计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100227

手语识别、翻译与生成综述

郭　丹　唐申庚　洪日昌　汪　萌
合肥工业大学计算机与信息学院　合肥 230601
大数据知识工程教育部重点实验室(合肥工业大学)　合肥 230601
智能互联系统安徽省实验室　合肥 230601

摘　要　手语研究是典型的多领域交叉研究课题,涉及计算机视觉、自然语言处理、跨媒体计算、人机交互等多个方向,主要包括离散手语识别、连续手语翻译和手语视频生成。手语识别与翻译旨在将手语视频转换成文本词汇或语句,而手语生成是根据口语或文本语句合成手语视频。换言之,手语识别翻译与手语生成可视为互逆过程。文中综述了手语研究的最新进展,介绍了研究的背景现状和面临的挑战;回顾了手语识别、翻译和生成任务的典型方法和前沿研究;并结合当前方法中存在的问题,对手语研究的未来发展方向进行了展望。

关键词:视频理解;机器翻译;离散手语识别;连续手语翻译;手语视频生成
中图法分类号　TP391.4

Review of Sign Language Recognition, Translation and Generation

GUO Dan, TANG Shen-geng, HONG Ri-chang and WANG Meng
School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China
Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230601, China
Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei 230601, China

Abstract　Sign language research is a typical cross-disciplinary research topic, involving computer vision, natural language processing, cross-media computing and human-computer interaction. Sign language research mainly includes isolated sign language recognition, continuous sign language translation and sign language video generation. Sign language recognition and translation aim to convert sign language videos into textual words or sentences, while sign language generation synthesizes sign videos based on spoken or textual sentences. In other words, sign language translation and generation are inverse processes. This paper reviews the latest progress of sign language research, introduces its background and challenges, reviews typical methods and cutting-edge research on sign language recognition, translation and generation tasks. Combining with the problems in the current methods, the future research direction of hand language is prospected.

Keywords　Video understanding, Machine translation, Isolated sign language recognition, Continuous sign language translation, Sign language video generation

1　引言

健全人可以使用口头语言便捷交流,而听障人士(失聪者或失语者等)则需要通过手语来传达自己的想法。由于大部分健全人未学习过手语,推广手语使之适用于正常社会的沟通存在障碍。手语识别与翻译技术为听障人士与健全人沟通提供了便利。手语研究不仅要让健听人读懂手语,还应该让听障人士看懂健全人说了什么。手语识别与翻译是前者,手语生成研究属于后者。对于听障人群来说,这种交互过程尤为重要。因此,手语识别翻译与手语生成研究有着重要的理论价值和应用价值以及社会意义。

近年来,随着视频采集、人机交互以及虚拟现实技术的快速发展,基于视频的手语识别、翻译与生成研究在国际上受到了越来越多的关注。手语实时通信已成为当前计算机视觉与模式识别领域的一个重要课题。由于手语研究涉及视频理解、手势识别、动作识别、视频描述生成、视觉生成等任务,其

到稿日期:2021-01-29　返修日期:2021-02-19
基金项目:国家重点研发计划(2018YFC0830103);国家自然科学基金(61876058);中央高校基本科研业务费专项资金(JZ2020HGTB0020)
This work was supported by the National Key Research and Development Program of China(2018YFC0830103), National Natural Science Foundation of China(61876058) and Fundamental Research Funds for the Central Universities of Ministry of Education of China(JZ2020HGTB0020).
通信作者:郭丹(guodan@hfut.edu.cn)

发展对视频处理、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都有借鉴意义。

手语研究主要包括离散手语识别（Isolated Sign Language Recognition[1-4]/Sign Language Interpretation[5-6]）、连续手语翻译（Continuous Sign Language Recognition[7-10]/Sign Language Translation[11-15]）和手语生成（Sign Language Generation[16-17]/Sign Language Production[18-21]/Sign Language Synthesis[22-24]）3个方面。如图1所示，手语识别与翻译旨在将手语视频转换成文本词汇或语句，而手语生成是根据自然语言语句或口语生成合成视频。手语识别翻译与手语生成可互为逆过程。

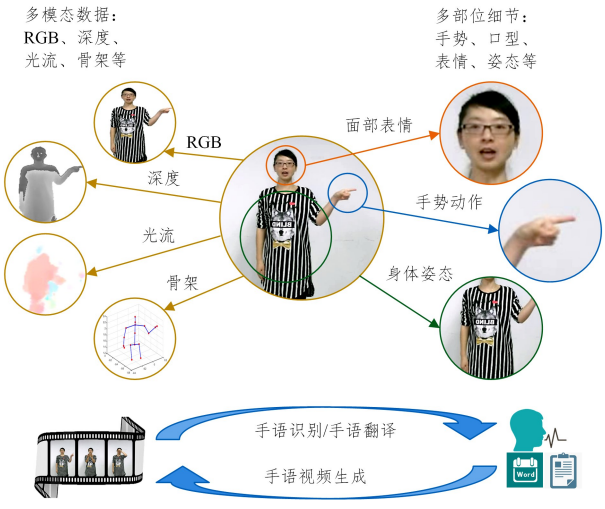

图1　手语研究内容以及手语样本示例

Fig. 1　Illustration of sign language research and sign language samples

本文基于手语研究的3个主要分支，介绍了手语研究的发展现状和挑战，并对主流方法进行了梳理分析，最后探讨了手语研究的未来发展方向。

2　手语研究的问题与挑战

手语研究面向的主要群体是听障人士（包括失语者、失聪者等），在考虑技术创新的同时，不可忽略当今手语AI技术在失语者社区中的适用性和实用性。成功的手语研究需要了解失语者的文化、背景和生活环境，创建符合用户地区、年龄、性别、教育程度、所使用手语语种的类别以及语言熟练程度等属性的手语应用系统。从研究技术上看，目前的手语研究面临着如下挑战。

2.1　在线复杂手语视频理解

视频是手语展示内容的主要载体。现有手语研究所采集的视频数据背景干净，只包含手语演示者在实验室环境下以白墙或者蓝布为背景的摄像头区域中演示的手语动作，这与真实场景下的手语应用环境相距甚远。复杂视频中前背景分割、光照差异、遮挡等噪声问题有待进一步的研究。同时，体态、手势的细微变化，尤其是手部联合面部的局部视觉差异普遍存在。此外，常规的视频分析可以通过抽取关键帧来描述整个视频故事，但在手语视频中，帧与帧之间的转换十分紧密，如果抽取少量关键帧，则会丧失连续转换的时序性，形成语义缺失。这些特性对在线实时手语翻译形成了挑战。

