基于深度学习的手语识别综述

张淑军* 张 群 李 辉
(青岛科技大学信息科学技术学院 青岛 266061)

摘 要：手语识别涉及计算机视觉、模式识别、人机交互等领域，具有重要的研究意义与应用价值。深度学习技术的蓬勃发展为更加精准、实时的手语识别带来了新的机遇。该文综述了近年来基于深度学习的手语识别技术，从孤立词与连续语句两个分支展开详细的算法阐述与分析。孤立词识别技术划分为基于卷积神经网络(CNN)、3维卷积神经网络(3D-CNN)和循环神经网络(RNN) 3种架构的方法；连续语句识别所用模型复杂度更高，通常需要辅助某种长时时序建模算法，按其主体结构分为双向长短时记忆网络模型、3维卷积网络模型和混合模型。归纳总结了目前国内外常用手语数据集，探讨了手语识别技术的研究挑战与发展趋势，高精度前提下的鲁棒性和实用化仍有待于推进。

关键词：深度学习；手语识别；卷积网络；循环神经网络；长时序建模
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2020)04-1021-12
DOI: 10.11999/JEIT190416

Review of Sign Language Recognition Based on Deep Learning

ZHANG Shujun ZHANG Qun LI Hui
(*College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao* 266061, *China*)

Abstract: Sign language recognition involves computer vision, pattern recognition, human-computer interaction, etc. It has important research significance and application value. The flourishing of deep learning technology brings new opportunities for more accurate and real-time sign language recognition. This paper reviews the sign language recognition technology based on deep learning in recent years, formulates and analyzes the algorithms from two branches - isolated words and continuous sentences. The isolated-word recognition technology is divided into three structures: Convolutional Neural Network (CNN), Three-Dimensional Convolutional Neural Network (3D-CNN) and Recurrent Neural Network (RNN) based method. The model used for continuous sentence recognition has higher complexity and is usually assisted with certain kind of long-term temporal sequence modeling algorithm. According to the major structure, there are three categories: the bidirectional LSTM, the 3D convolutional network model and the hybrid model. Common sign language datasets at home and abroad are summarized. Finally, the research challenges and development trends of sign language recognition technology are discussed, concluding that the robustness and practicality on the premise of high-precision still requires to be promoted.

Key words: Deep learning; Sign language recognition; Convolutional Neural Network (CNN); Recurrent Neural Network (RNN); Long-term temporal sequence modeling

1 引言

手语是一种重要的人类肢体语言表达方式，包含信息量多，是聋哑人和健听人之间沟通的主要方式。手语识别涉及视频采集和处理、计算机视觉、人机交互、模式识别、自然语言处理等多个研究领域，是一项具有高难度的挑战性课题。手语识别技术可用于手语翻译、日常交流、研发手语教学机器人，促进手语教学、培训和推广，可以拓宽到其他手势指令相关的领域，如交警手势识别、军事手势识别及智能家电控制等。由于手语语义丰富、动作幅度相比其他人体行为具有局部性和细节性，同时又受到光照、背景、运动速度等影响，使用传统模式识别或机器学习方法所能达到的精度与鲁棒性已

收稿日期：2019-06-06；改回日期：2019-11-20；网络出版：2020-01-18
*通信作者： 张淑军 lindazsj@163.com
基金项目：国家自然科学基金(61702295, 61672305)，山东省重点研发计划项目(2017GGX10127)
Foundation Items: The National Natural Science Foundation of China (61702295, 61672305), The Key Research & Development Plan Project of Shandong Province (2017GGX10127)

达到瓶颈期，往往局限于静态手势识别或粗粒度的动态手势识别。自2006年Hinton等人[1]提出深度学习方法以来，手语识别迎来了新的机遇。

手语识别是借助计算机自动将手语信号转换为文本或语音的过程[2]，其研究可以追溯到20世纪90年代。根据手语获取方式的不同，分为基于数据手套和基于视觉的手语识别，前者可以实时采集手势的3维运动信息和时序变化，然后应用识别算法进行处理，识别速度快，准确率高，但是设备复杂，价格昂贵，且对操作者有约束，佩戴不方便，因此基于视觉的手语识别成为主流。从历史沿革上，又可分为传统识别方法和基于深度学习的研究方法两大类，而从研究对象上又可分为孤立词和连续语句。2016年之前，基于视觉的传统手语识别技术研究较为广泛，详见文献[3]。传统方法能够解决一定数据规模下的手语识别问题，但算法复杂、泛化性不高，且面向的数据量与模式种类受限，无法将人类对于手语的智能理解完全表述。因此，在当前大数据飞速发展的时代背景下，基于深度学习、挖掘人类视觉与认知规律的手语识别技术成为必然。根据具体研究对象的不同，所采用的深度学习网络和算法有所不同，一般从数据输入、网络架构和融合方式3部分进行改进，其中网络架构对研究工作的区分度较大，不同的网络架构可有相同或近似的数据输入和融合方式。因此，本文的技术分类主要立足于主干网络架构，总体分类如图1所示。

本文第2节与第3节将分别阐述基于深度学习的孤立词手语识别与连续语句手语识别，第4节归纳总结了目前国内外常用手语数据集及评估标准，最后探讨了研究挑战与未来发展趋势。

2 基于深度学习的孤立词手语识别

孤立词手语识别的对象是以视频表达的单个孤立手语词汇相对连续语句而言，孤立词手语视频时长较短，语义简单明确，识别主要围绕如何更有效地描述手语的底层特征、降低误判率展开。从时序信息的处理上，将技术方法分为基于卷积神经网络、3维卷积神经网络或循环神经网络3种网络的主体框架。此方面的研究国内以中国科学院计算所、中国科学院自动化所、中国科学技术大学、西安电子科技大学等比较活跃，国外以亚琛工业大学、根特大学等成果较多。

2.1 卷积神经网络

由于卷积神经网络(CNN)在深度学习网络中的基础地位，自2013年至今一些研究团队进行了一系列基于CNN的孤立词手语识别研究，主要是加入多模态数据(包括深度、骨架、人体关键点等)、关注手部姿态特征、特征融合(卷积层的特征集合)等相关优化策略，实现了较好的识别效果。

数据输入主要涉及进入网络之前数据层的融合，该融合方式对于数据集的依赖性较强，预处理较为复杂，但也一定程度上提高了识别准确率。文献[4–7]提出基于多模态数据、利用多尺度捕获各层级图像特征的CNN网络自动学习手部图像特征的手语识别方法。关注手部姿态特征是手语识别中较为主要的特征设计部分，文献[8,9]提出了一种重点关注手型变化的CNN网络用于手语识别，将手型特征送入一个端到端的弱监督分类框架完成识别，并能对小规模孤立词手语数据集进行实时识别。联合目标检测网络对手部进行检测跟踪也是近年来手语识别研究中一个较为主要的方法，通过对手部检测跟踪，更好地定位手部。Kim等人[10,11]通过目标检测网络对手部关键特征进行ROI分割预处理，并联合原始手语视频特征送入CNN网络，在准确率提高的同时训练时间也减少了一半。Kopuklu等人[12]在CVPR2018会议上提出了一种将运动信息融合到静态图像的数据级融合策略，并将融合后的时空特征送入到CNN网络用于后续分类，取得了较好的识别效果。手部关节点及全身骨架数据为手语

