有手部出现的帧，特别是抬起和准备放下的动作。

4.4 手语识别结果及性能比较

为了验证框架各模块的作用，本文设计了一系列模型的对比试验，并将 RGB 通道的准确率总结在表4.3中。

本文预期目标是减少输入的预处理步骤，并保持一定的性能。所以，我们选择了以对齐后的数据输入无修改的 P3D 模型作为基础模型 (Baseline model) 进行对比。从表4.3可知，引入注意力机制和多任务学习使得性能相比基础模型有了

很大的提升。

表 4.3 各模型 RGB 通道的准确率

Table 4.3 RGB Accuracy of different models

模型	验证集准确率	测试集准确率
Baseline	52.58%	55.96%
仅有注意力机制，无多任务学习	50.43%	52.30%
串行多任务学习	59.99%	60.53%
并行多任务学习	**61.50%**	**63.37%**

如果仅使用时空注意力机制，而无多任务学习，由于注意力机制引入了过多的参数，使得模型更容易过拟合于训练集上，使得性能下降。而加入多任务学习，使得注意力的参数得到了限制，减轻了过拟合的现象，性能有了显著的提升。

本文比较了两种多任务学习的性能。由表4.3的比较可见，并行学习模式手语识别性能好于串行学习模式（两个模型参数数量一致）。分析有两种成因：1、并行模式下，注意力结果是逐通道的，更有利于修正各个卷积层的特征提取结果；2、手脸逐像素检测类似于语义分割任务，难度高于物体检测类任务，现有的训练数据量不够大，难以使网络学习到准确的结果，对识别产生影响。

表 4.4 主流模型性能对比

Table 4.4 Performance Comparison of mainstream models

No.	Model	Accuracy（Validation set）	Accuracy（Testing set）
1	Ours	66.30%	68.60%
2	Miao et al. (2017)	64.40%	67.71%
3	Liu et al. (2017) (Updated version)	66.49%	68.92%
4	Narayana et al. (2018) (Global size input)	61.40%	67.50%
5	Narayana et al. (2018)	**80.96%**	**82.67%**

本文将并行多任务学习模式下，RGB 和 Depth 通道融合后的准确率与现行主流方法进行了比较，具体结果见4.4。方法 2 为 2017 年 ICCV 孤立手语识别竞赛首位结果Miao et al. (2017)，方法 3 为Liu et al. (2017)（2017 年 ICCV 连续手语识别竞赛首位）的改进版本，方法 4 和 5 来自自当前最高性能方法Narayana et al. (2018)。方法 5 在数据大小上使用了左右截出图及完整图，数据来源上利用了原始 RGB 和深度，及其各自光流数据，总计 12 个通道，所以达到了极高的性能，

但其仅适用原图大小的方法 4 性能较本文方法低。如果本文也使用光流数据，并对超参进一步调节，有望提高性能。

第 5 章　结束语

本文设计了一种新型深度学习手语识别模型，利用注意力机制和多任务学习，将关键区域检测和手语识别相结合，减少数据预处理步骤，并在 ChaLearn 孤立词手势数据集上达到了与当前最高水平方法可比的性能水平。本文由于时间限制，对模型内层数等超参数还未进行有效调节，后期工作将关注于对模型超参的细致调优，结合光流通道数据及改进融合方法，希望能够减小与当前最高性能的差距。同时，后期工作将尝试把此模型迁移到手势连续词的识别上，使其应用更加广泛。

参考文献

BRADSKI G, KAEHLER A, 2008. Learning opencv: Computer vision with the opencv library[M]. [S.l.]: " O'Reilly Media, Inc.".

CARREIRA J, ZISSERMAN A, 2017. Quo vadis, action recognition? a new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE: 4724–4733.

CHEN L C, PAPANDREOU G, KOKKINOS I, et al., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 40(4): 834–848.

CORRADINI A, 2001. Dynamic time warping for off-line recognition of a small gesture vocabulary [C]//Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings. IEEE ICCV Workshop on. [S.l.]: IEEE: 82–89.

FU J, ZHENG H, MEI T, 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]//Conf. on Computer Vision and Pattern Recognition. [S.l.: s.n.].

GIRSHICK R, 2015. Fast r-cnn[J]. arXiv preprint arXiv:1504.08083.

GIRSHICK R, DONAHUE J, DARRELL T, et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.]: 580–587.

GUYON I, ATHITSOS V, JANGYODSUK P, et al., 2013. Results and analysis of the chalearn gesture challenge 2012[M]//Advances in Depth Image Analysis and Applications. [S.l.]: Springer: 186–204.

HE K, ZHANG X, REN S, et al., 2016. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.]: 770–778.

KARPATHY A, TODERICI G, SHETTY S, et al., 2014. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. [S.l.: s.n.]: 1725–1732.