2.2　弱监督语句翻译

由于大部分标注人员缺乏手语专业认识并且标注成本昂贵，现有的手语语句数据集通常只包含句子级标签，而不提供精细粒度的对齐注释，即词汇级标签。因此，连续手语语句翻译是典型的弱监督学习任务。手语强调各手势之间在时序上的分割关系与传递关系。分割关系指一个句子中独立单词或词组之间合理停顿的分割点；传递关系指在语法规则下，构成手势单词的子视觉元素合理的出现顺序。这对捕捉视频中的时空细节变化一致性也提出了更高的要求，例如，如何实现粗粒度动作语义单元建模；先针对见过的句子自动解析学习切割出来的词汇，再进行新短语组合以及新句子组合等。

2.3　特殊的语言学约束

与口头语言规则类似，手语由一系列动作按照语义约束规则组合而成。例如，手势的移动方向具有指示主语和宾语的语法功能；同种手语动作可能表达名词、动词等多种词性或语义。头部、手形和体态的变化均是手语动作的主要表现方式。关于头部的语义约束关系也有头部运动、面部表情、口型变化、耸肩和眼睛注视等微动作，这些都是手语表达的关键语义元素。由于跨语种的语言习惯和语法规则不同，各国手语的语言约束规则也存在着差异，例如同一动作在不同语种中表示不同的含义。此外，在实际应用场景中，手语动作也常对自然语言中的不同词性的词汇进行取舍等区别对待。由此可见，手语研究需要充分考虑手语语义约束的特殊背景知识，才能有效解决上述手语语言学中存在的语义问题和挑战。

2.4　多模态动态序列融合

单一模态数据存在一些天然的弊端，例如：彩色图像容易受到光线、角度等的影响；深度数据缺少手指和面部细节等。有效的多模态数据融合有助于弥补单模态输入的缺陷，利用多种模态的特异性和互补性来获得鲁棒的手语表征。目前，此研究方向在离散手语识别上开展的工作较多，在连续手语翻译上的相关工作较少。手语研究领域常见的多模态融合大多是通过特征融合（如前端特征拼接）或概率融合（如后端得分融合）实现的。如何在模型学习的中端层面动态地学习时序关联和多模态互补仍然是一个挑战。在序列学习模型中嵌入融合模块可以避免模态特异性在各种模态独立嵌入过程中过早被损耗掉，以尽可能地保留和学习有利线索。综合来看，探索动态的多模态序列融合方法的研究是必要的。

2.5　可扩充词汇学习

当前，手语研究正在从小型的人工词汇任务过渡到现实世界中包含大规模词汇的语句理解任务。为了将手语技术真正应用于各类实际生活场景，需要探究大规模词汇的手语识别与翻译方法。此外，良好的手语识别系统需要具备一定的灵活性和鲁棒性，能够对未见过的新词汇进行合理的近似语义推测，并且具备便捷的扩展补充途径。

2.6　自然流畅的高质量视频生成

由于手语视频生成任务发展较晚，目前手语领域对该任

务的研究仍处于起步阶段。传统的日常动作视频生成任务通常考虑人类的周期性动作(如吃饭或走路),而手语视频生成的难点在于手语动作具有变化细微且重叠度较低的特质。同时,手语生成需要同时建模身体姿态、手部动作、面部表情等多个部分,以保证各部分在时间上的同步。再者,生成的手语视频对于听障人士来说必须是可理解且可接受的,这对生成视频的画面自然度、手势细节准确度和手语动作之间的连贯性等都提出了更高的要求。

3 手语识别与翻译

现有手语识别与翻译工作的主要区别在于单词识别属于视频分类问题,语句翻译属于弱监督时序解码问题。鉴于手语视频在特征优化和时序学习技术上的共通性,下面我们不区分任务,主要从方法路线来介绍相关工作,如图 2 所示。

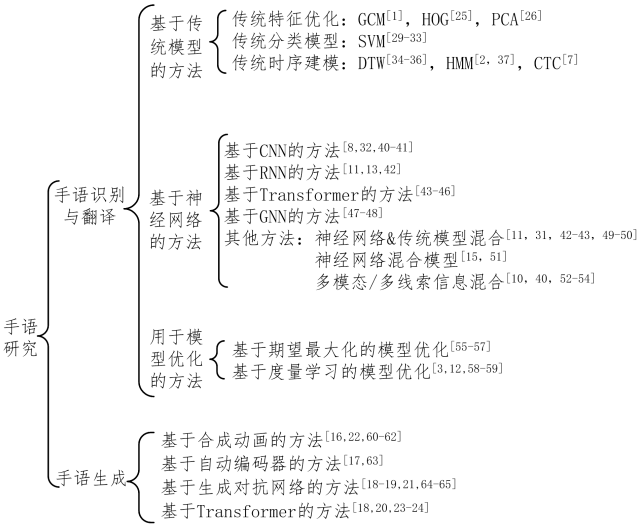

图 2 手语研究的相关工作

Fig. 2 Summary of different methods related to sign language research

3.1 基于传统模型的方法

传统的手语识别研究经历了关注特征抽取、辅以分类模型、侧重时序建模这 3 个阶段。

(1)传统特征优化。早期的手语识别方法较多使用手工特征,除了视觉特征,对基于传感器或穿戴设备采集的数据,如骨架深度坐标、肌电图(Electromyography,EMG)信号和手部传感信号等也进行了处理。Zheng 等[25]设计了一种基于三维运动图的方向梯度直方图(Histogram of Oriented Gradients,HOG)金字塔,用于表征不同尺度的手势外观信息。Oliveira 等[26]采用多阶主成分分析技术(Principal Component Analysis,PCA)压缩视觉特征,从而降低特征维度并提高识别效率。Hassan 等[27]采用均值、标准差、协方差等统计学习方法对手部位置和运动进行处理。Lin 等[28]提出了一种基于流形分析的曲线匹配方法,将人体动作的三维轨迹视为连续线段组合,并通过建模线段的几何特征来匹配曲线模式,从而识别出手语动作。Wang 等[1]提出分层的格拉斯曼协方差矩阵模型(Hierarchical Grassmann Covariance Matrix,HGCM),该模型能够获得更紧凑和更有区分性的手语特征。

(2)传统分类模型。支持向量机(Support Vector Machine,SVM)及其改进算法被广泛用于手语词汇的识别与分类[29]。Yin 等[30]提出一种基于片段的级联稀疏编码的 SVM 分类模型来获得识别结果。Pu 等[31]提取手形和轨迹特征并输入 SVM 分类器进行识别。Li 等[32]对比了 K 近邻分类器和 SVM 分类器在手语识别上的性能差异。Thang 等[33]研究了朴素 SVM、简化支持向量机(SimpSVM)和相关向量机(Relevance Vector Machine,RVM),并对识别精度和测试效率等性能进行了对比探究。

