中文图书分类号：TP391
UDC：004
学 校 代 码：10005

硕士学位论文

MASTER DISSERTATION

论 文 题 目：知识驱动的视觉手势识别技术研究

论 文 作 者：吕孟飞

学 科：软件工程

指 导 教 师：何坚 副教授

论文提交日期：2024 年 06 月

UDC：004　　学校代码：10005
中文图书分类号：TP391　　学　　号：S202175005

北京工业大学硕士学位论文

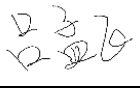
题　　目：知识驱动的视觉手势识别技术研究

英文题目：**Research on Knowledge-driven Visual Gesture Recognition Technology**

论 文 作 者：吕孟飞
学 科 专 业：软件工程
研 究 方 向：物联网与嵌入式智能计算
申 请 学 位：工学硕士
指 导 教 师：何坚 副教授
所 在 单 位：信息学部
答 辩 日 期：2024 年 5 月
授予学位单位：北京工业大学

独 创 性 声 明

本人声明所呈交的论文是我个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，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

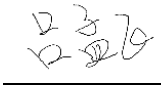
签　　名： 

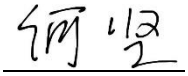
日　　期：　2024 年 6 月 1 日

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布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论文。

（保密的论文在解密后应遵守此规定）

签　　名：  　　日　　期：　2024 年 6 月 1 日

导师签名：  　　日　　期：　2024 年 6 月 1 日

摘 要

手势（Gesture）交互作为一种新兴的人机交互通道，被广泛应用于智能手机、电视等现代生活中常见的各类电子设备中。然而手势表达的含义具有一定的随意性，因此手势的识别与理解需要结合使用情境、语境上下文等因素进行综合考量。现有的手势识别研究关注于从姿态骨架的时序动态信息中提取手势的时域特征，作为语境上下文因素优化手势识别。对于场景信息的利用则主要通过网络结构进行抽象的提取，因此不具有应用场景上的可解释性和可扩展性。

而知识图谱作为存储概念性知识的有效方法，在知识检索、问答系统等领域已经有了广泛的研究，但其在场景的识别和划分任务中，尚没有实际的应用。本文构建了一套基于图像信息的手势知识图谱，将知识图谱的知识推理能力应用于对手势使用场景的分析中，从而提高手势识别在混合场景下的识别效率，并设计一套方法使得手势识别算法可以轻松、高效地适应到新的应用场景下，形成一套融合知识特征与姿态特征的视觉手势识别方法。

具体来说，本文的主要工作与成果如下：

（1）面向手势应用场景的分析，设计多模态手势知识图谱的本体结构，并研究多模态的构建方法以及对应的知识融合与存储方法，从而基于 Wikidata 开源知识图谱和开源手势数据集，构建出包含丰富实体和关系信息的多模态手势知识图谱。

（2）从轻量化的算法需求出发，设计实现快速高效的多头人体姿态估计算法，并根据算法需求，完成相应的损失函数设计、数据前后处理等工作，在迁移部分 MoveNet 网络参数的基础上，用 COCO 数据集对模型进行训练和测试，取得了 80.1%的平均精度和 42.4ms/帧的识别速率。

（3）设计基于图卷积神经网络（GCN）的特征提取算法，从知识图谱中提取知识特征，并从人体姿态骨架中推理出的姿态特征。然后设计手势识别算法，融合知识特征与姿态特征，用于混合场景下的手势识别任务，并在来自 5 个不同应用场景的混合数据集中对模型进行了训练和测试，取得了优于现有算法的识别效果，以及满足实际应用的增量训练与推理速度。

此外，本文利用所提出的方法，结合自动驾驶汽车的车载场景，设计并实现了一个可以用于车载环境的视觉手势识别系统原型。通过这个系统的验证，完善了知识驱动视觉手势识别算法的应用开发方法，证明了算法在实践中的有效性和

可行性。

关键词：人机交互；知识图谱；姿态估计；视觉手势识别；图神经网络

Abstract

Gesture interaction, as an emerging human-computer interaction channel, is widely used in various types of electronic devices that are common in modern life, such as smartphones and TVs. The meanings expressed by gestures are arbitrary, so the recognition and understanding of gestures should be comprehensively considered with respect to the context of use, contextual context, and other factors. Existing research on gesture recognition focuses on extracting the time-domain features of gestures from the temporal dynamic information of the gesture skeleton, which is used as a contextual factor to optimize gesture recognition. For the utilization of scene information, the abstraction is mainly extracted through the network structure, and thus does not have the interpretability and scalability on the application scene.

Whereas knowledge graphs, as an effective method for storing conceptual knowledge, have been widely researched in the fields of knowledge retrieval and question-answer systems, they have not yet been practically applied in the task of recognizing and delineating scenes. In this thesis, we construct a set of gesture knowledge graphs based on image information, apply the knowledge reasoning ability of knowledge graphs to the analysis of gesture usage scenarios, so as to improve the recognition efficiency of gesture recognition in mixed scenarios, and design a set of methods to make gesture recognition algorithms easily and efficiently adaptable to new application scenarios to form a set of visual gesture recognition methods integrating knowledge features and gesture features.

The main work and results of this thesis are as follows:

(1) Oriented to the analysis of gesture application scenarios, we design the ontology structure of multimodal gesture knowledge graph, and study the construction method of multimodality as well as the corresponding knowledge fusion and storage method, so as to construct a multimodal gesture knowledge graph containing rich entity and relationship information based on the Wikidata open-source knowledge graph and open-source gesture dataset.

(2) Starting from the lightweight algorithm demand, we design and implement a fast and efficient multihead human gesture estimation algorithm, and according to the

algorithm demand, we complete the corresponding loss function design, data pre- and post-processing, etc. Based on migrating part of the MoveNet network parameters, we use the COCO dataset to train and test the model, and achieve an average accuracy of 80.1% and a recognition 42.4ms per frame.

(3) A feature extraction algorithm based on graph convolutional neural network (GCN) is designed to extract knowledge features from the knowledge graph and gesture features derived from the human gesture skeleton. Then, the gesture recognition algorithm is designed to fuse the knowledge features with the gesture features for gesture recognition tasks in hybrid scenes, and the model is trained and tested on hybrid datasets from five different application scenarios, achieving recognition results superior to existing algorithms, as well as incremental training and inference speeds to meet real-world applications.

In addition, in this thesis, a prototype of a visual gesture recognition system that can be used in in-vehicle environments is designed and implemented using the proposed methodology in conjunction with the in-vehicle scenario of self-driving cars. The validation of this system refines the application development methodology of the knowledge-driven visual gesture recognition algorithm and proves the effectiveness and feasibility of the algorithm in practice.

Keywords: Human-computer interaction, Knowledge graph, Pose estimation, Visual gesture recognition, Graph neural network

目 录

第1章 绪论

本章节首先说明手势识别在自然人机交互中的发展潜力和重要意义；接着对国内外主要的场景上下文语义理解和视觉手势识别方法进行阐述，并分析各个方法的特点和缺陷；最后概况介绍本文的研究内容和论文组织结构。

1.1 研究背景及意义

手势（Gesture）是人们日常生活中常用的一种表达方式，包括作为语言辅助的辅助性手势、专业领域有特殊约定的专用手势、特殊群体使用的手语等等。通常可以按照持续时间分为静态手势和动态手势，其中，静态手势也可以被看作是姿态（Pose），动态手势也可以被看作是动作（Action），而后者动态手势更符合人们日常交流的使用习惯和需要。

而随着信息技术的逐步发展，人机交互技术也随之成熟、完善，如何使人与计算机更加流畅地完成协同工作，更加舒适地进行交互成为广受关注的研究方向。人机交互技术经历了早期命令行交互界面（Command Line Interface, CLI）阶段、图形交互界面（Graphical User Interface, GUI）阶段，发展到现在的自然人机界面（Natural User Interface, NUI）阶段，从以机器为中心逐渐转向以人为中心[1]。在自然人机交互中，用户不再严格的通过离散化的操作指令进行控制，而是使用连续、非确定的多模态数据来表示自身的目的与意图[2]。手势作为一种非常符合自然交互界面需求的交互方式，已经广泛应用于相关领域，逐步成为一个新的主流人机交互通道[3]。

手势交互系统通常包括四个部分：手势实施者、手势输入通道（摄像头、可穿戴设备等）、手势分析模块以及手势操作界面[4]。而手势分析模块及其使用的算法则是整个系统中最核心的组成部分，按照前一部分输入设备的不同，通常可以分为基于视觉和基于可穿戴设备两类。现代智能电子设备广泛配备的摄像头用于图像采集，这使得基于视觉的手势识别技术相对于其他方法更易于使用[4]，因此受到了广泛的关注和研究。而基于可穿戴设备的识别技术需要使用者穿戴特定的设备，对用户的存在一定的干扰和负担，但其相较于视觉方法有更高的准确度和详细的跟踪信息[5]。

手势作为一种自然的交互方式，通常用于表达人类的一些简单的、明确的意

图或对象，具有随意性的特点[6]。因此，利用手势进行沟通需要沟通双方基于某种预设好的协定，如：交警手势基于国家法律法规、日常通用手势基于一定范围内人们的约定俗成等。如果没有这样的协定或超出协定范围，手势则可能存在被误解的可能[6]。所以手势的识别与理解需要针对使用情境、语境上下文等因素进行综合考虑，才可以实现准确的识别。

知识图谱（Knowledge Graph）作为一种存储先验知识，建模关联关系的图结构，本质是一种具有有向图结构的知识库。其中，节点通常用于保存实体或概念，而边则用于表示这些实体、概念之间存在的各种语义关系[7]。知识图谱的建立通常需要从结构化数据或文本数据中抽取信息，并利用各种技术进行知识表示和建模。知识图谱因为其在存储知识和知识推理方面的优势，被广泛应用于诸如人机问答、翻译系统、自然语言理解等研究领域，取得了很好的效果。而随着多模态学习和计算机视觉的发展，单纯的文字信息已经满足不了知识图谱的需要，多模态知识图谱成为补全知识图谱，增强知识图谱推理能力的一个热点方向[8]。

为了将情境信息等融入手势识别任务，利用先验知识提高手势识别的泛用性和识别效率，本课题研究设计多模态知识图谱用于存储手势相关的先验知识，并在手势知识图谱上进行推理，将推理解结果用于视觉手势的分类任务，从而提高视觉手势识别方法的识别速度以及其跨场景识别的能力。

1.2 国内外研究现状概述

1.2.1 多模态知识图谱的构建

知识图谱最早是由 Google 的科研团队在 2012 年为优化搜索引擎搜索结果的智能性而提出，从早期自然语言处理任务中使用的语义网络（semantic network）发展而来。知识图谱是通过事实三元组的方式，精准地描绘了现实世界中的概念、实体以及它们之间的关联。在这种表示中，实体和概念被提炼为节点，而它们之间的关联则被提炼为边。这样的结构化建模方法，不仅继承了早期语义网络的精髓，更着重强调了实体间的内在联系以及实体的固有属性，使得知识的表达更为精准和丰富。知识图谱的构建通常以自动化的方法为核心，从百科页面、开源知识图谱为等结构化数据中进行知识抽取，再通过人工处理进行完善和补全[9]。其中涉及自然语言处理（NLP）、机器学习（ML）等多个领域的技术，并已经取得了卓有成效的进展，如卡耐基梅隆大学的 NELL[10]等。

从语义理解的角度出发，任何一条信息都可以通过多种媒体形式进行传递或

记录，因此同一条知识可以被表示为文字、图像、语音等多种媒体形式，且其互相之间是等价的，这样的信息就是多模态信息，其存储形式为多模态数据。但多模态数据因其结构特点，存在语义鸿沟和异构鸿沟[11]，在表示和度量等方面仍缺少高效、通用的解决方案。要想从多模态数据中构建知识图谱，就需要通过多模态学习，实现各个模态信息之间的交流与转换。随着深度学习的方法在自然语言处理、计算机视觉等领域的快速发展，多模态学习进入了多模态深度学习的阶段，构建深度神经网络将 2 个或多个模态的数据映射到同一个公共空间上进行表征[12]。

人类的感知来自不同的感官信息的融合，让机器提高理解与分析的能力也需要提高机器对于多模态信息的理解与处理。类比知识图谱在知识存储与检索中的作用，多模态知识图谱可以辅助计算机对多模态信息的处理。多模态知识图谱，即具有包括但不限于图像、音频、文本等类型实体和属性的知识图谱。这类知识图谱中利用多模态信息消除实体或属性之间可能存在的歧义，补充知识的多模态细节信息[13]，相关的研究主要关注于文字和图像两种模态信息[13]-[17]。

像 Wikidata[18]、DBpedia[19]等主流的大型知识库中存储的多模态信息指的是多媒体文件的作者、大小等元数据信息，而缺少关于多模态实体内容的语义结构。IMGpedia[20]是一个典型的多模态知识图谱，它的多模态信息来自于 Wikimedia Commons 数据集的图像数据，借此构建了 1500 万个视觉内容描述符，生成了 4.5 亿组视觉相似关系。不过，IMGpedia 依然存在关系的种类稀疏、图谱结构稀散、视觉类别模糊等缺陷，并且未对图像中包含的语义信息进行分析和抽取。

2017 年，Rubio 等人创建了一个包含约 15 万个实体及其之间关系的知识图谱 ImageGraph[17]，用以实现一种基于卷积神经网络（CNN）的机器学习方法来完成视觉查询任务。随后 Liu 等人在其基础上设计了一个知识图谱的集合（MMKG），其中包含了图谱中各个实体的特征与图像信息。MMKG 在数据表示上选用了 N-Triples 格式，并依托 FB15K 数据集作为基准，以实现与其他知识图谱的高效对齐与融合。与传统的在同一知识图谱内进行视觉推理任务的方式不同，MMKG 的设计初衷是跨越不同知识图谱的界限，对来自不同知识图谱的实体和图像进行深度的关系推理、多关系预测以及实体匹配。然而，值得注意的是，MMKG 目前主要聚焦于小数据集的应用，并且其图像处理策略依赖于相应的文本实体，即图像与文本实体紧密绑定，而非独立存在。此外，MMKG 在处理图像数据时，尚未充分考虑到图像自身的多样性特点。

以上这类知识图谱是由传统文本知识图谱扩展而出的，包含多模态信息的常

识知识图谱（Common Sense Knowledge Graph）。除此之外，结合计算机视觉领域技术的发展，Yang J 等人提出了一类新的多模态知识图谱，基于复杂场景图谱生成（Scene Graph Generation, SGG）的多模态知识图谱[21]，表示场景中多个目标实体之间的位置或语义关系。在场景知识图方法中，一般通过计算机视觉方法将图像转化为语义元素，进而进行其他的处理和分析[22]。Alberts 等人所实现的 VisualSem 便是将两者相结合后，依据标注好的图像数据，得到的一个高质量的多模态知就知道识图谱[23]。

多模态知识图谱自底向上的构建方法的研究已经相对完善，这样构建出的常识知识图谱可以广泛应用于视觉问答等任务中。但对于场景语义知识图谱或某个业务领域的细分知识图谱，使用该类方法会受到手动标记数据的数量限制，存在数据量不足等问题。

1.2.2 知识驱动的人类活动识别

传统的视觉检测识别方法通常是数据驱动的，即通过采集大量数据样本进行模型训练，从而获得聚焦任务功能的识别检测模型。但由于数据采集难度高，采集过程繁琐，以及真实场景复杂、不确定等方面的问题，这种方法并不能很好地用于复合人类活动识别任务（Composite Human Activity Recognition）[24]。因此，很多就研究者将注意力转移到了知识驱动的、结合多模态信息的技术框架中，通过对单个动作单元的分析和识别组成具有语义信息的多元复合活动，利用知识驱动的相关技术和方法进行语义层面的分析，从而更好地完成人类活动识别任务。

事实上，由于表情学相关的研究基础以及表情本身的强语义性，知识驱动的方法已经被广泛应用于面部动作单元（Facial action unit, AU）的识别任务中。Yanan Chang 等人在 2019 年提出了 TCAE 方法[25]，利用人脸动作编码系统(Facial Action Coding System, FACS)定义的人脸标记规则，设计了一种新的知识驱动的自监督表征学习框架，用于人脸识别。表示编码器使用大量没有 AU 注释的面部图像进行训练。从 FACS 中总结出 AU 标记规则，设计面部划分方式，确定面部区域之间的相关性。该方法利用骨干网络提取局部面部区域表示，并利用项目头部将这些表示映射到低维潜在空间中。在潜在空间中，对比学习组件利用区域间的差异来学习与区域相关的局部表示，同时保持区域内的实例区分。还探讨了从 AU 标记规则中总结的面部区域之间的相关性，以便使用预测学习组件进一步学习表征。在两个基准数据库上的测试表明，学习到的表示对于 AU 识别是强大的和高效的。

而在人体活动识别任务中，知识驱动的方法也已经被研究人员们广泛关注。视频作为载体的人体行为分析在多个领域展现出巨大的应用潜力。尽管郭萍等人深入探讨了利用隐式马尔科夫模型对人体行为进行分类的方法[26]，但深度学习技术的飞速发展已经验证了神经网络在人体行为特征提取和识别中的显著优势，表现出更高的效率和准确性[27]。在基于深度学习的识别方法中，为提升分类性能，研究人员通过增加网络层数以获取更多的特征信息。这一策略促使了更深层神经网络结构的诞生。Lecun Y 提出的 LeNet5，一个 5 层的卷积神经网络，为图像特征提取奠定了重要基础[28]。随后，Hinton 的 AlexNet 以更深的网络结构在 ImageNet 数据集上取得了显著成就，引起了广泛关注，进一步推动了深度学习在识别任务中的应用和发展[29]。随着深度学习技术的持续演进，新的神经网络模型结构层出不穷。Simonyan K 等人提出的 VGG 网络在深度学习识别任务中表现出色[30]。VGG 的深层次设计强调了深度神经网络在提升识别准确率中的关键作用。接着，Szegedy C 等人提出的 GoogLeNet 则通过融合不同尺寸的卷积核并进行网络拆分，实现了横向扩展，从而在保持较少网络参数的同时，在相同任务中取得了更佳的性能[31]。这些创新不仅展示了深度学习在人体行为分析领域的强大潜力，也为未来的研究提供了新的方向。

随着深度学习研究的不断深入，研究人员逐渐认识到在神经网络的层数增加时，信息在网络层中的传递过程中容易丢失，这限制了网络性能的持续提升。具体而言，持续增加神经网络的层数并不会一直带来性能的提升，反而可能因为过深的网络结构导致收敛速度变慢，甚至无法继续有效训练，最终降低识别的准确率。为了克服这一挑战，何凯明等人提出了残差网络（ResNet）概念[32]。

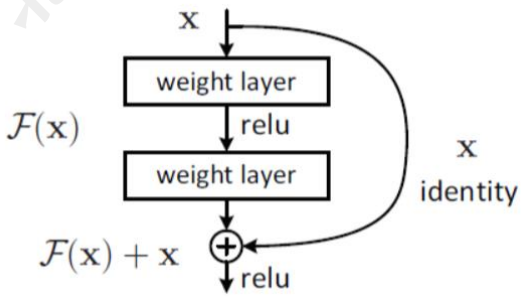


图 1-1 基本残差结构[32]

Fig. 1-1 Basic Residual Structure

ResNet 通过引入恒等映射的残差块设计，巧妙地解决了深层神经网络在训练过程中难以收敛的问题。这种设计允许网络直接跳过部分网络层，即通过如图 1-1 所示的残差快结构避免了由于网络层数过深而导致的性能瓶颈。残差块的结

构使得网络能够学习输入与输出之间的残差，而非直接学习整个映射，这在很大程度上提高了网络的训练效率和性能。ResNet 网络结构使得更深的神经网络能够有效收敛的同时，也带来了新的问题。随着网络深度而增加的计算成本与提高的准确率收益并不对等，即网络深度的增加耗费了大量的计算资源，但提高的模型效果却并不多。为了解决这一问题，Saining Xie 等人将 ResNet 与分组卷积相结合，提出了 ResNeXt 网络[33]。通过减少运算量的方法，来保证在不增加运算符但的前提条件下提高准确率。而 S. Zagoruyko 等人则通过优化残差块结构，增加每个残差块中卷积核的数量来减少网络参数，提出了 WideResNet 网络在公开数据集上得到了更好模型效果[34]。这两个团队提出的方法均是考虑在保证识别效果的情况下，如何通过拓宽神经网络的方式减少网络层数，从而减少网络参数和模型大小。而 Huang G 等人提出的模型则希望特征重复利用减少的问题来改进残差网络，具体的方法是将每层网络都与前面所有网络层相连接，保证特征信息的传递从而提高特征利用率[35]。

1.2.3 视觉手势识别

交互手势依照用途和使用情境的不同，可以分为交互性手势和操纵性手势[3]。交互性手势指的是与表达相关的有意识或无意识的姿态，如手语、日常沟通手势、演讲辅助手势等；操纵性手势通常有明确的交互对象，如敲击键盘的手势、拿取物品的手势。而视觉手势识别任务的流程，通常可以划分为手势检测（或分割）和手势识别两个部分，如图 1-2 所示。

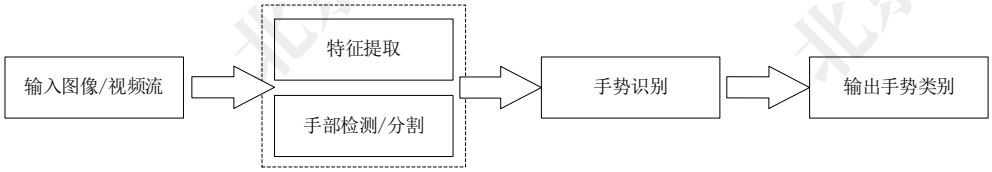


图 1-2 视觉手势识别的一般流程

Fig. 1-2 General Flow of Visual Gesture Recognition

通过视频采集设备获取输入数据，根据动态手势、静态手势的不同，输入的可能是图像或者视频流。对输入数据进行手势检测（或分割）和特征提取，这一步主要将背景信息从输入数据中剔除，提取出手势特征、动态手势起止位置等信息。然后对提取出的手势进行分类识别，并输出识别结果。按照手势识别时使用的方法不同，可以分为基于图像或视频的手势检测识别和基于姿态的手势检测识别两种。

基于图像或视频流的方法在检测出手势所处的空间和时间区间的基础上，采

用传统图像处理方法或深度学习方法，对输入图像进行特征提取并根据提取出的特征进行分类。常用的传统图像处理方法主要有动态时序规整（Dynamic Time Warping, DTW）[36]、隐马尔可夫模型（Hidden Markov Model, HMM）[37]、条件随机场（Conditional Random Fields, CRF）和随机森林（Random Forest, RF）[38]方法。DTW 是一种实现简单的模板匹配算法，不需要通过大量数据进行训练，但是需要设计适当的模板来进行匹配。HMM 和 CRF 方法都是基于统计概率模型的算法，这类算法可以更好地提取动态时序信息。而 RF 算法则是一种常用于决策分类任务的传统机器学习算法，通过集成多个树状分类器来实现特征分类。这些传统机器学习方法对训练数据和计算力需求都不太高可以应用于绝大多数计算平台，但是输出结果的准确率低于深度学习方法。尽管近年来对于传统机器学习方法的研究趋向减少，但在硬件资源受限等特定的应用情况下，传统的模式识别方法依然保持着其不可替代的地位。