KLASER A, MARSZAŁEK M, SCHMID C, 2008. A spatio-temporal descriptor based on 3d-gradients[C]//BMVC 2008-19th British Machine Vision Conference. [S.l.]: British Machine Vision Association: 275–1.

KONG W, RANGANATH S, 2008. Automatic hand trajectory segmentation and phoneme transcription for sign language[C]//Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on. [S.l.]: IEEE: 1–6.

LAPTEV I, 2005. On space-time interest points[J]. International journal of computer vision, 64 (2-3): 107–123.

LIU Z, CHAI X, LIU Z, et al., 2017. Continuous gesture recognition with hand-oriented spatiotemporal feature[C]//The IEEE International Conference on Computer Vision (ICCV) Workshops. [S.l.: s.n.].

MALGIREDDY M R, NWOGU I, GOVINDARAJU V, 2013. Language-motivated approaches to action recognition[J]. The Journal of Machine Learning Research, 14(1): 2189–2212.

MIAO Q, LI Y, OUYANG W, et al., 2017. Multimodal gesture recognition based on the resc3d network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.]: 3047–3055.

NAGI J, DUCATELLE F, DI CARO G A, et al., 2011. Max-pooling convolutional neural networks for vision-based hand gesture recognition[C]//Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on. [S.l.]: IEEE: 342–347.

NARAYANA P, BEVERIDGE J R, DRAPER B A, 2018. Gesture recognition: Focus on the hands [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.].

PASZKE A, GROSS S, CHINTALA S, et al., 2017. Automatic differentiation in pytorch[Z]. [S.l.: s.n.].

QIU Z, YAO T, MEI T, 2017. Learning spatio-temporal representation with pseudo-3d residual networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). [S.l.]: IEEE: 5534–5542.

REN S, HE K, GIRSHICK R, et al., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. [S.l.: s.n.]: 91–99.

SCOVANNER P, ALI S, SHAH M, 2007. A 3-dimensional sift descriptor and its application to action recognition[C]//Proceedings of the 15th ACM international conference on Multimedia. [S.l.]: ACM: 357–360.

SIMONYAN K, ZISSERMAN A, 2014. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. [S.l.: s.n.]: 568–576.

TRAN D, BOURDEV L, FERGUS R, et al., 2015. Learning spatiotemporal features with 3d convolutional networks[C]//Computer Vision (ICCV), 2015 IEEE International Conference on. [S.l.]: IEEE: 4489–4497.

WAN J, ZHAO Y, ZHOU S, et al., 2016. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.: s.n.]: 56–64.

WANG C, GAO W, SHAN S, 2002. An approach based on phonemes to large vocabulary chinese sign language recognition[C]//Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on. [S.l.]: IEEE: 411–416.

WANG F, JIANG M, QIAN C, et al., 2017. Residual attention network for image classification[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.].

WANG H, STEFAN A, MORADI S, et al., 2010. A system for large vocabulary sign search[C]// European Conference on Computer Vision. [S.l.]: Springer: 342–353.

WANG S B, QUATTONI A, MORENCY L P, et al., 2006. Hidden conditional random fields for gesture recognition[C]//Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on: volume 2. [S.l.]: IEEE: 1521–1527.

WILLEMS G, TUYTELAARS T, VAN GOOL L, 2008. An efficient dense and scale-invariant spatio-temporal interest point detector[C]//European conference on computer vision. [S.l.]: Springer: 650–663.

YAMATO J, OHYA J, ISHII K, 1992. Recognizing human action in time-sequential images using hidden markov model[C]//Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on. [S.l.]: IEEE: 379–385.

ZHANG Z, 2012. Microsoft kinect sensor and its effect[J]. IEEE multimedia, 19(2): 4–10.

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

谈清扬，江苏省南京市人，中国科学院大学计算机与控制学院本科生。

已发表 (或正式接受) 的学术论文:

[1] Mesh-based Autoencoders for Localized Deformation Component Analysis, AAAI Conference on Artificial Intelligence, 2018

[2] Variational Autoencoders for Deforming 3D Mesh Models, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

致　谢

感谢指导教师陈熙霖研究员，柴秀娟副研究员对本课题的指导。感谢刘志鹏学长对本课题的帮助。

感谢神经网络的先驱者为广大研究人员带来简单易用的深度学习平台。感谢英伟达 CEO 黄仁勋对 GPU 平台通用化的重视，为我们带来了计算能力的提升。感谢现代医学的进步，让我们能够情绪稳定、生理健康地工作生活。感谢我的母亲对我临近毕业阶段各种选择的理解。

感谢我曾经遇到的难缠的审稿人，部分申请学校冷漠的教授，激发了我继续学习、提升研究水平和 taste 的斗志。