(3)传统时序建模。对于视频来说,仅依靠优化空间结构表征来提升特征的性能是远远不够的,时序线索的挖掘也同样重要。经典的时序解码方法包括动态时间规整(Dynamic Time Warping,DTW)、隐马尔可夫(Hidden Markov Model,HMM)和联结时序分类(Connectionist Temporal Classification,CTC)算法。

DTW 算法基于动态规划的思想求解手语视频和句子序列之间的最小操作距离,以实现潜在空间中视觉与文本的匹配计算[34-35]。HMM 在早期主要用于语音识别领域,基于马尔可夫链推算来推断状态序列的变化[36]。Guo 等[2]发现 HMM 中的固有潜在状态与隐状态间的转换关系有关,提出了一种自适应 HMM 手语识别框架,利用近邻传播聚类对手语手势的隐状态进行分析与研究。Tornay 等[37]开发了一种基于 KL 散度 HMM 的手语处理方法,在特征观测值与状态分类分布之间使用基于 KL 散度的损失函数来设置 HMM 模型的参数。CTC 算法对序列之间多对一的映射关系进行建模,计算所有可能的映射路径代价之和,从而实现序列转换。Pu 等[7]设计了一种基于软动态时间规整(soft-DTW)和 CTC 的双解码器结构,两种解码器均通过带有对齐约束的最大似然准则进行联合优化。

3.2 基于神经网络的方法

得益于深度学习的快速发展,深度特征往往具有更好的空间时序表达能力,可以对更多手语动作外观和运动的变化细节建模[38]。神经网络模型包括卷积神经网络(Convolutional Neural Network,CNN)、循环神经网络(Recurrent Neural Network,RNN)和图神经网络(Graph Neural Network,GNN)等,它们均已被用于手语研究。

(1)基于 CNN 的方法。二维 CNN 模型善于提取图像特征[39]。根特大学手语团队很早就提出了一个包含双二维 CNN 的手语识别系统[40]来提取手部特征和上半身特征。三维卷积神经网络(3D CNN)则考虑到了视频的时间维度。针对视频数据,为了提取具有区别性的时空特征,Huang 等[41]将多源视频流组合成多通道数据输入 3D CNN 模型,以集成颜色、深度和轨迹信息。为了更好地学习手势细节,Li 等[32]利用 C3D 模型学习和提取定长视频的 RGB 和深度时空特征。Zhou 等[8]构建了一个 2D+1D CNN 特征提取器,将连续多个 2D 图像特征转换成视频片段特征。

(2)基于 RNN 的方法。RNN 的提出是为了处理序列数

据中的时序建模问题。Li 等[42]设计了金字塔型 BiLSTM 网络来分层搜索手语关键动作。Guo 等[13]提出了一种分层 LSTM 模型，其中顶层 LSTM 用于探索视频片段序列中的时序线索，并通过时间注意力加权机制来平衡视素源位置之间的内在关系；底层的两个 LSTM 分别用来编码视觉表征和解码单词嵌入。Cihan 等[11]利用带有注意力机制的两层 RNN 构建了编码器-解码器结构，使用类似于神经机器翻译的框架端到端地将连续手语视频翻译成口语化语句。目前，RNN 模型依旧是处理手语时序学习的主流方法。

(3)基于 Transformer 的方法。相比 RNN，Transformer 模型以注意力机制为核心，具有便于并行、计算效率高的特点。Niu 等[43]在得到手语视频帧特征序列后，使用带有相对位置编码的 Transformer 编码器作为上下文模型，进一步学习帧间的时序关联。Camgoz 等[44]提出一种包含手型、脸部和姿态多种异步信息的多通道 Transformer 结构，该结构能够建模通道内部和多通道之间的上下文关系，并保留通道中的特定信息。此外，他们还设计了一个兼具手语识别和口语翻译功能的 Transformer 框架[45]，将手语识别和口语语句翻译任务集成到一个统一网络结构中进行联合优化。Zhang 等[46]使用 Transformer 结构将特征序列转换为目标语句，并采用强化学习机制优化手语翻译网络。

(4)基于 GNN 的方法。目前 GNN 主要用于手语研究中对人体骨架坐标等非结构化数据的处理。Amorim 等[47]将时空图卷积网络引入手语识别任务，在连续多帧骨架数据上构建各骨架节点间的相互关系。Tunga 等[48]在使用 GNN 模型对骨架点坐标间的关系进行建模的基础上，使用 Transformer 结构探索帧间时序依赖，实现基于姿态的手语识别。目前手语识别与翻译中使用的 GNN 模型主要用于解决单一模态的数据建模问题，使用 GNN 模型处理多模态数据嵌入以及视觉和文本语义的跨模态交互是可以探索的方向之一。

(5)其他方法。研究者们充分考虑了上述不同方法的优势，并进行了一系列的混合使用，主要包括如下 3 个方面。

1)神经网络 & 传统模型混合。现有的一些方法将传统模型和神经网络模型进行串联或嵌入，以同时利用两类模型的优势。Pu 等[31]使用 LeNet 和 3D CNN 分别提取骨架轨迹特征和手形视觉特征，采用 SVM 进行分类识别。Koller 等[49]将 CNN 端到端的嵌入引入到 HMM 中，同时以贝叶斯的方式解释 CNN 的输出，CNN-HMM 混合模型结合了 CNN 强大的识别能力和 HMM 的序列建模能力。此外，他们构建了多流 CNN-LSTM-HMMs 混合模型[11]，在多流之间添加中间同步约束，以并行地学习手语、口型和手型序列表征。在文献[50]中，训练标签被视为弱标签，并在弱监督方式下快速改进标签和图像的对齐方式，深度网络 CNN-BLSTM 被嵌入到一个 HMM 模型中用于校正帧标签。Niu 等[43]基于 Transformer 编码器和 CTC 分类器对手语数据进行状态建模和语句翻译。Li 等[42]基于 RNN 和 CTC 设计了多路序列对齐模型，并通过联合训练来优化从视频表征到单词序列的翻译过程。

2)神经网络混合模型。另一种流行的做法是结合 CNN 在特征提取方面的优势和 RNN 在时序分析上的优势，以二者结合使用的方式共同完成手语识别与翻译。Song 等[51]首先利用 ResNet3D 网络提取手语视频特征，然后设计了一种基于 CNN 和 RNN 的并行时间编码器，从局部和全局的角度分别探索特征序中的时序线索。Wang 等[15]提出一种基于 TCN 与 BGRU 融合的手语翻译方法，其中 TCN 专注于捕获相邻片段特征之间的短期时间过渡，BGRU 对长时上下文关联建模，融合模块连接 TCN 和 BGRU 的特征嵌入，以学习多特征互补关系。

