计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009879] [CSTR: 32024.14.csa.009879]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 动态手语识别综述①

王哲楷, 冯云霞, 王佳文

(青岛科技大学 信息科学技术学院, 青岛 266061)
通信作者: 王哲楷, E-mail: 1040081939@qq.com

**摘　要**: 手语是用手势比量动作, 根据手势的变化模拟形象或者音节以构成的一定意思或词语, 手语是听力障碍者或无法用言语交流的人普遍采用的一种交际工具. 随着计算机视觉和深度学习的不断发展, 手语识别技术随之出现并不断发展, 使普通人与聋哑人士交流成为可能. 然而, 动态手语的复杂性和变化性使得对手语的精确检测和识别仍具挑战. 为了推动该领域的研究, 本文深入调研现有的动态手语识别方法和技术. 首先, 调研了动态手语识别技术的发展历程和研究现状、常用动态手语数据集以及手语识别方法的评价指标. 其次, 重点调研了动态手语识别常用的深度学习模型, 探讨了动态手语识别技术面临的问题以及对应的解决方案. 最后, 基于手语识别现状, 总结了当前动态手语识别面临的问题, 并对下阶段如何提升手语识别性能进行分析和展望.

**关键词**: 动态手语识别; 计算机视觉; 深度学习

引用格式: 王哲楷,冯云霞,王佳文.动态手语识别综述.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9879.html

## Survey on Dynamic Sign Language Recognition

WANG Zhe-Kai, FENG Yun-Xia, WANG Jia-Wen

(School of Information Science and Technology, Qingdao University of Science & Technology, Qingdao 266061, China)

**Abstract**: Sign language is a communication tool commonly used by people with hearing impairments or those who are unable to communicate verbally. It utilizes gestures to convey actions and simulate images or syllables that form specific meanings or words. With the continuous development of computer vision and deep learning, sign language recognition technology has emerged and continued to develop, making it possible for hearing individuals to communicate with the deaf or mute. However, the complexity and variability of dynamic sign language still pose challenges for its accurate detection and recognition. To promote research in this field, this study conducts an in-depth review of existing dynamic sign language recognition methods and technologies. First, the development history and current research status of dynamic sign language recognition technology, commonly used dynamic sign language datasets, and evaluation metrics for sign language recognition methods are reviewed. Second, deep learning models frequently used in dynamic sign language recognition are examined, and the challenges faced by dynamic sign language recognition technology, along with corresponding solutions, are discussed. Finally, based on the current status of sign language recognition, the challenges of dynamic sign language recognition are summarized, and an analysis and outlook are provided regarding the potential improvements to sign language recognition performance in the next stage.

**Key words**: dynamic sign language recognition; computer vision; deep learning

① 基金项目: 国家自然科学基金面上项目 (62171246); 山东省自然科学基金重大基础研究项目 (ZR2021ZD12)
收稿时间: 2024-08-16; 修改时间: 2024-09-24, 2024-11-07, 2025-01-02; 采用时间: 2025-01-21; csa 在线出版时间: 2025-04-01

手语是聋哑人融入社会并在社会中正常生活的重要手段. 它是聋哑人交际的一种手的语言, 还是聋哑人与非聋哑人建立交际的途径. 在手语中, 每个手势都有特定的意义, 每个手语词汇都用不同的手势来表示, 一段连贯的句子则包含多个词汇, 需用一系列手势动作来表示. 随着计算机视觉和深度学习的不断发展, 手语识别技术也逐渐发展, 为聋哑人提供了无障碍参与社会各项活动的可能性. 随之出现了基于传感器和基于视觉的手语识别方法. 基于可穿戴传感器的手语识别方法[1]主要通过数据手套等设备直接检测手部和各个关节的空间信息作为输入数据, 具有不依赖于背景和照明条件的优势, 识别精度也往往高于基于视觉的手语识别精度. 但因成本高、依赖硬件设备, 基于传感器的手语识别并未成为手语识别的主流方法[2]. 相反, 基于视觉的手语识别方法[3]因其成本低廉、不依赖硬件设备等优点, 已居于手语识别方法的主导地位[4]. 随着计算机硬件的进步和人工智能大模型的发展, 传统的基于视觉的机器学习方法在特征表示、变化性、可扩展性、上下文理解、对噪声的敏感性和实时处理等方面逐渐落后, 而蓬勃发展的深度学习技术能够有效解决这些问题. 且与传统机器学习方法相比, 深度学习在处理大规模数据和复杂模式识别任务方面表现出色[5]. 所以基于深度学习的手语识别技术成为当下研究热点. 虽然目前已经存在大量的手语识别检测算法, 且在静态手语识别方面取得了较好的实验结果. 但因动态手语的复杂性和变化性, 在动态手语识别领域尤其是连续语句识别领域的研究仍具挑战. 为了推动未来的研究, 本文将对近年来的动态手语识别方法和技术进行归纳整理, 主要从动态手语识别技术的发展历程和研究现状、常用动态手语数据集以及手语识别方法评价指标来介绍. 重点调研了动态手语识别中常用的深度学习算法和模型, 探讨当下基于深度学习的手语识别方法的不足、面临的挑战, 以及展望下阶段动态手语识别方法的发展方向.

## 1 动态手语识别技术发展和研究现状

按照手语识别研究来说, 手语识别可分为静态手语识别和动态手语识别, 动态手语识别则包括孤立词识别和连续语句识别, 其分类结构如图 1 所示. 静态手语识别是指从图像或照片中识别和理解手语表达的过程. 它将手势的位置、形状、动作等信息转化为对应的手语词汇. 所以静态手语识别本质上是一种图像分类问题. 动态手语识别是指从连续的手势动作序列中识别和理解手语表达的过程. 与静态手语识别不同, 动态手语识别关注的是手势的时间序列信息, 以捕捉手势的动态变化和运动轨迹. 相较于静态手语识别, 动态手语识别多了跟踪这一步骤. 因其灵活性和扩展性高于静态手语识别, 且手语主要为一系列连贯的动作, 所以动态手语识别更符合实际需求.

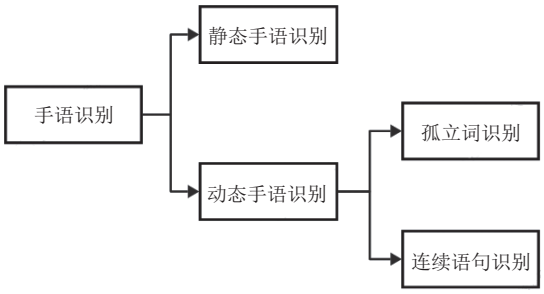


图 1 手势识别分类

在早期的动态手语识别研究中常使用隐马尔可夫模型 (HMM)[6]. Pashaloudi 等人[7]开发了可以识别希腊手语 (GSL) 的系统. 该论文提出利用手形几何特征的特征提取方法, 从每幅图像中提取描述性特征来表示 GSL 字母. 将特征向量以序列的形式输入到 HMM 中. Yoon 等人[8]则是通过手部定位和手部跟踪来生成手部轨迹, 并使用位置、角度和速度的组合特征生成的离散向量作为 HMM 输入. 但是上述的方法都是在非复杂背景上离线进行的. Binh 等人[9]开发的无约束环境下的实时手势识别系统由实时手势跟踪、训练手势和伪二维隐马尔可夫模型 (p2-dhMM) 这 3 个模块组成. 利用卡尔曼滤波和手部斑点分析方法对手部进行跟踪, 得到运动描述和手部区域. 该算法利用肤色进行手势跟踪和识别, 对背景聚类具有较强的鲁棒性. 但是, 这种方法研究的是手的姿势, 而不是动态的手部运动轨迹. 而相同的手势在形状、轨迹和持续时间上都会有所不同. 基于上述问题, Elmezain 等人[10]开发了基于时空特征提取的实时手势识别系统. 在预处理阶段, 使用颜色和 3D 深度图来检测和跟踪手部, 然后使用位置、方向和速度的 3D 组合特征训练. 最后使用左右带状拓扑 (LRB) 结合 Viterbi 路径来识别手势路径. 实验结果表明, 此系统在孤立的手势识别率达到 98.33%. HMM 在早期的动态手语识别中被广泛采用的原因之一是其对时序关系的建模能力. 手势通常具有一定的时间顺

