结合多模态数据的深度神经网络在手语识别中的应用研究

李富钢

西南交通大学计算机与人工智能学院 四川成都

【摘要】随着人工智能和计算机视觉技术的快速发展，手语识别成为人机交互和无障碍通信研究的重要方向。传统手语识别方法往往依赖单一模态数据，如图像或视频，存在信息丢失、识别精度受限等问题。多模态数据融合结合视觉、深度信息、肌电信号、IMU 等，能够丰富语义表达，提高识别准确性。本文探讨深度神经网络如何结合多模态数据提升手语识别性能，分析关键技术、挑战及其应用前景，以期为智能手语翻译系统的优化提供参考。

【关键词】多模态数据；深度神经网络；手语识别；应用研究

【收稿日期】2025 年 2 月 15 日 【出刊日期】2025 年 3 月 31 日 【DOI】10.12208/j.sdr.20250006

Application of deep neural networks combined with multi-modal data in sign language recognition

Fugang Li

Southwest Jiaotong University, Chengdu, Sichuan

【**Abstract**】With the rapid development of artificial intelligence and computer vision technology, sign language recognition has become an important direction of human-computer interaction and barrier-free communication research. Traditional sign language recognition methods often rely on single modal data, such as image or video, which has problems such as information loss and recognition accuracy limitation. Multi-modal data fusion combined with vision, depth information, EMG, IMU, etc., can enrich semantic expression and improve recognition accuracy. This paper discusses how deep neural networks combine multi-modal data to improve sign language recognition performance, analyzes the key technologies, challenges and application prospects, in order to provide reference for the optimization of intelligent sign language translation system.

【**Keywords**】Multi-modal data; Deep neural network; Sign language recognition; Applied research

前言

手语是听障人士交流的重要方式，其自动识别对于促进无障碍信息交流和智能人机交互具有深远意义。传统的手语识别方法主要依赖单一模态数据[1]，如 RGB 图像或视频序列，受光照变化、遮挡和背景干扰的影响较大，识别精度有限。近年来，多模态数据融合技术的兴起[2]，为手语识别提供了新的突破口。通过结合视觉信息、深度数据、肌电信号和 IMU 传感器数据[3]，可更全面地捕捉手势特征，提高识别的鲁棒性。与此同时，深度神经网络凭借强大的特征学习能力，在模式识别领域展现出卓越的性能。本文将探讨基于多模态数据的深度神经网络在手语识别中的应用，分析关键技术及其优化策略，以期为智能手语识别系统的发展提供理论支持和实践参考。

1 多模态数据与深度神经网络对手语识别活动的促进作用与应用意义

多模态数据与深度神经网络的结合,极大地推动了手语识别技术的发展。相比于单一模态数据，多模态数据能够综合利用视觉、深度信息、肌电信号及惯性传感器数据，使系统在复杂环境中具备更强的鲁棒性。例如，视觉信息提供手势形态，深度数据增强空间结构感知，肌电信号捕捉肌肉活动特征，而惯性传感器则可感知动态轨迹[6]。这种信息互补性有效降低了光照变化、遮挡及手势相似度高等问题对识别精度的影响。与此同时，深度神经网络凭借强大的特征学习和模式识别能力，能够深度挖掘多模态数据之间的相关性，提高手势分类的准确性

和泛化能力。因此，多模态数据与深度神经网络的融合为手语识别提供了更高效、精准的解决方案，推动了无障碍交流技术的发展。实验采用肌电信号与IMU信号相融合在大规模数据集上取得了较稳定的高识别率，证明了多模态研究的潜力。多模态特征级融合框架如图 1 所示。