3)多模态/多线索信息混合。除了对模型进行改进或混合，融合多路线索或多模态数据的方法也被广泛用于手语特征优化学习[52]。Pigou 等[40]构建了两个三层 CNN 网络，分别用于提取手语执行者的上半身特征和手部特征，两种特征串联后通过经典 ANN 模型进行进一步融合与分类。Wu 等[53]分别采用高斯-贝努利深度信念网络 3D CNN 来处理骨架动态和提取视觉特征，通过中间融合和后端融合两种策略证明了多通道融合的优越性。Zhou 等[10]提出了一种由空间多线索模块和时间多线索模块组成的时空多线索网络，以探究多线索时空特征的独特性和协同性。Yuan 等[54]设计了一种全局和局部空间注意力网络，该网络分别提取侧重于上下文的全局信息和侧重于手臂动作的局部信息后，自适应融合预测概率得分。

3.3　用于模型优化的方法

除了上述典型的代表性方法外，研究者们还采用了模型优化的一些框架，以提高手语模型的准确性和鲁棒性。

(1)基于期望最大化的模型优化。期望最大化(Expectation-Maximum，EM)算法在每一次迭代优化时分为两步，即期望步(E 步)和极大步(M 步)。具体而言，EM 算法首先估计缺失的隐含数据；接着基于观测数据和隐含数据极大化对数似然，以求解模型参数(M 步-特征提取模型优化)；最后然后基于当前模型参数重新估计隐含数据(E 步-时序学习模型优化)；反复迭代，直至收敛。借鉴 EM 算法思想，采用多阶段的伪监督学习策略，通过将伪标签作为补充监督信息来迭代优化连续手语翻译框架中的特征提取和序列对齐模型，该过程如图 3 所示。清华大学团队[55]设计了一个三阶段的手语翻译优化方案[56]，首先端到端地学习对齐方案，然后利用对齐方案训练特征提取器，最后使用改进的表征训练序列学习模块。通过在迭代 EM 算法中嵌入一个 CNN，Koller 等[57]提出了基于帧的分类器的学习方法。EM 算法利用 CNN 的判别能力，迭代地细化手语帧级标注和 CNN 后续训练。

(2)基于度量学习的模型优化。度量学习 (Distance Metric Learning，DML)的本质是相似度学习。在文献[3]中，一种参考驱动的度量学习方法被用来学习手势的通用表征模型，该模型包含一组手势参考；距离度量将样本及其同类别的参考手势拉近，使样本与异类别参考的距离疏远。基于这一思想，采用迭代方式交替更新手势参考和距离度量，以获得鲁棒的手语表征。

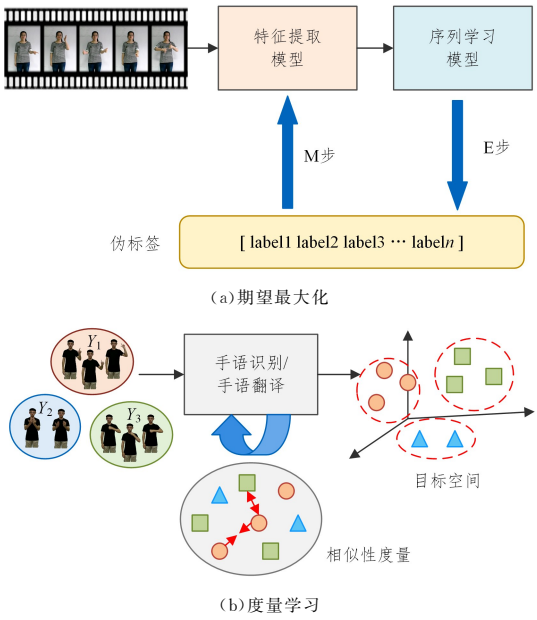

图 3 手语模型优化策略

Fig. 3 Optimization strategy of sign language model

由于手语词汇数量较大，以手语动作类别为区分的正负样本对的数量差异明显。Guo 等[12]设计了一种基于 batch 的改进三元损失距离度量，通过先分别计算正负样本对的距离损失再进行累加来衡量手语片段特征的相似度。Yin 等[58]提出了一种弱监督度量学习框架来解决非特定手语执行者的手语识别问题，该框架先通过不同执行者的标记数据来学习一个通用的距离度量，然后根据收集到的未标记样本，使通用度量适应新执行者。度量学习还被用于开放集合中的手势识别。Hazra 等[59]提出了一种基于距离度量的元学习方法，采用 K 近邻方法分类已知手势，采用聚类方法添加新的自定义手势，并将基于距离的三元损失相似度度量嵌入三维卷积神经网络以学习手势嵌入式表征。基于度量学习的方法更加关注衡量样本的相似性，从而降低了对分类类别数的敏感度，这为应对和解决开放手语识别场景中手势类别不断增加的问题提供了可行的思路。

4 手语视频生成

手语视频生成指输入口语或文本文字，由系统自动输出对应的手语视频。现有的手语视频生成任务分为手语动画视频合成、手语姿态视频生成和逼真手语视频生成 3 个子任务。图 2 和表 1 列举了近年来手语视频生成代表性的工作，其中涉及到动画合成方法、自动编码器、Transformer 和生成对抗网络等模型，具体技术说明如表 1 所列。

表 1 手语生成代表性工作

Table 1 Representative works of sign language generation

类型	方法	发表时间出处	技术特点	技术细节
人体姿态生成	NN-Based Synthesis	SPECOM 2019[24]	Transformer	将口语语句翻译成手语单词序列，然后生成骨架姿态
	Progressive Transformers	ECCV 2020[20]	Transformer	设计了 Symbolic Transformer 和 Progressive Transformer，分别用于口语到手语词序列翻译和词序列到姿态视频生成
	Adversarial Training	BMVC 2020[18]	生成对抗网络 Transformer	使用 JoeyNMT 方法将口语句子翻译成手语词序列，再使用条件 GAN 网络生成姿态视频
	Skeleton-based SLG	Neural Networks 2020[17]	自动编码器	使用聚类方法对骨架坐标数据进行预处理，基于 GAN 网络训练姿态视频生成网络
	Neural Sign Synthesis	WACV 2020[23]	Transformer	从原始手语视频中提取骨架，并利用该骨架数据训练一个手语生成网络
逼真手语视频生成	Text2Sign	BMVC 2018[64] IJCV 2020[19]	生成对抗网络	使用机器翻译方法先将口语句子翻译成手语词序列，然后采用条件 GAN 网络分两步生成手语视频
	Deep Gesture Generation	TMM 2019[63]	自动编码器	使用 DAE-LSTM 网络预测姿态序列，再利编解码结构网络合成逼真手语视频帧序列
	SignSynth	ACVR 2020[21]	生成对抗网络	使用 GAN 网络从手语词序列生成姿态视频，再将姿态视频转换成逼真手语视频
	Everybody Sign Now	SLRTP 2020[65]	生成对抗网络	使用 Everybody Dance Now 中的人体姿态迁移方法从人体姿态生成逼真手语视频