序和动态特征, HMM 可以捕捉手势序列中的时序关系, 使手语识别系统能够区分不同的手势. 另外, 许多手语识别方法是在其他学习算法上开发的. 例如 Yang 等人[11]使用时滞神经网络 (TDNN)[12]从提取的轨迹中学习运动模式, 识别了 40 种 ASL 手势, 证明了其在手语识别方面的可用性. 早期动态手语识别研究也常用支持向量机 (SVM)[13]、最近邻算法 (KNN)[14]等机器学习算法, 对提取的特征进行分类和识别.

随着深度学习的发展, 特别是卷积神经网络 (CNN)[15]和循环神经网络 (RNN)[16]的应用, 基于深度学习的动态手语识别取得了显著进展, 并在许多任务中超越了传统的方法. Yang 等人[17]提出了基于视频的中国手语识别方法. 该方法从视频中提取上身图像, 采用预训练的卷积网络模型 (CNN) 对图像中的姿态进行识别. 同时简化手形分割过程, 避免特征提取中的信息丢失. Huang 等人[18]提出一种新的三维卷积神经网络 (3D-CNN), 它可以在没有任何先验知识的情况下自动从原始视频流中提取具有区分性的时空特征, 避免了设计特征. 将包括颜色信息、深度线索和人体关节位置在内的多通道视频流输入到 3D-CNN 中, 整合颜色、深度和轨迹信息. 实验表明它比基于手工特性的传统方法更有效. Chai 等人[19]通过设计两个递归神经网络 (2S-RNN) 来处理连续动态手势识别问题. 该方法使用定位-识别策略, 首先将连续的手势分割成单独的手势, 利用 2S-RNN 对每个单独手势进行识别. 其中手势分割模块是基于 R-CNN 训练的手部检测器. 手势识别模块 2S-RNN 融合多模态特征, 即 RGB 和深度信道. 实验结果表明, 该框架具有良好的性能.

然而, 这样的动态手语识别方法仍存在一些缺陷, 如难以识别复杂的手势, 大多数动态手语的识别精度较低, 较大的视频序列数据训练存在潜在问题等. 因为时间信息在动态手语识别过程中起着关键作用, 所以同时学习时间和空间特征对手语识别来说更有效. 目前主要采取混合网络模型的方法, 如卷积神经网络与循环神经网络结合. Zhu 等人[20]提出一种基于三维卷积 (3D-CNN) 和 LSTM 的多模态手势识别方法. 该方法利用 3D-CNN 从输入视频中提取短时空特征, 利用 LSTM 进一步学习长期时空特征. 采用空间金字塔汇聚算法 (SPP)[21]对时空特征进行归一化. 通过 RGB 和深度网络分别进行训练, 并将预测结果进行融合, 得到最终的分类结果. Liao 等人[22]提出一种基于深三维剩余 ConvNet 和双向 LSTM 网络的多模态动态手语识别方法, 即 BLSTM-3D (B3D ResNet). 该方法由 3 个主要部分组成. 首先将手部对象定位在视频帧中, 降低网络计算的时间和空间复杂度. 然后 B3D ResNet 从视频序列中自动提取时空特征, 经过特征分析建立与视频序列中每个动作相对应的中间评分, 最后通过对视频序列进行分类, 识别动态手语.

根据动态手语识别的发展历程来看, 在神经网络盛行之前, 隐马尔可夫模型 (HMM) 在手语识别领域被广泛使用. 但 HMM 本质是统计分析模型, 无法考虑长序列信息, 时序表征能力与 RNN 网络相比逊色很多且无法处理上下文信息. 随着计算机硬件和深度学习的发展, 卷积神经网络和循环神经网络尤其是 3D-CNN 等混合网络的出现, 将丰富的语义表达、精确的词汇间隔以及长时依赖关系结合, 提升了模型性能, 逐渐成为动态手语识别的主流方法.

虽然目前针对动态手语识别的研究取得了一些进展, 但仍面临着诸多挑战, 例如数据稀缺性、模型泛化能力、实时性要求等. 当下的主流算法主要依靠纯视觉方案对手部进行定位和提取信息, 但动态手语的复杂性和变化性受到个体差异、环境干扰等因素的影响, 单纯依靠纯视觉方案不能够准确提取手部信息. 随着近几年姿态估计的兴起与发展, 研究人员开始将姿态估计和骨骼关键点提取方案应用于动态手语识别, 解决复杂条件下手部信息提取的问题. 经过调研发现, 此方案在面对手语动作多样性、姿态变化、光照变化、背景干扰和手部遮挡等问题具有一定优势. 目前姿态估计和骨骼关键点提取方案仍在初始阶段, 存在着巨大的研究空间和应用价值, 但大部分手语识别领域的综述文章, 如文献[1,23,24]等, 缺乏对该方法的综述和总结归纳. 文献[1]主要针对动态手势识别中的孤立词识别进行综述, 详细介绍了如 RGB、深度和光流数据的多模态输入. 重点分析了多模态信息融合的诸多方法以及多模态信息融合对于提高模型性能的优势. 文献[1]还介绍了基于机器学习和深度学习的动态手势识别算法. 重点介绍了基于深度学习的 3 种动态手势识别方法, 即基于双流网络、基于 3D 卷积神经网络和基于递归神经网络的动态手势识别方法, 分析现有方法的不足和未来发展方向. 文献[23]则主要对静态手语识别、孤立词识别以及连续语句识别的机器学习和深度学习方法进行概括. 介绍了手语识别的具体步骤, 分析

概括了基于动态时间规整、HMM、CNN、LSTM 以及混合网络的方法. 文献[24]则主要针对中国手语识别中的机器学习和深度学习方法进行概括, 分析中国手语识别的不足和未来发展方向. 相较于文献[1]针对的是范围更广的手势识别, 本文则是针对手势识别中的手语识别这一特定领域, 不只对孤立词识别进行调研, 还包括对连续语句识别的概括分析. 相较于文献[23]主要针对大部分主流方法的系统性概括和文献[24]主要针对中国手语识别方法的综述, 本文则是详细介绍各国动态手语识别中的深度学习算法, 包括对 CNN、LSTM 以及混合网络模型方法的介绍和优劣分析. 本文还对常用的动态手语数据集和评价指标进行概括分析. 另外相较于文献[1,23,24]等工作, 本文还对近几年应用于手语识别领域的姿态估计和骨骼关键点检测方法进行分析, 详细介绍了姿态估计和骨骼关键点检测方法的优劣以及未来发展前景, 希望可以为姿态估计和骨骼关键点应用于手语识别领域的发展提供动力.