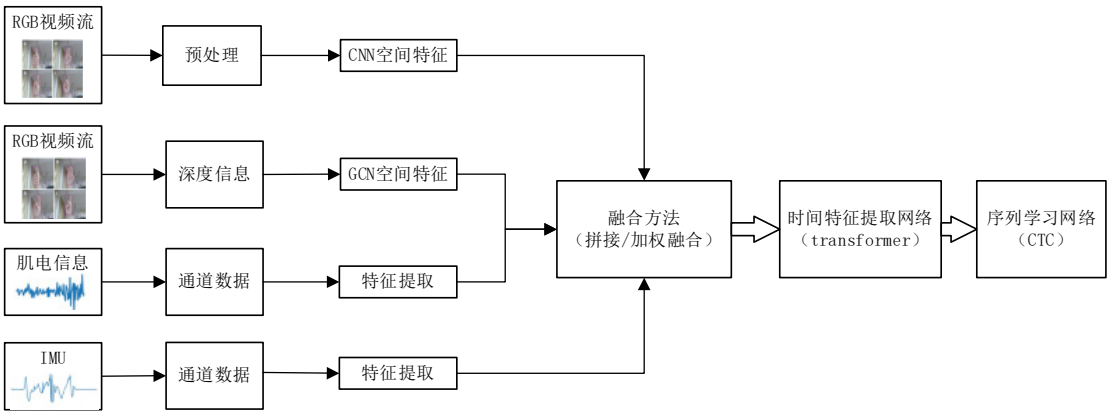

图 1　多模态特征级融合框架

2 多模态数据与深度神经网络在手语识别中的应用难点

2.1 数据多样性与标注难

手语识别涉及多种数据模态，包括 RGB 图像、深度数据、肌电信号和惯性传感器数据等，这些数据类型具有不同的采集方式和特性，导致数据预处理和融合面临较大挑战。此外，手语表达因手势、面部表情、身体动作等因素的组合变化而呈现高度多样性，不同国家和地区的手语体系也存在差异，使得数据采集和标准化工作更加复杂。同时，高质量的手语数据集往往需要大量的人工标注，特别是在多模态环境下，精准标注每一帧的手势信息耗时耗力，增加了数据准备的难度。因此，如何构建大规模、高质量的多模态手语数据集，提升自动化标注效率，是手语识别研究的基础难题。

2.2 模态融合与特征提取

多模态数据的有效融合是手语识别的关键问题之一。不同模态的数据具有不同的时空分布特性，例如视觉信息主要依赖手部形态，而肌电信号则反映肌肉活动模式，如何对这些异构数据进行合理融合以提升识别效果是一个重要挑战[7]。现有的融合方法包括数据级融合、特征级融合和决策级融合，但在实际应用中，模态间的时序同步、信息互补性及噪声干扰等问题仍然存在。此外，高效的特征提取也是影响识别精度的关键因素。传统手工设计的特征难以适应复杂多变的手势环境，而深度神经网络虽然能够自动学习特征，但如何设计适合多模态输入的神经网络结构，充分挖掘模态间的相关性，是亟待解决的问题。如图 2 所示 CNN 神经网络特征提取。

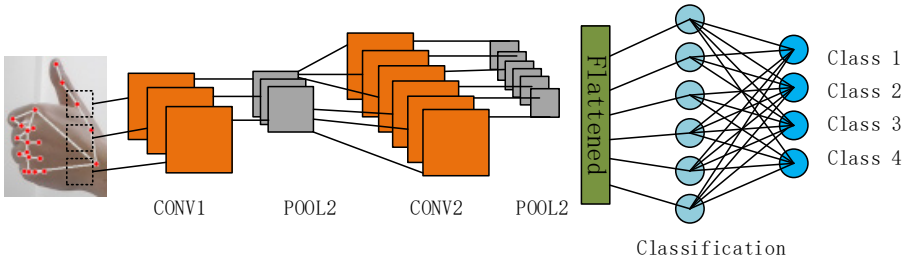

图 2　CNN 神经网络特征提取

2.3 模型训练与计算资源

深度神经网络的训练依赖于大量的数据和计算资源，而手语识别涉及时序信息和多模态数据，导致模型结构复杂，计算量庞大。例如，基于卷积神经

网络（CNN）和长短时记忆网络（LSTM）的混合模型，在长序列输入时容易出现梯度消失或梯度爆炸的问题，而基于自注意力机制的 Transformer 模型虽然在自然语言处理领域表现优异，但其计算复杂度较高，难以直接应用于实时手语识别[8]。此外，多模态数据的融合通常需要高维特征表示，导致计算资源需求进一步上升。因此，如何优化网络结构，提高模型的训练效率，降低计算开销，甚至在边缘设备上实现高效推理，是手语识别技术落地的重要挑战。