4.1 手语合成动画生成

早期的手语视频生成主要围绕基于统计模型和计算机图形学的动画合成技术展开，通过对手语词汇手势进行建模，将手语句子中的短语或词汇与手语库进行匹配等方式合成对应语义的手语动画视频。

Glauert 等[60]设计了一个语音手语助手系统，通过将语音或文本形式的输入与基于统计的语言模型进行比较，在此基础上从手语库中选出最匹配短语，并由基于 SiGML(一种手语手势标记语言)的手语动画转录技术展示出手语动画描述。Karpouzis 等[22]使用语法解析器解码书面文字的结构模式，并将其与手语的等效模式进行匹配，最后采用标准虚拟字符动画技术合成手语动画序列。为了便于手语教学，Sagawa 等[16]开发了一个包含手语识别与生成的教学系统，手语动作被视为手部描述和非手部描述的结合，学习三维计算机图形的一系列动作参数，并根据句子描述将这些参数连接起来生成同步的手语动画。Wang 等[61]基于运动跟踪原理和人体运动编辑方法构建了一个高质量的手语词运动数据库，在数据库中寻找与输入文本对应的手语运动片段并将其拼接合成，基于虚拟现实建模语言的方法生动显示视频。此外，Brock 等[62]采用 3 个循环神经网络分别推断身体、面部和手指的三

维位置数据，以获得高分辨率的手语骨架数据，随后获取逆运动学估计的关节时序角位移，并将其映射到形象化的虚拟手语动画。

综合来看，基于动画合成方法生成手语视频具有操作便捷和效率高的优势，但其依赖于大规模手语动画数据库的构建，且动画视频缺乏动作执行的逼真细节，合成动画的可理解性仍然受到人工设计的外观和动作的影响。因此，越来越多的研究开始探索更加灵活自然的手语生成方案。

4.2　手语姿态视频生成

近年来，随着计算机视觉研究的发展，跨模态转换和图像生成技术已逐渐趋于成熟，研究人员开始尝试从文本语言中直接挖掘语义信息，从而生成个性化的手语视频。由于研究的侧重点不同，不少工作着眼于解决从文本到手语骨架视频的精确转换和生成问题。

为了预测人体姿态的变化，Cui 等[63]将 DAE(Dropout Autoencoder)嵌入 LSTM。具体而言，LSTM 根据前一时刻的人体姿态给出当前时刻的姿态预测；而 DAE 作为过滤器，基于人体骨架的隐含约束来进一步优化预测的手语姿态。Xiao 等[17]提出了一种基于 VAE 的概率骨架序列生成方法，其中基于 VAE 构建的编解码模型被用于生成随机手势编码序列，并在此过程中保持序列顺序等基本模式不变。Zelinka 等[23]提出了一种包含前馈 Transformer 和循环 Transformer 的姿态序列生成框架，采用非单调的软注意力优化了基于 Transformer 的序列学习模型的性能。

Saunders 等[18]设计了一种由改进式 Transformer 生成器和条件判别器组成的对抗性多通道手语生成框架，将口语词序列和手语姿态序列同时输入判别器以鉴别姿态序列的真假。他们还提出了一种基于渐进式 Transformer 的手语生成模型[20]，该模型首次以端到端的方式将离散口语语句转换为连续手语姿势序列，其中构造的符号化 Transformer 和渐进式 Transformer 分别执行由口语语句到手语词序列的翻译和由词汇序列到人体骨架视频的转换。

4.3　逼真手语视频生成

从文本直达逼真手语视频生成的研究，现阶段基本都通过多阶段数据拟合来实现。这类方法基于前述工作中提到的手语姿态视频合成技术，在获得相对准确的人体关键点坐标后继续生成具有真实感的手语视频。典型的基于对抗生成模型 GAN 的手语视频生成框架如图 4 所示。

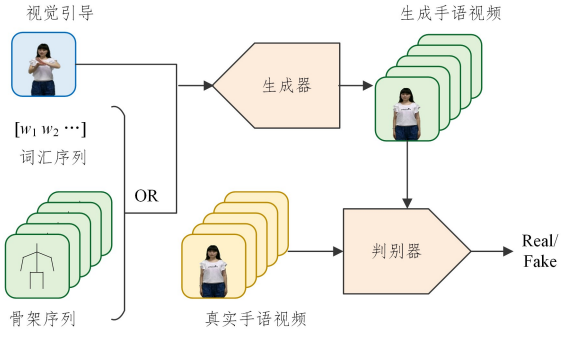

图 4　基于 GAN 的手语视频生成框架

Fig. 4　Framework of GAN-based sign language generation

Stoll 等[19,64]构建了一个 DCGAN 和一个卷积图像编码器，其中图像编码器将手语执行者外观编码为潜在表征，将人体姿态骨架和外观共同作为条件送入生成器，生成的逼真视频帧再由判别器进行评估。为了得到更加真实的手语视频，视频渲染技术也被应用于手语视频的生成。Ventura 等[65]致力于探索一种从 2D 姿态得到手语视频的方法，在对身体进行建模的同时，还引入了一个额外的 GAN 来对脸部进行精细化微调，以得到更为真实的手语动作视频。

5　数据集及评测指标

5.1　数据集介绍

随着手语研究技术的不断发展，对大规模、多语种手语数据集的需求也不断增加。目前，手语研究已涉及德国[66]、中国[34]、美国[67]、希腊[68]、波兰[69]、阿拉伯[70]、意大利[71]、韩国[72]、阿根廷[73]等近 30 个国家和地区的手语语言。常用的手语数据集如表 2 所列，包括手语单词识别(类型为“词汇”)以及手语句子翻译(类型为“语句”)。从数据规模来看，中小规模词汇的手语数据集总体占比较大；从数据特点来看，手语数据集主要包括 RGB 数据、深度数据、骨架数据、光流数据和黑白图像等。由于 RGB 数据采集的便利性，RGB 数据在不同规模词汇的研究任务中的使用非常普及，而深度数据、骨架数据等常常被作为 RGB 数据的辅助信息。数据集采集方式包括 RGB-D 视觉信息采集和传感器采集。前者有单目摄像头[74]、多目摄像头[75-76]和深度摄像头[41,77](如 Kinect、飞跃运动控制器等)，骨架坐标信息通过一些可穿戴的设备与用户建立物理连接，从而获取手部姿态和体态的位置、动作、轨迹等运动信息。此类设备包括数据手套[78]、表面肌电仪[79-80]和加速度传感器[81](如搭载微传感器的手环、臂环等设备[82])等。相比之下，RGB 视觉信息采集较为便利，深度数据和传感器采集需要更高性能的硬件设备和采集场地的支持。