## 2 动态手语识别数据集

数据集是深度学习中必不可少的组成部分, 数据集的质量影响着深度学习算法的成果, 丰富的数据集有助于开发、训练和改进算法. 本文对近几年常用的动态手语数据集及相关项目进行调研, 结果如表 1 所示.

表 1 手语常用数据集

| 数据集名称 | 国家 | 标签类型 | 标签数量 | 样本数量 | 数据类别 |
|---|---|---|---|---|---|
| SIGNUM | 德国 | 句子 | 780 | 33210 | RGB |
| ASLLVD | 美国 | 孤立词 | 3300 | 9800 | RGB |
| RWTH-PHOENIX-Weather | 德国 | 句子 | 1200 | 45760 | RGB |
| DEVISIGN | 中国 | 孤立词 | 700 | 30000 | RGB |
| WLASL | 美国 | 孤立词 | 2000 | 21083 | RGB |
| CSL | 中国 | 孤立词、句子 | 600 | 150000 | RGB、骨骼、深度 |

(1) SIGNUM

SIGNUM[25]项目是德国研究基金会资助的研究项目, 旨在开发一个基于视频的自动手语识别系统, 实现独立于签名者的连续识别. 作者因此创建了 SIGNUM 数据集, 其中包含来自 25 名当地手语使用者的视频数据. 这些使用者涵盖了不同的性别和年龄范围, 以确保数据的多样性. 数据集包含 450 个手语词汇和 780 个句子, 这些句子是基于德国手语 (DGS) 中的 450 个基本手势, 并结合手部和面部表达. 数据采集过程是使用一台高清摄像机, 捕捉使用者的手语表达. 为了保证数据的质量和一致性, 使用统一的拍摄环境和指令. 每个使用者被要求执行一系列句子, 涵盖不同的语法结构和手势组合. 在采集过程中, 使用者的手势和面部表达被同时记录. SIGNUM 数据集的建立解决了签名者之间的个体差异问题, 改变了识别系统在独立于签名者的任务中表现不佳的现状.

(2) ASLLVD (American sign language lexicon video dataset)

美国手语词典视频数据集 (ASLLVD)[26]是一个超过 3300 个美国手语 (ASL) 词汇的视频集合. 每个词汇由 1–6 名本地手语使用者进行表演, 总共约有 9800 个视频样本. 数据集包含多个同步视频, 使用了 4 台同步相机进行视频捕捉, 包括手语使用者的侧面视图、头部区域的特写视图、半速高分辨率的正面视图和全分辨率的正面视图. 注释包括词汇标签、手语开始和结束时间、两只手的手势标签, 以及对手势类型的形态和发音进行分类. 对于复合手势, 数据集还包括对每个形态要素的注释. 在面对姿态完全不同但意义相同的手势时进行恰当的区分、分类、标记和注释. 另外数据集还包括标志变体的数字 ID 标签、未压缩的原始视频和摄像机校准, 对在手部和面部区域产生高保真度视频进行自动皮肤区域分割, 并将自动皮肤区域分割应用于每个视频帧, 进行归一化处理以确保视频的亮度均匀. 此处理减小了不同数据捕捉会话之间捕捉设置差异引入的方差. ASLLVD 数据集的建立促进了基于计算机视觉的手语识别的发展, 目前被广泛使用.

(3) RWTH-PHOENIX-Weather

RWTH-PHOENIX-Weather 数据集[27]是从德国公共电视台 PHOENIX 的天气预报中获得的德语手语视频组成的视频集合. 目前包含由 9 个人提供的 45760 个视频样本, 影片的分辨率为 210×260 像素, 帧率为 25 帧. 其中包含 5356 个与天气预报相关的句子以及 1200 个德国手语词汇. 因为天气预报的局限性, RWTH-PHOENIX-Weather 数据集使用的词汇大部分为地理方面的, 如“阿尔卑斯山”“柏林”“莱茵河”等, 对特定手语 (如分类符号) 的使用也有限. 虽然它使用受限制的真实数据, 但与其他手语数据集相比, 它的规模相当大. 虽然数据集的视频没有在实验室条件下录制, 但是电视台控制灯光条件和手语人在摄像机前的位置, 并且

手语人需在灰色背景前穿深色衣服. 这使得数据集中的视频具有很高的质量, 很适合用来进行手语识别项目. 因此 RWTH-PHOENIX-Weather 数据集成为大部分手语识别任务中常使用的手语数据集.

(4) CSL-Daily (Chinese sign language corpus)

中国手语数据集 (CSL)[28]是由中国科学技术大学自 2015 年起利用 Kinect 采集的视频集合. 其中包含 25 000 个标记的视频实例, 共有超过 100 h 的视频. 由 50 个操作者拍摄, 每个操作者重复 5 次. 视频标签分为孤立词和连续语句. 其中单词有 500 类, 每类含 250 个样例, 包含 21 个骨架关节点坐标序列. 句子有 100 个, 共有 5 000 个视频, 每一个句子平均包含 4–8 个单词. 数据集主要围绕人们的日常生活 (例如旅行、购物、医疗) 展开. 数据集是使用 Microsoft Kinect 构建的, 因此具有 3 种数据模式: 分辨率为 1280×720 像素、帧速率为 30 f/s 的 RGB 视频; 分辨率为 512×424 像素、帧速率为 30 f/s 的深度视频; 每个手势的 25 个骨架关节位置. 作为中国手语数据集且拥有较为丰富的数据模式, CLS 被更多用来处理中国手语识别项目, 且经常被用于验证手语识别性能.

手语数据集是手语识别技术的基础, 手语识别方法本质上依靠数据驱动. 当下主流手语数据集环境大部分是实验室环境, 缺乏真实环境的数据, 会影响真实环境中的识别表现. 且像中国手语数据集 (CSL) 这样同时包含 RGB、骨骼、深度的多模态信息的数据集较少, 大部分数据集只包含 RGB 这一项数据类别. 另外大部分手语数据集只包含对手部的标注, 缺乏对肢体、面部的标注, 但手语的表达是多方面的, 多方面的信息表达有助于提高手语识别的准确率. 另外, 一个手语数据集只包含一个国家的手语, 不利于对多个国家的手语进行比较研究, 也不利于促进手语使用者的国际化交流.

## 3 手语识别方法评价指标

在手语识别领域, 评价指标对于衡量算法性能和效果至关重要. 目前手语识别方法的评价指标主要包括准确率、*F*1-score、混淆矩阵等.

(1) 准确率 (Accuracy): 准确率是最基本的评价指标, 表示模型在所有样本中正确分类的比例. 在手语识别中, 准确率可以帮助评估模型对手语动作正确分类的能力, 通常情况下模型的准确率越高, 识别性能越好. 但在某些情况下, 准确率高并不代表模型的性能好. 例如在面对不平衡数据集时, 某个类别的样本数量远远多于其他类别, 模型可能会更倾向于预测这个类别, 导致准确率高但对少数类别的分类效果较差. 另外准确率并不区分模型在不同类别上的表现, 无法反映模型对各类别的区分能力.

(2) 精确率 (Precision) 和召回率 (Recall)[29]: 精确率指模型预测为正例的样本中有多少真正的正例, 衡量了模型的预测准确性. 召回率指所有真正的正例中有多少被模型正确预测为正例, 衡量了模型的覆盖能力. 这两个指标通常是一对矛盾的度量. 精确率和召回率对模型的误差类型 (假正例、假负例) 敏感, 如果任务对不同类型错误的惩罚不同, 单独使用精确率或召回率可能不够准确. 所以在实际应用中, 需要根据具体任务需求来平衡精确率和召回率.