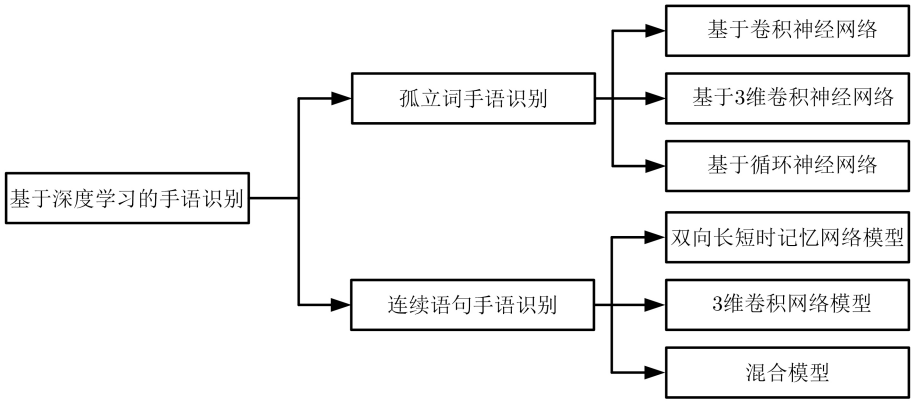

图 1 总体分类图

视频特征的提取起到辅助作用，Konstantinidis等人[13]提出了一种从RGB手语视频序列提取手部和全身骨架数据的手语识别方法，Devineau等人[14]提出了一种基于手部骨架数据的CNN用于3维动态手势的识别，采用并行卷积处理手部骨骼关节的位置序列，提取手部关节点，取得了较高的识别精度。

2.2 3维卷积网络

CNN虽具有较强的特征提取能力，但仅限于单帧图像数据的输入，手语识别还需要辅助一些挖掘帧间相关性的方法，3维卷积神经网络(3D-CNN)应运而生。3D-CNN可以同时对视频级的手语时空特征建模，以捕捉从多个连续帧得到的运动信息，实现更具全局意义的识别。

不同融合方式、时空注意力机制以及双流3D-CNN是基于3维卷积网络手语识别研究的3大突破点。不同于数据融合方式，特征融合是对多模态数据进行特征提取后的卷积层融合，该融合方式使用较多。随着近几年注意力机制在分类任务中的应用，3D-CNN中加入注意力机制的思想也越来越多，时空注意力机制能捕捉手语视频中最为重要的手语动作，获得较高的识别精度。双流3D-CNN是一种两路的网络组织形式，通过构建双流3D-CNN，实现对两路分类结果的得分融合。NVIDIA研究院Molchanov等人[15]在CVPR2015会议上首次提出了将3D-CNN用于动态手势识别，将多尺度数据作为网络输入，构建两路子网络用于提取手势的时空特征集合，最后对两路分类结果进行得分融合，在自动驾驶场景中实现了较好的识别效果。Wu等人[16]提出了一种基于多通道数据融合的3D-CNN用于动态手势识别，将RGB和深度数据堆叠并入一路送入3D-CNN，引入HMM模型进行短时序建模，在Charlearn数据集取得了不错的识别结果。

中国科学技术大学Huang等人[17,18]于2015年针对数据融合问题提出了一种基于多模态数据输入的3D-CNN网络。2018年又提出一种基于时空注意力机制的3D-CNN网络用于大词汇量的手语识别，在训练3D-CNN捕捉时空特征时，将空间注意力融入网络，聚焦感兴趣的手部特征，特征提取后，利用时间注意机制进行重要手语动作的分类。西安电子科技大学近几年对于手语识别的一系列研究都是围绕3D-CNN展开的，在多模态数据输入、数据增强及关键帧的选取等方面都取得了显著的研究成果，在2016年、2017年Chalearn LAP国际大规模独立手势识别竞赛中名列前茅。文献[19–21]在ICPR2016上，提出了一种基于3D-CNN网络和大规模RGB-D数据的手势识别算法，将RGB和深度视频分别送入到C3D模型中提取时空特征，最后用SVM进行分类，次年在Chalearn手势比赛中以56.9%的准确率取得第1名。在TCSVT2017上Miao等人[21]提出了一种将显著性特征与3D-CNN结合的手势识别框架，利用频域变换生成的显著图增强手部特征，在Chalearn数据集上识别精度达到59.43%。同年，他们在ICCV上提出了一种基于ResC3D网络的手势识别方法，其关键思想是找到一种紧凑而有效的视频序列表示方法，采用神经网络和中值滤波等视频增强技术来消除输入视频中的光照变化和噪声，并采用加权帧一致策略对关键帧进行采样。在此基础上，结合3D残差网络，提出了一种基于正则相关分析的特征融合方法用于视频特征的提取，在Chalearn数据集上准确率达到67.71%。

ElBadawy等人[22]提出了一种仅由RGB数据输入的3D-CNN网络用于阿拉伯手语孤立词的识别，并将其研发为手语识别系统，实现了一定识别率前提下的应用。Ye等人[23]提出了一种捕获手语视频片段中时间和空间信息的3DRCNN网络模型，能够识别和定位不同视频长度的语义信息，提升了识别精度。Liang等人[24]提出了一种基于多模态数据和3D-CNN网络的手语识别算法，并对多种数据进行卷积融合，在大规模数据集上验证了其有效性，但是在大多仅由RGB数据组成的手语数据集上泛化性不高。

2.3 循环神经网络

相比于以上所述的两种网络架构，循环神经网络(RNN)是一类用于处理序列数据的神经网络，更善于捕捉长时间的上下文语义信息。因此，近年来基于循环神经网络进行手语识别的研究也很广泛。

Cate等人[25]利用RNN对手语视频特征进行时序建模，实现了对95类手语孤立词的识别。Chai等人[26]于2016年提出了一种双流的RNN网络(2S-RNN)，首先将连续手势分割成孤立手势，将RGB数据用于网络的一路输入，并提取骨架数据和梯度直方图特征，然后将提取的特征融合送入另一路RNN网络，两路网络进行得分融合，在当年Chalearn手势识别挑战赛中排名第一。

中国科学技术大学于2016年起基于RNN网络进行了一系列手语识别的研究，主要是加入轨迹数据、骨架关节点数据、建立手型描述符以及关键帧筛选等方法。文献[27]针对手工设计特征的不确定性，将用于表示上下文信息的长短时记忆(LSTM)网络加入到手语识别的研究中，并将4个骨架关节点的运动轨迹作为网络输入，在中国孤立词手语数据集上取得了不错的识别效果。针对手语识别的单