使用深度学习方法的手势检测与识别主要使用的有长短期记忆网络（Long Short Term Memory Networks, LSTM）和卷积神经网络（Convolutional Neural Networks, CNN）两种方法。LSTM 能够很好的提取出视频流的时序特征，对于动态手势识别任务有较好的效果。Molchanov 等人利用 C3D 来提取每个视频片段的特征，然后将这些特征输入到卷积神经网络（Recurrent Neural Network, RNN）中提取时序特征，最后将输出矩阵激活后输入到 CTC（Connectionist temporal classification）层中得到手势的出现时间和类别[39]。对于分辨率较大的图像无法直接输入到 LSTM 中，而 CNN 在计算机视觉领域被广泛应用于从图像中提取特征。在手势识别任务中，有研究者将光流信息作为一个新的通道与 RGB 图像组合后作为卷积网络的输入提取特征[40]，在特征层融合两个通道的特征后，输入到分类器中进行手势分类。LSTM 和 CNN 各有优势，将两者结合可以取得更好的识别效果。Zhang 等人提出了一个第一人称视角下的 RGB-D 手势识别数据库 EgoGesture，并且提出了一种循环 3D 卷积神经网络[41]。该网络使用 3DCNN 提取视频切片的特征，然后通过时空特征转换模块 STTM 缓解镜头视角变换的影响，最后利用 LSTM 进行分类。

基于图像的方法在提取特征时往往无法轻易筛选出背景信息，造成对姿态估计和手势识别的干扰，因此不能取得更好的效果。不过，随着 RGB-D 摄像头深度数据的加入，该类方法在背景较为简单，运算资源有限的环境下，仍有很大的应用前景。

基于姿态的手势检测与识别是一种潜力更大，效果更好的手势识别与检测方

法。其对于肢体手势和手部手势的方法略有不同，但大体一致。其核心是特征提取部分按照生理结构提取肢体或手部的关节关键点[42]，估计姿态信息，以消除背景信息对手势识别的影响，能更好的关注于手部的位置与运动等信息。提取出的姿态特征一般表示为关节图网格的结构，因此手势识别部分通常采用循环神经网络（RNN）[4],[43]或图神经网络（Graph Neural Network, GNN）[44]的方法进行分类。

对于姿态的识别，即图像中关键点位置的确定，一般可以分为自顶向下和自底向上两种方式。自顶向下的方法主要是针对单一个体的姿态估计，其深度学习方法主要有两种：直接预测每个关键点坐标和预测每个关键点的热力图（Heatmap）。Google 在 2014 年发布的 Deeppose 就是基于 regressing 的方式，直接使用深度神经网络估计所有关键点的坐标[45]。而基于 Heatmap 的方法在检测效果上大多优于基于关键点回归的方式，比如，Sun 等人设计的 HRNet 先从输入图像中找出人体位置，再在裁剪出的图像中估计每个关键点的 Heatmap[46]。自顶向下的方法在目标检测失灵时无法进行估计，同时时间复杂度也与人数正相关，而自底向上的姿态估计方法可以解耦这个问题，更好地完成多人情境下的姿态估计任务。Openpose 就是自底向上方法中的一个典型[47]，2017 年 Cao 等人提出了一种多人姿态检测方法通过亲和力场（Part Affinity Fields, PAF）来学习身体的各部分关键点之间的关联性，利用全局纹理信息，先检测出全部的关键点再划分不同个体，达到高实时性和高精度。

以上的姿态估计方法大多都是 2D 的姿态估计，但对于一些特殊手势，如交警手势，手势主体的朝向等信息也是手势所传达意图的重要判断依据，所以需要使用 3D 体态估计模型，其中最典型的肢体模型是多人线性蒙皮模型（Skinned multi-person linear model, SMPL）[48]以及同系列的 SMPL-X[49]等。SMPL 模型使用 pose 和 shape 两个参数定义了人体的形状和姿态，依据估计出的关节点坐标使用 3D mesh 对估计的人体姿态进行了蒙皮，从而生成人体的 3d 模型。而 SMPL-X 则是在 SMPL 的基础上细化手部姿态和面部表情。同时，Pavlakos 等人还提出了一种基于 Openpose 实现的算法 SMPLify-X，在估计出 2D 姿态的情况下拟合 3D 模型，并通过改进损失函数和优化方法以提高推理效率[49]。对于手部姿态，也有许多使用类似思路实现的手部 3D 姿态估计，取得了很好的效果，如 Chen 等人基于关节热力图实现的相机空间中的手部 3D Mesh 拟合[50]，Huang 等人参考 Transformer 结构设计的 Hand-Transformer[51]等。

手势识别部分的主要任务是根据估计出的姿态和提取出的特征进行推理和分类。Chen 等人基于手势的时序骨架信息，通过双向循环神经网络处理运动特

征实现手势识别[52]。近年来，注意力机制在计算机视觉领域受到了越来越广泛的关注，据此有研究者将其修改后用到了手势识别任务中。Hou 等人提出了 Spatial-temporal Attention Res-TCN[53]，对主干卷积网络中每一步输出的特征生成一个权重，从而实现了注意力机制。

在此基础上还有一些应用驱动的手势识别方法，针对多模态、情境先验知识等问题进行了针对性的设计和优化[54],[55]。但这些方法都存在上下文情境信息考虑不足，涉及手势或应用场景单一，泛用性差等问题，难以推广到日常手势应用场景或其他复杂手势识别任务中。

1.3 本文主要研究内容

本文主要致力于解决在手势识别过程中如何有效利用领域先验知识的问题，从手势知识图谱的设计与构建、基于关键点的视觉手势检测与姿态估计算法、知识驱动的视觉手势识别算法三个方面开展研究。通过设计通用性的手势知识本体，从而自顶向下构建知识图谱。同时，本文基于关键点的视觉手势识别算法，设计一种在各应用场景下通用的识别算法。最终，将知识图谱中的先验知识特征与手势识别算法进行融合优化，利用先验知识提高手势识别的效率与准确率。

具体来说，本文的主要工作包括：

（1）提出具有参考意义的自然交互手势本体的基本结构，并设计了针对性的知识抽取算法。

（2）基于开源数据集及常识知识图谱，应用本文设计的算法构建具有丰富实体节点和关系的手势知识图谱。

（3）设计基于关键点的自底向上的多头姿态估计网络，对其进行轻量化调整，并设计相应的前后处理算法以及损失函数等，从而高效快速的推理出目标的姿态骨架。

（4）设计网络模型提取手势知识特征与姿态骨架特征，并使用集成学习的方法，将融合特征用于混合场景下的视觉手势识别任务，提升识别的准确率以及算法的泛用性。

（5）基于本文提出的算法，结合无人驾驶车载场景，开发一个视觉手势识别系统，验证算法在实际应用场景下的可行性以及开发过程。

1.4 本文组织结构

本文一共包含六个章节，每个章节的内容安排如下：

第 1 章：绪论。介绍了手势识别技术的研究背景、意义，重点介绍了手势识别在自然人机交互中意义与前景，以及关于场景上下文语义理解和视觉手势识别的国内外研究现状和研究趋势，并对本文的整体研究内容与结构进行了概要性的描述。

第 2 章：知识驱动的视觉手势识别关键技术。围绕本文将要讨论的主要技术架构进行介绍，对知识图谱技术、视觉手势识别技术以及自然人机交互技术展开了详细的讨论。

第 3 章：多模态手势知识图谱的构建技术。设计了手势知识图谱的结构，采用多模态构建方法，融合 Wikidata 知识图谱和手势数据集，创建了包含丰富实体和关系信息的多模态手势知识图谱。

第 4 章：基于关键点的视觉手势检测算法。以轻量化算法需求为出发点设计快速有效的多头人体姿态估计算法，通过设计损失函数和数据预处理，并在部分迁移 MoveNet 网络参数的基础上，用 COCO 数据集进行了训练和测试。

第 5 章：知识驱动的手势识别方法。提出了基于图卷积神经网络的特征提取算法，分别推理知识与姿态特征，融合后用于混合场景下的手势识别任务，并在来自不同应用场景的混合数据集中进行训练和测试，获得了优于现有算法的识别效果。

第 6 章：车载手势识别系统设计与实现。基于本文所设计的手势识别算法，结合无人驾驶车载场景，从需求分析出发，对手势系统进行概要以及详细设计，进而在基于安卓系统的模拟平台上开发了具有完整功能的原型系统。

第2章 知识驱动手势识别的关键技术

通过对基于视觉的手势识别技术算法框架与主要弊端的分析和研究，结合国内外相关的研究现状，本文通过知识图谱补足先验知识与语义分析的部分来完善手势的识别与检测任务。本章对知识驱动手势识别的关键技术中主要涉及的关键技术，包括知识图谱和视觉手势识别的相关技术进行介绍，为本文后续采用合适的算法构建视觉手势识别算法框架提供技术支撑。

2.1 知识图谱相关技术

本小节将对本文中将要使用的知识图谱相关技术进行简单的介绍，主要包括知识图谱的构建和知识的表示学习算法两个部分。

2.1.1 知识图谱的构建

在知识图谱概念问世之前，研究人员就展开了对语义网络技术的探索。Berners-Lee 提出一种数据链接的思想，使用该思想的方法可以对技术标准 RDF (resource description framework)进行了进一步的完善，此外还有 URI (uniform resource identifier) 和 OWL (Web ontology language) 等技术标准也得到了扩展和完善，为后来的知识图谱提供了技术基础条件。知识图谱由实体、属性和关系三部分组成，其中实体代表现实世界中的具体或抽象概念，属性代表实体的特征或属性，关系则表示实体之间的连接和交互。通过将这些三元组导入到计算机系统中，就可以通过知识图谱中的实体、属性和关系之间的关联关系来建立各种类型的语义关系[7]。

知识图谱架构包含两部分内容：知识图谱的逻辑架构及体系架构。

知识图谱的逻辑架构可以进一步细分为三个层次：底层、中间层和顶层。底层是数据层，其中包含了海量的实体、关系和属性值等基础信息。中间层是模式层，其中包含了本体库和推理引擎。本体库用于规范实体及实体属性、关系和关系类型等，推理引擎则用于进行逻辑推理和知识发现。顶层是应用层，基于中间层的知识和底层数据进行开发和应用的层次，主要包括搜索引擎、推荐系统、机器翻译、自动驾驶、医疗诊断等应用场景。

知识图谱的体系架构分为三个主要部分：获取源数据、知识融合和知识计算

与应用[8]。构建知识图谱的首要步骤是数据获取，其中数据来源包括公共数据库、网络爬虫和人工标注等。知识融合是将来自不同数据源的多种形式的知识整合在一起，消除冗余和矛盾，并对不同来源的实体和关系进行对齐和映射。知识计算与应用是指利用知识图谱的存储特性和图形查询进行推理、计算和应用的过程。

对知识图谱的构建通常可以选择两种方式进行，包括自顶向下和自底向上两种方法。在自顶向下的方法中，通过对先验知识的了解和人工构建的本体，可以对特定领域的实体和关系进行抽取和填充，以形成知识图谱。通过这种方式构建的知识图谱结构清晰、语义准确性高，但是需要大量的专家指导和知识体系参考，要花费大量的时间和精力。另一种自底向上的方法，在大量的非结构化数据中对实体和关系进行自动化的抽取，再进行本体的构建和知识图谱的组织。这种方法的优点是可以处理大规模的数据，节省了人力成本，但是缺点是知识图谱的准确性可能会受到非结构化数据噪声和错误的影响。知识抽取指从大规模文本数据中自动识别、提取、转换并导入到知识库中的过程，是知识图谱构建中的关键技术。通过知识抽取可以将各种不同形式的数据源整合起来，并将它们转化为结构化的知识图谱，为知识的理解和推理提供更好的基础。

知识抽取的过程通常包括以下三个步骤：实体识别、关系抽取和事件抽取。其中，实体识别是指从文本数据中自动识别出命名实体，如人名、时间、地名等，并将其标注为预定义的实体类型。实体识别通常使用基于统计的机器学习算法或基于规则的方法，以提高准确率。关系抽取是指从文本中挖掘实体之间存在的各种关系，例如基于模板的关系抽取需要人工指定一些模板或规则，从而自动提取出相应的关系。而基于机器学习的关系抽取则是从已标注的数据中学习关系抽取模型，以识别新的关系。事件抽取是指从文本中自动识别出事件，并提取事件涉及的实体及其关系，是关系抽取的一种扩展形式，其目的是识别出事件的发生和事件的影响因素。

除了上述关键技术之外，还有一些附加的技术，如实体链接、关键词提取和情感分析等。实体链接是指在知识图谱中匹配与文本中的实体对应的实体，进行链接，以增强知识库的实体信息。关键词提取则是从文本中自动识别出重要的关键词，以帮助人们更好地理解文本。情感分析则是指从文本中自动识别出情感极性，如正面、负面或中性等，以帮助人们更好地理解情感信息。

2.1.2 知识表示学习

在构建好的手势知识图谱后，可以通过一系列操作完成知识的检索，从而进

一步分析推理得到手势的知识特征。而其中关键问题是为手势知识图谱中存储的知识与多模态实体，确定一个合适的、可以用于知识特征表示的编码方法，用于统一表示融合后的多模态知识特征。本质是将知识图谱中存储的文本、图像实体以及之间的关系转换为低维、稠密的实值向量，提高计算效率，便于迁移和融合。

常见的模型有：距离模型、翻译模型、语义匹配模型等，本节分别介绍其中具有代表性的一个模型。

（1）距离模型：基于距离的模型理解上非常直接，通过数学模型直接计算实体间的距离。TransE（Translation-based Embedding Model）是一种经典的知识图谱表示学习模型[56]，它基于一个简单的假设：关系可以被表示为实体之间的平移（translation）。该模型将实体和关系映射到低维空间中的向量，如图 2-1 所示，然后通过最小化头、尾实体之间的关系向量与关系向量的距离来学习实体和关系的表示。

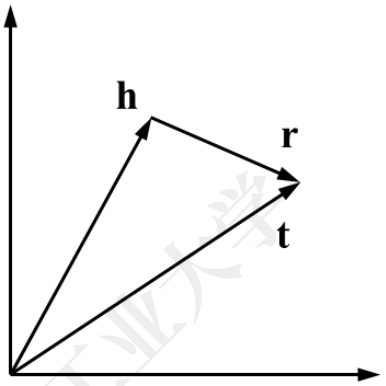


图 2-1 实体和关系向量空间示意图

Fig. 2-1 Schematic Representation of Entity And Relation Vector Spaces

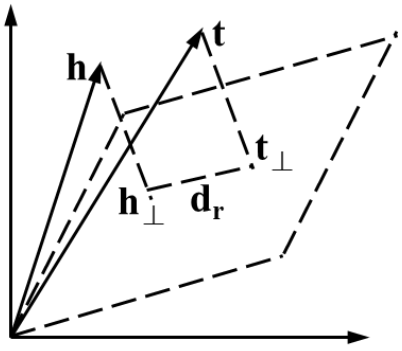


图 2-2 实体向量映射到投影矩阵

Fig. 2-2 Mapping Entity Vectors to Projection Space

（2）翻译模型：TransH（Heterogeneous Space Translation Model）是另一个知识图谱表示学习模型，旨在克服 TransE 的一些缺点，特别是在处理不同类型

的关系时[57]。TransH 引入了关系特定的投影矩阵，将实体从原始空间映射到关系特定的空间，如图 2-2 所示，以更好地捕捉不同关系之间的语义信息，这样的映射过程也被称作翻译（Transfer）。

（3）语义匹配模型：这类模型不关注模型的可解释性，通常使用深度学习的方法。例如 ConvE（Convolutional Embeddings for Link Prediction）是一种利用卷积神经网络进行知识图谱表示学习的模型[58]。它将实体和关系表示为向量，并将它们组合成三元组，然后通过卷积操作来学习实体和关系之间的语义匹配。ConvE 模型在处理知识图谱中的链接预测任务时表现出色。

2.2 实体抽取相关技术

实体抽取是自然语言处理领域中一项关键的信息抽取任务，旨在从给定的文本语料中识别和提取出具有特定类型和语义含义的命名实体。这些命名实体可以是人名、地名、组织机构名、日期时间等具有独特身份和指代意义的实体，在文本中具有重要的信息价值。实体抽取技术基于自然语言处理和机器学习方法，通过利用语法、语义、统计模型等手段，对文本进行结构化处理和分析，从而准确地识别和提取出文本中出现的命名实体，并将其分类到预定义的实体类别中。通过对文本进行实体抽取，可以有效识别和提取出这些类别的命名实体，有助于深入理解文本内容并支持各种自然语言处理任务的进行。

传统的文本实体抽取解决方法中，序列标注是一种常用的技术，通过为文本序列中的每个单词标注其所属的命名实体类别来实现实体抽取。在实际的实体抽取任务中，为了帮助模型准确判断命名实体的开始和结束，通常会使用特殊字母来标记这些位置。常用的标注方法包括“BIO 标注”和“BIOES”标注。在 BIO 标注中，每个标记单元（通常是单词）被标记为三种类型之一：B（Beginning）、I（Inside）、O（Outside）。具体而言，如果一个标记单元是某个实体的起始单元，则被标记为 B-实体类型；如果一个标记单元是某个实体的内部部分，则被标记为 I-实体类型；如果一个标记单元不属于任何实体，则被标记为 O。BIOES 标注是在 BIO 标注基础上的扩展，引入了额外的两种标记：E（End）和 S（Single）。在 BIOES 标注中，最后一个单元用 E-实体类型标记，如果一个实体仅包含一个单元，则用 S-实体类型标记。

随着应用中对多模态知识图谱的需求逐步显露，图像实体抽取技术的重要性日益凸显。图像实体抽取技术能够从复杂的图像信息中准确地识别和提取出关键实体，为多模态知识图谱的构建提供丰富的视觉信息。这种技术通常包括两个主

要方面：目标检测和实例分割。目标检测旨在定位图像中特定类别的目标，并提供其边界框的位置信息。而实例分割则将目标检测进一步扩展。图像数据实体抽取的关键挑战之一是对图像中复杂场景和多样目标的准确识别，这需要深度学习模型具备强大的特征表示和语义理解能力。

实体抽取的方法主要有以下三种：

（1）规则驱动的实体抽取：这种方法基于人工编写的规则来识别和抽取实体。规则可以基于词性、语法结构、关键词匹配等来确定实体位置，是一种传统的实体抽取方法。这种方法适用于特定领域或任务，但需要耗费大量人力和时间来设计和维护规则。

（2）基于机器学习的实体抽取：这种方法使用传统机器学习算法来训练模型从文本中识别实体，如支持向量机（SVM）、随机森林（Random Forest）、神经网络等。特征可以包括词性、词向量、上下文信息等。传统的序列标注模型如隐马尔可夫模型（HMM）和条件随机场（CRF）也常被用于实体抽取任务。

（3）基于深度学习的实体抽取：随着深度学习技术的发展，基于深度学习的文本实体抽取方法取得了显著进展。这种方法利用深度神经网络，如循环神经网络（RNN）、长短时记忆网络（LSTM）、注意力机制和预训练模型（如 BERT、GPT）等，可以更好地捕捉文本中的语义和上下文信息，从而提高实体抽取的准确性和泛化能力。在图像实体抽取任务中，基于深度学习的目标检测算法，如 Faster R-CNN、YOLO（You Only Look Once）、SSD（Single Shot MultiBox Detector）等，可以用于实体物体的检测和定位。这些算法能够在图像中准确地标定出物体的位置并进行分类，从而实现图像中实体的抽取。

本节将逐个介绍本文涉及的实体抽取的相关技术。

2.2.1 循环神经网络 RNN

循环神经网络（Recurrent Neural Network，RNN）是一种专门用于处理序列数据的神经网络结构[39]。RNN 具有记忆功能，能够保持前一时刻的状态，并将其作为当前时刻的输入，从而更好地处理序列数据中的时间相关性信息。

RNN 的基本结构包括一个循环连接，使得信息可以在网络内持续传递。在每个时间步，RNN 会接受当前输入和前一时刻的状态作为输入，输出当前时刻的状态，并将其传递到下一个时间步。这种设计使得 RNN 能够对不定长度的序列数据进行建模，并具有一定的记忆能力。

然而，标准的 RNN 存在着梯度消失和梯度爆炸的问题，导致难以处理长期

依赖关系。为了解决这一问题，有一些改进型的 RNN 结构被提出，如长短期记忆网络（Long Short-Term Memory，LSTM）和门控循环单元（Gated Recurrent Unit，GRU）等。这些改进型的 RNN 结构通过引入门控机制来更好地捕捉长期依赖关系，提高了 RNN 在处理序列数据中的性能。

2.2.2 长短期记忆网络 LSTM

长短期记忆网络（LSTM，Long Short-Term Memory）[40]是一种特殊的循环神经网络（RNN）结构，专门设计用来解决 RNN 中的梯度消失和梯度爆炸问题，以更好地捕捉长期依赖关系。相比于传统的 RNN 结构，LSTM 具有三个关键的门控单元，分别是遗忘门（forget gate）、输入门（input gate）和输出门（output gate），以及一个细胞状态（cell state）。这些门控单元能够有效地控制信息的流动，从而实现对长期依赖关系的建模和记忆。模型结构图如图 2-3 所示。

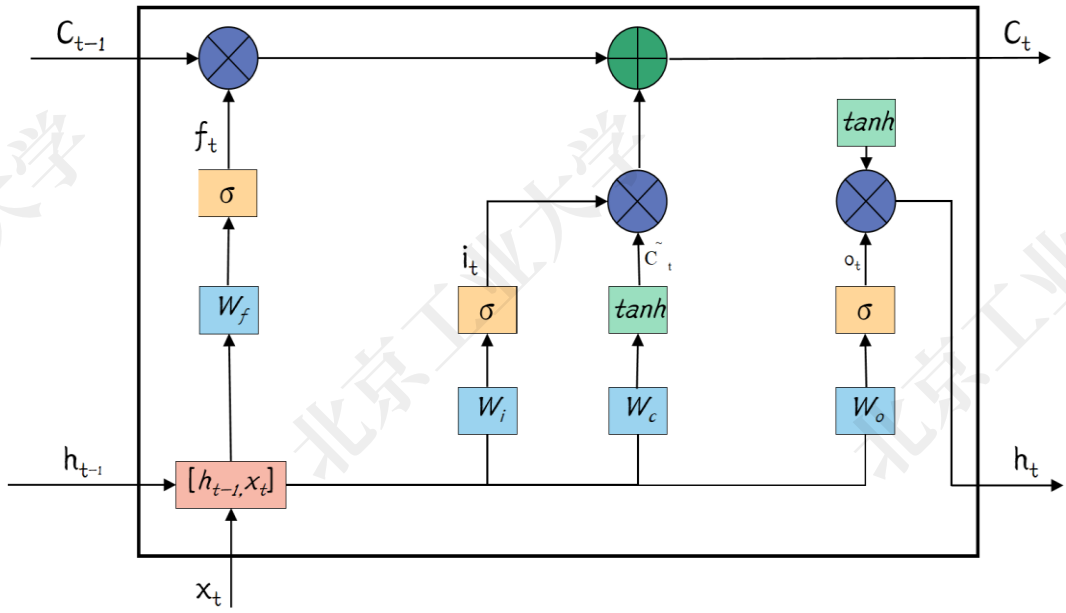


图 2-3 LSTM 网络模型结构图

Fig. 2-3 Structure of LSTM Network Unit

尽管 LSTM 在处理时间序列数据时能够维持信息的长期记忆，但并不能够充分地考虑上下文信息。LSTM 仅依赖于当前时刻之前的信息，无法充分利用序列中后续时刻的信息，导致模型的表示能力受限。为了克服这一缺点，引入了双向长短时记忆网络（BiLSTM）。BiLSTM 在 LSTM 的基础上进一步增强了模型对上下文信息的获取能力。通过同时考虑正向和反向的序列信息，结合了前向和后向 LSTM 结构的双向上下文的特征，BiLSTM 能够更全面地捕获序列中时间