(3) *F*1-score[30]: *F*1-score是精确率和召回率的调和平均值, 综合考虑了精确率和召回率, 对于不平衡数据集具有较好的表现, 可以帮助评估分类模型在不同数据集上的性能表现. 但是, *F*1-score 因为假设了二分类情况, 所以对于多类别分类或回归等问题, *F*1-score 可能不是最合适的评价指标.

(4) 混淆矩阵 (confusion matrix): 混淆矩阵可以展示模型在每个类别上的分类表现, 包括真正例、假正例、真负例和假负例的数量, 有助于评估模型的分类准确性. 通过混淆矩阵, 可以直观地了解模型在不同类别上的表现, 帮助发现模型的弱点和改进空间. 在处理类别不平衡的数据集时, 混淆矩阵能够提供更全面的性能评估. 但混淆矩阵通常用于二分类问题, 对于多类别分类或回归等问题, 混淆矩阵的应用可能受限. 且混淆矩阵主要适用于分类问题, 不适合应用于回归、聚类等其他类型的机器学习问题.

(5) ROC 曲线和 AUC 值: ROC 曲线是根据不同的分类阈值绘制的真正例率与假正例率之间的曲线, AUC 值表示 ROC 曲线下的面积, 即 ROC 曲线与横轴之间的面积, 取值范围在 0–1 之间. AUC 值越接近 1, 表示模型的性能越好; AUC 值为 0.5 时, 则表示模型预测的效果与随机猜测相当. 但 ROC 曲线和 AUC 值并未直接考虑模型的阈值选择问题. 在实际应用中模型的阈值选择会对 ROC 曲线和 AUC 值产生影响, 有时会导致评估结果不够准确.

(6) mAP (mean average precision): mAP 是 Precision-

Recall 曲线下的平均精度, 通过计算平均精度, 能够全面评估目标检测算法在多类别检测任务中的表现. 通常情况下, mAP 数值越高, 识别效果越好. mAP 主要考虑检测物体的准确性, 但对于目标定位的精度可能存在一定依赖. 在某些情况下, 模型可能会有较高的检测精度但定位不准确. 另外 mAP 通常无法区分模型的误检和漏检情况.

通过对评价指标的分析, 当数据集中各个类别的样本数量存在明显不平衡时, 传统评价指标如准确率可能无法准确反映模型在少数类别上的性能. 精确率、召回率和 $F1$-score 等指标在评估模型性能时对误差类型敏感, 所以有时会忽略不同误差类型的影响. 混淆矩阵、ROC 曲线和 AUC 值等主要用于二分类问题, 对于多类别分类或回归等问题可能不适合. 评价指标的准确性和有效性也受数据质量和标注准确性的影响, 低质量的数据集和错误的标注会对评估结果造成影响. 所以在使用评价指标时, 需要综合考虑实际问题, 根据具体情况选择合适的指标或者结合多个指标来全面评估模型的性能和效果.

## 4 基于深度学习算法的动态手语识别

基于深度学习的手语识别算法借助深度神经网络的强大表示学习和模式识别能力, 能够更好地捕捉手势的特征和动态变化, 更好的处理时序数据. 相较于传统算法, 它们通常具有更高的准确性和更好的泛化能力. 随着技术的进步和数据的积累, 基于深度学习的动态手语识别方法在实际应用中越来越受到重视. 本文所涉及的深度学习方法如表 2 所示.

表 2 基于深度学习的动态手语识别方法

| 深度学习算法 | 方法 | 年份 | 数据集 | 准确率 (%) | 操作 | 优缺点 |
|---|---|---|---|---|---|---|
| 卷积神经网络(CNN) | CNN[31–36] | 2019 | 自建数据集 | 98.76 | 时空数据增强、批量归一化 | 优点: 强大的特征提取能力和适应性, 能够有效处理手势图像数据.<br>缺点: 仅适用处理单帧图像数据, 无法捕捉手语运动过程中帧间相关信息. |
| | 3D-CNN[37] | 2018 | 自建Kinect数据集 | 91.23 | Kinect分析追踪手部姿势 | 优点: 引入新的维度信息, 能够捕捉空间和时间维度上的判别特征.<br>缺点: 数据质量需求高, 需要复杂的预处理工作. 在处理时序数据时, 缺乏显式的序列建模能力. |
| | 3D-CNN[18] | 2015 | 自建Kinect数据集 | 94.2 | 彩色、深度、骨骼多维度输入 | |
| | 3D-CNN[38] | 2022 | 自建2D-Camera数据集 | 98.49 | 背景噪声处理、注意力机制 | |
| 图卷积神经网络(GCN) | GCN[39] | 2023 | 自建骨架KSL数据集<br>标准 KSL-77 数据集 | 100<br>99.87 | 双流GCN网络、与通用深度学习模型集成、通道注意力机制 | 优点: 减少数据复杂度且更具有抗噪声性, 解决复杂环境问题保留关键点的空间结构信息, 能够实现对运动轨迹的追踪.<br>缺点: 主要关注空间特征, 缺乏对时间维度的建模, 无法很好地捕捉动作序列中的时序信息. |
| | ST-GCN[40] | 2021 | 中国手语数据集 (CSL) | 87.02 | Mask R-CNN模型与指数滤波识别校准手部关键点、ST-GCN构建时空图 | 优点: 处理具有时空关系的复杂数据, 相较于GCN能够更好地处理动态数据.<br>缺点: 骨骼关键点的获取需要借助其他算法或外部设备图构建的复杂性、计算复杂度问题. |
| 长短期记忆网络(LSTM) | LSTM[41] | 2022 | 自建ISL数据集 | 92.68 | 双手的坐标位置赋值获取手部数据 | 优点: 建模时间序列信息, 具备高度记忆能力, 能够捕捉序列数据上下文关系.<br>缺点: 不适合捕捉空间信息, 无法从空间数据中有效学习局部特征. |
| | LSTM[42] | 2021 | 自建ASL数据集 | 99.44 | 球半径、手指间角度、手指位置间距等多特征输入 | |
| 混合网络模型 | Inception-RNN[43] | 2019 | 美国手语数据集 (ASL) | 90 | Inception识别空间特征, RNN训练时间特征 | 优点: 综合利用不同网络优势、提取时空特征和多模态信息融合等.<br>缺点: 结构的复杂性面临计算资源增加、实时处理挑战、模型解释性下降等问题. |
| | 3DCNN-LSTM[20] | 2017 | 孤立手势数据 (IsoGD)<br>手势数据集 (SKIG) | 51.02<br>98.89 | 3D-CNN学习短期时空特征, LSTM学习长期时空特征 | |
| | ST-GCN/BiLSTM[44] | 2022 | PH2014T<br>中国手语数据集 (CSL) | 21.34<br>GER | ST-GCN与BiLSTM结合模型捕获短期和长期的动态信息 | |