一数据输入问题，Li等人[28]于2017年提出了新的手型描述符，并对这些描述符进行基于LSTM的时序建模，在100类中国手语孤立词上实现了准确的识别结果，但是骨架数据的获取比较困难。Huang等人[29]于2018年针对冗余信息影响识别精度的问题，提出了基于关键帧视频序列的手语识别算法，将关键帧算法嵌入到RNN网络中，从而允许对输入数据进行不同关注，取得了显著的识别效果。北京工业大学于2017年对中国手语识别进行了相关研究，主要是利用RNN对手语视频进行时序建模，在自己构建的数据集上取得了不错的识别效果。Yang等人[30,31]提出了将CNN与LSTM相结合，并将RGB和光流数据作为两路输入，最后进行全连接层的融合，输出分类结果。在构建的小规模手语数据集上进行了评估，并满足小规模手语识别系统的实时性要求。

Lin等人[32]于2018年提出了一种将带有掩膜的Res-C3D网络和用于骨架数据建模的LSTM网络相结合，同时处理RGB-D视频数据，在Chalearn数据集上取得了68.42%的识别精度，本文将分割算法与分类网络结合在一起用于手语识别的研究，提高了识别精度，但是前期掩码的建立过程较为耗时。Halim等人[33]提出一种基于SIBI词形变化的双向RNN网络，用于印尼语词根孤立词的识别，并将识别模型扩展到多个设备，可拓展性较强。文献[34]和文献[35]提出了一种基于Inception模型的RNN网络用于手语识别。Konstantinidis等人[36]提出了基于RGB和骨架数据的RNN网络的手语识别框架，并将脸部表情融入到手语特征中，探讨分析了不同的数据融合方案，验证了骨架数据对于手语识别的鲁棒性。文献[37]将CNN和LSTM结合用于美国手语孤立词的识别，并利用一种缩放、平滑等的数据增强技术来改进网络模型的泛化性。Liao等人[38]于2019年提出了一种基于BLSTM网络的手语识别框架，首先使用检测网络对手部进行分割，然后将分割后的手部特征与原始RGB数据一起送入到LSTM，实现动态长时序特征建模，最后输出分类结果，实现了在两个大型中国手语孤立词数据集上的准确识别。

基于深度学习的孤立词手语识别技术及代表性工作如表1所示。

3 基于深度学习的连续语句手语识别

相比于孤立词手语识别，连续语句的识别需要建立更为可靠的长期时序依赖。最初的连续语句识别是在单个孤立词识别的基础上进行研究，需要用到时序分割的相关算法，但由于时序分割过程复杂、误判率高等问题，近年来学者们逐渐绕开了时序分割，将语音识别的时序对齐算法及编解码网络用于本领域的研究，其中包括基于CTC时序算法和用于长时序建模的编解码网络，在此基础上实现连续语句的手语识别。

3.1 双向长短时记忆网络模型

连续语句的手语识别更为复杂，长期的时序性要求更强，而近年来由于双向长短时记忆网络(BLSTM)能更好地对手语长时序序列进行上下文语义信息的建模，因此较为广泛地应用于连续语句手语识别的研究。

亚琛工业大学Camgoz等人[39]在ICCV2017会议上提出了一种利用BLSTM网络来解决端到端的手语识别问题，将LSTM分解为一系列子网络，然后对这些子网络之间进行基于BLSTM网络的时序关系建模，并将用于语音识别的CTC算法用于本研究任务，在RWTH数据集上实现了不错的效果。本方法借鉴语音识别的时间连接时序分类(CTC)算法用于手语连续语句识别的标签对齐，绕开了时序分割，提高了识别精度，但是算法实现较为复杂。清华大学Cui等人[40]于2018年提出了一种基于BLSTM的手语识别框架，并融入光流、RGB-D数据，在基准的连续语句手语数据集上取得了不错的效果。Shi等人[41]于2018年提出了一种在低帧率和视角不同条件下的基于注意力的LSTM网络用于连续手语识别，在美国连续手语数据集上取得了较好的效果。此算法对于数据集的泛化性较强，可以接受多视角下的数据，识别效果较好。Ko等人[42]提出了一种基于LSTM的人体关键点检测的手语识别方法，利用人脸、手部和身体部位提取人体关键点，开发了一套手语识别系统，提取的人体关键点向量通过关键点的均值进行标准化，然后将提取的关键点特征输入到LSTM网络中。结果表明了其识别算法的鲁棒性。Zhang等人[43]、Mittal等人[44]提出一种改进的BLSTM模型，用于连续手语句子的识别。该方法将连续句子分解为单词向量，在连续语句的手语数据集上实现了准确的识别结果。