2.4 文化差异与适应问题

不同国家和地区的手语体系存在较大差异，如美式手语（ASL）、中国手语（CSL）和英国手语（BSL）等在手势形态、语法结构甚至词汇表达上均有所不同。因此，手语识别系统在不同手语体系间的泛化能力成为一大难点。此外，手语表达具有个体化特征，不同手语使用者在手势幅度、速度、习惯动作等方面存在较大差异，传统的固定模型难以适应所有用户。此外，在实际应用场景中，手语识别系统还需要考虑多模态交互，如口型、表情及身体语言的结合，这进一步增加了系统的复杂性[9]。

3 结合多模态数据的深度神经网络在手语识别中的应用策略

3.1 融合多模态数据提升手语识别精度

手语识别的准确性依赖于高质量的数据输入，多模态数据融合能够有效弥补单一模态的局限性，提高识别精度。数据级融合通过对原始数据进行同步采集和处理，使不同模态的信息保持完整性，从而减少信息丢失。两个模态的数据,视觉模态 V 和运动模态 M。

$$V_{fused} = \alpha \cdot V + (1 - \alpha) \cdot M \quad \text{-式 1}$$

其中，a 是权重系数，用于平衡两个模态的贡献。

特征级融合采用深度学习模型对各模态数据进行独立编码，并在中间层进行特征对齐和联合学习，以增强模态间的互补性。

$$F_{fused} = Concat(F_V(V), F_M(M)) \quad \text{-式 2}$$

其中，F_V 和 F_M 分别是视觉模态和运动模态的特征提取函数,Concat表示特征拼接。

决策级融合则通过独立训练各模态识别模型，并利用加权投票、置信度分析等策略进行最终预测，提高系统的稳定性和泛化能力。

$$P_{fused} = \beta \cdot P_V + (1 - \beta) \cdot P_M \quad \text{-式 3}$$

其中，P_V和P_M分别是视觉模态和运动模态的预测结果，β 是权重系数。

在多模态融合过程中，时序同步、数据尺度匹配和信息冗余控制是关键问题，需要采用时空注意力机制、自动加权学习等方法优化模态间的信息整合，确保模型能够充分利用多源数据，提高手语识别的准确性。

为了提升系统在复杂环境下的手语识别准确率。研究采用多模态数据融合策略，结合 RGB 图像、深度信息和肌电信号，以减少单一模态带来的信息缺失问题。

RGB 图像用于捕捉手势形态，深度信息增强空间结构感知，肌电信号提供手部肌肉运动模式。在数据处理阶段，系统利用时序对齐技术，确保不同模态数据的同步性，并通过特征级融合方法，将 CNN 提取的视觉特征、LSTM 捕捉的时序信息以及肌电信号转换后的频谱特征进行联合建模。为了优化模态权重，引入自适应加权机制，使系统能够根据不同环境动态调整模态贡献。实验表明，Mediapipe 骨骼特征融合[2]，及双通道特征融合[3]，都比单模态识别率分别高出 9%及 7%。该融合策略显著提高了手语识别的准确率，尤其在遮挡、光照变化等复杂条件下的识别性能得到有效提升，如表 1，表 2 所示。

3.2 优化神经网络结构增强模型计算能力

手语识别模型的计算效率与神经网络结构密切相关，合理优化网络设计能够降低计算复杂度，提高实时性能。轻量级神经网络结构通过减少网络层数、降低参数规模，实现高效推理，同时避免过拟合问题。时序建模优化利用基于注意力机制的 Transformer 架构或改进型长短时记忆网络（LSTM），增强模型对动态手势变化的捕捉能力，提高时间序列预测的精度。

表 1 单模态与多模态对比实验数据 1

模型	识别率
Signspeaker	61.1
CG-Recognizer	75.3
多模态	82.1

表 2 单模态与多模态对比实验数据 2

分类方法	P	R	ACC
YOLOv5	0.987	0.981	0.992
Mediapipe	0.916	0.907	0.904
Mediapipe 多模态	0.998	0.997	0.997