### 4.1 基于卷积神经网络的动态手语识别方法

Ciregan 等人[31]指出对于图像分类任务采用多个并行网络的多列深度 CNN 可以将单个网络的识别率提高 30%–80%. 同样, 对于大规模视频分类, Karpathy 等人[32]观察到, 将训练的 CNN 与原始视频帧和空间裁剪视频帧两个独立流相结合的结果更好. Krizhevsky 等人[33]通过实验发现使用大量不同的训练示例对 CNN 的训练效果有显著影响, 他们提出数据增强策略以解决 CNN 在使用有限个不同数据集训练时过度拟合的问题. 对训练和测试图像进行平移、水平翻转和 RGB 抖动, 将其分为 1000 个类别, 以此来丰富数据集. Simonyan 等人[34]则是在每个视频帧上采用空间增强方法训练 CNN. 然而这些数据增强方法仅限于空间变化. 为了给包含动态运动的视频序列添加变化, Pigou 等人[35]除了应用空间变换外, 还对视频帧进行了时间平移. 2019 年, Zhan 等人[36]提出一种手势识别系统, 通过提取图像中的手部组件, 并使用二维卷积神经网络进行学习和预测. 为了减少潜在的过拟合, 提高手势分类器的泛化水平, 作者结合现有的空间增强技术并提出一种变形手势输入体积的时空数据增强方法. 作者利用批量归一化(BN)[45], 在训练过程中通过对每一层每批数据进行归一化. 训练结束后, 数据最后一次通过网络, 以分层的方式计算和统计数据并在测试时固定. 通过实验得出, 加入 BN 可以更快地完成训练, 同时实现更好的系统精度和正则化. 实验证明低分辨率和高分辨率的结合可以大大提高分类精度, 进一步证明所提出的数据增强技术在提高性能方面发挥着重要作用. 作者提出的 CNN 在由 9 个手势和每个手势 500 张图像组成的数据集上实现了 98.76% 的平均准确率.

CNN 虽然具备强大的特征提取能力, 但仅适用处理单帧图像数据, 无法捕捉手语运动过程中帧间相关信息, 3D-CNN 则解决了这个问题. 3D-CNN 是一种具备时间维度的卷积神经网络, 主要用于处理图像序列数据. 与传统的 CNN 相比, 3D-CNN 引入新的维度信息, 能够捕捉空间和时间维度上的判别特征. 在手语识别中, 手势的运动过程涉及帧间的相关性. 通过 3D-CNN, 可以将手语动作的时间序列作为输入, 并在卷积操作中同时考虑空间和时间信息, 从而提高手语识别的性能. Soodtoetong 等人[37]通过使用 3D-CNN 识别运动, Kinect Sensor 采集 RGB-D 图像, 用 5 个词评估运动识别的效率. 结果表明, 3D-CNN 算法能够高精度识别手势运动, 识别准确率最高为 91.23%. Huang 等人[18]开发一种以多种类型数据作为输入的 3D-CNN. 该模型同时将彩色图像、深度图像和身体骨骼图像作为输入, 通过对相邻视频帧执行卷积和子采样来集成颜色、深度和轨迹信息, 处理图像光照变化、背景干扰和手部遮挡等的问题. 实验结果表明, 3D-CNN 在 25 个手语词汇上准确率达到 94.2%, 明显优于 GMM-HMM 混合模型. Ma 等人[38]提出一个具有注意力机制的 3D-CNN 实时数字手语识别算法. 在预处理阶段先后进行背景减法[46]、图像边缘侵蚀[47]、边缘侵蚀图像与原始图像之间做“和”运算、灰度和高斯滤波[48], 去除手势数据提取过程中的背景噪声, 引入注意力机制[47]使模型能够从角度信息增强特征表达. 实验结果表明, 系统对数字手语的实时识别性能较好, 准确率达到 98.49%.

基于 CNN 的动态手语识别方法具有强大的特征提取能力和适应性, 能够有效处理手势图像数据. 研究人员结合其他深度学习方法和时空数据增强技术训练 CNN, 进一步提升动态手语识别的性能和实用性. 文中提到的多列深度 CNN、数据增强和批量归一化等方法在改善图像分类和手势识别任务方面都具有一定优势. 但也存在一些缺点, 如对于时间信息的处理和关注单纯依靠 CNN 无法有效实现. 3D-CNN 则解决了 CNN 无法捕捉手势运动中帧间相关信息的问题. Huang 等人[18]采用多种输入模态的方法, 增强模型的表达能力, 提高识别准确率, 但也增加了数据的维度和复杂性, 导致模型训练和推理的计算成本增加. 此外, 对于某些类型的输入数据, 如身体骨骼图像, 其获取和预处理可能需要额外的传感器或技术支持. Ma 等人[38]引入注意力机制增强模型对关键特征的关注, 使模型更加关注重要的信息. 分析发现, CNN 和 3D-CNN 虽然具备强大的特征提取能力, 但研究人员需要通过一系列方法, 如空间增强、Kinect 提取手部信息、多类型数据输入、预处理等, 实现对手部运动的定位跟踪, 削弱复杂条件(手语动作多样性、姿态变化、光照变化、背景干扰和手部遮挡) 对识别准确率的影响. 所以动态手语的复杂性和变化性以及环境干扰等因素是限制卷积神经网络方法性能的重要原因, 如何定位跟踪手部运动和分析手部姿态也成为重要的研究方向.

### 4.2 基于图卷积神经网络的动态手语识别方法

虽然结合计算机视觉的手语识别研究取得了重大进展, 但在基于骨骼的视频分类领域在很大程度上仍

未被探索. 近年来 一些研究人员使用图卷积神经网络(GCN)[49]进行基于骨架的动态动作识别, 如 GSTCAN[50]、GSTCAN[51]、ASGCN[52]和 GSCAN[53]等. 发现相较于原始图像数据, 骨架数据通过提取关键点信息, 减少了数据复杂度且更具有抗噪性, 有助于减少复杂环境对手语识别准确率的影响. 骨架数据保留关键点的空间结构信息, 捕获不同关键点之间的空间关系, 且能够实现对运动轨迹的追踪, 从而提高识别准确性. Shi 等人[52]利用基于关节和关节运动骨架信息的双流 GCN 网络, 试图提高性能, 但模型未能实现手语识别的高精度. 主要原因在于没有考虑非连接的骨架关节和关节运动特征. 针对以上问题, Shin 等人[39]使用双流深度学习网络进行基于关节骨架的动态手语识别. 其中每个流都是由图卷积网络 (GCN) 和基于注意力的通用神经网络方法构建. 通过通道注意力增强解决与非连接关节骨架数据相关问题, 通过标准 CNN 模块进一步完善特征, 丰富时间上下文. 最后将两个流中的特征信息连接起来, 并输入到分类模块中进行手语识别. 通过对自建的基于骨架的 KSL 数据集和标准 KSL-77 数据集进行实验, 分别实现了 100.00% 和 99.87% 的准确率, 有效解决了因连续手语的复杂性和变化性带来的性能精度欠佳和计算复杂性增加的问题.

时空图卷积网络 (ST-GCN)[54]是一种用于处理时空数据的深度学习模型. 它结合了图卷积网络 (GCN) 和时间序列数据的特点, 用于建模和分析具有时空关系的复杂数据, 相较于 GCN 能够更好地处理动态手语中的时间特征. Wang 等人[40]为了消除环境干扰, 提出一种基于动态骨骼的手势识别匹配方法. 该方法使用掩码 R-CNN 和指数滤波识别和校准手部关键点. 对实时帧图像分割和特征提取, 利用组合网络获得手部关键点并构建手部骨骼关键点数据集, 发送到 ST-GCN 网络进行训练. 最后利用模板匹配实现手势识别. 实验结果表明, 该方法能够最大程度地消除环境干扰、传统数据集的不完整以及特殊样品不足导致的模型精度缺陷. 中国手语数据库的识别准确率达 87.02%. 与之前的手势识别研究相比, 解决了复杂环境下准确率下降的问题, 同时提高了模型的鲁棒性.