3.2 3维卷积网络模型

相比于BLSTM网络模型的复杂度，基于3维卷积网络模型的连续手语识别避开了BLSTM网络的复杂建模，在同样能进行时序建模的基础上节省了复杂的计算量。

亚琛工业大学Camgoz等人[45]提出了一种基于3维卷积神经网络的连续手语识别模型，该网络通过3维卷积，从RGB数据中提取时空相关特征，通

表 1　基于深度学习的孤立词手语识别技术及代表性工作

作者/单位	年份	技术特点	准确率(%)	数据集	样本大小
Tang Ao, Li HouQiang, Huang Jie, Li Xiaoxu, Huang Shiliang/中国科学技术大学	2013	卷积神经网络(基于RGB-D并对手部进行分割与追踪)[4]	98.12	American Sign Language(ASL)	50700帧
	2015	3维卷积神经网络(多模态输入)[17]	94.20	Chinese Sign Language(CSL)	25类
	2016	循环神经网络(加入轨迹数据)[27]	85.60		500类
	2017	长短时记忆网络(加入手型描述符)[28]	86.20		100类
	2018	循环神经网络(关键帧视频序列筛选)[29]	91.18		310类
		3维卷积网络(基于注意力机制)[18]	88.70		500类
Pigou L/根特大学	2014	卷积神经网络[5]	91.70	Chalearn	20类
	2016	3维卷积网络(多模态数据的特征融合)[16]	81.00	2014	
Molchanov P，Garcia B，Hardie Cate/斯坦福大学	2015	3维卷积网络(多尺度数据)[15]	77.50	VIVA Dataset	
		循环神经网络[25]	90.80	南威尔士大学数据集	95类
	2016	卷积神经网络[9]	91.63	ASL fingerspelling	
Kang B /加州大学	2015	卷积神经网络[6]	99.99	ASL fingerspelling	31类
Miao Qiguang /西安电子科技大学	2016	3维卷积神经网络(基于RGB-D)[19]	56.90	Chalearn	
	2017	(基于显著性特征和RGB-D)[20]	59.43		
		(基于多模态数据和手部特征增强)[21]	67.71		
Koller O/亚琛工业大学	2016	卷积神经网络(关注手型变化)[8]		Danish Sign Language	分辨率4730×22
Chai Xiujuan/中科院计算所	2017	改进的RNN(对手部分割定位)[26]	99.00	Chinese Sign Language(CSL)	40类
Yang Su/北京工业大学	2017	RNN和CNN相结合[30]	98.43	CSL	40类
		RNN(数据预处理)[31]	99.00	CSL	40类
Hossen M A /特斯瓦拉工程学院	2017	卷积神经网络[7]	100.00	Kinect录制	10类
ElBadawy M /埃及埃因萨姆斯大学	2017	3维卷积网络[22]	98.00	阿拉伯数据集	25类
Kim S /韩国首尔大学	2017	卷积神经网络(帧间采样)[10]	86.00	摄像头采集	20类
	2018	卷积神经网络(手部分割)[11]	98.00		12类
Kopuklu O/德国慕尼黑大学	2018	卷积神经网络(时空特征融合)[12]	96.28 57.40	Jester Chalearn	
Konstantinidis D /希腊大学	2018	卷积神经网络(RGB和骨架数据)[13]	98.09	阿根廷数据集LSA64	
		循环神经网络(多模态数据融合)[36]	89.50	印度手语数据集(IIT)	
Devineau G /巴黎圣米歇尔研究大学	2018	卷积神经网络(骨架数据、加入手部关节点位置序列)[14]	84.35	DHG Dataset	28 类
Ye Yuancheng /纽约城市大学	2018	3维卷积网络(特征融合)[23]	69.20	American Sign Language	27类
Liang Zhijie /华中师范大学	2018	3维卷积网络(骨架、轮廓、深度数据)[24]	83.60	Chalearn	
Lin Chi/中国科学院自动化所	2018	带有掩膜的ResC3D网络与RNN相结合[32]	68.42	Chalearn	
Halim K /印尼大学	2018	循环神经网络(基于SIBI词性变化手势的特征集)[33]	96.15	印尼手语数据集	
Masood S /新德里大学	2018	循环神经网络和卷积神经网络相结合[34]	95.20	阿根廷数据集LSA64	46类
Bantupalli K /美国肯尼索州立大学	2018	循环神经网络和卷积神经网络相结合[35]	93.00	American Sign Language(ASL)	100类
Hernandez V /东京农业大学	2019	卷积神经网络与长短时记忆网络相结合[37]	89.30	American Sign Language(ASL)	19类
Liao YanQiu/南昌大学	2019	循环神经网络和3维卷积网络相结合[38]	86.90	Chinese Sign Language(CSL)	500类

过汇聚每层提取的特征进行时空不变性编码。在Chalearn2016手势识别挑战赛和德国手语数据集上都达到了较高的识别精度。中国科学技术大学Pu等人[46]于2018年提出了一种基于3维残差网络和膨胀卷积网络的连续语句手语识别框架，并提出了一种基于CTC算法的迭代优化策略，在德国基准数据集上证明了此方法的有效性和优越性。针对以往研究工作的不足，Huang等人[47]在AAAI2018会议上提出了一种新的无需时间分割且带有潜在空间的分层注意力网络(LS-HAN)用于手语识别，将双流3维卷积神经网络用于连续语句的识别中，并设计分析了不同联合损失函数对识别的影响，在中国手语数据集上验证了其方法的有效性。同年Wang等人[48]提出了一种由时域卷积模块、双向递归单元模块和融合层模块组成的混合深层结构，设计了一种基于CTC损失的联合优化方法和一种基于深度分类分数的融合策略来生成最终的语句，通过融合时域卷积模块捕获的邻近视频片段特征上的短期时序信息和递归单元模块跨时间维度上的长期上下文时序信息，得到了很好的识别效果，但是网络架构较为复杂，对硬件以及时间的要求比较高。

3.3 混合模型

相对于以上两种主干网络模型架构，基于混合网络模型的连续语句手语识别研究最为广泛，充分利用了卷积神经网络的特征提取能力和循环神经网络时序分类优势，实现更为准确的识别。

Koller等人[49]为代表的德国亚琛工业大学在2009～2012年基于德国手语数据集进行了一系列的连续手语识别的研究，并在2016年联合萨里大学提出了一种基于CNN和HMM的混合模型用于德国连续手语的识别，在两个公开的大规模基准手语数据集上取得了很好的识别效果，但是容易过拟合。针对数据过拟合问题，他们于2017年提出了一种基于CNN+BLSTM+HMM混合模型，用于连续手语识别的研究[50]，该文提出了一种迭代的重新对齐方法用于解决视频序列和标签的对应关系，首先是用CNN+BLSTM训练一个初始化模型，然后通过提出的Re-alignment算法不断进行参数调整，最后接入HMM模型用于分类结果的输出。并于2018年发表在计算机视觉国际期刊上[51]，实现了手语连续语句的自动识别。

Pigou等人[52]在ICCV2017会议上提出了一种将3维残差网络与LSTM相结合用于连续手语的识别，在ChaLearn手势识别大赛中取得了优异的成绩。Cui等人[53]在CVPR2017会议上针对视频序列和标签序列的对齐问题，提出了一种基于CNN+BLSTM+CTC+DetectionNet网络的连续手语识别3阶段优化模型。Ariesta等人[54]提出了将3D-CNN与RNN网络相结合，用于视频级的连续手语识别，3D-CNN从图像帧中提取特征，双向RNN从视频帧的序列行为中提取时序特征。Guo等人[55]于AAAI2018会议上提出了一种分层的HLSTM编解码模型，通过传输帧、剪辑和视素单元之间的时空转换来处理不同粒度的手语识别。首先利用3维卷积网络挖掘视频片段的时空特征，然后利用自适应可变长在线关键片段挖掘方法选取关键帧序列，并提出了一种时间注意加权机制来平衡视频时序位置之间的内在关系。最后，利用两层LSTM进行解码网络。该算法既保留了原始视频特征，又得到了高级特征，在中国手语数据集中取得了不错的识别效果，但模型复杂，不易实现。

基于深度学习的连续语句手语识别技术及代表性工作如表2所示。

4 手语数据集的发展及简述

手语识别问题本质上属于人工智能范畴，人工智能的3大核心驱动力包括：算力、算法和大数据。没有大规模的数据，深度学习方法就无法发挥其优势作用。因此，基于深度学习的手语识别技术的发展也必然依附于大规模的手语数据集，以进行算法研究、对比分析与标准评估。随着手语识别研究的不断深入，手语数据集的需求也不断扩大，国内外各种规模和特点的手语数据集陆续推出。中国科学技术大学于2015年起自制中国手语数据集，包括单个孤立词和连续语句，并于2017年公开中国手语数据集；国外以亚琛工业大学为代表的德国手语数据集自2009年公布，为国内外众多手语识别研究团队提供了数据支持。手语数据集根据数据特点可以分为RGB数据、深度数据、骨架数据、光学数据和混合数据等。从历年来手语识别的研究中，大部分手语识别的研究都是基于中国手语(CSL)、德国手语(RWTH)等。手语数据集详细分类如表3所示。接下来将重点讲述[47,56,57]所述数据集，并简要介绍手语识别的评估标准。