离散手语数据集的语种分布比较广泛，但规模差异明显，其中代表性数据集包括中文词汇数据集 USTC-CSL500[83]和 DEVISIGN[84]，以及英文单词数据集 RWTH-Boston-50[67]。上述数据集均包含多源数据，例如 USTC-CSL500 和 DEVISIGN 包含 RGB、深度、骨架 3 种模态的数据，RWTH-Boston-50 则带有多视角的 RGB 和黑白图像。这些多源输入数据能够相互补充以优化手语词汇识别效果，也适用于探究不同环境条件(模态或视角差异)下的手语识别。目前公开可用的大规模连续手语的代表性数据集为德国天气手语数据集 RWTH-Phoenix-2014[12,43,56,66]和中科大中国手语数据集 USTC-CSL[14,34,42,92]，这两个数据集也成为连续手语翻译研究发展的主要推动力。另外，在划分训练集和测试集时，一些数据集有意地使两个子集中包含的语句不重合，以此来测试模型在翻译未见过的手语语句时的性能。在表 2 所列的数据集中，仅有 RWTH-Phoenix-2014 来源于天气类电视节目，其他数据集均是在实验室环境下采集的。对于真实场景的手语应用研究来说，更切合生活实际的手语数据收集也是未来需要完善的方向之一。

表 2 常用的手语数据集

Table 2 Commonly used sign language datasets

数据集类型	数据集名称	语种	词汇量	数据特点	测试类别是否见过	是否来自真实场景
词汇	Montalbano V2[71]	Italian	20	RGB-D/骨架	是	否
	PSL Kinect 30[69]	Polish	30	RGB-D/骨架	是	否
	Purdue RVL-SLLL[85]	English	39	RGB	是	否
	DGS Kinect 40[86]	German	40	RGB-D	是	否
	RWTH-Boston-50[67]	English	50	RGB/黑白图	是	否
	LSA64[73]	Argentine	64	RGB	是	否
	PSL ToF[69]	Polish	84	RGB-D	是	否
	GSL isol.[68]	Greek	310	RGB-D	是	否
	USTC-CSL500[83]	Chinese	500	RGB-D/骨架	是	否
	KSL[72]	Korean	1 229	RGB/光流	是	否
	WLASL2000[87]	English	2 000	RGB	是	否
	LSE-Sign[88]	Spanish	2 400	RGB	是	否
	DEVISIGN-G/D/L[84]	Chinese	36/500/2 000	RGB-D/骨架	是	否
	ASLLVD[89]	English	3 300	RGB	是	否
语句	RWTH-Boston-104[67]	English	104	RGB/黑白图	否	否
	USTC-CSL100-Split1[34]	Chinese	178	RGB-D/骨架	是	否
	USTC-CSL100-Split2[34]	Chinese	178	RGB-D/骨架	否	否
	GSL SD[68]	Greek	310	RGB-D	否	否
	GSL SI[68]	Greek	310	RGB-D	是	否
	SIGNUM[90]	German	455	RGB	是	否
	RWTH-Phoenix-2014[66]	German	1 231	RGB	否	是
	RWTH-Phoenix-2014T[11]	German	3 000	RGB	否	是
	How2Sign[91]	English	16 000	RGB-D/骨架	否	否

5.2 评测指标

考虑到手语识别、翻译和生成 3 种任务的特性，研究人员结合视频分类、语音识别、机器翻译和视觉生成等领域的评价指标，对手语研究中词汇准确性、句子语义误差和视频自然度等方面的性能进行了评测。

(1)手语识别评测。准确率(Acc)指被正确分类的样本数占所有样本数的比例；召回率(Recall)指实际标签为正的样本被预测为正样本的概率。常使用 Recall@K 表示按照分类得分排序后在前 K 个词汇中出现正确结果的比例，以此衡量手语词汇识别的性能。

(2)手语翻译评测。错词率(Word Error Rate，WER)是目前手语翻译任务中使用得最普遍的评价指标。其通过计算在将翻译语句转换成真实标签语句的过程中进行删除、插入和替换操作的最小次数之和来衡量手语翻译性能，该计算过程表示为：$WER=(DEL+INS+SUB)/Num$。其中，DEL，INS 和 SUB 分别表示删除、插入和替换操作的次数；Num 表示一个语句中的单词个数。在英文语句的识别中，错词率又可分为字符级 WER 和单词级 WER。语义评估指标来源于机器翻译任务，包括评测指标 BLEU，METEOR，ROUGE 和 CIDEr。BLEU 用于分析待测语句和标签语句中共有的 n 元组(即 n-gram)在标签语句中出现的频率。METEOR 使用 WordNet 计算序列匹配、同义词、词根和词缀以及释义之间的匹配关系。ROUGE 通过计算待测语句和标签语句两者中重叠元组的数目和待测语句长度的比值，来评价待测语句的质量。CIDEr 通过 TF-IDF 计算各个 n-gram 的权重来度量待测语句和标签语句的相似性。

(3)手语生成评测。对生成视频的定性判断包括对生成骨架、图像和视频质量进行评价，具体表现为：1)骨架质量评测，基于生成骨架数据，计算其与真实骨架坐标的均方差(Mean Squared Error，MSE)和平均绝对误差(Mean Absolute Error，MAE)；2) 图像质量评测，利用结构相似性(Structural Similarity，SSIM)、图像特征相似性(Fréchet Inception Distance，FID)来学习感知图像块相似度(Learned Perceptual Image Patch Similarity，LPIPS)，对比生成图像和标签图像，利用峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)度量图像的失真率；3) 视频质量的评估，利用视频特征相似性(Fréchet Video Distance，FVD)计算真实视频和生成视频的特征距离。

6 总结与展望

6.1 真实场景数据集创建

创建能够反映真实场景的大规模数据集将极大地促进手语技术的落地与生活化应用。手语研究工作的数据采集、技术方法、实施系统都应服务于实际生活的场景。目前，现有数据集基本上都来源于实验室环境下的人工采集，大规模、公开可用的真实场景手语数据集屈指可数。有必要借助 Web 信息平台或广播电视平台等网络电视资源扩宽更多元化的搜集途径，从而对数据库进行扩充，如搜集带手语标注的新闻联播等视频来源等。另外，搜集离散的手语单词太过于苛刻，应侧重于对日常句子数据库的收集，并利用现有句子样本，采用现有研究中语义对齐、分割与解析的理论算法和模型，对已有句子进行语义切割和视频重组，生成新的样本补入，以达到数据扩增的目的。