基于图卷积神经网络尤其是基于时空图卷积网络的动态手语识别方法能够捕捉手势序列中的空间和时间关系. 且相较于原始图像信息, 骨骼关键点信息减少了数据的复杂度并融合多模态信息, 有效解决了动态手语复杂环境下准确率低的问题. Shin 等人[39]引入通道注意力机制, 自动学习关节骨架数据中与非连接关节相关的重要特征, 提升了识别准确率. ST-GCN 能够充分利用手势的时空特征, 并提供多尺度特征提取, 实现更准确的手语识别. Wang 等人[40]提出的基于动态骨骼的手势识别匹配方法通过骨骼关键点数据集消除环境干扰, 进一步提高了准确性和鲁棒性. 分析发现, GCN 和 ST-GCN 处理的是骨骼关节点数据, 相较于处理原始图像, 处理骨骼关节点数据能够有效解决光照变化、背景干扰和手部遮挡等问题. 同时 GCN 和 ST-GCN 能够有效捕获骨骼关节点之间的空间关系并实现手部追踪, 有效解决手语动作的多样性、姿态变化等问题. 但基于图卷积神经网络的手语识别方法也存在一些问题, 例如手语数据集需要包含骨骼关键点等多模态数据类型, 骨骼关键点的获取可能需要借助其他模型或外部设备, 图构建的复杂性、计算复杂度也是需要考虑的问题.

### 4.3 基于长短期记忆网络的动态手语识别方法

长短期记忆 (long short-term memory)[55]是循环神经网络 (RNN) 的变体, 用于处理序列数据和时间序列数据, 捕捉时间序列中的长期依赖关系. 相较于 CNN, LSTM 在处理长序列时表现较好, 对于动态手语识别这类需要考虑时间顺序的任务较为适用. 印度手语 (ISL) 手势复杂, 现有的基于 CNN 的模型很难完整检测这些手势. Sharma 等人[41]提出一种地标方法, 通过对双手坐标位置赋值获取手部数据. LSTM 模型是处理这类数据的最佳选择. 因此作者创建了一个在 Mediapipe 和 LSTM 上工作的系统. 系统使用 3 种 LSTM 模型: 简单 LSTM 模型、双向 LSTM 模型和堆叠 LSTM 模型. 创建一个由 7 个手势和 26 个字母组成的合成数据集来训练模型并进行实验对比, 结果得出简单的 LSTM 和双向 LSTM 模型优于堆叠 LSTM 模型. Lee 等人[42]提出一个美国手语 (ASL) 学习应用系统. 手语字母既有静态符号也有动态符号, 作者采用 LSTM 和 K 最近邻法作为分类方法. 提取球半径、手指间角度、手指位置间距等特征作为分类模型的输入. 模型训练了 2 600 个样本, 每个字母采取 100 个样本. 实验结果显示, 26 个手语字母的识别率平均达到 99.44% .

基于长短期记忆网络的动态手语识别方法能够建模时间序列信息, 具备高度记忆能力, 更好地捕捉序列数据中的上下文关系. 而 CNN 主要用于处理静态数

据, 对于时序信息的处理能力相对较弱, 难以捕捉时间序列中的长期依赖关系. 因此在面对时序问题时LSTM 表现更为出色. 但是, LSTM 的结构设计并不适合捕捉空间信息, 在空间数据中无法有效学习局部特征. Sharma 等人[41]、Lee 等人[42]都是借助其他算法或设备提取数据中的空间信息并转化为 LSTM 的输入. 另外 LSTM 涉及复杂的门控机制, 训练和推理过程的计算复杂度相对较高, 尤其在处理大规模数据时可能需要更多的计算资源, 所以在面对复杂条件下的手语识别任务时, 单纯依靠 LSTM 可能无法取得良好的实验结果.

### 4.4 基于混合网络模型的动态手语识别方法

CNN 具备强大的特征提取能力, 适合处理空间特征, 对于手语图像等静态特征的提取较为有效. LSTM 和 RNN 适合处理时序数据, 能够捕捉时间序列中的长期依赖关系, 对于需要考虑时间顺序的任务较为适用. 所以在面对复杂性和变化性的动态手语识别任务时, CNN 与 RNN 结合的混合网络模型具有一定优势. Ahuja 等人[43]使用 Inception 和 RNN 相结合的模型对视频序列处理并提取时间和空间特征. 通过在美国手语数据集上的实验表明, 网络在 Softmax 层中产生了 90% 的准确率. 三维卷积神经网络 (3D-CNN) 可以直接学习时空特征, 但 LSTM/RNN 更适合学习长期的时间信息. 因此对于长期依赖的情况, 通过 3D-CNN 学习短期时空特征, LSTM/RNN 学习长期时空特征会更合理. Zhu 等人[20]提出一种基于 3D-CNN 和 LSTM 的多模态手势识别方法. 首先通过 3D-CNN 学习手势的短期时空特征, 然后将提取的短期时空特征通过 LSTM 学习手势的长期时空特征. 该方法在 ChaLearLAP 大规模孤立手势数据集 (IsoGD) 和 SheffieldKinect 手势数据集 (SKIG) 取得了较高的识别准确率 (IsoGD 验证集的准确率为 51.02%, SKIG 验证集的准确率为 98.89%). 手语识别需要丰富的时空结构, 而 CNN 和 RNN 在其原生形式下无法很好地捕获时空信息. 基于上述原因, Parelli 等人[44]提出一种新颖的连续手语识别方法, 该方法通过 ST-GCN 从视频中获取空间和时间信息捕获手语者的姿势、形状、外观和运动信息. 引入 ST-GCN 与 BiLSTM 结合模型捕获短期和长期的动态信息. 最后提出一种融合方案, 将 3 个 ST-GCN/BiLSTM 模块在不同的特征流上并行运行, 通过 CTC 方法对齐进行光泽度预测. 通过实验表明其在 RWTH-PHOENIX Weather 2014T 数据集和 CSL 数据集上都具有很好的性能.

基于混合网络模型的动态手语识别方法具有综合利用不同网络优势、提取时空特征和多模态信息融合等优点. 文中提到的 Inception 和 RNN 的结合模型、3D-CNN 和 LSTM 的结合模型、ST-GCN 和 BiLSTM 的结合模型都是充分发挥了卷积神经网络的空间处理优势和循环神经网络的时间处理优势, 获得良好的时空数据, 进而取得了较高的识别准确率. 但构建的混合网络通常比单一网络更复杂, 需要更多的计算资源和时间来训练优化模型. 复杂的模型会影响推理速度, 尤其是在需要实时处理的场景中, 需要权衡模型的准确性和推理速度之间的关系. 另外结构的复杂性还可能导致模型的解释性下降, 难以解释模型在动态手语识别任务中的预测依据.