4.1 RWTH-PHOENIX-Weather

历时6年多(2009～2014)，德国亚琛工业大学录制了RWTH手语数据集，用于德国凤凰公共电视台每日新闻及天气预报节目的手语解说。所录制的手语视频每秒25帧，分别于2012年、2014年共录制两个版本，2014年是对2012年的数据扩充。2014年版本总共有190个样例，包含965940帧，共有1558个词汇组合成6861个连续语句，其中369帧

表 2　基于深度学习的连续语句的手语识别技术及代表性工作

作者/单位	年份	技术特点	评估标准(%)	数据集	样本大小
Camgoz NC, Koller O/亚琛工业大学	2016	3维卷积网络(从RGB数据提取时序特征)[45]	Jaccard系数：26.9	Chalearn	
	2016	基于卷积神经网络和HMM的混合模型[49]	WER:39.7		
	2017	基于CNN、HMM、CTC[50]	WER:38.8	RWTH-PHOENX-Weather	分辨率：5000×90
	2017	双向长短时网络-BLSTM(基于CTC算法)[39]	WER:43.1		
	2018	基于CNN、HMM及RNN的混合模型[51]			
Pigou L /根特大学	2017	基于3维网络和LSTM混合模型(RGB-D)[52]	Jaccard系数：31.6	Chalearn	
Cui Runpeng/清华大学	2017	基于CNN和BLSTM(基于CTC算法)[53]	WER:38.7	RWTHPHOENIX-Weather	分辨率：16000×20
	2018	双向长短时网络-BLSTM(多模态数据)[40]	WER:46.9		
Shi B /美国芝加哥大学	2018	基于注意力机制的长短时网络[41]	WER:41.9	AmericanSign Language (ASL)	
Ko S K /韩国电子研究所	2018	循环神经网络(加入骨架关节点数据)[42]	Acc:89.5	KETI韩国手语数据集	100类
Zhang Qian/上海交通大学	2018	双向长短时网络-BLSTM[43]	Acc:93.1	AmericanSign Language(ASL)	100类
Li Houqiang, Huang Jie /中国科学技术大学	2018	3维卷积网络(时间分类的对齐算法)[46]	WER:37.3	RWTH-PHOENIX-Weather	
		双流3维卷积网络(加入LSTM)[47]	Acc:82.7	ChineseSign Language	100类
Guo Dan/合肥工业大学，中国科学技术大学	2018	3维卷积神经网络(时域卷积、CTC算法、后融合策略)[48]	WER:37.8	RWTH-PHOENIX-Weather	
		3维卷积网络和RNN相结合(自适应变长在线关键片段挖掘关键帧)[55]	Acc:92.9	ChineseSign Language(CSL)	100类
Ariesta M C /雅加达大学	2018	3维卷积网络和RNN相结合(基于CTC)[54]		SIBI	30类
Mittal A /印尼科技大学	2019	改进的长短时记忆网络[44]	Acc:72.3	印度手语数据集(ISL)	942类

表 3　手语数据集分类

名称	所属国家	类别	场景	样本	数据特点	数据类型	可用性
RWTH-PHOENIX-Weather[56]	德国	1200	9	45760	RGB	句子	公开
Chalearn[57]	美国	249	7	50000	RGB/深度	单词	部分公开
DGS Kinect 40[58]	德国	40	15	3000	多视角	孤立词	
CSL[47]	中国	500/100	1	25000	深度/骨架/RGB	孤立词/句子	公开
SIGNUM[59]	德国	450	25	33210	RGB	句子	公开
GSL 20[60]	希腊	20	6	840	RGB	单词	
Boston ASLLVD[61]	美国	3300+	6	9800	RGB	单词	公开
PSL Kinect 30[62]	波兰	30	1	300	RGB/深度	单词	公开
LSA64[63]	阿根廷	64	10	3200	RGB	单词	公开
DEVISIGN-G[64]	中国	36	8	432	RGB	单词	
DEVISIGN-D[64]		500		6000			
DEVISIGN-L[64]		2000		24000			
CUNY ASL[65]	美国		8		RGB	句子	
SignsWorld Atlas[66]	阿拉伯	32	10		RGB	单词	公开
ASL Fingerspelling[67]	美国	24	5	131000	RGB/深度	单词	公开

对脸部表情进行了标注，如图2所示。该数据集具体参数如表4所示。

4.2　CSL数据集

中国手语数据集(CSL)是由中国科学技术大学自2015年起利用Kinect采集的中国手语数据集，包含25 K标记的视频实例，共有超过100个时长的视频，50个操作者拍摄，每个操作者重复5次，包含RGB、深度以及骨架关节点数据，分为孤立词和

图 2 RWTH德国手语数据样例

表 4 RWTH-PHOENIX-Weather参数

参数	2012年版	2014年版
# 操作者数量	7	9
# 样例	190	645
# 帧数	293077	965940
# 语句数量	1980	6861
# 词汇量	911	1558
# 分辨率	210×260	720×576

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5000个视频，每一个句子平均包含4～8个单词。每一个视频实例都由专业的中国手语老师进行标注。具体的CSL数据集参数如表5所示，数据集样例如图3所示。

4.3 Chalearn数据集

不同于以上2种标准手语数据，Chalearn是由一系列连续的动态手势组成的手势数据集，但由于技术方法的通用性，常把Chalearn也视为手语识别研究的测试数据集。该数据集总共47933个视频，包含RGB和深度数据，每个RGB-D视频只代表一个手势，由21个操作者执行。该数据集某帧的视觉方式如图4所示。

表 5 CSL数据集参数

参数名称	数值
RGB分辨率	1920×1080
深度数据分辨率	512×424
视频时长(s)	10～14
平均样例数	7
总样例	25000
# 操作者数量	50
词汇量	178
骨架关节点数	21
fps	25
总时长	100+