多任务学习策略在模型训练过程中同时优化手势分类和语义解析任务，提升识别的综合表现[10]。模型剪枝、知识蒸馏和量化技术可有效减少模型参数，提高推理速度，使其适应边缘设备部署需求。在计算资源有限的环境下，异构计算框架结合云端计算和本地计算，实现高效的手语识别推理，提高系统的可用性。

为了确保系统能够在移动设备上高效运行，同时保证识别精度。韩晓冰研究团队针对模型计算复杂度问题，采用轻量级 YOLOv7tiny 神经网络架构[4]，用于高效提取空间特征，减少计算开销，增强模型对手势实时变化的理解能力。此外，可以应用知识蒸馏技术，通过训练一个较大规模的教师模型，并将其知识传递给轻量级学生模型，从而在降低计算资源需求的同时保持识别性能[11]。为了进一步优化推理速度，研究人员利用模型剪枝和量化方法减少冗余计算，提高推理效率。实验结果表明，该优化策略使得系统在移动端设备上的运行速度提升了30%—40%以上，同时识别准确率仍保持在高水平，满足了实时手语识别的应用需求。

3.3 跨模态特征提取提升多样性适应性

手语识别涉及多种模态数据，其特征存在异构性，为了提高模型的适应性，需要构建统一的跨模态特征提取框架[5]。自监督学习方法通过无标签数据进行特征预训练，使模型能够自主学习多模态之间的关联，提高对不同手语体系的适应性。共享特征空间构建利用对比学习、跨模态嵌入等技术，使不同模态的数据映射到同一高维特征空间，从而降低模态间的信息偏差，提高识别的泛化能力。多尺度特征融合策略在网络不同层级提取局部与全局特征，并采用动态权重分配机制，使模型能够同时关注手势的细节变化和整体运动模式[12]。通过提升跨模态特征提取能力，模型可以有效应对不同手势风格、个体差异以及文化背景对手语识别的影响，提高系统的适应性。

为构建一个能够适应不同手语体系的识别系统，以支持多语言、多文化背景下的手语转换。采用跨模态特征提取策略，通过自监督学习方法，使模型在无标签数据上进行预训练，提升其对不同手势的泛化能力。同时，利用跨模态嵌入技术，将 RGB 图像、深度信息和肌电信号投射到同一高维特征空间，减少模态间的信息偏差。研究人员还设计了一种多尺度特征融合机制[5]，使模型能够同时关注局部手势细节和整体动态模式，增强对不同手语风格的适应性。在实验过程中，该系统在美式手语、英国手语和中国手语等多个手语数据集上均取得了较高的识别精度，并表现出较强的跨文化适应能力，为全球化手语识别提供了可行的技术方案。

3.4 应用数据增强技术提升识别鲁棒性

手语识别系统在实际应用中需要具备较强的抗干扰能力，数据增强技术能够有效提升模型的泛化能力，使其在复杂环境下仍能保持较高的识别精度。几何变换增强策略包括手势旋转、缩放、镜像翻转等，使模型适应不同角度和尺度的手势变化。时序数据增强采用随机裁剪、时序扰动和时间掩码，以提升模型对手势动态变化的鲁棒性[13]。模态变换增强策略利用 GAN（生成对抗网络）生成逼真的手势数据，以扩展训练集，提高模型对低质量数据的适应能力。多环境数据模拟结合噪声注入、背景变化等方式，使系统具备更强的抗干扰能力。通过多种数据增强技术的应用，手语识别系统能够有效适应复杂场景，提升在不同环境下的稳定性和鲁棒性。

通过提高模型在不同光照、背景和个体差异条件下的识别鲁棒性。引入数据增强策略，通过几何变换增强，使手势图像在旋转、缩放、翻转等情况下仍能被准确识别。同时，利用时序数据增强技术，采用随机裁剪、时序扰动和时间掩码等方法，提高模型对不同手势变换的适应能力。此外，研究人员使用生成对抗网络（GAN）生成多样化的手语数据，以扩展训练集规模，提高系统对低质量数据的识别