通过对 4 种动态手语识别算法的介绍和分析, 当前的深度学习算法能够有效提高手语识别性能, 但仍存在着诸多问题. 例如改进的 3D-CNN 网络与 CNN 网络相比, 其优点在于 3D-CNN 可以直接处理视频等时空数据, 同时考虑时间和空间信息, 更好地捕捉视频中物体的运动和变化.但 3D-CNN 对数据质量需求高, 通常需要借助其他算法或设备完成复杂的预处理工作. 另外 3D-CNN 在处理时序数据时缺乏显式的序列建模能力, 无法处理上下文信息方面的任务. 基于长短期记忆网络的动态手语识别方法虽然能够有效处理长时依赖关系但缺少卷积网络特征提取的优势. 所以提出了如 3DCNN-LSTM 的混合网络模型, 结合了卷积网络和循环网络的优点. 但混合网络模型通常包含多个不同类型的神经网络结构, 导致模型复杂度较高, 增加了训练和推理的计算成本, 其实时性和可嵌入性也受到挑战. 另一方面, 采用 CNN 作为特征提取依靠的是纯视觉方案, 在面临手语动作的多样性、姿态变化、光照变化、背景干扰和手部遮挡等问题时没有较好的解决方案. 但姿态估计和骨骼关键点检测方案的提出为手语识别方法提供了新的思路. GCN 和 ST-GCN 通过处理骨骼关节点数据能够有效解决光照变化、背景干扰和手部遮挡等问题. 同时 GCN 和 ST-GCN 能够有效捕获骨骼关节点之间的空间关系并实现手部追踪, 有效解决手语动作的多样性、姿态变化等问题. 但目前包含骨骼关键点数据的手语数据集较少, 可能需要借助其他算法和外部传感器来实现. 另外, 目前的手语识别方法大部分只聚焦于手部动作姿态, 手语识别语义

单一, 缺少对肢体、面部、词义、句法结构和情感表达等方面的研究.

## 5 总结与展望

本文对动态手语识别技术的发展历程和常用手语数据集进行回顾和总结, 强调了深度学习技术在提高手语识别性能方面的吸引力. 详细介绍了动态手语识别中常用的深度学习算法和模型, 调研了姿态估计和骨骼关键点检测应用于手语识别领域的方法, 尤其介绍并分析了图卷积神经网络特别是时空图卷积网络在手语识别领域的应用和优势. 探讨了深度学习技术在特征学习、复杂变化处理、上下文理解、鲁棒性、实时性能和可扩展性等方面的优势. 展示了深度学习技术如何针对不同问题提升动态手语识别的性能, 使其在处理大规模数据或复杂识别任务等方面获得出色表现. 手语识别研究在计算机视觉、人机交互等领域有着巨大的研究和应用潜力, 对于改善聋哑人生活质量也具有重要意义. 目前手语识别整体上取得了良好的综合评价指标, 但仍面临着诸多问题: 例如手语数据集的规模相对较小且受限于特定地区或特定手语方言的问题; 性能和泛化能力相对较差的问题; 手语识别系统的性能和实时性问题; 手语识别系统语义单一的问题等.

数据集的规模相对较小且受限于特定地区或特定手语方言, 且不同的人可能有不同的手势风格和习惯. 为此可以通过数据增强技术, 扩充手语数据集的规模并增加数据的多样性. 另外可以考虑使用合成技术生成虚拟手语数据, 以进一步增加数据的多样性和丰富性. 未来的研究也可致力于开发可适应不同手语系统的通用手语识别模型, 以促进跨语言和跨文化的交流. 针对性能和泛化能力方面可以利用预训练的深度学习模型和迁移学习的方法, 将已有的知识和特征迁移到手语识别任务中, 以提高模型的性能. 领域自适应技术可以帮助模型在不同手语方言或手势风格之间进行适应和泛化. 针对手语动作所涉及的手部和手指姿态变化以及光照变化、背景干扰和手部遮挡等问题, 可以结合多种传感器 (如深度相机、惯性传感器等) 的数据, 提供更丰富的手语信息, 提升模型的稳定性和鲁棒性进而改善手语识别的性能. 另外多模态融合技术可以将不同传感器的数据进行融合, 提供更全面的手语特征. 针对实时通信或手语翻译系统的实时性和延迟问题. 未来的研究可以探索更加智能和自适应的用户交互方式, 比如可以通过引入强化学习技术, 使系统在与用户的交互中逐步优化手语识别性能. 强化学习可以通过与用户的反馈进行模型调整和优化, 逐步提高系统的准确性和适应性, 进而提供更直观、自然和高效的交互体验. 目前的动态手语识别主要关注手势的形状、姿势和运动等低级特征. 但手语中还包含丰富的语义信息, 如词义、句法结构和情感表达等. 针对手语识别系统语义单一的问题, 未来可以将深度学习与自然语言处理相结合, 实现对手语的深度语义理解, 更准确地理解手语中的含义和意图.

虽然目前的手语识别研究和在实际生活中的应用仍面临着诸多问题, 但随着数据增强和合成、通用手语识别模型、迁移学习和领域自适应、多模态融合、强化学习、模型架构优化、深度语义理解等方案的提出和发展, 将为手语识别系统的性能提升和实际应用提供更多可能性.

## 参考文献


1 Lu ZY, Chen X, Li Q, *et al*. A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. IEEE Transactions on Human-machine Systems, 2014, 44(2): 293–299. [doi: 10.1109/thms.2014.2302794]

2 Shi YY, Li YN, Fu XL, *et al*. Review of dynamic gesture recognition. Virtual Reality & Intelligent Hardware, 2021, 3(3): 183–206.

3 Konečný J, Hagara M. One-shot-learning gesture recognition using HOG-HOF features. In: Escalera S, Guyon I, Athitsos V, eds. Gesture Recognition. Cham: Springer, 2017. 365–385.

4 Kudrinko K, Flavin E, Zhu XD, *et al*. Wearable sensor-based sign language recognition: A comprehensive review. IEEE Reviews in Biomedical Engineering, 2021, 14: 82–97. [doi: 10.1109/RBME.2020.3019769]

5 Noor TH, Noor A, Alharbi AF, *et al*. Real-time Arabic sign language recognition using a hybrid deep learning model. Sensors, 2024, 24(11): 3683. [doi: 10.3390/s24113683]

6 Poritz AB. Hidden Markov models: A guided tour. Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing. New York: IEEE, 1988. 7–13.

7 Pashaloudi VN, Margaritis KG. On feature extraction and sign recognition for Greek sign language. Proceedings of the

2003 International Conference on Artificial Intelligence and Soft Computer. 2003. 93–98.

8 Yoon HS, Soh J, Bae YJ, *et al*. Hand gesture recognition using combined features of location, angle and velocity. Pattern Recognition, 2001, 34(7): 1491–1501. [doi: 10.1016/S0031-3203(00)00096-0]

9 Binh ND, Shuichi E, Ejima T. Real-time hand tracking and gesture recognition system. Proceedings of the 2005 International Conference GVIP. Cairo, 2005. 362–368.

10 Elmezain M, Al-Hamadi A, Pathan SS, *et al*. Spatio-temporal feature extraction-based hand gesture recognition for isolated American sign language and Arabic numbers. Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis. Salzburg: IEEE, 2009. 254–259. [doi: 10.1109/ISPA.2009.5297719]

11 Yang MH, Ahuja N, Tabb M. Extraction of 2D motion trajectories and its application to hand gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(8): 1061–1074. [doi: 10.1109/TPAMI.2002.1023803]

12 Waibel A, Hanazawa T, Hinton G, *et al*. Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37(3): 328–339. [doi: 10.1109/29.21701]

13 Dardas NH, Georganas ND. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Transactions on Instrumentation and Measurement, 2011, 60(11): 3592–3607. [doi: 10.1109/TIM.2011.2161140]

14 Gupta B, Shukla P, Mittal A. K-nearest correlated neighbor classification for Indian sign language gesture recognition using feature fusion. Proceedings of the 2016 International Conference on Computer Communication and Informatics. Coimbatore: IEEE, 2016. 1–5. [doi: 10.1109/ICCCI.2016.7479951]