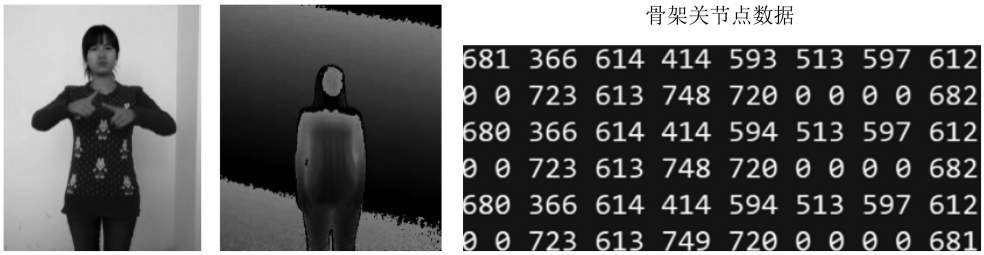

图 3 CSL中国手语数据样例

图 4 每帧的视觉方式

4.4 手语数据集评估标准

基于上述数据集，学者们提出了评估手语识别方法性能的指标，作为各种技术和算法的评估标准。

(1) 孤立词识别：现阶段基于准确率进行评估，较为单一。准确率是指被分对的样本数占所有样本数的比例，如式(1)所示。通常来说，准确率越高，识别结果越好。

$$Acc = \frac{(\mathrm{TP}+\mathrm{TN})}{(P+N)} \tag{1}$$

其中TP为被正确地划分为正例的个数，即实际为正例且被分类器划分为正例的实例数(样本数)；TN为被正确地划分为负例的个数，即实际为负例且被分类器划分为负例的实例数；P为正样本数，N为负样本数。

(2) 连续语句识别：数据集标签复杂多样，为使识别出来的手语词序列和标签词序列保持一致，需要进行替换、删除或者插入某些单词。因此借鉴语音识别领域的相关评估指标进行评价，其中包括单词错误率(WER)[68]、跟踪错误率(TER)[69]等。

单词错误率(WER)是由Levenshtein距离导出的某种性能度量，代表将一个单词转换为另一个单词所需的单字符编辑(插入、删除或替换)的最小数量，如式(2)所示。通常来说，WER越小，识别性能越优。

$$\mathrm{WER} = \frac{I+D+S}{L} \tag{2}$$

其中L为标准序列单词的总个数，I, D和S分别代表插入、删除和替换的词的总个数。

跟踪错误率(TER)即跟踪误差，指所标注的真实目标位置与自动跟踪发现的位置之间的平均偏差，如式(3)所示。通常来说，TER越大，识别效果越差。

$$\begin{aligned}&\mathrm{TER}=\frac{1}{T}\sum_{t=1}^{T}\delta_\tau\left(l_t,l^t\right),\\&\delta_\tau\left(l,m\right)=\begin{cases}0,\|l-m\|<\tau\\1,\text{其它}\end{cases}\end{aligned}\tag{3}$$

其中，(l,m)为视频帧中的2维坐标，τ为偏差阈值，l_t为t时刻视频帧横向坐标。

5　总结与展望

5.1　已有研究存在的问题

手语识别在计算机视觉、模式识别、人机交互、虚拟现实等相关领域有着重要的研究价值，尽管近年来深度学习技术已经很大程度上提高了手语识别的精度与速度，但距离真正实时、鲁棒、精准的手语识别与翻译的应用目标，还有一定的空间。主要挑战表现在：

(1) 手语行为本身的灵活性与细节性：手语是由上肢和手部动作形成的行为序列，手部是人体最灵活的肢体，其内外、正反、距离上肢的远近、角度以及五指的动作等，都对手语语义有影响。部分手语还涉及嘴唇及面部表情的配合。因此，识别准确率与实时性仍是手语识别追求的目标。

(2) 手语行为受背景干扰、光照、观察角度及操作者规范程度等的影响：目前的数据集中，操作者通常都是整体站立不动、只有上肢和手部运动，但在现实应用中存在复杂背景、多人遮挡、光照条件变化、操作者全身运动、打手语不够标准等各种情况，为识别带来更大的难度。

(3) 连续语句中的长时时序关系及孤立词之间过渡帧的自由性：连续语句手语识别需要挖掘长时期的时序依赖，以便建立语义结构，同时需要适应空间信息的复杂度以及孤立词之间过渡帧的自由性与随意性。

高精准、可扩展性、鲁棒性、实时性及用户无关性仍然是未来手语识别研究中所面临的重要挑战，同时，如何将已有成果应用于实际生活、实现跨平台部署等，也是亟待解决的问题。

5.2　展望

由于深度学习本质上属于数据驱动，基于深度学习的手语识别技术随着大数据量的积累与深入挖掘，必定会有进一步的发展。未来手语识别将随着手语本身的特性、基准数据集、识别算法以及高效算力等多方面的推进，取得新的突破。

(1) 单个孤立词的手语识别主要在同时提高准确率与速度上深化，而连续语句的识别将会更多借助于语音识别、自然语言理解、机器翻译等领域的原理和方法、结合视频本身的特点来进行创新。

(2) 创建良好的数据集与评价标准仍然迫在眉睫，完整且规范的数据标签能够大幅度促进技术方法的优化。尽管现阶段已经有一些相关的基准手语数据集，但真实场景下的手语视频数据集仍然非常匮乏，而且数据集标签的制作是一项费时费力的昂贵任务。

(3) 随着深度学习理论研究的深入，视频理解与分析问题上的本质创新将带来手语识别算法上的突破，以上所述的不同网络架构也需要继续探索，对于时序性较强的连续语句识别，混合网络模型将是未来的主流网络算法；而对孤立词手语识别而言，3维卷积神经网络计算量大、强依赖于所使用的内存及GPU等性能，在受限的硬件条件下，使用循环神经网络仍有较大的发展空间。

(4) 深度学习技术相对传统方法而言，与网络速度、GPU等硬件性能的关系更为紧密，计算机硬件设备的更新换代及网络升级，将会缩短实验、测试与评估时间，加速成果转化，从而更好地推动理论研究与算法创新。

未来手语识别随着不同领域的交叉融合将会得到更大的发展。期待更多的学者加入手语识别的研究中，使得手语识别的研究成果能够真正服务于大众，提高整个社会的智能信息化水平。

参 考 文 献

[1] HINTON G E, OSINDERO S, and TEH Y W. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527–1554. doi: 10.1162/neco.2006.18.7.1527.

[2] 周宇. 中国手语识别中自适应问题的研究[D].[博士论文], 哈尔滨工业大学, 2009.
ZHOU Yu. Research on signer adaptation in Chinese sign language recognition[D].[Ph.D. dissertation], Harbin Institute of Technology, 2009.

[3] CHEOK M J, OMAR Z, and JAWARD M H. A review of hand gesture and sign language recognition techniques[J]. *International Journal of Machine Learning and Cybernetics*, 2019, 10(1): 131–153. doi: 10.1007/s13042-017-0705-5.

[4] TANG Ao, LU Ke, WANG Yufei, *et al*. A real-time hand posture recognition system using deep neural networks[J]. *ACM Transactions on Intelligent Systems and Technology*, 2015, 6(2): 1–23. doi: 10.1145/2735952.

[5] PIGOU L, DIELEMAN S, KINDERMANS P J, *et al*. Sign language recognition using convolutional neural

networks[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014: 572–578.

[6] KANG B, TRIPATHI S, and NGUYEN T Q. Real-time sign language fingerspelling recognition using convolutional neural networks from depth map[C]. The 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 2015: 136–140.