6.2 手语知识图谱构建

知识图谱为手语语义提供了辅助信息和拓展关联信息。构建手语实体的知识图谱有助于增强对手语识别与翻译的语

义理解，并且借助通用知识库（如语言知识图谱 WordNet，常识知识图谱 ConceptNet、HowNet，百科知识图谱 DBpedia、Freebase、Wikidata、YAGO 以及 Knowledge Vault 等）以及特定领域知识库，来建立手语知识元和通式知识的关联，从而构建全面的手语知识图谱。构建知识图谱有望实现手语机器翻译中的知识理解，从而推动手语识别与翻译技术的发展。

6.3　跨领域手语知识迁移学习

在现实世界中，虽然可以直接利用的手语训练数据（目标域数据）不足，但通常也有大量跨域的相关数据（源域数据）可以利用，如手语与手势素材、特定应用领域知识-新闻手语与天气预报手语等。如果能消除源域和目标域的分布差异，则可以大大缓解数据缺乏对网络训练的影响。例如，Li 等[93]考虑到新闻手语和离散手语数据具有相似的视觉概念，通过学习领域不变的视觉概念，将带有字幕的新闻手语知识迁移到手语识别数据中以优化手语识别模型。除了基于样本的跨域知识迁移外，基于模型的迁移方法在手语研究中的应用也是可以探究和延伸的方向之一。

6.4　少样本/零样本词汇可扩展研究探索

在实际应用中，手语系统需要应对随时新增的词汇。这就要求模型能够对少样本孤僻词和未见过的新词汇进行合理的近似语义推测，具备进化学习的能力。少样本学习（Few-Shot Learning）的目标是利用模型良好的泛化能力和学习能力来增强对孤僻词的理解；零次学习（Zero-Shot Learning）的目标是识别出测试集中未经训练的类别的样本，这对识别和理解未训练过的手语词汇或语句是有启发的。为了建立训练类别和未见过类别之间的联系，零次学习通常会选择与二者相关的属性作为语义链接。最近提出的一些方法将零次学习引入手势识别任务[94]，但其在手语研究领域的应用依然寥寥无几。少样本以及零样本的手语研究是解决词汇可增加的手语识别问题的有益探索。

6.5　多样丰富的手语应用系统

研发更多样化且具实用价值的手语系统，能够为真实场景下健全人与失语人士的交流以及手语教学等日常应用提供更多的便利。该系统应涵盖手语自动标注、手语翻译和手语生成等多种功能，能够处理文本、语音、视频等多类型输出，从而实现自然场景下的人机交互和在线手语互译。

为了提高生成手语视频系统的实用价值，其应该同时具有语音识别和自然语言翻译功能，实现适用于自然对话场景下的口语到手语视频的直接转换。同样，研发手语标注系统对于训练识别与翻译模型以及为生成手语视频提供输入也是必不可少的。开发标注支持软件将有助于提高手语数据标注的准确性、可靠性，并降低成本。

结束语　本文对手语研究的背景和挑战进行了介绍，并结合当前的技术发展，分别对手语识别、翻译和生成任务上经典的、最新的方法进行了总结分析。此外，本文还从真实大规模数据集的构建、知识图谱、知识迁移、词汇可扩展、在线丰富应用等方面对未来手语研究的发展方向进行了展望，以期为相关领域的研究提供参考和启发。

参 考 文 献

[1] WANG H，CHAI X，CHEN X. A Novel Sign Language Recognition Framework Using Hierarchical Grassmann Covariance Matrix[J]. IEEE Transactions on Multimedia，2019，21(11)：2806-2814.

[2] GUO D，ZHOU W，LI H，et al. Online Early-Late Fusion Based on Adaptive HMM for Sign Language Recognition[J]. ACM Transactions on Multimedia Computing Communications and Applications，2018，14(1)：1-18.

[3] YIN F，CHAI X，CHEN X. Iterative Reference Driven Metric Learning for Signer Independent Isolated Sign Language Recognition[C]// European Conference on Computer Vision. Springer，Cham，2016：434-450.

[4] WANG Q，CHEN X L，WANG C L，et al. A Data-Deficiency-Tolerated Method for Viewpoint Independent Sign Language Recognition[J]. Chinese Journal of Computers，2009，32(5)：953-961.

[5] YUAN T，SAH S，ANANTHANARAYANA T，et al. Large Scale Sign Language Interpretation[C]// IEEE International Conference on Automatic Face & Gesture Recognition. IEEE，2019：1-5.

[6] KUSHWAH M S，SHARMA M，JAIN K，et al. Sign language interpretation using pseudo glove[C]// International Conference on Intelligent Communication，Control and Devices. Singapore：Springer，2017：9-18.

[7] PU J，ZHOU W，LI H. Iterative Alignment Network for Continuous Sign Language Recognition[C]// Computer Vision and Pattern Recognition. 2019：4165-4174.

[8] ZHOU M，NG M，CAI Z，et al. Self-Attention-based Fully-Inception Networks for Continuous Sign Language Recognition[C]// European Conference on Artificial Intelligence. 2020：8.

[9] PU J，ZHOU W，LI H. Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition[C]// International Joint Conference on Artificial Intelligence. 2018：885-891.

[10] ZHOU H，ZHOU W，ZHOU Y，et al. Spatial-Temporal Multi-cue Network for Continuous Sign Language Recognition[C]// AAAI Conference on Artificial Intelligence. 2020：13009-13016.

[11] CIHAN C N，HADFIELD S，KOLLER O，et al. Neural sign language translation[C]// Computer Vision and Pattern Recognition. 2018：7784-7793.

[12] GUO D，TANG S，WANG M. Connectionist Temporal Modeling of Video and Language：A Joint Model for Translation and Sign Labeling[C]// International Joint Conference on Artificial Intelligence. 2019：751-757.

[13] GUO D，ZHOU W，LI H，et al. Hierarchical LSTM for Sign Language Translation[C]// AAAI Conference on Artificial Intelligence. 2018：6845-6852.