依赖的结构，并提高了模型对输入序列的理解和表征能力。BiLSTM 模型的结构图如图 2-4 所示。

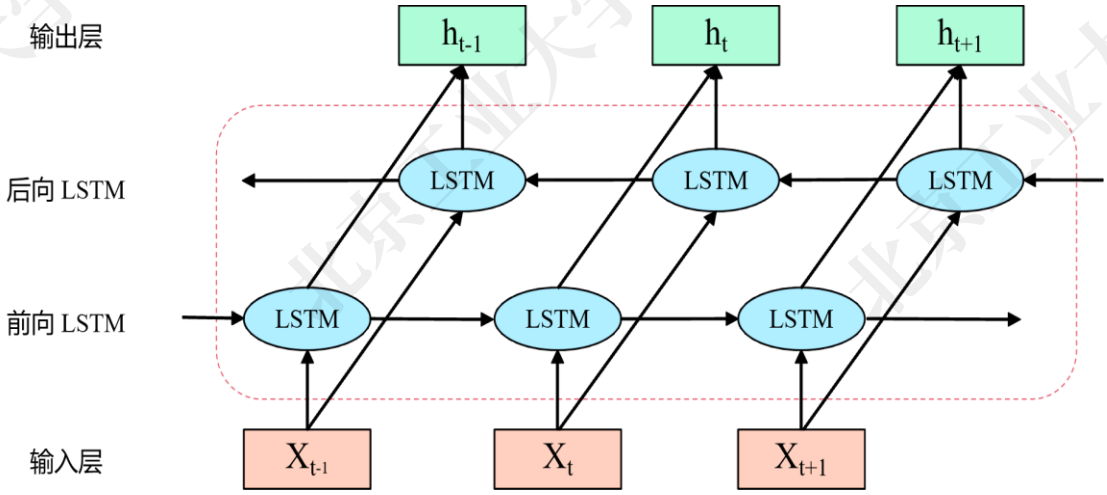


图 2-4 BiLSTM 网络结构图

Fig. 2-4 Structure of BiLSTM Network

具体的，LSTM 单元处理和更新数据的流程如下：

（1）输入门的作用是选择性地更新细胞状态，用于捕捉当前时刻的重要信息。通过 sigmoid 激活函数控制将新信息添加到细胞状态中的程度。输入门的计算公式如下：

$$i_t = \sigma\left(W_i \bullet \left[h_{t-1},\ x_t\right] + b_i\right) \tag{2-1}$$

输入门接收上一个时间步的隐藏状态 ht-1 和当前时间步的输入 xt 作为输入，经过权重矩阵 Wi 的线性变换和偏置项 bi 的加权求和后，得到一个介于 0 和 1 之间的门控参数 i t，i t 作为输入门的输出。

（2）遗忘门控制细胞状态中哪些信息需要被舍弃，以适应序列数据中的变化。通过 sigmoid 激活函数输出介于 0 和 1 之间的值来决定细胞状态中哪些信息需要保留，哪些旧的信息需要被遗忘。遗忘门决定了哪些信息需要被遗忘掉。它类似于输入门，也包括一个 sigmoid 激活函数，用于输出介于 0 和 1 之间的值来决定细胞状态中哪些信息需要保留。遗忘门的计算公式如下：

$$f_t = \sigma\left(W_f \bullet \left[h_{t-1},\ x_t\right] + b_f\right) \tag{2-2}$$

遗忘门接收上一个时间步的隐藏状态 ht-1 和当前时间步的输入 xt 作为输入，经过权重矩阵 W_f 的线性变换和偏置项 b_f 的加权求和后，得到遗忘门的输出 f t。

（3）细胞状态是 LSTM 网络中用来传递信息的主干，负责在不同时间步之间传递信息，并且经过适当的门控单元调节信息的传递和处理。

（4）输出门控制着从当前时间步的细胞状态到隐藏状态的信息流，使用一个 Sigmoid 激活函数来过滤细胞状态，并结合 Tanh 激活函数对于当前时间步的细胞状态信息进行编码。输出门和当前时刻的隐藏状态的计算公式如下：

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (2\text{-}3)$$

$$h_t = o_t * tanh(C_t) \quad (2\text{-}4)$$

输出门接收当前时间步的输入数据 xt 和前一时间步的隐藏状态 ht-1 作为输入，经过权重矩阵W_o和偏置项b_o的线性变换后，通过 Sigmoid 激活函数产生一个[0,1]之间的门控值。根据门控值过滤细胞状态信息，并结合 tanh 激活函数输出当前时间步的隐藏状态。

2.2.3 条件随机场 CRF

条件随机场（Conditional Random Field，简称 CRF）是一种概率图模型，在自然语言处理领域被广泛应用[59]。CRF 模型能够在推断时考虑整个标注序列的依赖关系，从而保证标签序列的一致性，减少标签错误，提高了标注的准确性。另外，CRF 模型允许设计多样化的特征函数来表达标注序列的特征信息，包括观察特征和转移特征等。通过充分挖掘数据中的各种特征关系，CRF 模型能够更好地适应不同任务的需求，并提升标注准确性和泛化能力。

在序列标注任务中，鉴于输入和输出序列均呈线性结构，因此广泛采用线性链条件随机场（Linear Chain Conditional Random Fields）进行建模。设$X = (X_1, X_2, \dots, X_n)$和$Y = (Y_1, Y_2, \dots, Y_n)$均为线性链表示的随机变量序列。若在给定随机变量序列 X 的条件下，随机变量序列 Y 的条件概率分布 $P(Y|X)$构成条件随机场，即满足马尔可夫性，即 $P(Y_i|X, Y_1, Y_2, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) = P(Y_i|X, Y_{i-1}, \dots, Y_{i+1})$，其中$i = 1,2, \dots, n$（在$i = 1$ 或n时，只考虑单边情况），则称 $P(Y|X)$为线性链条件随机场，链式 CRF 图结构如图 2-5 所示。

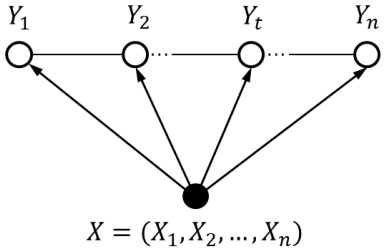


图 2-5 链式 CRF 图结构

Fig. 2-5 Structure of the Chain CRF Diagram

在序列标注任务中，X 通常表示输入观测序列，Y 表示对应的输出标记序列

或状态序列。在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率具有如下形式：

$$P(Y=y|X=x)=\frac{1}{Z(x)}exp\left(\sum_{l,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \quad (2\text{-}5)$$

其中 $Z(x)$是归一化因子，t_k 和 s_l 是特征函数，λ_k 和 μ_l 对应的权重。

归一化因子 Z(x)的公式如下所示。

$$Z(x)=\sum_y exp\left(\sum_{l,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \quad (2\text{-}6)$$

其中求和范围为所有可能的标注序列 Y。归一化因子 $Z(x)$是所有可能的标注序列的联合概率之和，用于确保条件概率 $P(Y|X)$的总和为 1。

对于输入来说，CRF 模型的实现的功能就是搜索概率最大的 y*满足如下公式。

$$y^*=argmax\,P\,(y/x) \quad (2\text{-}7)$$

2.2.4 目标检测算法 YOLOv8

YOLOv8 是 Ultralytics 公司最新推出的 YOLO 系列目标检测算法[60]，具有广泛的应用领域，包括图像分类、物体检测和实例分割等任务。它建立在 YOLO 系列历史版本的基础上，并引入了新的功能和改进点，以进一步提升性能和灵活性。

具体而言，YOLOv8 中创新性的结构包括骨干网络、Ancher-Free 检测头和损失函数，这些创新使得 YOLOv8 能够在 CPU 到 GPU 的多种硬件平台上高效运行。此外，YOLOv8 还具有可扩展性，不仅仅能够用于 YOLO 系列模型，而且能够支持非 YOLO 模型以及分类分割姿态估计等各类任务。YOLOv8 还融合了 BiFPN 网络，这种融合显著提升了性能。BiFPN 网络的主要思想是高效双向跨尺度连接和加权特征融合，这有助于更好地表示和处理多尺度特征，从而提高目标检测的准确性。YOLOv8 模型的网络结构如图 2-6 所示。

在实际应用中，YOLOv8 展现出了显著的性能优势。相较于之前的版本，它在物体检测的精度上有所提升，能够更准确地检测和定位目标物体。同时 YOLOv8 仍然能够保持较快的检测速度，这对于实时应用非常重要。此外，YOLOv8 还表现出强大的通用性，在不同场景和数据集上都表现出色，具备较强的泛化能力。

然而，YOLOv8 也存在一些局限性。由于具有更深的网络结构和更多的参数，其训练时间相较于之前的版本会更长。同时，由于网络结构更复杂，YOLOv8 需

要更高的计算资源才能实现较好的性能。为了达到更好的性能，YOLOv8 通常需要更多的标注数据进行训练。

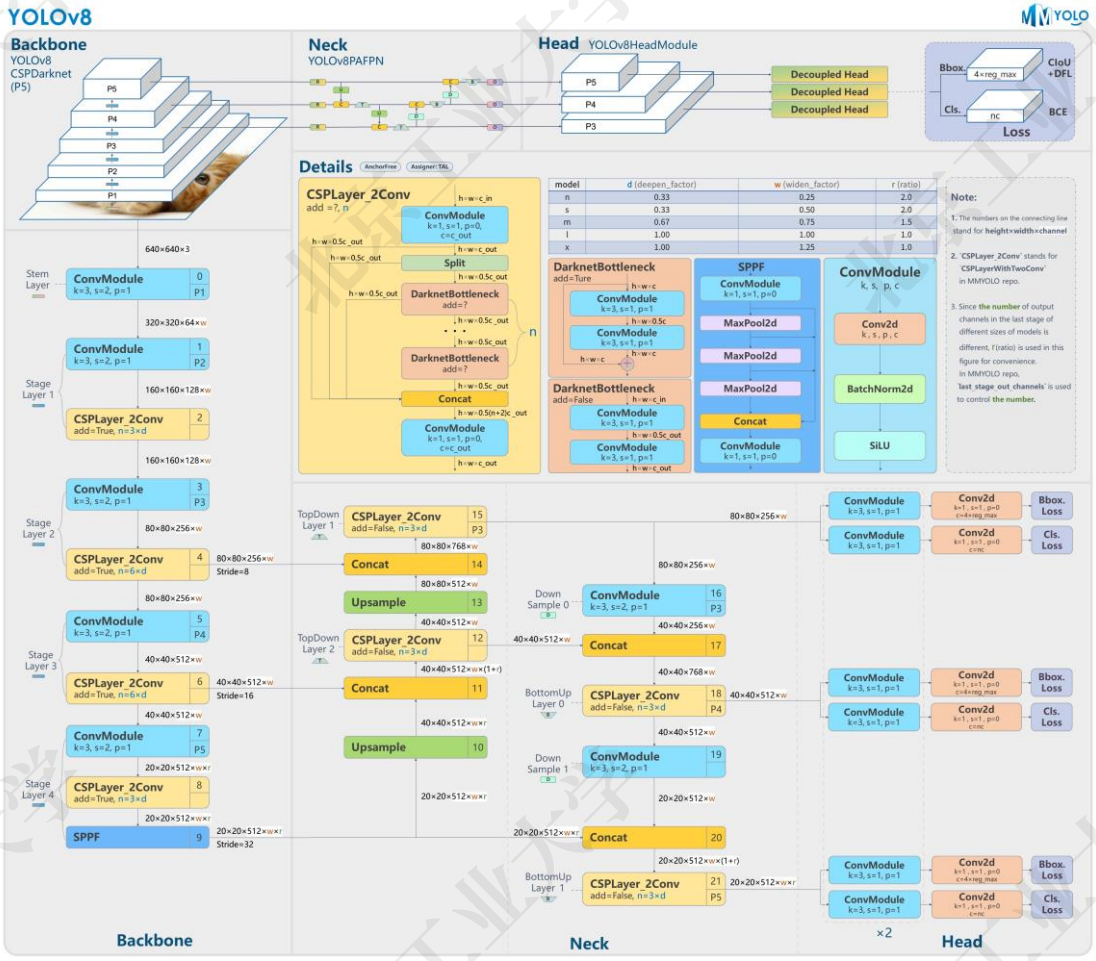


图 2-6 YOLOv8 网络结构图[60]

Fig. 2-6 Structure of YOLOv8

总的来说，YOLOv8 是一款功能强大、性能卓越的目标检测算法，适用于各种实际应用场景。本文将其用于知识图谱的构建，通过对场景图像进行目标检测，抽取其中的对象实体存储到知识图谱中。

2.3 视觉手势识别技术

一般的视觉手势识别方法主要包括姿态估计算法和手势分类方法两个关键部分，本小节将分别对这两部分的方法技术进行介绍。

2.3.1 基于关键点检测的姿态估计

姿态估计算法的主要目的是将从摄像头输入的图像流分析转化为人体关键点与姿态骨架图，用于手势动作的表示以及手势意图的分类任务。因此，姿态估计方法的输入通常为摄像头等输入设备采集到的图像信息，经过一系列算法模型处理，输出预设的各类关节点（如手、肘、肩等）的坐标，以及关节之间的连接关系（如小臂、大臂、躯干等，隐含于关节类别信息中），这种关节点的预设通常由所使用的数据集决定。

常见的姿态估计算法主要分为自顶向下和自底向上两个大类。其中，自顶向下的姿态估计算法出现的比较早，先进行目标检测找到被检测的人体（或面部、手部等），再通过卷积神经网络（CNN）等模型进行黑盒回归，得到目标的各类关节点坐标。这一类方法依赖于高效、准确的目标检测算法，且对于多人场景、人体遮挡等问题处理起来非常烦琐吃力。

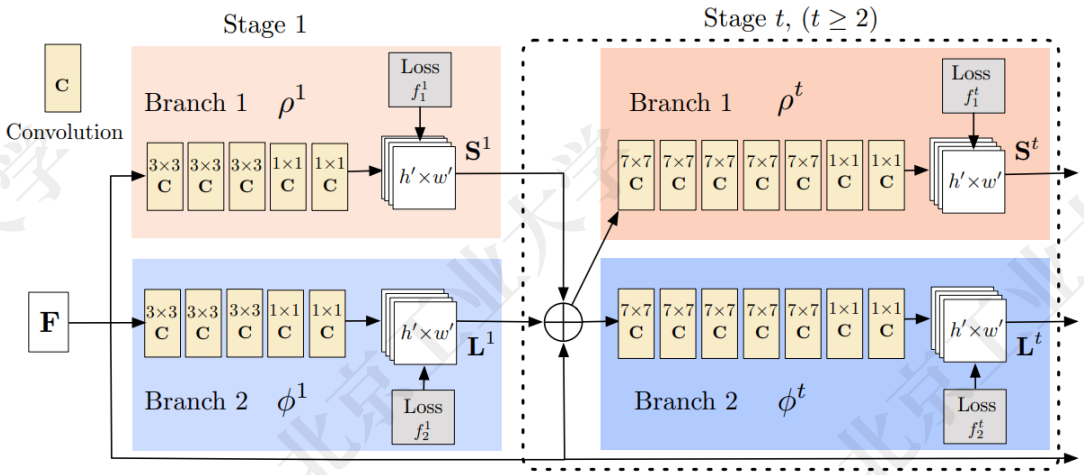


图 2-7 OpenPose 网络结构图[47]

Fig. 2-7 Structure of Openpose

而自底向上的姿态估计算法在 2017 年的 OpenPose 中被首次提出，具有端到端、高效、快速等优点，在绝大多数应用场景下都取得了优于自顶向下方法的结果[47]。自底向上的方法，即先推理出输入图像中所有目标的各类关节坐标，再将这些关节坐标按照类别和位置等信息进行分组组装，最后得到多人关节骨架图。

具体来说，OpenPose 以 VGG19 作为主干网络提取图片特征，然后迭代地进行推理，如图 2-7 所示，分别得到输入人体图像的关节亲和场（PAF，图中蓝色部分）和关节热力图（PCM，图中橙色部分）。在 PAF 和 PCM 之间，通过使用中间监督减少因层数过多导致的训练过程中的梯度消失。

2.3.2 基于图卷积网络的手势分类

在取得每一帧图像中的人体骨架后，按照时序进行拼接，可以得到一个连续的动态人体骨架网络图。通过对这个图结构的分析和推理，可以对其中包含的手势动作进行分类，识别出用户的交互意图。

对于像时序人体骨架图这样的图结构数据，通常采用图卷积神经网络（GCN）来提取其中包含的特征信息[61]，这是最早提出的一种可以在图结构上执行卷积操作的方法，本质思想与图像卷积相似，都是从周围提取信息然后通过执行某种操作（如计算平均数），进而获得新信息。GCN 模型以全部节点特征构成的特征矩阵X和表示图结构的邻接矩阵A作为输入数据，在一个基于分层传播规则的 L 层 GCN 中层与层之间的传播方式如下式所示。

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{2-8}$$

其中，$\tilde{A} = A + I_n$，I_n是一个 N 维单位矩阵，用于在邻接矩阵中添加对角线元素表示指向自身的边，从而在与特征矩阵相乘时保留节点自身特征；$\tilde{D}$是$\tilde{A}$的度矩阵（degree matrix），其对角线元素为邻接矩阵对应行 1 的个数，即$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$，其目的是消除不同顶点度不一样导致的权重不同，故而对$\tilde{A}$做标准化处理；$W^{(l)}$表示第l层的权重矩阵；$\sigma(\cdot)$表示非线性激活函数，通常使用 ReLU 等；$H^{(l)} \in \mathbb{R}^{N*D}$为第$l$层的激活矩阵，即层与层之间传递的参数，特殊的$H^{(0)} = X, H^{(L)} = Z$。

需要注意的是，GCN 中的一层操作，会将图中一阶邻域的信息传递给当前节点。因此，GCN 的层数不宜过多，通常选用 2~4 层的图卷积操作。

2.4 本章小结

本章介绍了知识图谱、实体抽取和视觉手势识别中的关键技术要点和相关研究基础。其中，第一小节重点说明了知识图谱的体系架构、知识图谱的构建和抽取以及知识的表示学习，这将为本文第三章关于手势知识图谱的构建和应用提供技术基础。第二小节对实体抽取过程的常见方法进行描述，并详细介绍了与实体抽取相关技术，主要有：循环神经网络、长短期记忆网络、条件随机场和目标检测算法 YOLOv8。第三小节的视觉手势识别技术详细介绍了以 OpenPose 为代表的自底向上的姿态估计算法和基于图卷积网络 GCN 的手势分类技术，为本文手势识别框架的设计和搭建提供了技术支撑和理论基础。

第3章 多模态手势知识图谱的构建技术

知识驱动的视觉手势识别技术的基础任务是获取手势的知识特征，而知识图谱是一项被广泛验证并使用的，存储先验知识数据的有效方法，因此本文选择知识图谱存储视觉手势先验数据。

3.1 多模态手势知识图谱本体设计

本体（Ontology）设计是对知识图谱结构进行设计的关键步骤，是对知识图谱模式层的简要说明。本节将以领域、类别为单位，设计统一的或有公共结构的高泛用性手势本体，再由本体自顶向下进行知识图谱的构建。需要特殊说明的是，为提高手势知识图谱的泛用性，以及本文方法架构的普适性，本文不对手势（Gesture）、动作（Action）、姿势（Pose）做详细区分，三者的具体实例或类似行为均视为一种手势。

结合对手势识别任务的具体分析，本文将使用手势（gesture）、手势主体（subject）、观察者（observer）、手势交互目标（object）以及所处情境（situation）五类实体来描述一个独立的手势或手势类别，实体间的关系如图 3-1 所示。

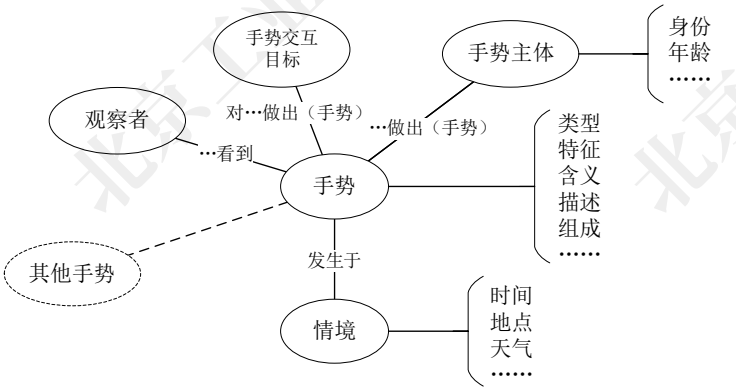


图 3-1 手势知识本体

Fig. 3-1 The Ontology of Gestures Knowledge

手势知识本体在此结构之上扩展和完善各个实体各自的属性，以及各实体之间的具体关系，用以描述手势详细的情况，并以 web 本体语言（Web Ontology Language, OWL）表示为资源描述框架（RDF）。其中最核心的 SPO 三元组是手势主体、“做出”关系、以及手势实体。

手势主体指做出手势的人，在第三视角数据中或肢体手势数据中表现为核心

被检测目标；而在第一人称视角中则被隐藏，需要以动作意图或预设类别进行代指补全。手势实体即为预期检测识别结果提炼出的抽象对象，其属性中应至少包含类型、含义、文字描述等基本信息，以及从数据集中收集到的所有可以表征手势行为的信息。除此之外，观察者、情景、交互目标等实体主要采用预设参数、目标检测等方法，从数据集中自动采集对应手势有可能出现的关联关系。

3.2 数据的获取与预处理

3.2.1 数据的获取

基于手势本体的设计，本文从 Wikidata 知识图谱中提取手势相关实体及关系，并选择 Something-something、Charades、FPHA 三个公开手势数据集组合为混合手势数据集，用于本章节知识抽取算法的设计和实施。

Wikidata 是一个由维基媒体基金会维护的免费开放的联机数据库项目，以语义网络三元组的格式存储来自 Wikipedia 或用户上传的大量数据。Wikidata 的内容涵盖了各种领域，包括人物、地点、事件、艺术作品、科学概念等，以及它们之间的关系。因此需要从中查询出与本文所述的手势相关的实体和关系，并与本文所设计的手势知识图谱相融合，整合为适用于本文后续所述算法的知识语义网络。

```
SELECT DISTINCT ?entity ?entityLabel ?description WHERE {
  {
    ?entity (wdt:P31/wdt:P279*) wd:Q11410 ;  # 手势
            rdfs:label ?entityLabel .
    OPTIONAL { ?entity schema:description ?description FILTER(LANG(?description) = "en") }
  }
  UNION
  {
    ?entity (wdt:P31/wdt:P279*) wd:Q4026292 ;  # 动作
            rdfs:label ?entityLabel .
    OPTIONAL { ?entity schema:description ?description FILTER(LANG(?description) = "en") }
  }
  UNION
  {
    ?entity (wdt:P31/wdt:P279*) wd:Q15117302 ;  # 姿态
            rdfs:label ?entityLabel .
    OPTIONAL { ?entity schema:description ?description FILTER(LANG(?description) = "en") }
  }
}
```