[7] HOSSEN M A, GOVINDAIAH A, SULTANA S, *et al*. Bengali sign language recognition using Deep Convolutional Neural Network[C]. The 7th Joint International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 2018: 369–373.

[8] KOLLER O, BOWDEN R, and NEY H. Automatic alignment of hamNoSys subunits for continuous sign language recognition[C]. The 10th Edition of the Language Resources and Evaluation Conference, Portorož, Slovenia, 2016: 121–128.

[9] GARCIA B and VIESCA S A. Real-time American sign language recognition with convolutional neural networks[J]. *Convolutional Neural Networks for Visual Recognition*, 2016, 2: 225–232.

[10] JI Y, KIM S, and LEE K B. Sign language learning system with image sampling and convolutional neural network[C]. The 1st IEEE International Conference on Robotic Computing (IRC), Taichung, China, 2017: 371–375.

[11] KIM S, JI Y, and LEE K B. An effective sign language learning with object detection based ROI segmentation[C]. The 2nd IEEE International Conference on Robotic Computing (IRC), Laguna Hills, USA, 2018: 330–333.

[12] KÖPÜKLÜ O, KÖSE N, and RIGOLL G. Motion fused frames: Data level fusion strategy for hand gesture recognition[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, USA, 2018: 2103–2111.

[13] KONSTANTINIDIS D, DIMITROPOULOS K, and DARAS P. Sign language recognition based on hand and body skeletal data[C]. 2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), Helsinki, Finland, 2018: 1–4.

[14] DEVINEAU G, MOUTARDE F, WANG Xi, *et al*. Deep learning for hand gesture recognition on skeletal data[C]. The 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xian, China, 2018: 106–113.

[15] MOLCHANOV P, GUPTA S, KIM K, *et al*. Hand gesture recognition with 3D convolutional neural networks[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition workshops, Boston, USA, 2015: 1–7.

[16] WU Di, PIGOU L, KINDERMANS P J, *et al*. Deep dynamic neural networks for multimodal gesture segmentation and recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(8): 1583–1597. doi: 10.1109/TPAMI.2016.2537340.

[17] HUANG Jie, ZHOU Wengang, LI Houqiang, *et al*. Sign language recognition using 3D convolutional neural networks[C]. 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 2015: 1–6.

[18] HUANG Jie, ZHOU Wengang, LI Houqiang, *et al*. Attention-based 3D-CNNs for large-vocabulary sign language recognition[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, 29(9): 2822–2832. doi: 10.1109/TCSVT.2018.2870740.

[19] LI Yunan, MIAO Qiguang, TIAN Kuan, *et al*. Large-scale gesture recognition with a fusion of RGB-D data based on the C3D model[C]. The 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2016: 25–30.

[20] LI Yunan, MIAO Qiguang, TIAN Kuan, *et al*. Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, 28(10): 2956–2964. doi: 10.1109/TCSVT.2017.2749509.

[21] MIAO Qiguang, LI Yunan, OUYANG Wanli, *et al*. Multimodal gesture recognition based on the resc3d network[C]. 2017 IEEE International Conference on Computer Vision Workshops, Venice, Italy, 2017: 3047–3055.

[22] ELBADAWY M, ELONS A S, SHEDEED H A, *et al*. Arabic sign language recognition with 3d convolutional neural networks[C]. The 8th International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2017: 66–71.

[23] YE Yuancheng, TIAN Yingli, HUENERFAUTH M, *et al*. Recognizing American sign language gestures from within continuous videos[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, USA, 2018: 2064–2073.

[24] LIANG Zhijie, LIAO Shengbin, and HU Bingzhang. 3D convolutional neural networks for dynamic sign language recognition[J]. *The Computer Journal*, 2018, 61(11): 1724–1736. doi: 10.1093/comjnl/bxy049.

[25] CATE H, DALVI F, and HUSSAIN Z. Sign language recognition using temporal classification[EB/OL]. http://arxiv.org/abs/1701.01875v1, 2017.

[26] CHAI Xiujuan, LIU Zhipeng, YIN Fang, *et al*. Two streams recurrent neural networks for large-scale continuous gesture recognition[C]. The 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2016: 31–36.

[27] LIU Tao, ZHOU Wengang, and LI Houqiang. Sign language recognition with long short-term memory[C]. 2016 IEEE

International Conference on Image Processing (ICIP), Phoenix, USA, 2016: 2871–2875.

[28] LI Xiaoxu, MAO Chensi, HUANG Shiliang, *et al*. Chinese sign language recognition based on SHS descriptor and encoder-decoder LSTM model[C]. The 12th Chinese Conference on Biometric Recognition. Shenzhen, China, 2017: 719–728.

[29] HUANG Shiliang, MAO Chensi, TAO Jinxu, *et al*. A novel chinese sign language recognition method based on keyframe-centered clips[J]. *IEEE Signal Processing Letters*, 2018, 25(3): 442–446. doi: 10.1109/LSP.2018.2797228.

[30] YANG Su and ZHU Qing. Continuous Chinese sign language recognition with CNN-LSTM[J]. *SPIE*, 2017, 10420.

[31] YANG Su and ZHU Qing. Video-based Chinese sign language recognition using convolutional neural network[C]. The 9th IEEE International Conference on Communication Software and Networks (ICCSN), Guangzhou, China, 2017: 929–934.

[32] LIN Chi, WAN Jun, LIANG Yanyan, *et al*. Large-scale isolated gesture recognition using a refined fused model based on masked Res-C3D network and skeleton LSTM[C]. The 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 2018: 52–58.

[33] HALIM K and RAKUN E. Sign language system for Bahasa Indonesia (Known as SIBI) recognizer using TensorFlow and Long Short-Term Memory[C]. 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Yogyakarta, Indonesia, 2018: 403–407.

[34] BHATEJA V, COELLO C A C, and SATAPATHY S C. Intelligent Engineering Informatics[C]. The 6th International Conference on FICTA. Singapore: 2018: 623–632.

[35] BANTUPALLI K and XIE Ying. American Sign Language recognition using deep learning and computer vision[C]. 2018 IEEE International Conference on Big Data (Big Data), Seattle, USA, 2018: 4896–4899.

[36] KONSTANTINIDIS D, DIMITROPOULOS K, and DARAS P. A deep learning approach for analyzing video and skeletal features in sign language recognition[C]. 2018 IEEE International Conference on Imaging Systems and Techniques (IST), Krakow, Poland, 2018: 1–6.

[37] VINCENT H, TOMOYA S, and GENTIANE V. Convolutional and recurrent neural network for human action recognition: Application on American sign language[EB/OL]. http://biorxiv.org/content/10.1101/535492v1, 2019.

[38] LIAO Yanqiu, XIONG Pengwen, MIN Weidong, *et al*. Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks[J]. *IEEE Access*, 2019, 7: 38044–38054. doi: 10.1109/ACCESS.2019.2904749.