15 Koller O, Ney H, Bowden R. Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3793–3802. [doi: 10.1109/CVPR.2016.412]

16 Avola D, Bernardi M, Cinque L, *et al*. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Transactions on Multimedia, 2019, 21(1): 234–245. [doi: 10.1109/TMM.2018.2856094]

17 Yang S, Zhu Q. Video-based Chinese sign language recognition using convolutional neural network. Proceedings of the 9th IEEE International Conference on Communication Software and Networks. Guangzhou: IEEE, 2017. 929–934. [doi: 10.1109/ICCSN.2017.8230247]

18 Huang J, Zhou WG, Li HQ, *et al*. Sign Language Recognition using 3D convolutional neural networks. Proceedings of the 2015 IEEE International Conference on Multimedia and Expo. Turin: IEEE, 2015. 1–6. [doi: 10.1109/ICME.2015.7177428]

19 Chai XJ, Liu ZP, Yin F, *et al*. Two streams recurrent neural networks for large-scale continuous gesture recognition. Proceedings of the 23rd International Conference on Pattern Recognition. Cancun: IEEE, 2016. 31–36. [doi: 10.1109/ICPR.2016.7899603]

20 Zhu GM, Zhang L, Shen PY, *et al*. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access, 2017, 5: 4517–4524. [doi: 10.1109/ACCESS.2017.2684186]

21 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]

22 Liao YQ, Xiong PW, Min WD, *et al*. Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. IEEE Access, 2019, 7: 38044–38054. [doi: 10.1109/ACCESS.2019.2904749]

23 米娜瓦尔·阿不拉, 阿里甫·库尔班, 解启娜, 等. 手语识别方法与技术综述. 计算机工程与应用, 2021, 57(18): 1–12. [doi: 10.3778/j.issn.1002-8331.2104-0220]

24 蒋贤维, 孙计领, 张艳琼, 等. 中国手语识别方法及技术综述. 现代特殊教育, 2024(6): 47–58. [doi: 10.3969/j.issn.1004-8014.2024.06.009]

25 von Agris U, Kraiss KF. SIGNUM database: Video corpus for signer-independent continuous sign language recognition. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. Valletta: European Language Resources Association, 2010. 243–246.

26 Athitsos V, Neidle C, Sclaroff S, *et al*. The American sign language lexicon video dataset. Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Anchorage: IEEE, 2008. 1–8.

27 Forster J, Schmidt C, Hoyoux T, *et al*. RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and

translation corpus. Proceedings of the 2012 International Conference on Language Resources and Evaluation. 2012. 3785–3789.

28 Pu JF, Zhou WG, Li HQ. Iterative alignment network for continuous sign language recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4160–4169. [doi: 10.1109/CVPR.2019.00429]

29 Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7(2): 179–188. [doi: 10.1111/j.1469-1809.1936.tb02137.x]

30 Powers DMW. Evaluation: From precision, recall and *F*-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2011, 2(1): 2229–3981.

31 Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3642–3649.

32 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1725–1732.

33 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

34 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 568–576.

35 Pigou L, Dieleman S, Kindermans PJ, *et al*. Sign language recognition using convolutional neural networks. Proceedings of the 2014 Computer Vision —ECCV 2014 Workshops. Zurich: Springer, 2015. 572–578.

36 Zhan F. Hand gesture recognition with convolution neural networks. Proceedings of the 20th IEEE International Conference on Information Reuse and Integration for Data Science. Los Angeles: IEEE, 2019. 295–298. [doi: 10.1109/IRI.2019.00054]

37 Soodtoetong N, Gedkhaw E. The efficiency of sign language recognition using 3D convolutional neural networks. Proceedings of the 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. Chiang Rai: IEEE, 2018. 70–73. [doi: 10.1109/ECTICon.2018.8619984]

38 Ma Y, Xu TP, Kim K. A digital sign language recognition based on a 3D-CNN system with an attention mechanism. Proceedings of the 2022 IEEE International Conference on Consumer Electronics-Asia. Yeosu: IEEE, 2022. 1–4.

39 Shin J, Miah ASM, Suzuki K, *et al*. Dynamic korean sign language recognition using pose estimation based and attention-based neural network. IEEE Access, 2023, 11: 143501–143513. [doi: 10.1109/ACCESS.2023.3343404]

40 Wang JY, Yu NG, Firdaous E. Gesture recognition matching based on dynamic skeleton. Proceedings of the 33rd Chinese Control and Decision Conference. Kunming: IEEE, 2021. 1680–1685.

41 Sharma K, Aaryan KA, Dhangar U, *et al*. Automated Indian sign language recognition system using LSTM models. Proceedings of the 2022 International Conference on Computing, Communication, and Intelligent Systems. Greater Noida: IEEE, 2022. 461–466. [doi: 10.1109/ICCCIS56430.2022.10037711]

42 Lee CKM, Ng KKH, Chen CH, *et al*. American sign language recognition and training method with recurrent neural network. Expert Systems with Applications, 2021, 167: 114403. [doi: 10.1016/j.eswa.2020.114403]

43 Ahuja R, Jain D, Sachdeva D, *et al*. Convolutional neural network based American sign language static hand gesture recognition. International Journal of Ambient Computing and Intelligence (IJACI), 2019, 10(3): 60–73. [doi: 10.4018/IJACI.2019070104]

44 Parelli M, Papadimitriou K, Potamianos G, *et al*. Spatio-temporal graph convolutional networks for continuous sign language recognition. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 8457–8461. [doi: 10.1109/ICASSP43922.2022.9746971]

45 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 448–456.

46 Kalsotra R, Arora S. Background subtraction for moving object detection: Explorations of recent developments and challenges. The Visual Computer, 2022, 38(12): 4151–4178. [doi: 10.1007/s00371-021-02286-0]

47 Niu ZY, Zhong GQ, Yu H. A review on the attention mechanism of deep learning. Neurocomputing, 2021, 452:

48–62. [doi: 10.1016/j.neucom.2021.03.091]

48 Yan JJ, Xu ZY, Wu ZQ, *et al*. Edge detection method of laser cladding pool image based on morphology. Proceedings of the 2021 Advanced Laser Technology and Applications. Beijing: SPIE, 2021. 246–253. [doi: 10.1117/12.2606710]

49 Liu JX, Xu C, Yin C, *et al*. K-core based temporal graph convolutional network for dynamic graphs. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(8): 3841–3853. [doi: 10.1109/TKDE.2020.3033829]

50 Yan SJ, Xiong YJ, Lin DH. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018. 7444–7452.

51 Li MS, Chen SH, Chen X, *et al*. Actional-structural graph convolutional networks for skeleton-based action recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3590–3598.

52 Shi L, Zhang YF, Cheng J, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12018–12027.

53 Egawa R, Miah ASM, Hirooka K, *et al*. Dynamic fall detection using graph-based spatial temporal convolution and attention network. Electronics, 2023, 12(15): 3234. [doi: 10.3390/electronics12153234]

54 de Amorim CC, Macêdo D, Zanchettin C. Spatial-temporal graph convolutional networks for sign language recognition. Proceedings of the 28th International Conference on Artificial Neural Networks Artificial Neural Networks and Machine Learning. Munich: Springer, 2019. 646–657.

55 Aly S, Aly W. DeepArSLR: A novel signer-independent deep learning framework for isolated Arabic sign language gestures recognition. IEEE Access, 2020, 8: 83199–83212. [doi: 10.1109/ACCESS.2020.2990699]

(校对责编: 张重毅)