图 3-2 在 Wikidata 中检索手势相关数据的 SPARQL 语句

Fig. 3-2 SPARQL Statements for Retrieving Gestation-Related Data in Wikidata

具体来说，Wikidata 官方提供了基于 SPARQL 语言的查询工具，通过如图 3-2 所示的查询语句可以得到与手势相关的所有实体或关系。这里需要再次强调

的是，本文为增强整体算法的泛用性，不对手势、动作、姿态三者进行详细区分，三者都可以视为本文所指的手势。

Something-something 数据集包含了密集标注的短视频剪辑，每个剪辑都包含与手势相关的动作，数据集的注释是基于描述性标签，描述了在每个剪辑中发生的事情。Charades 数据集中的每个剪辑都是家庭场景下的日常行为，由志愿者根据给定的关键词（动词、名词）造句，再按照描述进行表演，筛选其中涉及人体动作和手势的视频、用于手势知识的抽取。FPHA（First-Person Hand Action Benchmark）是第一人称下手部与社交、厨房、办公室三种生活场景的一些物品的交互动作，虽然不包含完整的文字描述，但可以从对应标签中按照模式提取实体关系。

3.2.2 图像预处理

图像主要包括手势数据集中的视频和 Wikidata 中筛选出的手势实体中包含的手势图像，前者将用于对手势实体的多模态标注以及实体抽取任务，而后者以结构化数据的形式直接与构建的知识图谱进行融合，具体知识融合的处理操作将在 3.4.1 中进行介绍。

由于手势数据集中的图像数据以视频形式存储，因此需要先从中提取有代表性的图像作为手势的图像属性。由于数据集中的手势视频是划分好的手势动作切片或具有起止时间标注，因此不用对数据集中的手势进行检测。考虑到视频切片不一定精准，提取图像时在视频前后各间隔总时长的 10%，在余下的部分中取开头、结尾以及中点三帧图像作为对应手势的图像属性。同时为了保证尽可能完整地保存动态手势的相关细节，将手势视频同样作为手势实体的属性存储到知识图谱中，与图像属性相区别。

另外作为知识抽取的输入图像，同样需要一定的预处理，主要包括缩放至统一尺寸以及将视频流转化为图像序列。为了节省存储资源，这部分的预处理操作在知识抽取的过程中动态完成，不保存处理后的图像。

3.3 多模态手势知识图谱构建

用于构建知识图谱的数据主要包括视频和文字两个部分，因此使用多模态知识抽取方法，过程如图 3-3 所示。这一节将分别针对图像和文本数据的知识抽取进行介绍。

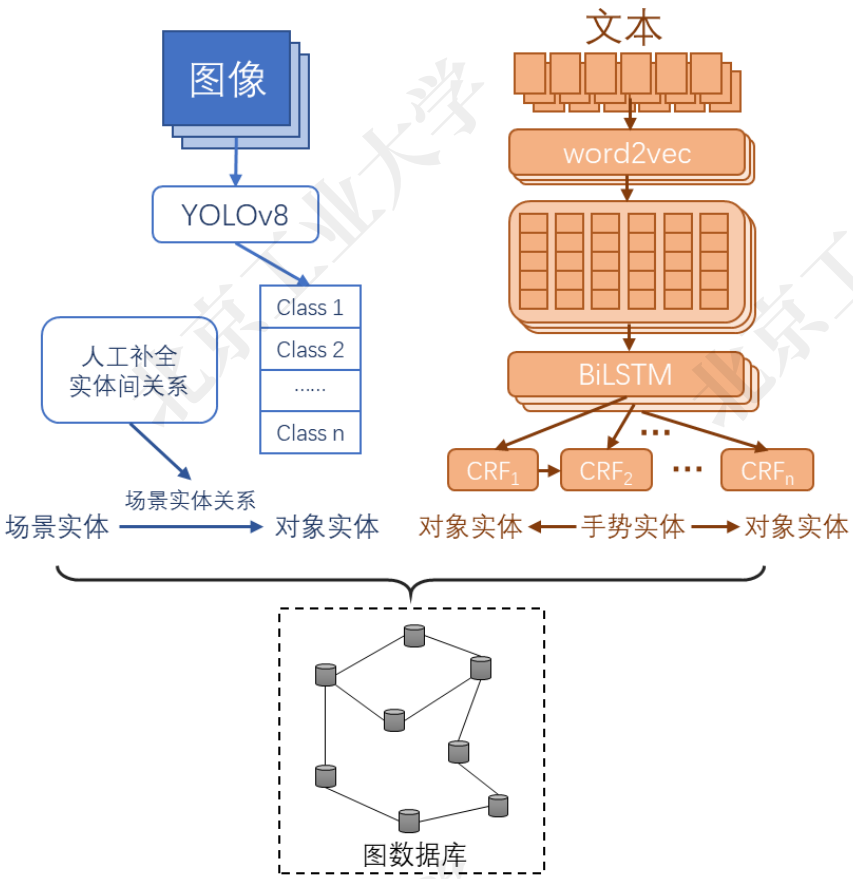


图 3-3 多模态手势知识抽取方法过程示意图

Fig. 3-3 Process of Multimodal Gesture Knowledge Extraction Method

3.3.1 基于文本数据的知识抽取

构建知识图谱所用的文本数据主要来源于手势数据集的标签数据，手势标签通常是以最简练的文字描述视频、图像中的手势动作内容，这种描述通常是由简单的主谓短语、动宾短语构成。因此使用 Word2Vec-BiLSTM-CRF 模型进行命名实体识别，即可提取出这部分文本数据中的实体与关系信息，模型结构如图 3-4 所示。

模型主要由词嵌入层、双向 LSTM 层和 CRF 层组成。

具体来说，词嵌入层使用 Word2Vec 将输入文本转换为词嵌入（Word Embedding），即用特征向量表示各个单词用于后续模型推理。

然后，将词嵌入层输出的特征向量输入到 BiLSTM 层，使用 BiLSTM 提取词嵌入中的特征信息。BiLSTM 即双向长短期记忆网络。相对于单向的 LSTM，双向的网络可以同时捕捉正向信息和反向信息，使得对文本信息的利用更全面，有更好的特征提取效果。

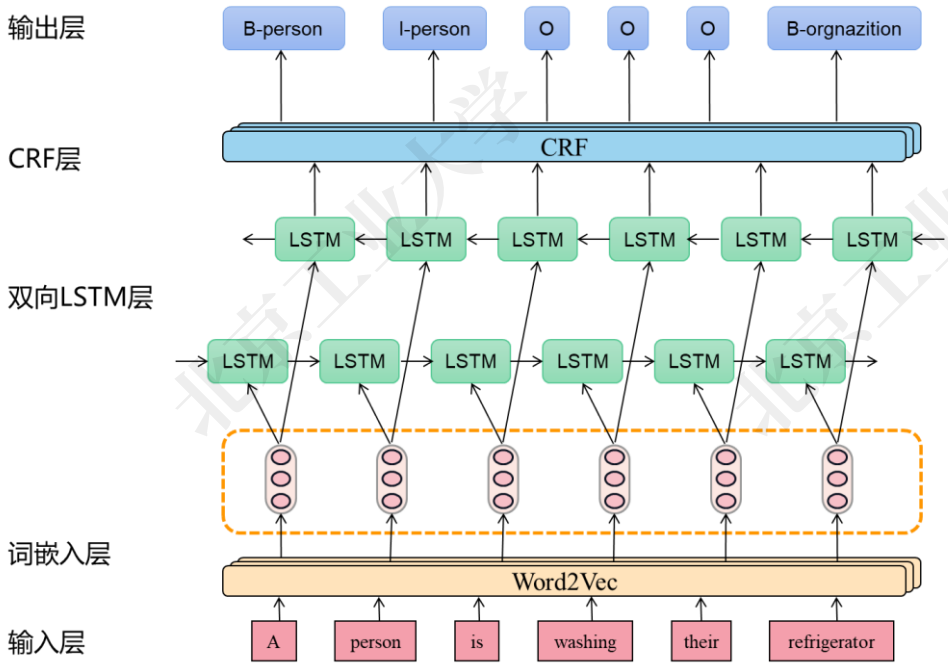


图 3-4 BiLSTM-CRF 模型结构图

Fig. 3-4 Structure Diagram of BiLSTM-CRF Model

BiLSTM 输出的词性标签序列可以直接作为实体识别结果，但其中存在不符合语法逻辑的标签序列，可以使用 CRF 对输出结果进行逻辑约束。经过 CRF 层处理可以获取到输入文本的概率预测值，通过将概率预测值中输出的最大值作为输入文本的预测标签可以大幅提升实体抽取的准确性。

由于文本结构简单，在识别到动词前后的命名实体后，剩余的部分即可作为手势动作的名称，存储为对应的手势实体。

3.3.2 基于图像数据的知识抽取

对于公开手势数据集中的视频和图像，使用深度神经网络模型提取其中的场景对象序列，并通过人工标注的方式补全其中的关系信息。

场景对象序列部分，利用在 COCO 数据集上预训练的 YOLOv8 模型对图像内容进行分析，识别其中出现的实体对象添加到知识图谱中并添加与相应的场景实体之间的关系。

预训练的 YOLOv8 一共可以识别 80 种不同的目标，具体目标种类如表 3-1 所示。在从数据中标签文本数据中提取出相应手势实体和目标对象实体的基础上，通过对图像数据的知识抽取，补全其他对象实体以及各个实体之间的各类关系。在按照名称添加实体以及关系的同时，还要保存实体的类别信息，如交通、动物、运动、日常、水果、食物等。

表 3-1 基于 YOLOv8 的图像实体识别器的输出标签

Tab. 3-1 Output Labels of YOLOv8-Based Image Entity Recognizer

序号	实体名称	序号	实体名称	序号	实体名称	序号	实体名称
0	person	20	elephant	40	wineglass	60	diningtable
1	bicycle	21	bear	41	cup	61	toilet
2	car	22	zebra	42	fork	62	tv
3	motorcycle	23	giraffe	43	knife	63	laptop
4	airplane	24	backpack	44	spoon	64	mouse
5	bus	25	umbrella	45	bowl	65	remote
6	train	26	handbag	46	banana	66	keyboard
7	truck	27	tie	47	apple	67	cellphone
8	boat	28	suitcase	48	sandwich	68	microwave
9	trafficlight	29	frisbee	49	orange	69	oven
10	firehydrant	30	skis	50	broccoli	70	toaster
11	stopsign	31	snowboard	51	carrot	71	sink
12	parkingmeter	32	sportsball	52	hotdog	72	refrigerator
13	bench	33	kite	53	pizza	73	book
14	bird	34	baseballbat	54	donut	74	clock
15	cat	35	baseballglove	55	cake	75	vase
16	dog	36	skateboard	56	chair	76	scissors
17	horse	37	surfboard	57	couch	77	teddybear
18	sheep	38	tennisracket	58	pottedplant	78	hairdrier
19	cow	39	bottle	59	bed	79	toothbrush

对于实体之间的关系，由于没有与手势图像数据分布相似的场景图数据集，且现有的场景图识别方法考虑的是实体间的位置关系，但这不适用于本文推断场景置信度的应用目的。因此，本文采用基于规则的手工标注的方法补全场景中实体间的关系。

为了更好地配合后续知识特征的提取，本文围绕“出现场景相似”、“出现于”、“依赖从属”三类关系，设计相应的抽取规则。“出现场景形似”关系标注与同类图像实体之间，如猫和狗、鼠标与键盘等，是一种双向关系。对于办公用品、车辆等与出现场景单一的目标类别，添加在类别内任意两个实体之间；而对于动物、食物等出现场景存在差异的类别，则以 YOLOv8 识别到的共现为基础手工补全或调整。“出现于”关系则添加在直接检测到的对象实体与相应场景实体，但即使 YOLOv8 是一个经过广泛验证的高效的目标检测算法，但其依然不可避免地会出现误识别的现象，因此需要选取一个合适的阈值作为构建出现关系的判断标准。经过实践测试本文使用 0.85 作为阈值，即在目标检测的预测结果中筛选置信度大于 0.85 的对象，与相应场景间构建“出现于”关系。“依赖从

属”关系是为了表示一些目标实体间在场景空间层面的上下级关系，如餐具与餐桌、各类衣服之间等，其主要目的是在知识特征提取时为对象识别置信度较低的情况补足场景信息。

3.3.3 知识抽取结果

经过对 Wikidata 中包含的手势数据与多个公开手势数据集的多模态知识抽取，我们总计得到了 181 个实体，476 条实体间关系。其中，实体包含 62 个手势实体以及共计 119 个交互目标、场景、观察者等相关节点实体，实体间关系包含手势实体与场景间的关系 142 条、手势实体与交互目标间的关系 133 条以及其他与手势无关的实体间关系 201 条。

手势实体少部分来自于从 Wikidata 知识图谱中的迁移或手工补全，大部分来自于对公开手势数据集标签文本的知识抽取。这样的抽取方法确保了手势实体的准确性，适用于本文的手势识别任务。其余实体则是在图像知识抽取方法的基础上人工补全而来，即以目标检测算法可以识别到的 80 类目标为基础人工补全相关的场景、观察者等实体，补全的方法是通过对数据集进行整体标注从而在知识抽取的过程中融入知识图谱。例如，根据数据集类别可以标注数据发生的场景是在于厨房、办公室等，根据数据集的视角属性可以标注观察者为手势主体、车辆、无关第三人等。

而在实体间关系中，手势与场景、手势与目标间的关系都来自于在整体标注数据集的基础上对标签文本的知识抽取，而其他实体间关系则在对图像数据进行知识抽取时按照目标类别进行手工添加。这类与手势无关的实体关系具体可以细分为对象实体之间的关系、对象实体与场景之间的关系。后者可以通过图像知识抽取方法自动提取，是实体间关系的主要组成部分。对象实体之间的关系不是通常意义上场景图结构中，场景图像中各个对象间的空间位置关系，而是对象属性中潜在的语义关系，需要通过人工分析和判断来添加。

总而言之，文本知识抽取方法可以准确的获取到半结构化的知识，仅需要简单的甄别和判断即可用于知识图谱的构建，是知识图谱构建过程的基础；而图像知识抽取由于数据来源是非结构化的图像数据，因此要依靠人工构建抽取规则，再通过基于规则的方法进行抽取，相对于文本中抽取出的知识，图像知识数量更多、更加广泛，包含更多的细节内容。两者之间相互补全，可以更全面地表达泛用场景下的手势先验知识，进而为本文提出的视觉手势识别方法提供数据辅助与支持。

3.4 多模态手势知识融合与存储

3.4.1 知识融合

本文所构建的多模态手势知识图谱所需要进行的知识融合过程主要包含属性对齐和实体消歧两个部分。

其中，属性对齐主要是对手势实体的图像属性进行规范化，同时对在 Wikidata 中存储为属性的一些实体进行类型的转换。手势的应用场景在 Wikidata 中以文本描述的形式被存储为手势的属性，而在本文所设计的手势知识图谱中应作为情景实体；按照本文方法创建的实体中，为了回避图数据库的保留关键字，图像属性的 key 为 img，而 Wikidata 中筛选出来的实体中图像属性的 key 包括 page banner、imge 等需要整合进 img 属性。

实体消歧部分的工作较为复杂，由于手势实体的收录主要来源于对各个视觉手势数据集中标签的命名实体识别。因此相同手势在不同数据集中可能在表述上存在一些不同，同时与 Wikidata 中的部分实体也可能存在重叠的情况。这种表述歧义与实体重叠包含两种形式：

（1）在手势名称以外存在多余的介词。这种问题造成的实体间名称属性不同的现象，通过设计正则表达式进行模式匹配的方式，去除多余的语素，只保留统一的动词名称；

（2）一些动作相同但在不同场景下表述不同的手势动作。对于这部分实体虽然可以通过算法对比手势骨架序列的相似性来判断，但其相似度阈值难以确定，因此本文采用人工筛选标注的办法进行实体消歧和链接。

3.4.2 知识存储

知识存储是指将信息、数据、经验和概念等以某种形式记录下来，并存储在特定的媒介或系统中，以便后续检索、利用和共享。知识存储可以采用多种形式，包括文本文档、数据库、图形、多媒体文件等。其中，数据库是一种常用的知识存储方式，特别是非关系型数据库（如 NoSQL 数据库）。它们能够有效地组织和存储大量的结构化或非结构化数据，并支持高效的数据检索和查询，非常适合存储类似于知识图谱中的知识数据。

知识图谱的结构可以用属性图模型来表示，类似于语义网络，由节点（实体）和边（关系）构成。每个节点可以包含多个属性，用来更全面地描述节点的特征，

如图 3-5 所示。

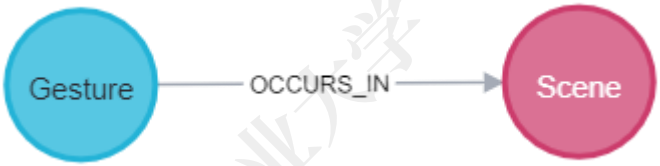


图 3-5 知识图谱三元组可视化示意图

Fig. 3-5 Visualization of Knowledge Graph Triples

针对这种图数据结构，Neo4j 是一种非常好的选择，它是一种基于节点和边的 NOSQL 图数据库系统，能够存储丰富的关系数据，并且遍历性能并不受数据量大小的影响，具有快速、高效的优点。Neo4j 让关联实体之间的联系更加直观清晰，因此是目前最流行的图数据库之一。因此本文使用 Neo4j 图数据库来存储知识组并以图结构的形式展示，有利于后续进行知识推理和检索。Neo4j 提供了将存储在本地三元组集合转化为知识图谱的功能，具体实现步骤如下所示：

（1）数据批量导入

数据导入有多种方式，本文使用了批量导入 CSV 文件的方式。为避免繁琐的 Cypher 语句创建和插入节点与关系，将实验结果以(实体，关系，实体)形式的三元组形式保存为 utf-8 格式的“手势实体.csv”文件，从“手势实体.csv”文件中读取实体及关系数据。然后，遍历所有实体并将它们的名称和类型作为节点添加到知识图谱中。接着，遍历每个实体的关系，将关系类型和相关实体添加到知识图谱中。最后，将知识图谱保存到文件“knowledge_graph.json”中，利用 python 编写脚本存入 neo4j 数据库中。

（2）数据查询

Neo4j 提供一种特殊的图形查询 Cypher 语言，用于从 Neo4i 图形数据库中检索和操纵数据。Cypher 具有类似 SQL 的语法，但专为处理图形数据而设计。Cypher 查询语句由三部分组成：一个匹配子句，一个返回子句，和一个可选的最终子句。

（3）知识图谱可视化

将数据导入图数据库后，利用 Neo4j 提供的可视化功能可以直观的看到其中存储的数据以及知识图谱的结构。图 3-6 展示了本文构建的手势知识图谱的部分节点和关系，可以看出图谱中包含了丰富的节点和关系。

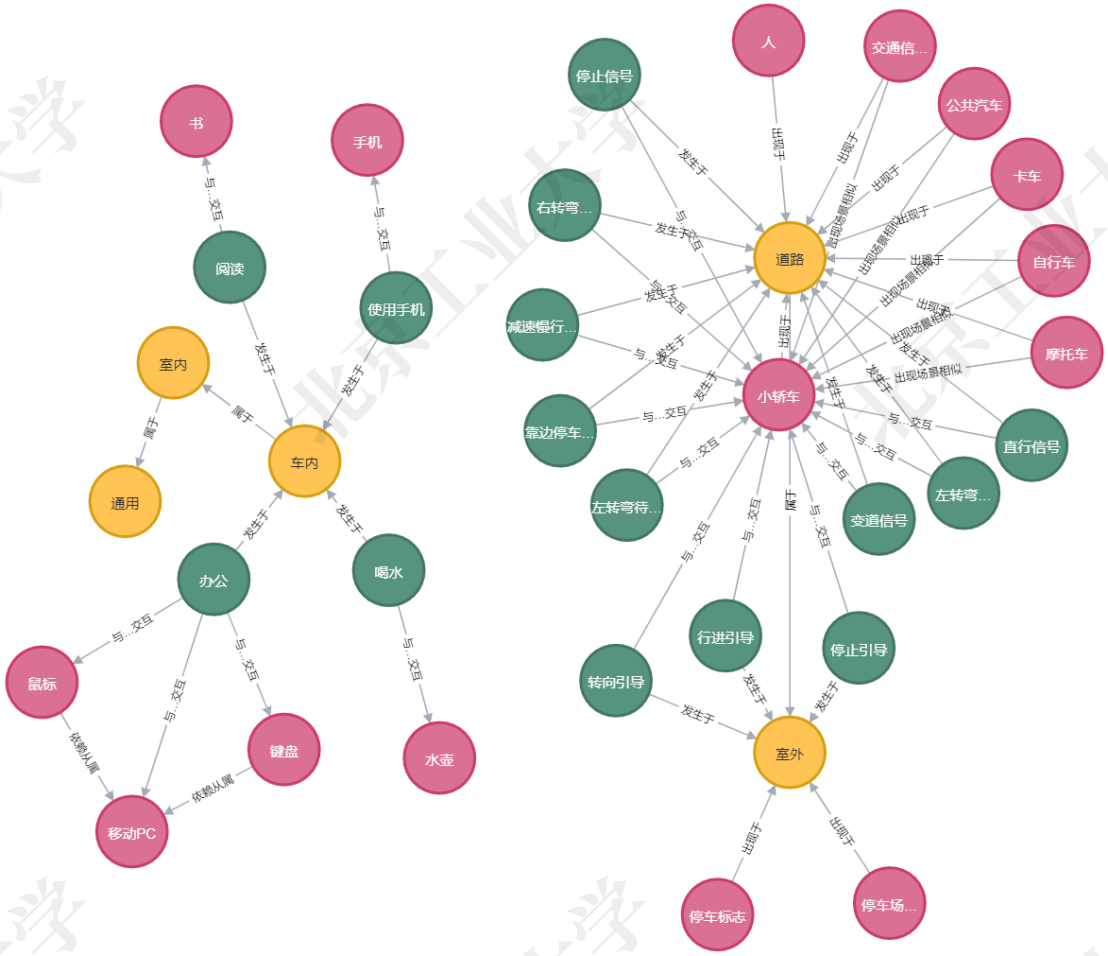


图 3-6 局部手势知识图谱可视化

Fig. 3-6 Visualization of Local Gesture Knowledge Graph

而节点中包含了一些从 Wikidata 中的实体合并过来的属性或者随着数据集及标签自动化添加的实体属性，主要是对手势节点和场景节点的一些描述，从而为知识图谱的后续扩展提供一些参考信息，防止在维护中出现实体歧义的问题。本文所构建的手势知识图谱中部分实体节点属性如图 3-7、图 3-8 所示。

手势实体包含 id、特征、描述、图像、名称和类型 6 项基础属性。其中，id 属性由数据库自动生成，每个实体节点具有唯一的 id 编号。描述、图像、名称属性都用来对手势实体进行清晰的定义。而特征属性对应实体的图像属性中的姿态骨架坐标，类型属性则表示了手势类别。