[39] CAMGOZ N C, HADFIELD S, KOLLER O, *et al*. SubUNets: End-to-end hand shape and continuous sign language recognition[C]. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 3075–3084.

[40] CUI Runpeng, LIU Hu, and ZHANG Changshui. A deep neural framework for continuous sign language recognition by iterative training[J]. *IEEE Transactions on Multimedia*, 2019, 21(7): 1880–1891. doi: 10.1109/TMM.2018.2889563.

[41] SHI Bowen, DEL RIO A M, KEANE J, *et al*. American Sign Language fingerspelling recognition in the wild[C]. 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018: 145–152.

[42] KO S K, SON J G, and JUNG H. Sign language recognition with recurrent neural network using human keypoint detection[C]. 2018 Conference on Research in Adaptive and Convergent Systems, Honolulu, USA, 2018: 326–328.

[43] ZHANG Qian, WANG Dong, ZHAO Run, *et al*. MyoSign: Enabling end-to-end sign language recognition with wearables[C]. The 24th International Conference on Intelligent User Interfaces, Marina del Ray, USA, 2019: 650–660.

[44] MITTAL A, KUMAR P, ROY P P, *et al*. A modified LSTM model for continuous sign language recognition using leap motion[J]. *IEEE Sensors Journal*, 2019, 19(16): 7056–7063. doi: 10.1109/JSEN.2019.2909837.

[45] CAMGOZ N C, HADFIELD S, KOLLER O, *et al*. Using convolutional 3d neural networks for user-independent continuous gesture recognition[C]. The 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2016: 49–54.

[46] PU Junfu, ZHOU Wengang, and LI Houqiang. Dilated convolutional network with iterative optimization for continuous sign language recognition[C]. The 27th International Joint Conference on Artificial Intelligence, Wellington, New Zealand, 2018: 885–891.

[47] HUANG Jie, ZHOU Wengang, ZHANG Qilin, *et al*. Video-based sign language recognition without temporal segmentation[C]. The 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 2257–2264.

[48] WANG Shuo, GUO Dan, ZHOU Wengang, *et al*. Connectionist temporal fusion for sign language translation[C]. The 26th ACM International Conference on Multimedia, Seoul, Korea, 2018: 1483–1491.

[49] KOLLER O, ZARGARAN O, NEY H, *et al*. Deep sign: Hybrid CNN-HMM for continuous sign language recognition[C]. 2016 British Machine Vision Conference, York, UK, 2016: 1–2.

[50] KOLLER O, ZARGARAN S, and NEY H. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017: 4297–4305.

[51] KOLLER O, ZARGARAN S, NEY H, *et al*. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs[J]. *International Journal of Computer Vision*, 2018, 126(12): 1311–1325. doi: 10.1007/s11263-018-1121-3.

[52] PIGOU L, VAN HERREWEGHE M, and DAMBRE J. Gesture and sign language recognition with temporal residual networks[C]. 2017 IEEE International Conference on Computer Vision Workshops, Venice, Italy, 2017: 3086–3093.

[53] CUI Runpeng, LIU Hu, and ZHANG Changshui. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 7361–7369.

[54] ARIESTA M C, WIRYANA F, SUHARJITO, *et al*. Sentence level Indonesian sign language recognition using 3D convolutional neural network and bidirectional recurrent neural network[C]. 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), Jakarta, Indonesia, 2018: 16–22.

[55] GUO Dan, ZHOU Wengang, LI Houqiang, *et al*. Hierarchical LSTM for sign language translation[C]. The 32nd AAAI Conference on Artificial Intelligence, the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, USA, 2018: 6845–6852.

[56] FORSTER J, SCHMIDT C, HOYOUX T, *et al*. RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus[C]. The 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, 2012: 3785–3789.

[57] ESCALERA S, BARÓ X, GONZÀLEZ J, *et al*. Chalearn looking at people challenge 2014: Dataset and results[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014: 459–473.

[58] ONG E J, COOPER H, PUGEAULT N, *et al*. Sign language recognition using sequential pattern trees[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 2200–2207.

[59] VON AGRIS U, ZIEREN J, CANZLER U, *et al*. Recent developments in visual sign language recognition[J]. *Universal Access in the Information Society*, 2008, 6(4): 323–362. doi: 10.1007/s10209-007-0104-x.

[60] EFTHIMIOU E and FOTINEA S E. GSLC: Creation and annotation of a Greek sign language corpus for HCI[C]. The 4th International Conference on Universal Access in Human-Computer Interaction, Beijing, China, 2007: 657–666.

[61] NEIDLE C, THANGALI A, and SCLAROFF S. Challenges in development of the American Sign Language lexicon video dataset (ASLLVD) corpus[C]. The 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, Istanbul, Turkey, 2012: 1–8.

[62] OSZUST M and WYSOCKI M. Polish sign language words recognition with Kinect[C]. The 6th International Conference on Human System Interactions (HSI), Sopot, Poland, 2013: 219–226.

[63] RONCHETT F, QUIROGA F, ESTREBOU C A, *et al*. LSA64: An Argentinian sign language dataset[C]. The 22nd Congreso Argentino de Ciencias de la Computación (CACIC 2016), San Luis, USA, 2016: 794–803.

[64] CHAI Xiujuan, WANG Hanjie, and CHEN Xilin. The DEVISIGN large vocabulary of Chinese sign language database and baseline evaluations[R]. Technical Report VIPL-TR-14-SLR-001, 2014.

[65] LU Pengfei and HUENERFAUTH M. Collecting and evaluating the CUNY ASL corpus for research on American sign language animation[J]. *Computer Speech & Language*, 2014, 28(3): 812–831. doi: 10.1016/j.csl.2013.10.004.

[66] SHOHIEB S M, ELMINIR H K, and RIAD A M. Signsworld atlas; a benchmark Arabic sign language database[J]. *Journal of King Saud University-Computer and Information Sciences*, 2015, 27(1): 68–76. doi: 10.1016/j.jksuci.2014.03.011.

[67] PUGEAULT N and BOWDEN R. Spelling it out: Real-time ASL fingerspelling recognition[C]. 2011 IEEE International Conference on Computer Vision workshops (ICCV Workshops), Barcelona, Spain, 2011: 1114–1119.

[68] PRABHAVALKAR R, SAINATH T N, WU Yonghui, *et al*. Minimum word error rate training for attention-based sequence-to-sequence models[C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018: 4839–4843.

[69] KOLLER O, FORSTER J, and NEY H. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers[J]. *Computer Vision and Image Understanding*, 2015, 141: 108–125. doi: 10.1016/j.cviu.2015.09.013.

张淑军：女，1980年生，副教授，研究方向为计算机视觉.
张 群：女，1994年生，硕士生，研究方向为计算机视觉.
李 辉：男，1984年生，副教授，研究方向为计算机视觉.