能力。为了进一步增强环境适应性，研究团队在训练过程中加入背景噪声模拟，使模型能够在复杂背景下保持稳定识别性能。实验结果表明，经过数据增强训练的系统在多变环境下的识别准确率提升了15%以上，并能有效应对光照变化、遮挡和不同手势个体特征的干扰。

4 结语

综上所述，本文主要研究了结合多模态数据的深度神经网络在手语识别中的应用，探讨了多模态数据融合策略、神经网络结构优化、跨模态特征提取及数据增强技术对识别精度和鲁棒性的提升作用。研究表明，多模态数据能够有效弥补单一模态的局限性，深度神经网络的优化可提升计算效率，而跨模态特征提取和数据增强策略能增强系统适应性。未来工作将进一步探索更高效的多模态融合算法，提高模型在不同手语体系中的泛化能力，并推动手语识别技术在实际应用场景中的落地，为无障碍信息交流提供更智能的解决方案。

参考文献

[1] 韩国军, 马晨, 王战备, 尹继武. 基于视觉的手指语识别系统设计[J]. 实验技术与管理, 2023, 40(4): 119-124.

[2] 倪广兴, 徐华, 王超. 融合改进 YOLOv5 及 Mediapipe 的手势识别研究[J]. 计算机工程与应用, 2024, 60(7): 108-118.

[3] 曹一丹, 王青山, 王琦. 一种双路并行的大规模手势识别模型[J]. 合肥工业大学学报(自然科学版), 2024, 47(5): 585-589.

[4] 韩晓冰, 胡其胜, 赵小飞, 等. 改进 YOLOv7-tiny 的手语识别算法研究[J]. 现代电子技术, 2024, 47(1): 55-61.

[5] 郭乐铭, 薛万利, 袁甜甜. 多尺度视觉特征提取及跨模态对齐的连续手语识别[J/OL]. 计算机科学与探索, 2024 , 18 (10) :2762-2769.

[6] 常龙飞, 牛清正, 宋伟, 等. 压阻式柔性应变传感纤维的手指姿态识别装置[J]. 西安交通大学学报, 2020, 54(8): 116-123

[7] 张金, 冯涛. 基于改进的 Faster RCNN 的手势识别[J]. 信息通信, 2019(1): 44-46.

[8] 胡宗承, 周亚同, 史宝军, 等. 结合注意力机制和特征融合的静态手势识别[J]. 计算机工程, 2022, 48(4): 240-246.

[9] 梁华刚, 王亚茹, 张志伟. 基于 Res-Bi-LSTM 的人脸表情识别[J]. 计算机工程与应用, 2020, 56(13): 204-209.

[10] 周舟, 韩芳, 王直杰. 面向手语识别的视频关键帧提取和优化算法[J]. 华东理工大学学报(自然科学版), 2021, 47(1): 81-88.

[11] 路飞, 韩祥祖, 程显鹏, 等. 基于轻量 3D CNNs 和 Transformer 的手语识别[J]. 华中科技大学学报(自然科学版), 2023, 51(5): 13-18.

[12] 杨光义, 丁星宇, 高毅, 等. 基于注意力机制的复杂背景连续手语识别[J]. 武汉大学学报(理学版), 2023, 69(1): 97-105.

[13] 郭丹, 唐申庚, 洪日昌, 等. 手语识别、翻译与生成综述[J]. 计算机科学, 2021, 48(3): 60-70.

[14] 张淑军, 张群, 李辉. 基于深度学习的手语识别综述[J]. 电子与信息学报, 2020, 42(4): 1021-1032.

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

[16] SONG J, AREIAS P M A, BEIYTSCHKO T. A method for dynamic crack and shear band propagation with phantom nodes[J]. International Journal for Numerical Methods in Engineering, 2006, 67(6): 868-893.

[17] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C].European Conference on Computer Vision. Heidelberg: Springer, 2016: 21-37.

[18] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

[19] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C].2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 7464-7475.

[20] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.