场景实体除了唯一的 id 属性外，还包含功能、主要手势类别以及场景名称属性。这些属性都来自于手工录入，用于定义场景实体的详细信息。

图 3-7 手势实体的节点属性示例

Fig. 3-7 An Example of A Gesture Entity's Attributes

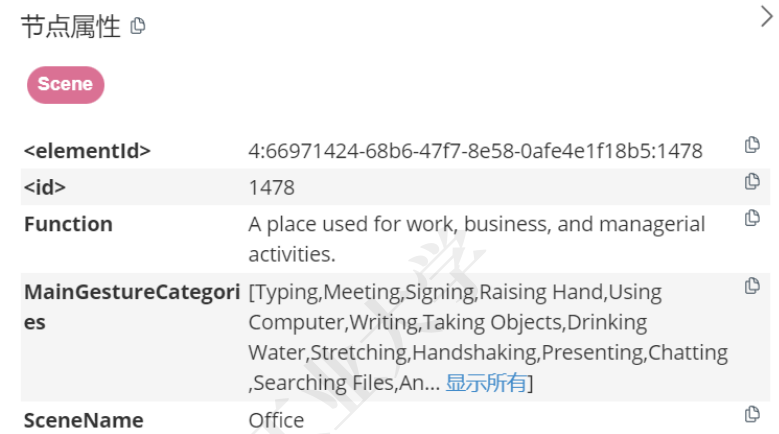


图 3-8 场景实体的节点属性示例

Fig. 3-8 An Example of A Scene Entity's Attributes

3.5 本章小结

本章节介绍了一套手势知识图谱的构建方法，包括多模态手势知识图谱本体模式的设计、多模态手势知识的提取、融合与存储。该方法使用基于 YOLOv8 的目标检测算法对场景对象序列进行分析和提取，使用基于关键点的视觉手势检测算法对手势姿态骨架进行估计，同时使用 Word2Vec 结合 BiLSTM-CRF 模型进行命名实体识别，用于提取手势数据集标签文本中的实体与关系信息。本章还基于 Wikipedia 开源知识图谱，以及 3 个常用开源手势数据集，构建了一个包含 5 种典型手势应用场景，62 个手势实体以及共计 119 个交互目标、观察者等相关节点的手势知识图谱，并通过 Python 代码将其导入存储至 Neo4j 图数据库中。

第4章 基于关键点的视觉手势检测算法

广义上的手势指的是人们在日常生活、交流沟通中，肢体或手部的动作，通常包含一些语义信息，用于表达或者辅助表达。通过对比学习相关研究与前期调研，本文确定手势的形式具有一般性，即不同含义的手势的元动作具有一般性。因此，在对手势进行分类识别前先将图像信息解读为一般性的关节骨架图，对于适应不同语义上下文场景中的手势识别是具有重大意义的。本章节将提出一个轻量高效的姿态估计算法，用于本文所描述的视觉手势识别框架。

4.1 自底向上的姿态估计方法

相比于先检测目标人体的自顶向下手势识别方法，自底向上的方法先识别各个关节位置，再将其进行匹配。这样的操作顺序可以有效解决多人、遮挡等常见问题，以更少的参数实现更好的效果。

4.1.1 轻量化的多头姿态估计网络

由于姿态估计功能模块将在整个算法框架中被频繁调用，因此希望所用的姿态估计算法尽可能地轻量化以降低运算成本。通过对比常见的算法，本文设计并实现了基于目标检测算法 CenterNet[62]的自底向上的姿态估计算法，参考 Google 在其商业化深度学习产品 Media Pipe 中使用的闭源姿态估计网络 MoveNet 的思想，对模型进行复现及改进，模型结构如图 4-1 所示。

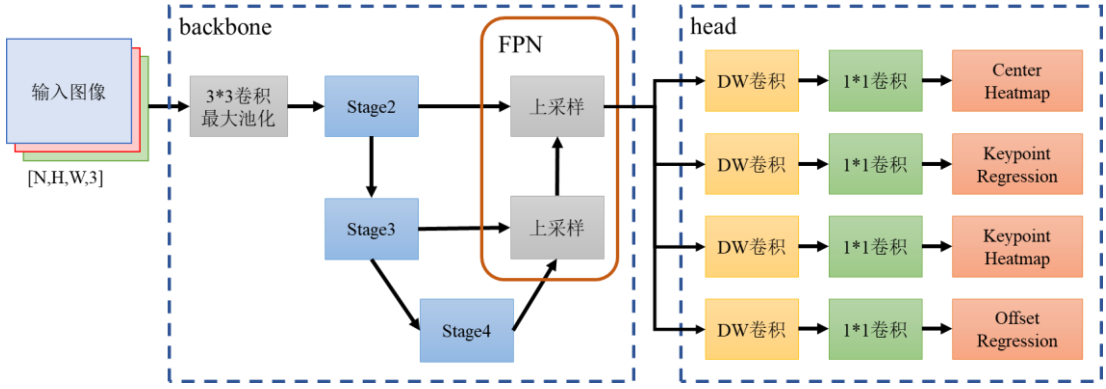


图 4-1 模型结构示意图

Fig. 4-1 Schematic of the Model Structure

模型整体结构上依旧采用与 Openpose 相似的，自底向上的姿态估计算法设

计思路，使用多个预测头分别估计关节关键点位置、关节亲和力场等组件，然后再通过后处理操作进行组装。其关键结构可以划分为 backbone、head 以及后处理三个主要部分。

backbone 部分用于特征提取，使用 ShuffleNet v2[63]结合特征金字塔（FPN），实现高分辨率且语义丰富的特征图。ShuffleNet v2 通过分析现有的轻量级图像特征提取算法中各种运算操作的内存读写次数，进而用更高效地操作替换高运算量操作，在保证模型容量的基础上提高运算效率。具体来说，在使用 DW 卷积（Depthwise Convolution，即为每个通道分配一个 3x3 卷积，通常通过分组卷积实现）替换 3x3 卷积的基础上，构造如图 4-2 所示的下采样与基本单元结构实现下采样特征提取，图中省略了必要的 BN 层与激活层。

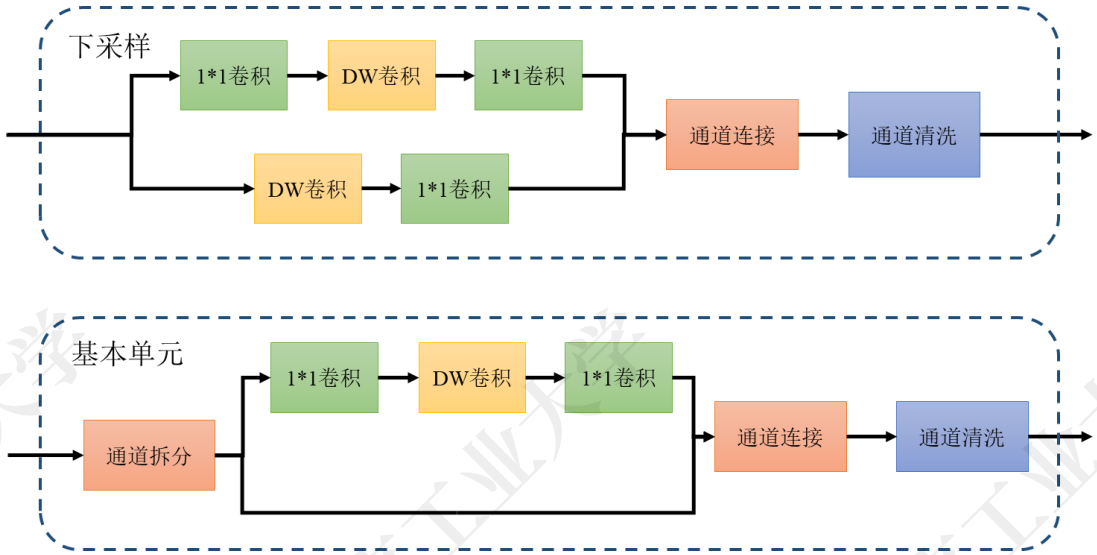


图 4-2 下采样与基本单元结构示意图

Fig. 4-2 Schematic of the Down Sampling and the Basic Unit

模型中每一个 Stage 模块都由一个下采样操作和多个基本单元组成，其中下采样操作由步长为 2 的 DW 卷积实现，而在基本单元中，则通过通道拆分、操作后再拼接的方式实现与 ResNet 残差块结构相似的训练稳定性，并降低模型运算量，减少对运算资源的需求。模型可以通过调整每个 Stage 的输出通道数，从而调整模型参数量和运算速度，本文设置其分别输出 48，96，192 通道，使得模型需要的运算量尽可能小。

而 FPN 在目标检测算法中比较常见，用于从图像中提取多尺度特征，以适应由于透视造成的目标尺度差异。而在本文中由于使用自底向上姿态估计框架，无法统一输入图像中被检测人体的尺度大小，因此在本文使用的模型中添加两层双线性上采样，使用 FPN 结构进行多尺度特征的提取和融合。

head 部分使用四组不同的预测头，每个预测头都包含一个 DW 卷积和以一

个 1*1 卷积对 backbone 计算出的特征图进行预测，分别预测目标几何中心热力图（Center Heatmap）、相对于中心点的关节（关键点）偏移坐标值（Keypoint Regression）、每种类型的关节（关键点）的热力图（Keypoint Heatmap）、关节热力图中各高斯核中心跟真实坐标的偏移值（Offset Regression）。四个预测头详细的数据构造方式将在 4.2.2 中进行说明。训练过程中将同时对这四个预测头计算损失函数，并通过超参数设置的权重比例进行加和，计算最终权重，具体损失函数计算方式将在 4.1.2 进行详细介绍。

为减少模型参数量，提高推理效率，通过增加后处理操作来减少模型的计算负担。具体过程如下首先，在 Center Heatmap 推理出的全部几何中心里，选择距离图像中心最近的目标作为被检测目标。具体操作上，使用反距离加权法来实现这个目的，即为热力图中的每个高斯中心P_i设置一个与其到中心点P_{center}的距离相关的权重$W_{i-center}$，计算方式如下式所示。

$$W_{i-center} = d(P_i, P_{center})^{-n} \tag{4-1}$$

在确定目标几何中心后结合 Keypoint Regression 和 Keypoint Heatmap 的结果可以得到该目标的全部关键点坐标。具体来说，通过上一步确定的目标中心坐标与关键点回归结果解算出对各个关节的回归坐标，随后再选择距离回归预测坐标最近的热力图高斯中心作为目标关节，选择方法与上一步操作类似，通过反距离加权实现。最后，参考无偏数据处理方法（Unbiased Data Processing，UDP），通过 Offset Regression 修正因缩放输入导致的量化误差，得到在原图中位置准确的人体骨架坐标。

4.1.2 多头损失函数定义

一般来说，对于热力图类的预测输出通常使用 L2 loss，即均方误差（Mean-square Error, MSE），比如 OpenPose。而多头网络 CenterNet 及其同类网络中使用的是优化后的 Focal loss，根据类别加权处理类别不平衡的问题，根据置信度加权处理难易样本的问题。

结合两者特点，本文对两个热力图预测输出使用加权均方误差（KMSE），通过加权平衡正负样本。权重的计算方式是将标签矩阵按照等比例映射到$[1, k+1]$区间内，以此作为权重模板与均方误差相乘，如下式所示。

$$W_{mask}^{ij} = target_{ij} * k + 1 \tag{4-2}$$

$$\mathcal{L}_{KMSE} = \frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(pred^{ij} - target^{ij}\right)^2 * W_{mask}^{ij} \tag{4-3}$$

其中，W_{mask}表示权重模板，$pred^{ij}$和$target^{ij}$分别表示模型预测结果和标签矩阵第 i 行第 j 列的取值，$i \in [1, H]$，$j \in [1, W]$。

最终 Center Heatmap 和 Keypoint Heatmap 采用了 KMSE 损失函数，Keypoint Regression 和 Offset Regression 则参考 CenterNet 使用 L1 Loss，分别计算损失。最后，还需要考虑不同损失函数之间的权重如何设置，由于关键点回归的结果是关节坐标与中心坐标的绝对插值，与其他三个结果在数量级上存在差异，因此进行适当缩放，如下式所示。

$$\mathcal{L}_{total} = \mathcal{L}_{KMSE}^{\text{center}} + \mathcal{L}_{KMSE}^{keypoint} + \alpha * \mathcal{L}_{1}^{keypoint} + \mathcal{L}_{1}^{offset} \tag{4-4}$$

其中，α是关键点回归损失的缩放比例，实验中取经验值 0.1。

4.2 模型训练与实验验证

上一节已经对本章所实现的姿态估计模型进行了全面的介绍，这一小节将对该模型展开训练和评估。

4.2.1 数据集选择及数据预处理

参考 Google 在 MoveNet 中使用的训练数据，本文将使用 COCO 数据集中包含少量目标者的图像作为训练数据集，再从 Bilibili 网站上爬取的一些舞蹈、瑜伽、健身等视频并进行标注，从而补充 COCO 数据集中缺少的部分少见动作、姿态，优化数据平衡。

数据预处理部分采取图像旋转、翻转、剪裁等技术进行数据增强。其中需要注意的是，在进行图像翻转时只进行水平方向的翻转以保证手势数据的合理性，另外，需要在翻转图像的同时对标注数据进行翻转，同时对于左右关节也需要进行相应替换。否则若只翻转关节坐标，得到的骨架与原关节骨架间是绕 z 轴旋转 180 度，而不是水平翻转。

除此之外，为匹配模型输出以进行训练，需要将标签中的关节点坐标构造为与 4 个预测头数据格式一致的目标结果。对于 Center Heatmap 和 Keypoint Heatmap 对应的标签数据，使用高斯核生成，关键点 j 在 P 点的热力图取值$S_j(P)$如下式所示。

$$S_j(P) = max_k \ exp\left(-\frac{\|P - x_{j,k}\|_2^2}{\delta^2}\right) \tag{4-5}$$

其中，k表示第 k 个目标$x_{j,k}$表示第 k 个目标的第 j 个关键点的位置，δ表示高斯核的标准差，在实验中取经验值 1.2。

对于 Keypoint Regression 和 Offset Regression 对应的标签数据，将对应回归标签标记在矩阵中对应的中心点位置及其周围，以提高模型训练速度。数据构造的工作是在模型训练时，随训练过程一同完成的。

4.2.2 姿态估计网络训练

模型输入图像按照 COCO 数据集图像大小进行缩放、裁剪对齐，最终都以 192*192 的大小输入到网络模型中。下面将详细介绍模型四个输出预测头。

Keypoint Heatmap 的维度是[N,K,H,W]。N 是 batchsize，训练时设定为 32，预测时一般为 1。K 代表关键点数量，在本章实验中为 17。H、W 分别为特征图的长和宽，本章使用的模型输入图像的大小是 192*192，经过模型 4 倍降采样后得到大小为 48*48 特征图与输出相对应。Keypoint Heatmap 代表的意义是当前图像上所有人的所有关键点的 heatmap。

Center Heatmap 的维度是[N,1,H,W]，表示当前图像上所有人中心点的 heatmap。Google 在 MoveNet 中使用每一个人的所有关键点的算术平均数，但是在经过实验发现这样的方法对像交警手势这样的伸展度较高的姿态效果不好，因此本文使用所有关键点的外接矩形的中心点作为目标中心。

Keypoint Regression 的维度是[N,2K,H,W]，包含 K 个关键点的横纵坐标共 2K 个通道。实现过程中在每个目标人体的中心坐标位置上，按 2K 通道顺序依次赋值 x1,y1,x2,y2,...，此处的 x、y 代表的是同一个人的一个关键点相对于其中心点的绝对偏移值。

Offset Regression 的维度是[N,2K,H,W]，通道含义和关键点回归相似，不过这个预测头预测的是偏移值，用于修正在原尺寸下（本章模型中是 192*192）关节坐标的偏移误差值。

对于网络训练过程中使用的超参数，本章设置批大小（batch size）为 32，即一次输入网络 32 张图。使用 Adam 优化器调整网络参数，同时设置权重衰减（Weight decay）为 5e-4。基本学习率设置为 0.001，并使用 MultiStepLR 进行学习率调整，在第 70 和第 100 个 epoch 将学习率下调 10 倍。总共迭代 120 个 epoch 完成网络训练。

4.2.3 训练与评估结果

本文在搭载了 GTX3090 的 Windows11 平台上对本章所介绍的模型进行了训练和评估，详细的平台配置在表 4-1 中展示。

表 4-1 实验环境配置

Tab. 4-1 Preparation for the Experiment Environment

条目	型号/版本号
系统	Windows 11
CPU	Intel CORE i7 12700k
GPU	NVIDIA GeForce RTX 3090
CUDA 版本	11.4
Python 版本	3.9.10
深度学习框架	PyTorch 1.10.0

评价指标方面，由于目标尺度不同以及不同类别关节的误差容忍程度不同等原因，不能直接通过关键点与标签之间的欧式距离判断估计结果的好坏。在 2D 姿态估计任务中最常用且直观的指标是关键点相似度（OKS），通过在距离计算中加入关键点归一化因子实现，计算方法如下式所示。

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2\sigma_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)} \tag{4-6}$$

其中，d_i表示当前检测的一组关键点中 ID 为 i 的关键点与相应标签的距离；s 表示尺度因子，其值为行人检测框面积的平方根；σ是关键点归一化因子，是和数据及相关的常数，本文使用 COCO 数据集按照表 4-2 所示取值；$\delta(v_i > 0)$表示该关键点是否存在标注，有标注的取 1，无标注取 0。OKS 的取值范围在 0 到 1 之间。

表 4-2 COCO 数据集中的关键点归一化因子取值

Tab. 4-2 Values of the Keypoint Normalization Factor in the COCO Dataset

关键点类型	归一化因子σ
鼻子	0.026
眼睛	0.025
耳朵	0.035
肩膀	0.079
手肘	0.072
手腕	0.062
臀部	0.107
膝盖	0.087
脚踝	0.089

本文通过$OKS > 0.75$来计算平均精度（AP）作为模型评价指标，并对比 Google 闭源模型 MoveNet、Openpose 等常见姿态估计模型。同时由于对模型效率的需求，使用仅在 CPU 上推理单张图像的时间作为标准，评价模型运行速度。

模型中 backbone 部分主要用于图像特征地提取，按照 google 在其公开发表的描述中，使用了 MobileNet v2[64]作为特征提取模块，但当前已经有很多轻量级特征提取网络的效果要优于 MobileNet v2，因此本文选择了 MoblieNet v2，MobileNet v3[65]以及 ShuffleNet v2，三个轻量高效的特征提取网络进行对比，结果如表 4-3 所示，使用 ShuffleNet v2 的效果相较于其他两者在精度和速度上都有一定的提升，因此本文选择 ShuffleNet v2 作为特征提取模块。其中，使用 MobileNetv2 的效果没有达到谷歌官方公布的 MoveNet 的水平，可能是在模型实现及训练优化方面有所差异。

表 4-3 特征提取模块的对比实验结果

Tab. 4-3 Comparative Experimental Results of the Feature Extraction Module

特征提取模块	$AP^{0.75}$（%）	Speed（毫秒）
MobileNet v2[64]	73.6	58.7
MobileNet v3[65]	78.9	48.6
ShuffleNet v2[63]（本文）	**80.1**	**42.4**

此外，本节还对所使用的特征金字塔（FPN）、多头网络结构的有效性进行了效融实验验证。与 FPN 相对应的空白对照是直接使用特征提取模块的输出进行预测；与多头网络结构相对应的空白对照是只保留关键点热力图和中心点热力图两个预测头，并相应调整后处理操作。实验结果如表 4-4 所示，表明本章所使用的 FPN 以及多头预测的方法对姿态估计任务是有效的，且两者组合有助于提升姿态估计的识别精度。

表 4-4 主要模块的消融实验结果

Tab. 4-4 Ablation Experiment Results for the Main Module

模型设置	$AP^{0.75}$（%）
ShuffleNet v2 + 单预测头	45.9
ShuffleNet v2 + FPN +单预测头	58.3
ShuffleNet v2 +多预测头	72.0
ShuffleNet v2 + FPN +多预测头（本文）	**80.1**

最后，将本章所设计的算法与常见的姿态估计算法进行比较。由表 4-5 所示结果可以看出，本章所实现的模型在测试数据集上的表现结果略优于 MoveNet，速度上略慢，但相较于其他常用模型均有一定程度上的提高。综合来看，可以满足本文知识驱动的视觉手势识别任务的需要。

表 4-5 姿态估计算法评估结果

Tab. 4-5 Evaluation Results of Multiple Pose Estimation Algorithms

模型	AP	Speed（毫秒）
YOLOv8-pose	60.1	131.8
Openpose	74.4	168.7
MoveNet	78.7	**39.0**
本章所使用的模型	**80.1**	42.4

4.3 本章小结

本章考虑到手势的特点、自底向上的姿态估计方法的优越性和姿态估计算法的使用频率等因素，设计并实现了一种轻量高效的多头姿态估计算法模型。详细介绍了模型各个部分的结构、模型的具体工作流程，并根据模型特点定义了多头损失函数。叙述了模型实验所用到的数据集、数据预处理、模型训练参数、模型评价指标、实验软硬件配置，通过对比试验和模型结构消融实验验证模型设计的有效性，并对模型实验结果进行分析与展示，验证模型在手势识别任务上的性能表现良好，能够准确地识别出不同的手势，并且在面对躯干、手部等多尺度姿态估计任务和多种手势时，能够保持较高的识别准确率。

第5章 知识驱动的手势识别算法

通过第三、四两章的研究，本文完成了具有可扩展性的手势知识图谱构建以及轻量高效的手势检测算法的设计。本章节将基于第三、四两章的成果，设计用于融合知识特征与一般姿态骨架特征进行手势识别的分类器算法，增强手势分类器在泛化应用场景下的识别能力。

5.1 知识驱动的手势识别算法框架

环境上下文信息作为手势识别任务中普遍存在的先验知识，要想学习到一个高效的表示方法，从中提取出具有泛用性的一般特征对数据集的分布、数据的采集等具有极高的要求和难度。因此本文在现有数据的范围内，设计了一套可行的算法思路，如图 5-1 所示。该方法不对场景信息进行表示学习，而是直接通过对输入图像的信息提取得到各个场景的权重向量，再由各个场景下的手势分类器进行加权多数投票，最终选取得分最高的手势。

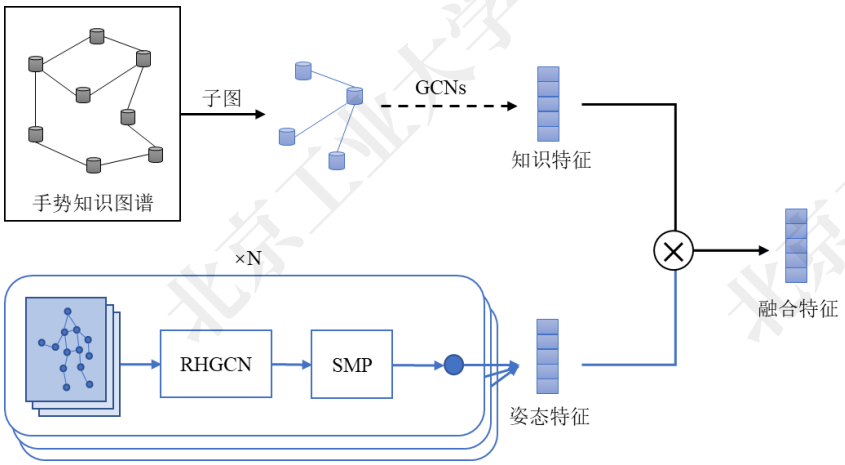


图 5-1　知识驱动的手势识别算法框架示意图

Fig. 5-1 The Framework Illustration of Gesture Recognition Algorithm with Fision Knowledge

具体来说，算法通过场景识别器提取输入图像中不同场景的概率权重作为知识特征，再通过 N 个轻量的手势姿态分类器结合姿态估计算法快速推理 N 个不同场景下的手势类别作为姿态特征，两者结合进行多数投票得到最终的分类结果。这样的方法可以通过迭代训练，向现有的场景中添加手势，也可以通过向知识图谱中添加新的场景或手势类别并训练新的手势分类器以达到灵活扩展应用场景的目的。

5.2 特征提取及融合

5.2.1 知识特征提取

在第三章的工作中已经完成了手势知识图谱的构建，其中存储了所有手势和相关实体（其他节点）之间的显式关系（边）。

首先，使用基于 YOLOv8 的图像实体识别算法检测输入图像中的实体序列，结合定义于算法系统的观察者信息，从构建好的手势知识图谱中筛选对应的实体节点及其二阶邻域内的实体和关系，剪枝后作为输入图像的场景语义知识网络图。通过筛选子图、邻域限制以及剪枝这一系列操作，可以有效减少用于知识推理的子图中的节点数量

然后在其上利用图卷积网络进行推理，从而得到各个场景类别的可能性分数，用于表示输入手势的场景知识特征。

为了方便后续描述，用符号$\mathrm{W} *_{\mathcal{G}} X$表示一个独立的单层图卷积操作，计算方法如下式所示。

$$\mathrm{W} *_{\mathcal{G}} X \triangleq Z = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}XW \tag{5-1}$$

其中，X 是输入节点特征，$X_i \in \mathbb{R}^k$表示图中第 i 个节点的特征表示。$\mathrm{W} *_{\mathcal{G}}$表示一个基于关系图$\mathcal{G}$和可训练操作 W 的单层图卷积操作。$\hat{A} = A + I_n$是具有自连接的知识图谱$\mathcal{G}$的邻接矩阵，$I_n$表示单位矩阵。$\hat{D}$是$\hat{A}$的度矩阵，即$\hat{D}_{ij} = \sum_j \hat{A}_{ij}$。总的来说，第 i 个图卷积层输入特征$X^i \in \mathbb{R}^{m\times k}$经过 $\mathrm{W} \in \mathbb{R}^{k\times c}$后，被转换为一组新的特征$\mathrm{X}^{i+1} \in \mathbb{R}^{m\times c}$。

具体来说，通过 3.3.2 所述的图像实体识别方法对输入图像进行处理后，可以得到图像中包含的各个类别目标的数量和置信度，使用 One-Hot 编码对目标检测所对应的节点进行编码作为节点特征，并用数量和置信度作为节点的权重系数，用于描述对应节点在输入图像中重要程度。将所有节点特征与各自对应的权重系数相乘，可以得到输入视频的知识特征表示X，每个节点特征X_i对应关系子图$\mathcal{G}'$中的一个确定的节点。然后使用一个三层的 GCN 构建特征提取网络$\mathcal{W}^{knowledge}$，操作如下式所示。

$$\mathcal{W}^{knowledge} = \phi\left(W^3 *_{\mathcal{G},} \phi\left(W^2 *_{\mathcal{G}}' \phi\left(W^1 *_{\mathcal{G}}' X\right)\right)\right) \tag{5-2}$$

其中，输入特征X的大小为$D_1 \times D_2$，D_1表示$\mathcal{G}'$中的节点数量，D_2表示输入特

征的维度，与实体识别算法所识别到的目标种类数量相对应。$\phi(*)$是一个非线性激活函数，后文将对不同的激活函数的效果进行对比。W^1, W^2, W^3标识三个不同尺寸的可训练操作，W 的第二维大小表示对应 GCN 层的特征维度，最后一层 GCN 的输出维度为 1 维用来表示实体节点的置信度。

最后，将知识图谱中S个场景实体所对应的置信度特征k_i提取出来，取不在子图中的场景实体置信度为 0，取通用场景占位符对应的置信度为 1，把它们按顺序组合为$S+1$维场景知识特征向量$\boldsymbol{k}=[1, k_1, k_2, \ldots\ldots, k_S]^T$，与下文所述姿态特征相融合。

此外，通过优化特征表示方法和图卷积网络模型，有一些研究实现了仅通过知识推理的方法实现人体活动识别、行为解析的任务。但本文认为这样的方法只通过上下文场景语义信息进行手势或行为分识别，不符合一般常识的逻辑，因此其结果缺乏可在泛用场景下推广的可能性。不过，这类方法也间接证明了上下文场景语义信息对人体行为、手势识别的重要辅助意义。因此本文在场景知识特征的基础上增加了姿态特征，从而完善泛用性手势识别方法。

5.2.2 姿态特征提取

姿态特征部分，本文结合基于相对高度的图卷积网络（RHGCN）与空间域平均预测器（SMP）构建基于姿态特征的手势识别器。RHGCN 将手势姿态骨架时空图G作为输入，其中顶点集合V作为输入特征，边集合E作为符合相对高度划分策略的邻接矩阵，输出如下式所示手势特征集Y_G。

$$Y_G = \mathcal{N}(V, E; W_G) \tag{5-3}$$

其中，W_G表示 RHGCN 的模型参数，Y_G表示输出手势特征。

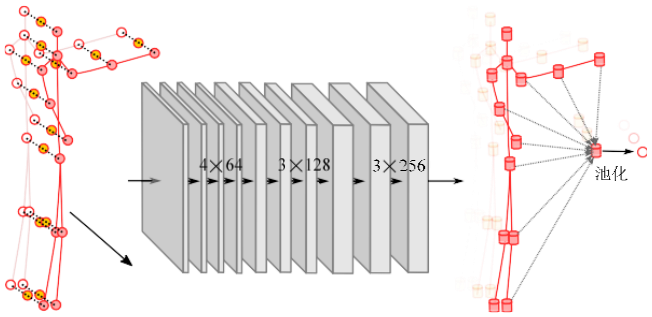


图 5-2 RHGCN 的体系结构

Fig. 5-2 Structure of RHGCN

RHGCN 的体系结构如图 5-2 所示，图中每个块表示一个时空图卷积层

（STGCL），数字表示前一层的通道数。每个 STGCL 由 7 个部分组成，如图 5-3 所示，依次为残差层起始位置、空间图卷积层、注意力层、ReLU 激活层、时间卷积层、残差层结束位置、ReLU 激活层。空间图卷积层依照相对高度划分策略，使用 3*3 的卷积核进行图卷积计算。对每条边乘上注意力权重后，再对计算结果使用 ReLU 激活函数进行激活，将结果输入到时间卷积层，在时间维度上使用 3*3 的卷积核再进行图卷积计算。残差层将以上从空间图卷积层到时间卷积层的计算部分作为残差部分，与直接映射部分相加，以提高网络训练效率。最后使用 ReLU 激活函数激活残差层求和的结果，输出到下一个 STGCL。

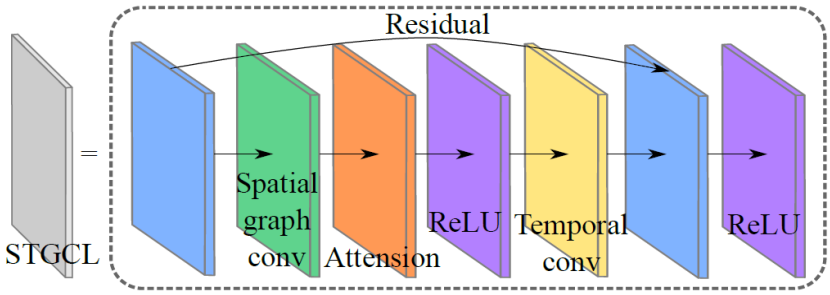


图 5-3 单个时空卷积层结构

Fig. 5-3 Stucture of a Single Spatial-temporal Convolution

SMP 与 RHGCN 相连，如图 5-2 所示，用于从手势特征Y_G中对手势进行分类。Y_G可以用标量y^{G}按下式开。

$$Y_G = \{y_{c,j,t}^G | c \in C^G, j \in 1, \dots, N_J, t = 1, \dots, T\} \quad (5\text{-}4)$$

其中，C^G代表输出中的通道数，N_J和T代表关节数和持续时间。由此，空间平均预测器(SMP)的计算如下式所示。

$$y_{c,t}^M = \frac{1}{N_J} \sum_{j=1}^{N_J} y_{c,j,t}^G \quad (5\text{-}5)$$

其中，$y_{c,t}^M$表示通道 c 上的空间特征值y^G在时间点 t 的平均值，再以公式（7）获取代表每类手势得分的向量。

$$\mathbf{o}_t = \mathcal{F}(y_t^M; W_F) \quad (5\text{-}6)$$

其中，向量$y_t^M = \{y_{c,t}^M | c \in C^G\}$，$\mathcal{F}$表示一个全连接网络，$W_F$为该网络参数，$\mathbf{o}_t \in \mathbb{R}^K$是一个向量，表示每个手势类别的得分，$K$表示手势类别数量。

5.2.3 特征融合

经过前两节的工作，已经提取出了输入图像中场景知识特征 $\boldsymbol{k}$ 和手势姿态特征 $\boldsymbol{o}$，将两者相乘作为手势的得分 $\boldsymbol{h}$。t时刻的手势类别h_t如下式所示。

$$h_t = \underset{s,c}{\operatorname{argmax}}\left(o_{c,t}^{s} \cdot k_s\right) \tag{5-7}$$

其中，$o_{c,t}^{s}$表示场景 s 的姿态特征手势识别器对在 t 时刻对 c 类手势的评分，k_s表示基于语义知识特征的场景 s 的置信度。可以注意到知识特征部分与手势类别 c 无关，因此可以单独计算最大值，将对场景 s 和手势类别 c 的最大化求解分别进行，从而减少计算量，计算方式如下式所示。

$$h_t = \underset{s}{\operatorname{argmax}}(\underset{c}{\operatorname{argmax}}\left(o_{c,t}^{s}\right) \cdot k_s) \tag{5-8}$$

5.3 实验设计

5.3.1 数据集预处理

本章节的设计和实现依托于第三章工作所完成的手势知识图谱，因此在数据选取方面与 3.2.1 小节所述相同。但是，现有的开源数据集无法直接用于本文所设计的算法框架，因此需要先对数据进行预处理操作。

在本文的算法框架中，知识特征与姿态特征是分别提取的，为了保证算法框架后续的可扩展性，因此将两个网络设置为独立模块分别进行训练。对于姿态特征提取模块，即特定场景下手势识别器的训练，其预处理操作主要在于依照场景进行的手势数据的筛选。由于在构建知识图谱时，保存了与手势对应的数据记录实体，通过形如图 5-4 所示的代码可以在 python 中调用 Cypher 语句，从图数据库中查询到所有指定场景的手势数据，组成训练数据集。再使用第四章所设计的姿态估计算法提取躯干和手部姿态特征，即各个关节点坐标，用于手势识别器的训练。

```
def run_query(query):
    with driver.session() as session:
        result = session.run(query)
        return result

# Cypher查询语句
cypher_query = """
MATCH (s:Scene {scene_name: 'Kitchen'})-[r1:OCCURS_IN]->(g:Gesture)-[r2:IS_INSTANCE_OF]->(v:VIDEO)
RETURN i
"""

# 执行查询
results = run_query(cypher_query)
```

图 5-4 查询厨房场景下的手势视频数据

Fig. 5-4 Query Gesture Video Data of Kitchen Scene

而对场景知识特征提取模块的训练，在从知识图谱查询出图像数据的基础上，使用前文设计的基于 YOLOv8 的图像实体提取器，找到所包含的实体节点作为

训练数据，并且按照知识图谱结构构造对应的标签信息。具体来说，场景识别器的输出 $\boldsymbol{k}$ 是一个 S+1 维的特征向量，表示 1 个通用场景占位符和知识图谱中包含的 S 类场景的置信度。因此对于场景 i 的识别结果中，k_i的值设置为 1。此时若设置其余场景的置信度均为 0，经实验发现效果并不好，分析后认为这是因为没有考虑到相似场景之间的相关性，例如社交场景和交通场景都属于室外场景。因此，需要根据每个场景实体与目标实体之间的关系设置标签值。考虑到场景实体间的“上级分类”关系是一个具有指向性的关系，因此对于目标场景实体的上级场景和下级场景处理不同。因此，本文设置场景的上级场景置信度为 0.8，下级场景置信度为 0.3，距离为 2 及以上的场景实体置信度为 0，开展后续实验。

5.3.2 网络训练

经过整理，本文在厨房（Kitchen）、办公室（Office）、社交（Social）、户外（Outdoor）、交通（Traffic）一共 5 个有足够数据的场景中实施了所设计的算法，数据集比例及分布如表 5-1 所示。可以发现，各个场景的数据并不平均，数量上存在一定差异。

表 5-1 各场景手势数据分布情况

Tab. 5-1 Distribution of Gesture Data in Each Scene

场景	各类手势数据总量（视频帧数）	包含手势/动作类别数
厨房	487	13
办公室	515	9
社交	498	8
户外	1004	25
交通	750	8

对于知识特征提取器的训练，本文使用 YOLOv8 提取图像实体后使用 L2 损失作为损失函数，计算预测标签与目标标签之间的均方误差（MSE）作为两者间的距离用来优化模型。循环训练网络模型，以数据集内全部的数据作为一个 epoch，直到连续 3 个 epoch 的损失函数均不再降低时停止训练，此时可以得到基于当前数据集的最优化网络参数。

而在训练各场景手势识别器的时候，不需要考虑场景数据间的不平衡性。将特定场景下的姿态骨架关节点坐标输入模型，通过 RHGCN 网络和 SMP 网络预测手势结果，以数据集手势标签作为真实结果，使用交叉熵测量二者之间的距离，以反向传播算法优化网络参数，降低交叉熵距离。同样训练至连续 3 个 epoch 的

损失函数均不再降低的时候为止。

5.4 实验结果

本节实验通过实例展示融合知识特征推理、识别手势的有效性，并且对本章的研究内容进行实验分析，得出相关结论。

5.4.1 知识驱动视觉手势识别实例

本小节将结合实例展示知识驱动的视觉手势识别方法的具体推理过程，图 5-5 是用于手势识别的一段视频中的一帧切片图像，接下来将以其为例展示本章所设计的算法的处理流程。

图 5-5 输入视频第 150 帧切片图像

Fig. 5-5 Single Frame of The Input Video

将图像输入算法模型后，会并行的调用目标检测和姿态估计模块，分别完成图像实体的抽取和手势骨架关节点的估计，结果如图 5-6、图 5-7 所示。

目标检测的结果将作为筛选条件，用于从知识图谱中提取子图进行后续的知识特征提取，从而计算各个场景的置信度。由图 5-6 可以看出，由于从图像中识别出了键盘、鼠标等办公场景的特有实体，场景知识特征中办公室的置信度远高于其他场景。而厨房场景的置信度也不低，有可能是因为 YOLOv8 从图像中检测到多个瓶子，从而加强了与厨房场景相关的权重。而户外类的三个场景的置信度均远低于办公室。

图 5-6 YOLOv8 实体识别结果

Fig. 5-6 Entity Recognition Results from YOLOv8

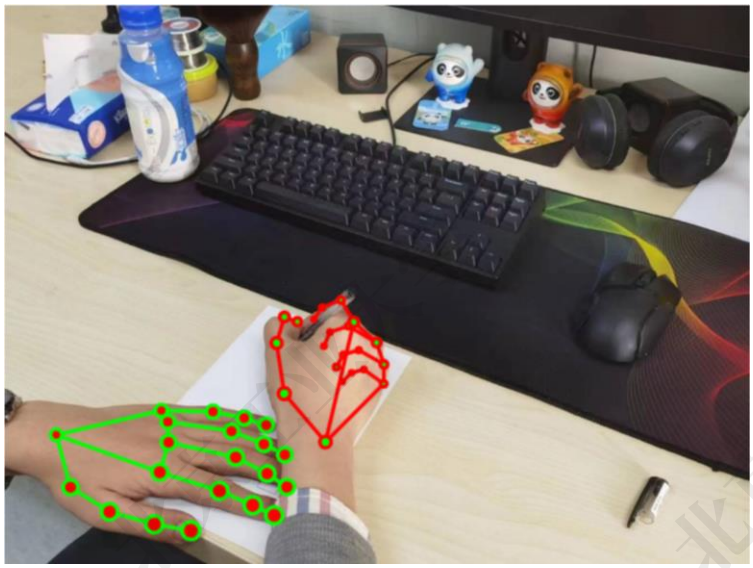


图 5-7 手部姿态估计结果

Fig. 5-7 Result of Hand Pose Estimation

手势姿态估计的关键点坐标按照时间顺序组成时间序列，通过不同场景下训练的手势识别器，得到各个场景中得分最高的手势及对应的分数，然后将其与知识特征相乘得到最终的融合特征，如表 5-2 所示。可以看到，虽然厨房场景下的 Pour 手势的得分要高于办公室场景下的 Use pen（Write），这可能是因为厨房场景下的手势动作相对单一，区分度更高。但通过知识特征的加权处理，在得到的融合特征中得分最高的是办公室场景下的 Use pen 手势，因此算法模型最终输出结果为 Use pen。

表 5-2 视频第 150 帧中知识与姿态特征的融合

Tab. 5-2 Example of the Fusion of Gesture Knowledge and Pose Features

场景	知识特征	姿态特征	融合特征
厨房	0.28	0.93 (Pour)	0.26 (Pour)
办公室	0.87	0.91 (Use pen)	**0.79 (Use pen)**
社交	0.15	0.75 (Handshake)	0.11 (Handshake)
户外	0.02	0.82 (Use pen)	0.02 (Use pen)
交通	0.01	0.46 (Go forward)	0.00 (Go forward)

5.4.2 实验结果分析

对于训练好的姿态特征提取器，在公开数据集上测试了其效果，并与其他常用方法进行了比较。本节在 FPHA 和 Something-something 两个数据集上进行了测试评估，选择了基于 2d 骨骼长度和角度[66]、ST-GCN[67]、扩大 GCN[68]以及数据集发布者提供的基线方法进行对比，结果如表 5-3 所示，可以看出本文所设计的手势识别器效果要优于其余模型。

表 5-3 不同模型/算法手势识别对比

Tab. 5-3 Comparison of Different Models/Algorithms for Gesture Recognition

模型/算法	在 FPHA 中的 mAcc（%）	在 SSv1 中的 mAcc（%）
基线方法	84.32	72.80
基于 2d 骨骼长度和角度	85.42	76.28
ST-GCN	89.36	71.73
扩大 GCN	92.18	73.47
本文	**94.75**	**78.91**

此外，本文还对提出的算法进行了消融实验，对算法各个组件的有效性进行了分析。将本文采集的全部场景的手势数据按照 7:3 的比例划分训练集和测试集，通过模型在测试集上的平均准确率（mAcc）评估模型效果。其中，基于知识图谱的场景知识特征模块与不使用该模块时的效果相对比；基于 RHGCN+SMP 的姿态特征模块用简单的图卷积网络代替，从而进行对比。实验结果如表 5-4 所示。

表 5-4 消融实验结果

Tab. 5-4 Results of Ablation Experiments

知识特征模块	姿态特征模块	mAcc（%）
无	基于 RHGCN+SMP 的手势识别器	69.2
基于知识图谱的场景识别器	3 层全连图神经卷积网络	55.3
基于知识图谱的场景识别器	基于 RHGCN+SMP 的手势识别器	**76.7**

可以看到，本文的方法有效提高了在混合场景下的手势识别准确率，可以提

升手势识别系统在一般混合场景中，对基础手势的识别效率，实现了可以包容多种不同应用场景的可扩展的手势识别方法。

同时还对算法扩展到新场景的性能进行了评估测试，主要是对知识图谱中增加新的场景后，知识特征提取器适应的速度，以及在新场景下训练手势识别器的速度进行了实验和测试。由于缺少足够数量的新场景的手势数据，本文采用已有场景中的 5 个场景，分别作为新增场景进行测试。为了统一测试标准，实验中从各个场景随机抽取来自其中 8 个随机手势的 300 条视频切片数据用于模型的训练和测试。表 5-5 展示了在各个场景中，场景知识特征提取模型的适应速度，以及手势识别模型的训练速度。

表 5-5 扩展性测试结果

Tab. 5-5 Results of Scalability Evaluation

场景	场景识别器训练速度（秒）	手势识别器训练速度（秒）
厨房	1207	689
办公室	814	866
社交	1010	331
户外	746	51
交通	1622	777
平均	1079.80	635.80

可以发现，场景识别算法适应到新应用场景的速度较慢，时间在 10 到 30 分钟之间，手势识别算法由于模型的特征提取部分功能相近，预训练模型的效果较好，平均训练时间在 10 分钟左右。这个扩展适应的过程直只在需要向系统添加新的模型的时候需要进行，因此在实际应用中并不是频繁调用的功能，本文认为，平均 10 至 20 分钟的调整时间是可以接受的，后续研究中可以通过适当的迁移学习算法，对模型迭代训练中的速度等进行优化。

而在算法整体的推理速度上，姿态估计算法和目标检测算法的速度都能达到 15FPS 以上，因此推理速度的瓶颈在于本章节所提出的特征提取和融合方法上。虽然本文的方法需要用到多个深度学习模型进行预测，但是基于并行运算的架构且在模型设计时上针对性地考虑了轻量化和运算性能等问题，不会对算法识别速率造成显著影响，当前速度主要还是受模型复杂度的影响。尽管如此，算法在本文所使用的实验平台上可以达到 5 到 10FPS 的识别速率，足以满足应用需求。

5.5 本章小结

本章在前文构建的知识图谱和姿态估计算法的基础上，设计了一套基于集成

学习方法的知识驱动手势识别算法，包含知识特征提取、姿态特征提取和特征融合三个模块。在知识抽取模块中，使用基于 YOLOv8 目标检测算法对第三章构建的知识图谱进行图像实体识别并根据预配置的手势主体信息从知识图谱中筛选出图像的场景语义知识网络图，之后输入图卷积网络进行推理，从而得到各个场景类别的可能性分数，用于表示输入手势的场景知识特征。在姿态特征提取部分，结合基于相对高度的图卷积网络（RHGCN）与空间域平均预测器（SMP）构建基于姿态特征的手势识别器，在 FPHA 和 SSv1 公开数据集上分别取得了 94.75%和 78.91%的平均准确率。通过场景识别器提取输入图像中不同场景的概率权重作为知识特征，再集成多个轻量的手势姿态分类器结合姿态估计算法快速推理不同场景下的手势类别作为姿态特征，两者融合得到最终的分类结果。本章通过手势识别实例和对比实验分析验证了所设计算法的效果和准确率，最终在本文制作的手势知识图谱数据集上取得了 76.7%的平均准确率和 5~10FPS 的推理速率。

第6章 车载手势识别系统设计与实现

本章基于前文所设计的算法框架，结合车载场景搭建原型系统，验证该方法在在实际应用中的可行性。车载手势识别系统将使用基于视觉的方法，对与车辆相关的交警手势、停车指挥手势、轿厢内乘客手势等手势类别进行自适应的识别，从而统一手势识别系统，降低维护和更新成本。下面将从系统的分析、设计以及实现三个方面，进行详细的介绍。

6.1 系统分析

本小节将结合车载应用场景的实际情况，分析手势识别系统的基本结构及设计思路，并进行需求分析。

6.1.1 系统概要分析

在现有的无人驾驶车辆中，对于各种交互手势分别使用不同的识别器进行独立的检测和预测。这样的设计方式，适用于定义严谨的手势类型，如交警手势、预定义的交互手势等。但在实际应用中还包含很多没有随意的交互手势，对于这类手势的识别和理解，有助于提高用户进行手势交互时的舒适性和易用性。出于这样的目的，基于本文所提出的算法设计原型系统作为无人驾驶车辆系统的子系统，实现跨场景一致、泛化能力强、扩展性强的手势识别系统，兼容车载场景下各种类别的手势识别任务。

对于无人驾驶应用，需要识别的手势主要分为车内乘客手势和车外指挥手势。对于车内、车外手势，我们需要使用不同的交互媒介，并以不同的方式进行响应。车外指挥人员只需要通过车载摄像头采集图像数据进行手势的检测识别，再通过特定的指令生成模块实现对车辆的控制。车内乘客作为车辆使用者除了输入手势交互指令以外，还应该可以从车载交互终端上查看手势识别系统的一些实时情况、修改系统参数及设置。

另外，为保留手势识别系统的可扩展性，需要保留系统开发维护人员对整个系统或其中部分模块更新修复的接口，由于这种更新操作通常是远程无线连接下完成的，所以这个接口预留为与系统更新服务器的交互接口。本系统的部署示意图如图 6-1 所示。

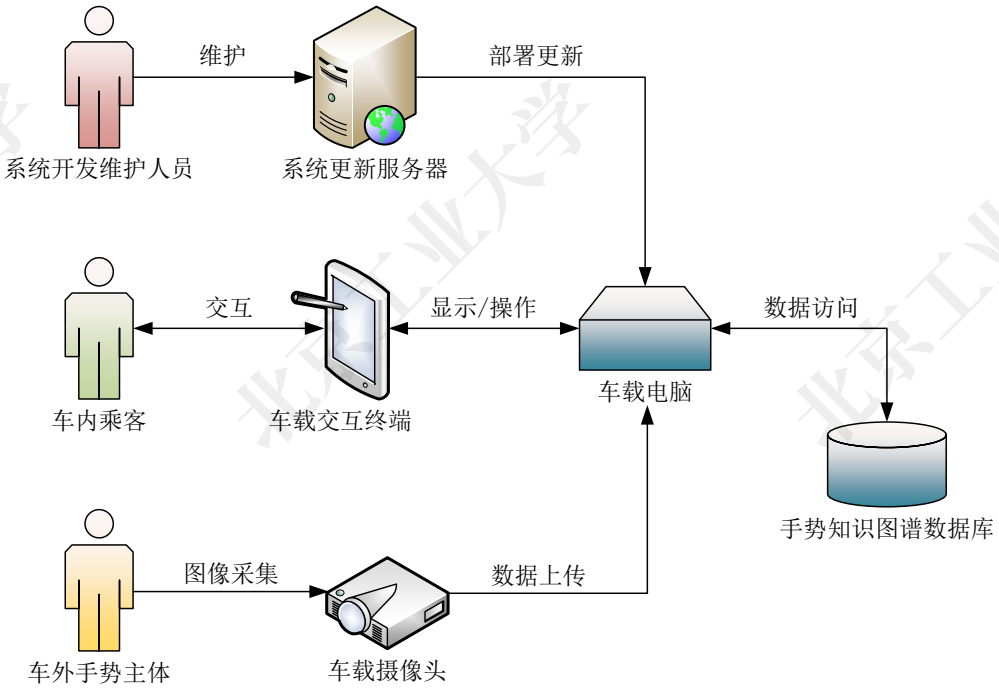


图 6-1 系统部署示意图

Fig. 6-1 System deployment diagram

6.1.2 功能性需求分析

下面将从车外手势主体、车内乘客以及系统开发维护人员三个参与者的角度对手势识别系统的功能性需求进行分析。

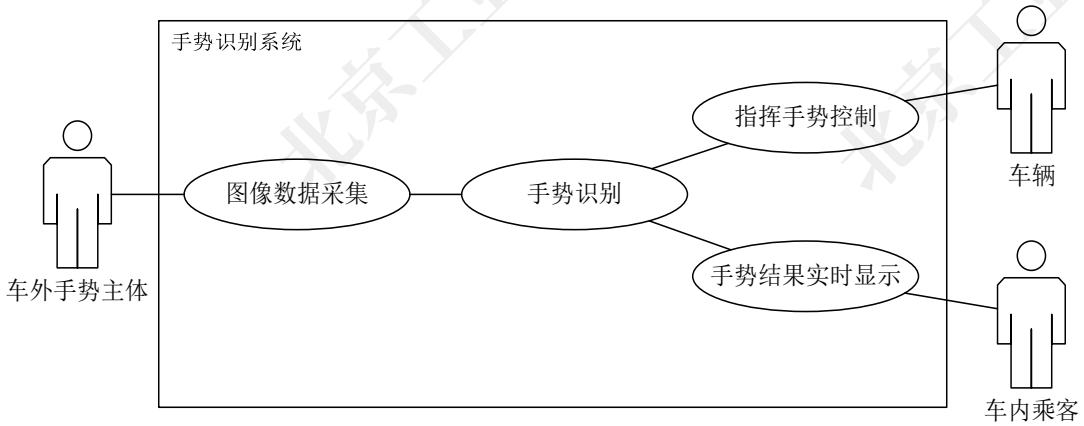


图 6-2 车外手势主体参与用例图

Fig. 6-2 Use Case Diagram of Outside-car Gesture Subject

车外指挥手势根据手势主体不同，可以分为交警指挥手势和其他指挥手势。两者在出现场景、手势动作和指令等级上都有所不同，手势识别系统需要在识别过程中对两者进行区分。两者参与系统的需求分析如图 6-2 所示，对应的用例描述如下：

（1）图像数据采集：调用车载摄像头，采集车身周围的图像信息，提取其中可能的手势视频流上传给手势识别模块。

（2）手势识别模块：手势识别系统的核心功能模块，使用本文所涉及的算法对多场景多目标的手势进行分析识别，预测手势的类别、内容等信息。

（3）手势结果实时显示：将手势识别模块的输出结果与相应的手势主体实时显示在车内的交互终端上，供用户查看。

（4）指挥手势控制：与无人驾驶车辆的控制模块相连，依据指挥手势的类别、内容等，按照符合无人驾驶安全标准的生成规则，将手势转换成相应的车辆控制指令。

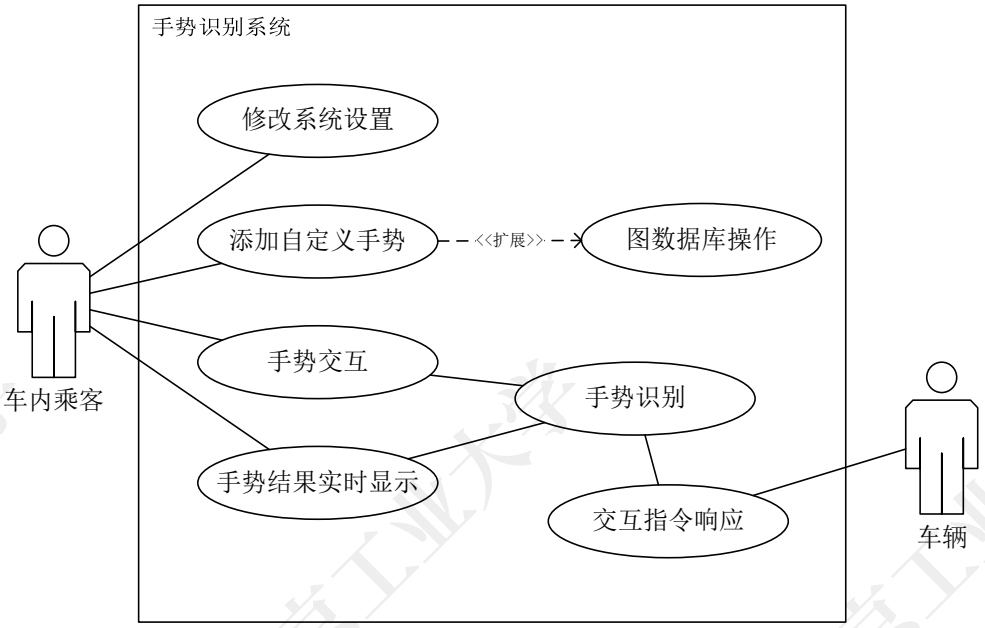


图 6-3　车内乘客参与用例图

Fig. 6-3 Use Case Diagram of Passengers

车内乘客的手势包括系统预定义的交互手势动作和车主按照自身使用习惯定义的手势。通过这些手势以及车载交互终端，乘客可与手势识别系统进行交互，其功能需求如图 6-3 所示，相应的用例描述如下：

（1）修改系统设置：乘客可以通过车载交互终端，对手势识别系统的一些基础参数进行设置，如系统界面风格，手势通知提示等级、弹窗提示的手势类别等。

（2）添加自定义手势：向手势知识图谱中添加新的手势实体或者修改现有的某个手势实体的类别、含义等信息，从而调整手势交互的方式。

（3）图数据库操作：按照知识图谱的设计，调用图数据库的相应接口，实现对知识图谱插入和修改。

（4）手势交互：车内乘客通过语音、终端等方式唤醒手势交互功能，并完成手势输入。

（5）交互指令响应：系统根据手势识别模块的输出结果，映射相应的系统功能，并调用底层系统接口相应乘客的手势交互，具体功能内容取决于车载电脑提供给手势识别系统的功能接口。

（6）手势识别与手势结果实时显示：这两个功能模块与车外手势主体的相应用例相同。

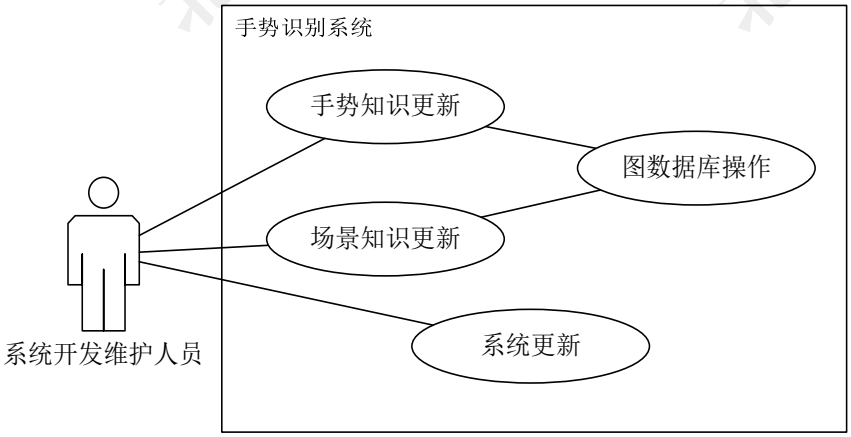


图 6-4 系统开发维护人员参与用例图

Fig. 6-4 Use Case Diagram of System Development Maintainer

对于系统开发维护人员，通过更新服务器完成系统的更新、优化以及其他维护工作的需求分析用例图如图 6-4 所示，相应的用例描述如下：

（1）手势知识更新：用于调整、部署知识图谱中存储的与手势相关的知识，如交警手势规定更新、综合分析发现的新的常用手势等，同时需要调整相应的指令生成、指令响应模块的规则模式。

（2）场景知识更新：用于调整、部署知识图谱中存储的与场景相关的知识，同时自动完成更新后的场景识别器的迭代训练。

（3）系统更新：用于部署整个手势识别系统级别的更新，如更新算法模块、更新预训练参数、调整交互界面等。

（4）图数据库操作：与车内乘客添加自定义手势时的相应用例相同。

6.1.3 非功能性需求分析

非功能性需求是在满足系统功能性需求的前提下，对系统的性能等方面的额外要求。非功能性需求能够保障系统的质量，提高系统的实际应用能力。对本系

统的主要非功能性需求分析如下：

（1）系统的手势识别速度以及指令响应时间在可接受范围内：对于系统处理的手势，其对应的指令具有不同的紧急程度，其响应具有实时性需求。但考虑到自动驾驶等车辆行驶场景中，指挥手势不是一种紧急指令。因此，这种实时性需求并不高，持平或者高于正常人的反应速度即可。根据相关研究以及经验数据，手势识别系统的总体响应速度应当在 600ms 以内，是可以保障安全且被用户接受的。结合前文对算法模型的分析，单帧图像的手势识别的速度在 150~200ms，同时完成车内和车外的手势识别需要双倍的识别时间，约为 300~400ms。因此在运算资源充足的情况下，只需保证系统其余的处理操作耗时在 200ms 以内，即可满足对实时性的需求。

（2）系统易用性：对于车内乘客的终端交互界面需要设计的尽可能简单，将功能尽可能集成化设计，以保障用户可以快速理解系统的使用方式，降低学习成本。

（3）平台限制：由于本章设计的系统将应用于车载场景，为了尽可能充分地利用车载电脑的运算性能，系统需要基于车载电脑安卓操作系统进行开发与优化。

6.2 系统设计

这一小节将以上节对于系统需求的分析为基础，开展系统的架构设计以及详细设计。

6.2.1 系统架构设计

本系统的整体架构如图 6-5 所示，主要包括基础层、支撑层、功能层以及交互层三个部分，下面我将对每一层的功能内容进行简要介绍。

（1）基础层包含一些第三方的应用软件和算法框架用于实现一些基础功能，或者作为接口工具，完成手势识别系统对外部功能以及硬件的调用。基础层的实现与操作系统和硬件配置相关，在进行系统间迁移时仅需修改这一部分。

（2）支撑层主要是通过调用基础层的接口或方法，实现系统的各个基础组件。支撑层组件作为完成系统功能的最小单位，被从各个应用功能中独立出来，将基础层工具适配成便于功能层调用的接口。各个组件互相之间没有调用，互不影响。

（3）功能层是面向系统的各个需求，通过调用支撑层组件，实现手势知识图谱维护、实时手势识别、识别结果显示、车辆手势控制等系统功能。各个系统功能直接对应于用户的功能性需求，互相之间存在数据的传递和共享。

（4）交互层实现了展示给乘客和系统开发维护人员之间界面接口，通过简单易懂的界面，为用户提供可视化的系统功能使用方式，完成与用户间的数据交互。

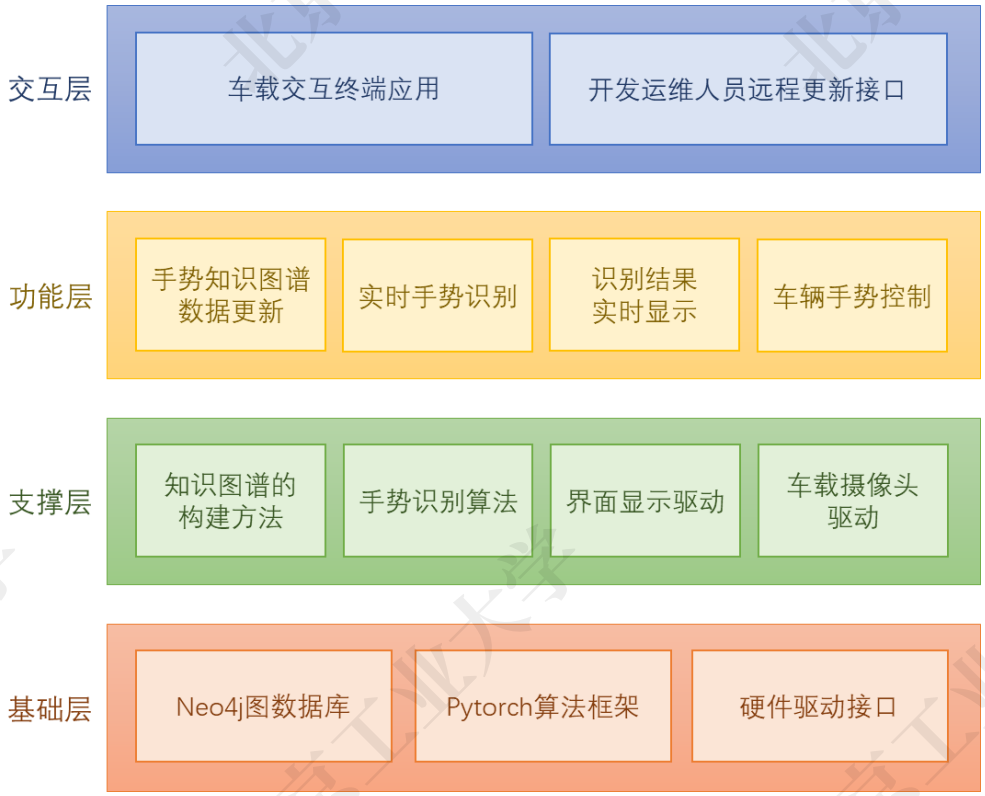


图 6-5 系统架构设计图

Fig. 6-5 System Architecture Design Drawing

6.2.2 系统详细设计

在系统架构设计的基础上，针对其中复杂的功能进行详细设计。支撑层和交互层的功能组件都是按照相应的功能或算法设计构造的顺序结构，不再单独进行介绍。接下来，将主要围绕手势知识图谱数据更新、实时手势识别、识别结果的实时显示、车辆手势控制四个核心系统功能进行设计。

（1）手势知识图谱数据更新功能

该功能可以被用户在添加或修改自定义手势动作时调用，或者被系统开发维护人员在部署系统更新时延迟调用。因此这个功能需要能循环处理多个各种类

型的实体，同时在完成知识图谱的插入和补全后，需要自动化地完成场景识别器、特定场景手势识别器的验证，并根据验证机结果判断是否需要进行优化训练。具体处理过程的活动图如图 6-6 所示。

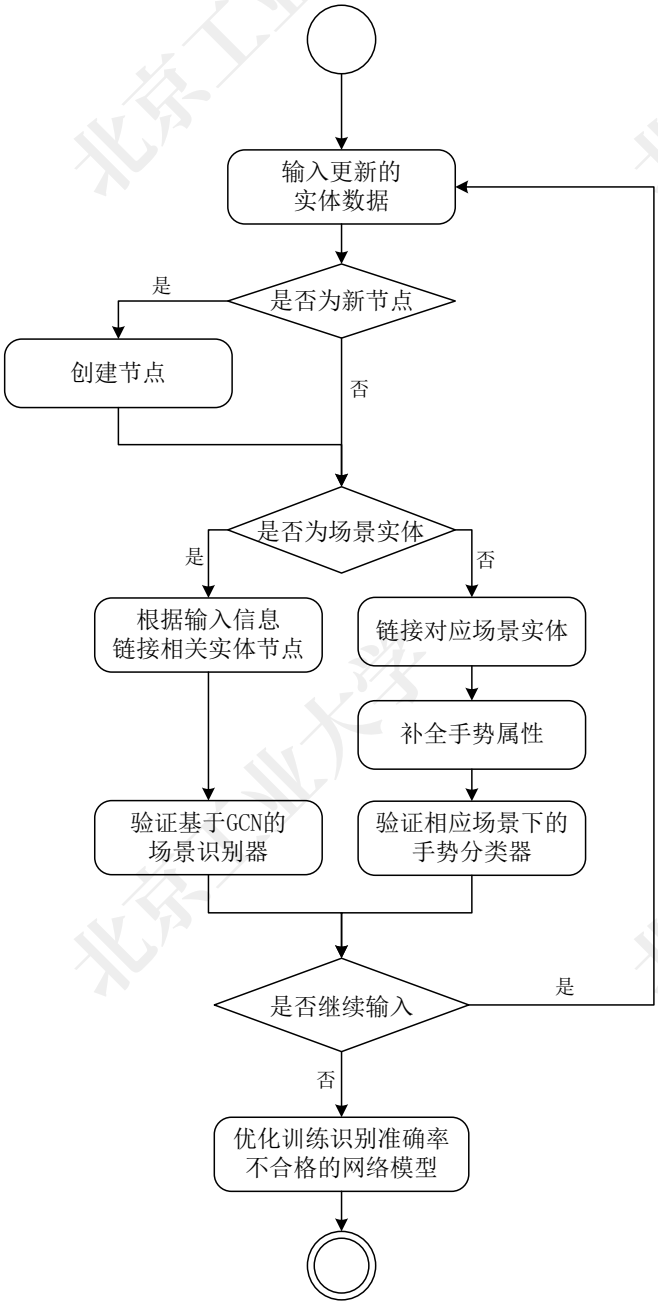


图 6-6 手势知识图谱数据更新功能的活动图

Fig. 6-6 Activity Diagram of the Gesture Knowledge Graph Data Update Function

（2）实时手势识别

该功能作为系统服务对于通过各个视频图像采集设备上产的视频进行切片并识别其中的手势，按照手势类别生成不同的指令数据包，进而调用对应的外部

功能。具体来说，对于各类指挥手势，应当按照指挥指令优先级提交给车辆手势控制模块，完成手势指挥指令到车辆控制指令的转换；而对于车内交互指令则直接调用车载系统的相应功能接口，如播放音乐、调节音量、开关窗等。这一功能的执行过程的活动图表示如图 6-7 所示。

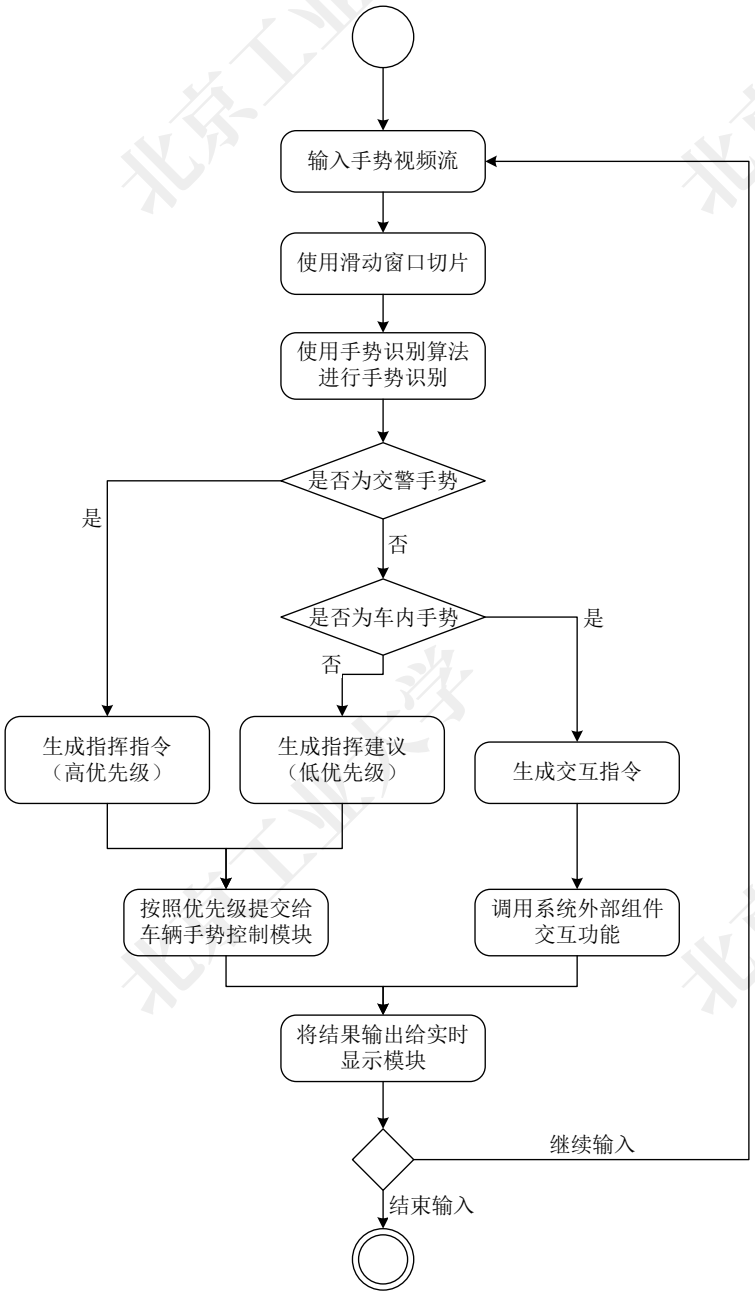


图 6-7 实时手势识别功能的活动图

Fig. 6-7 Activity Diagram of the Real-Time Gesture Recognition Function

（3）识别结果的实时显示

该功能比较简单主要是完成与手势识别模块的数据交互以及前端可视化界面的数据更新逻辑。由于显示实时手势视频图像的功能将占用系统内存，为节约

系统资源，仅在用户查看前端界面时，对界面内容进行更新维护。另外对于交警手势或其他重要手势，将以弹窗或者其他用户预设的方式，如提示音等，进行加强提示。用活动图表示其执行过程如图 6-8 所示。

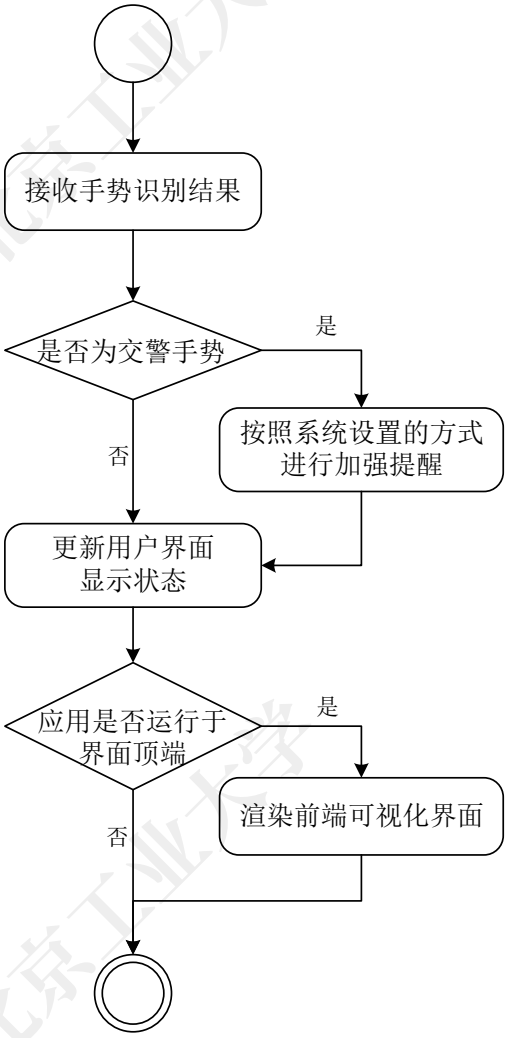


图 6-8 识别结果实时显示功能的活动图

Fig. 6-8 Activity Diagram of the Recognition Result Real-Time Display Function

（4）车辆手势控制

该功能主要用于分担手势识别算法的部分工作量，将与车辆控制相关的逻辑功能单独设计。这部分根据手势识别功能模块中识别到的控制指令，结合道路交通安全保障的原则，讲手势控制指令转化为车辆控制指令。其中，交警指令按照道路交通法规的要求，直接取代交通信号灯的指令；其他指挥指令中遇到停止或避让指令时，优先执行并核对避障等安全性判断；如果是行进指令的话，则仅作为参考辅助，或直接忽略。表示这一功能的活动图如图 6-9 所示。

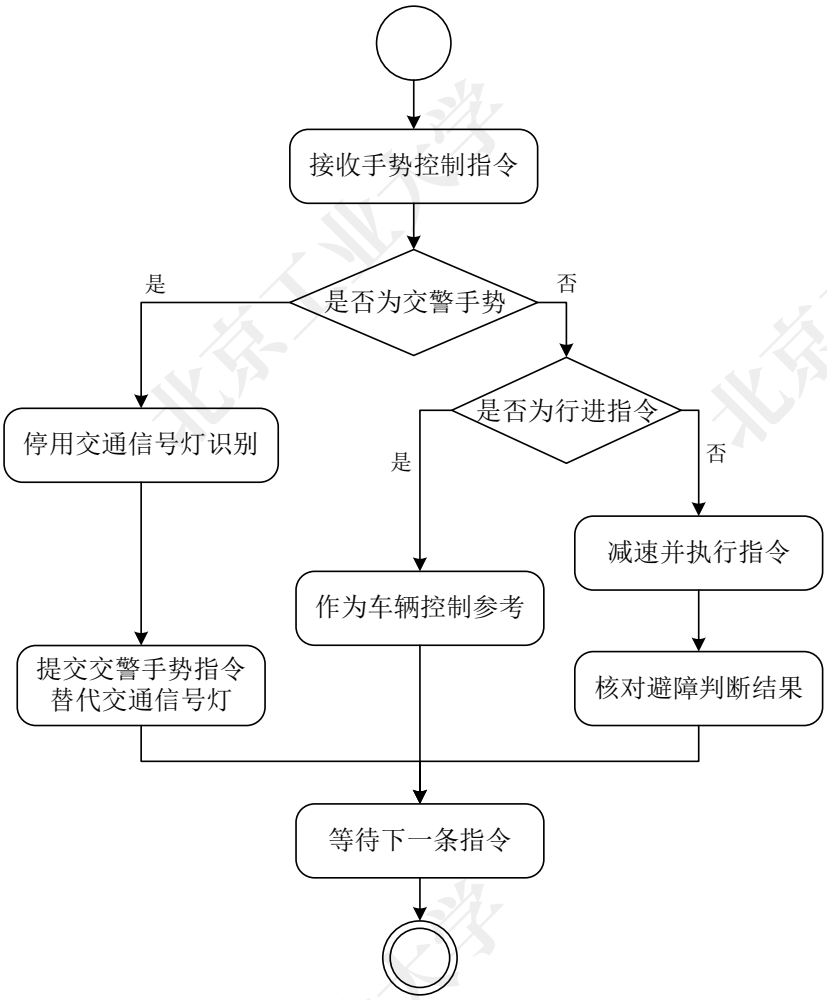


图 6-9 手势控制功能的活动图

Fig. 6-9 Activity Diagram of Gesture-based Controller Function

6.3 系统实现

本小节将对系统的实际运行环境进行介绍，并对系统进行详细的测试，展示原型系统的实现结果。

6.3.1 系统环境搭建

系统面向车载场景开发，但由于系统仅为原型阶段，为保证行车安全使用个人电脑模拟车载电脑的环境及性能配置等。具体来说，本章设计的原型系统部署在安卓平台上，调用外接摄像头和前置摄像头分别模拟车辆内外的图像采集模块，车辆控制部分使用软件接口进行模拟。

软件方面，需要配置用于存储知识图谱的图数据库 Neo4j 以及深度学习算法

框架 Pytorch，并基于 python 开发实现整个手势识别系统，通过 pyinstaller 进行封装打包。

6.3.2 系统功能实现结果

本章从需求分析和原型设计出发，对车载手势识别系统的功能逻辑及交互界面进行了开发，实现了可以区分车载场景下的不同手势，并生成与外部系统间的交互指令的原型系统。下面将对原型系统的实现结构进行展示。

车载手势识别系统的主界面如图 6-10 所示。由于本系统应用于车载场景，帐号通过从车载系统获取车辆唯一编码进行创建，绑定车辆。系统主界面即为手势识别结果的实时显示，包含三个部分：指挥手势图像显示；手势识别结果及中间置信度；系统功能按钮。图像显示将当前识别到的手势播放在界面中，供用户参考、确认。手势识别结果区域分别显示识别到的手势、场景识别结果以及手势识别器的中间结果，并根据手势类别进行高亮显示。系统功能按钮共有三个，分别是系统设置、手势输入和新增手势，其中手势输入及手动唤醒车内摄像头采集乘客手势，系统设置可以更改界面风格与高优先级手势的通知方式，新增手势则可以跳转到新的界面用于添加新的自定义手势并绑定到车内场景。

图 6-10 车载手势识别系统主界面

Fig. 6-10 The Main Interface of the Vehicle Gesture Recognition System

新增手势的用户界面如图 6-11 所示。包含功能按钮、显示窗口以及用于输入手势描述的文本框。手势描述的内容包括手势名称、动作描述以及映射的系统功能，手势识别系统将使用实体抽取方法处理描述文本，从而得到相应的实体和

关系。图像显示窗口回显标示姿态估计结果的手势图像，从而辅助用户确认是否正常完成了手势输入。功能按键包括返回、录制、播放以及下一步，对应添加自定义手势过程中必要的控制选项。

图 6-11 新增手势功能的用户界面

Fig. 6-11 User Interface of Adding Interaction Gesture

除了这两个主要界面以外系统还实现了对高优先级手势的增强警示，可选的方式包括播放声音和交互终端置顶弹窗等，效果如图 6-12 所示。

图 6-12 设置界面及置顶弹窗提示效果

Fig. 6-12 User Interface of Setting and the Effect of Top Pop-Up Prompts

6.4 本章小结

本章结合自动驾驶汽车的车载应用场景，使用本文所设计的算法，针对车载场景中交警手势、其他指挥手势以及用户交互手势，设计实现了跨场景一致、泛化能力强、可扩展性强的手势识别系统。从场景分析和需求分析出发，进行系统的整体架构和具体细节的详细设计，并在基于安卓系统模拟的车载电脑平台上开发了系统原型，展示了本文所设计的算法在实际应用场景下的可行性，以及使用该算法的基本设计开发过程。

结 论

本文研究了知识驱动的视觉手势识别算法，通过知识图谱的语义网络结构辅助知识特征的提取。首先，围绕手势本体的分析和设计，构建了有丰富节点数据的手势知识图谱；其次，参考 MoveNet 的设计思路，设计实现了轻量高效的视觉姿态估计方法，可以用于手部或躯干的骨架特征提取；最后，基于手势知识图谱和姿态估计模块，设计特征提取算法，融合知识和姿态特征用于混合场景下的手势监测与识别任务，并在开源数据集上进行了验证与分析。

本文的主要工作和成果如下：

（1）根据手势识别任务的需要，设计具有参考意义的自然交互手势本体，并提出了针对性的多模态知识抽取方法，使用基于 YOLOv8 目标检测算法的图像实体提取器结合场景模式构建三元组，并结合对标签文本的分析进行知识补全和完善。

（2）基于开源知识图谱 Wikidata 和三个开源手势数据集，使用所提出的方法抽取知识三元组，并处理在实际构建过程中遇到的实体歧义、属性对齐等问题，构建了一个包含 5 种典型手势应用场景，62 个手势实体以及共计 119 个交互目标、观察者等相关节点的手势知识图谱，为知识特征的提取提供了基础。

（3）设计了基于关键点的自底向上的多头姿态估计网络，实现相应的前后处理算法及损失函数设计，并对模型进行轻量化调整，从而高效快速的推理出被检测目标的姿态骨架，在开源数据集 COCO 上对设计的网络模型进行测试，取得了 80.1%的平均精度和 42.4ms/帧的识别速率，并分别训练躯干和手部姿态估计模块用于后续算法。

（4）设计了特征提取算法，分别提取知识图谱中与输入相关的场景知识特征和人体姿态骨架结构特征，并将两者融合，用于对泛应用场景下的手势进行检测识别，在多数据集混合场景的数据中对模型各方面性能进行了实验测试，取得了 76.7%的平均准确率，并针对新添加场景时算法模型进行迭代训练以适应新的应用场景的速度等性能进行了实验与评估，平均训练时间在 10 分钟左右，推理速率在 5~10FPS 左右，可以满足应用的需求。

（5）基于本文提出的算法框架，从车载场景的需求分析和系统架构设计出发，设计并实现了一个适用于无人驾驶车载场景的视觉手势识别系统原型，验证了该算法在实际应用中的实用性和可行性。

本文利用知识图谱中存储的手势先验知识，优化了手势识别算法的效果，在混合场景手势识别以及向新场景的扩展上表现出了优于现有方法的效果。但仍然存在一些有待改进完善的方面，值得后续进一步深入研究。具体的问题和改进点如下：

（1）由于多模态手势的构建在人工方面的巨大工作量，本文只按照所设计的手势识别算法中所需的知识数据，构建了最基础的手势知识图谱。在未来的研究工作中可以进一步完善多模态手势知识图谱，从而可以进一步实现在手势识别任务中融合更多的情景上下文先验知识，更好地提高手势识别算法的效果和可解释性。

（2）对于本文所设计的融合知识的视觉手势识别方法，在添加新的应用场景时，模型迭代训练的速度虽能满足应用需求，但仍有提升空间，可以结合迁移学习方法，对模型训练、数据预处理等过程进行优化和调整，进而提升算法适应于新应用场景的速度，从而可以完成更加复杂的应用任务。

参考文献

[1] 史元春.做好人机交互[J].中国计算机学会通讯,2022,18(3):40-47.

[2] 黄进, 张浩, 田丰. 面向动态交互场景的计算模型[J]. 图学学报, 2021, 42(3): 359.

[3] Vuletic T, Duffy A, Hay L, et al. Systematic literature review of hand gestures used in human computer interaction interfaces[J]. International Journal of Human-Computer Studies, 2019, 129: 74-94.

[4] 张维,林泽一,程坚,柯铭雨,邓小明,王宏安.动态手势理解与交互综述[J].软件学报,2021,32(10):3051-3067.

[5] Seneviratne S, Hu Y, Nguyen T, et al. A survey of wearable devices and challenges[J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2573-2620.

[6] yycoding. Kinect for Windows SDK 开发入门(十)：手势识别 上：基本概念[EB/OL]. [2022-11-17]. https://www.yycoding.xyz/post/2012/4/21/kinectsdk-gesturesdetection-part1-basicconception.

[7] 漆桂林,高桓,吴天星.知识图谱研究进展[J].情报工程,2017,3(1):004-025.

[8] 陈烨,周刚,卢记仓.多模态知识图谱构建与应用研究综述[J].计算机应用研究,2021.DOI:10.19734/j.issn.1001-3695.2021.05.0156.

[9] Xu Z, Sheng Y P, He L R, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4): 589-606.

[10] Carlson A， Betteridge J， Kisiel B , et al. Toward an Architecture for Never-Ending Language Learning[C]// Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010. AAAI Press, 2010.

[11] Wang K, Yin Q, Wang W, et al. A comprehensive survey on cross-modal retrieval[J]. arXiv preprint arXiv:1607.06215, 2016.

[12] 刘建伟, 丁熙浩, 罗雄麟. 多模态深度学习综述[J]. 计算机应用研究, 2020, 37(6):14.

[13] Liu Y, Li H, Garcia-Duran A, et al. MMKG: multi-modal knowledge graphs[C]//European Semantic Web Conference. Springer, Cham, 2019: 459-474.

[14] Wang M, Qi G, Wang H F, et al. Richpedia: a comprehensive multi-modal knowledge graph[C]//Joint International Semantic Technology Conference. Springer, Cham, 2020: 130-145.

[15] Wilcke W X, Bloem P, de Boer V, et al. End-to-end entity classification on multimodal knowledge graphs[J]. arXiv preprint arXiv:2003.12383, 2020.

[16] Sun R, Cao X, Zhao Y, et al. Multi-modal knowledge graphs for recommender systems[C]//Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 1405-1414.

[17] Oñoro-Rubio D, Niepert M, García-Durán A, et al. Answering visual-relational queries in web-

extracted knowledge graphs[J]. arXiv preprint arXiv:1709.02314, 2017.

[18] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

[19] Lehmann J, Isele R, Jakob M, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia[J]. Semantic web, 2015, 6(2): 167-195.

[20] Ferrada S, Bustos B, Hogan A. IMGpedia: a linked dataset with content-based analysis of Wikimedia images[C]//International Semantic Web Conference. Springer, Cham, 2017: 84-93.

[21] Yang J, Lu J, Lee S, et al. Graph r-cnn for scene graph generation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 670-685.

[22] Zareian A, Karaman S, Chang S F. Bridging knowledge graphs to generate scene graphs[C]//European conference on computer vision. Springer, Cham, 2020: 606-623.

[23] Alberts H, Huang T, Deshpande Y, et al. VisualSem: a high-quality knowledge graph for vision and language[J]. arXiv preprint arXiv:2008.09150, 2020.

[24] Chen K, Zhang D, Yao L, et al. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-40.

[25] Chang Y, Wang S. Knowledge-driven self-supervised representation learning for facial action unit recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 20417-20426.

[26] 郭萍. 基于视频的人体行为分析[D]. 北京: 北京交通大学交通运输学院, 2012.

[27] 叶旭庆. 基于 3D 卷积神经网络的人体行为识别[D]. 西安: 西安电子科技大学, 2015.

[28] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document re cognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[29] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[31] 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.

[32] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[33] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[34] Zagoruyko S, Komodakis N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.

[35] Huang G, Sun Y, Liu Z, et al. Deep networks with stochastic depth[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016,

Proceedings, Part IV 14. Springer International Publishing, 2016: 646-661.

[36] Arici T, Celebi S, Aydin A S, et al. Robust gesture recognition using feature pre-processing and weighted dynamic time warping[J]. Multimedia Tools and Applications, 2014, 72(3): 3045-3062.

[37] Elmezain M, Al-Hamadi A, Appenrodt J, et al. A hidden markov model-based continuous gesture recognition system for hand motion trajectory[C]//2008 19th international conference on pattern recognition. IEEE, 2008: 1-4.

[38] Dong C, Leu M C, Yin Z. American sign language alphabet recognition using microsoft kinect[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 44-52.

[39] Molchanov P, Yang X, Gupta S, et al. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4207-4215.

[40] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27.

[41] Zhang Y, Cao C, Cheng J, et al. EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition[J]. IEEE Transactions on Multimedia, 2018, 20(5): 1038-1050.

[42] 郭小爽.人机交互中的动态手势识别及应用研究[D]. 西安: 西安电子科技大学. 2014.

[43] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1110-1118.

[44] Zhou J, Cui G, Hu S, et al. Graph neural networks: A review of methods and applications[J]. AI Open, 2020, 1: 57-81.

[45] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1653-1660.

[46] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.

[47] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.

[48] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model[J]. ACM transactions on graphics (TOG), 2015, 34(6): 1-16.

[49] Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3d hands, face, and body from a single image[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10975-10985.

[50] Chen X, Liu Y, Ma C, et al. Camera-space hand mesh recovery via semantic aggregation and

adaptive 2d-1d registration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13274-13283.

[51] Huang L, Tan J, Liu J, et al. Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation[C]//European Conference on Computer Vision. Springer, Cham, 2020: 17-33.

[52] Chen X, Guo H, Wang G, et al. Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition[C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017: 2881-2885.

[53] Hou J, Wang G, Chen X, et al. Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition[C]//Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.

[54] Lang X, Feng Z, Yang X, et al. HMMCF: A human-computer collaboration algorithm based on multimodal intention of reverse active fusion[J]. International Journal of Human-Computer Studies, 2023, 169: 102916.

[55] Chen Y, Li Q, Kong D, et al. Yourefit: Embodied reference understanding with language and gesture[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1385-1395.

[56] Bordes A , Usunier N , Garcia-Duran A ,et al.Translating Embeddings for Modeling Multi-relational Data[C]//Neural Information Processing Systems.Curran Associates Inc. 2013. .

[57] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the AAAI conference on artificial intelligence. 2014, 28(1).

[58] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2d knowledge graph embeddings[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

[59] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Icml. 2001, 1(2): 3.

[60] Ultralytics. Ultralytics YOLOv8 Docs [EB/OL]. https://docs.ultralytics.com/. 2024-01-15.

[61] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[62] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[63] Ma N, Zhang X, Zheng H T, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design[C]//European Conference on Computer Vision. 2018.

[64] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[65] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-1324.

[66] He J, Zhang C, He X, et al. Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features[J]. Neurocomputing, 2020, 390: 248-259.

[67] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

[68] Xiong X, Wu H, Min W, et al. Traffic police gesture recognition based on gesture skeleton extractor and multichannel dilated graph convolution network[J]. Electronics, 2021, 10(5): 551.

攻读硕士学位期间所取得的科研成果

一、 参与发表的论文

[1] Hu Y, Wu T, Zhang J, Sun W, **Lv M**, et al. Perovskite-based photodetector for real-time and quantitative monitoring of sports motion[J]. Iscience, 2023, 26(11).

[2] Ren X, He J, Han T, Liu S, **Lv M**, et al. Exploring the effect of fingertip aero-haptic feedforward cues in directing eyes-free target acquisition in VR[J]. Virtual Reality & Intelligent Hardware, 2024, 6(2): 113-131.

[3] Zhang J, Xie H, Hu Y, Sun W, **Lv M**, et al. Printed 1d Perovskite Photodetector for Indoor/Outdoor Non‐Contact and Real‐Time Sports Training Monitoring[J]. Advanced Sensor Research, 2024: 2300158.

二、 参与发表的专利

[1] 何坚, **吕孟飞**, 张丞, 熊哲波. 基于时序线性人体蒙皮模型和图卷积网络的四方向交警手势识别方法, 国家发明专利. 已通过初审.

[2] 何坚, 魏鑫, 宋雪娜, **吕孟飞**. 基于目标检测和语义分割融合的障碍物检测方法, 国家发明专利. 已通过初审

[3] 何坚, 周睿, **吕孟飞**. 智能盲杖手柄, 外观设计专利. 已授权.

致　谢

在硕士研究生的学习生涯即将结束之际，我衷心感谢所有在我求学路上给予我帮助、支持和关心的人。正是你们的陪伴与指导，使我在学术道路上不断前行，在人生的舞台上不断成长。

首先，我要感谢我的研究生导师何坚老师。您严谨的学术态度、深厚的学术造诣和宽广的胸怀让我受益匪浅。在您的悉心指导下，我不仅学会了如何进行科学研究，更学会了如何面对困难与挑战，您的言传身教将是我未来人生道路上的宝贵财富。

同时，我也要感谢软件所的韩腾老师和孙伟老师。在您的耐心教导下，我掌握了更多的专业知识和科研方法，为我的研究工作奠定了坚实的基础。您对我的严格要求使我在学术上不断进步，您的鼓励和支持让我在困境中勇往直前。

感谢软件所的师兄师姐们，你们的热情帮助和无私分享让我少走了很多弯路。在与你们的交流中，我不仅学到了专业知识，更学到了为人处世的道理。你们的优秀品质和学术成就一直是我学习的榜样。

此外，我还要感谢我的室友和实验室的同学们。在求学路上，我们相互扶持、共同进步，一起度过了许多难忘的时光。你们的陪伴让我的生活更加丰富多彩，你们的友谊是我人生中最宝贵的财富。

特别感谢我的女朋友李志惠，在我求学期间给予我无尽的理解和支持。你的陪伴让我在面对困难时更加坚定，你的鼓励让我在追求梦想的路上更加勇敢。你的存在让我的人生更加美好，愿我们携手共度未来的岁月。

最后，我要感谢我的父母。是您们辛勤的付出和无私的爱让我能够专心学业，追求梦想。您们的期望是我前进的动力，您们的支持是我克服困难的勇气。在未来的日子里，我将更加努力，以优异的成绩回报您们的养育之恩。

再次感谢所有在我硕士研究生生涯中给予我帮助和支持的人。我将带着你们的期望和祝福，继续前行，为实现自己的人生价值而努力奋斗。

谨以此致谢，表达我对你们的感激之情。愿我们都能在未来的道路上继续前行，共同创造更加美好的明天。