[14] GUO D，WANG S，TIAN Q，et al. Dense Temporal Convolution Network for Sign Language Translation[C]// International Joint

Conference on Artificial Intelligence. 2019:744-750.
[15] WANG S,GUO D,ZHOU W,et al. Connectionist Temporal Fusion for Sign Language Translation[C]// ACM International Conference on Multimedia. 2018:1483-1491.
[16] SAGAWA H,TAKEUCHI M. A Teaching System of Japanese Sign Language Using Sign Language Recognition and Generation[C]// ACM International Conference on Multimedia. 2002:137-145.
[17] XIAO Q,QIN M,YIN Y. Skeleton-Based Chinese Sign Language Recognition and Generation for Bidirectional Communication between Deaf and Hearing People[J]. Neural Networks, 2020,125:41-55.
[18] SAUNDERS B,CAMGÖZ N C,BOWDEN R. Adversarial Training for Multi-Channel Sign Language Production[C]// British Machine Vision Conference. 2020:1-15.
[19] STOLL S,CAMGOZ N C,HADFIELD S,et al. Text2Sign:Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks[J]. International Journal of Computer Vision,2020:128(4):891-908.
[20] SAUNDERS B,CAMGOZ N C,BOWDEN R. Progressive Transformers for End-to-end Sign Language Production[C]// European Conference on Computer Vision. 2020:687-705.
[21] STOLL S,HADFIELD S,BOWDEN R. SignSynth:Data-driven Sign Language Video Generation[C]// Assistive Computer Vision and Robotics. 2020:353-370.
[22] KARPOUZIS K,CARIDAKIS G,FOTINEA S E,et al. Educational Resources and Implementation of A Greek Sign Language Synthesis Architecture [J]. Computers & Education, 2007, 49(1):54-74.
[23] ZELINKA J,KANIS J. Neural Sign Language Synthesis:Words Are Our Glosses[C]// IEEE Winter Conference on Applications of Computer Vision. 2020:3395-3403.
[24] ZELINKA J,KANIS J,SALAJKA P. NN-based Czech Sign Language Synthesis[C]// International Conference on Speech and Computer. Springer,Cham,2019:559-568.
[25] ZHENG L,LIANG B. Sign Language Recognition Using Depth Images[C]// International Conference on Control,Automation, Robotics and Vision. 2016:1-6.
[26] OLIVEIRA M,SUTHERLAND A,FAROUK M. Two-stage PCA with Interpolated Data for Hand Shape Recognition in Sign Language[C]// IEEE Applied Imagery Pattern Recognition Workshop. 2016:1-4.
[27] HASSAN M,ASSALEH K,SHANABLEH T. User-dependent Sign Language Recognition Using Motion Detection[C]// International Conference on Computational Science and Computational Intelligence. 2016:852-856.
[28] LIN Y,CHAI X,ZHOU Y,et al. Curve Matching from the View of Manifold for Sign Language Recognition[C]// Asian Conference on Computer Vision. 2014:233-246.
[29] MIAO Y W,LI J Y,LIU J Z,et al. Hand Gesture Recognition Based on Joint Rotation Feature and Fingertip Distance Feature [J]. Chinese Journal of Computers,2020,43(1):78-92.
[30] YIN F,CHAI X,ZHOU Y,et al. Semantics Constrained Dictionary Learning for Signer-Independent Sign Language Recognition[C]// IEEE International Conference on Image Processing. 2015:3310-3314.
[31] PU J,ZHOU W,LI H. Sign Language Recognition with Multimodal Features[C]// Pacific Rim Conference on Multimedia. 2016:252-261.
[32] LI Y,MIAO Q,TIAN K,et al. Large-scale Gesture Recognition with A Fusion of RGB-D Data Based on The C3D Model[C]// International Conference on Pattern Recognition. 2016:25-30.
[33] THANG P Q,THUY N T,LAM H T. The SVM,SimpSVM and RVM on sign language recognition problem[C]// IEEE International Conference on Information Science and Technology. 2017:398-403.
[34] HUANG J,ZHOU W,ZHANG Q,et al. Video-based Sign Language Recognition without Temporal Segmentation[C]// AAAI Conference on Artificial Intelligence. 2018:2257-2264.
[35] AHMED W,CHANDA K,MITRA S. Vision Based Hand Gesture Recognition Using Dynamic Time Warping for Indian Sign Language[C]// IEEE International Conference on Information Science. 2016:120-125.
[36] FANG G L,GAO W,CHEN X L,et al. A Signer-Independent Continuous Sign Language Recognition System Based on SRN/HMM FANG[J]. Journal of Software,2002(11):2169-2175.
[37] TORNAY S,RAZAVI M,DOSS M M. Towards Multilingual Sign Language Recognition[C]// IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2020:6309-6313.
[38] HUANG J,ZHOU W,LI H,et al. Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018,29(9):2822-2832.
[39] 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.
[40] PIGOU L,DIELEMAN S,KINDERMANS P J,et al. Sign Language Recognition Using Convolutional Neural Networks[C]// European Conference on Computer Vision. 2014:572-578.
[41] HUANG J,ZHOU W,LI H,et al. Sign Language Recognition Using 3D Convolutional Neural Networks[C]// International Conference on Multimedia and Expo. 2015:1-6.
[42] LI H,GAO L,HAN R,et al. Key Action and Joint CTC-Attention based Sign Language Recognition[C]// IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE, 2020:2348-2352.
[43] NIU Z,MAK B. Stochastic Fine-grained Labeling of Multi-state Sign Glosses for Continuous Sign Language Recognition[C]// European Conference on Computer Vision. 2020:172-186.
[44] CAMGOZ N C,KOLLER O,HADFIELD S,et al. Multi-channel transformers for multi-articulatory sign language translation [C]// European Conference on Computer Vision. Springer, Cham,2020:301-319.

[45] CAMGOZ N C，KOLLER O，HADFIELD S，et al. Sign Language Transformers：Joint End-to-end Sign Language Recognition and Translation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2020：10023-10033.

[46] ZHANG Z，PU J，ZHUANG L，et al. Continuous Sign Language Recognition via Reinforcement Learning[C]// IEEE International Conference on Image Processing. 2019：285-289.

[47] DE AMORIM C C，MACÊDO D，ZANCHETTIN C. Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition[C]// International Conference on Artificial Neural Networks. Springer，Cham，2019：646-657.

[48] TUNGA A，NUTHALAPATI S V，WACHS J. Pose-based Sign Language Recognition using GCN and BERT[C]// IEEE Winter Conference on Applications of Computer Vision. 2020：31-40.

[49] 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.

[50] KOLLER O，ZARGARAN S，NEY H. Re-sign：Re-aligned End-to-end Sequence Modelling with Deep Recurrent CNN-HMMs[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017：4297-4305.

[51] SONG P，GUO D，XIN H，et al. Parallel Temporal Encoder for Sign Language Translation[C]// IEEE International Conference on Image Processing. IEEE，2019：1915-1919.

[52] YANG Q，PENG J Y. Chinese Sign Language Recognition Method Based on Depth Image Information and SURF-BoW[J]. Pattern Recognition and Artificial Intelligence，2014，27(8)：741-749.

[53] WU D，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.

[54] YUAN Q，WAN J，LIN C，et al. Global and Local Spatial-Attention Network for Isolated Gesture Recognition[C]// Chinese Conference on Biometric Recognition. Springer，Cham，2019：84-93.

[55] CUI R，LIU H，ZHANG C. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training[J]. IEEE Transactions on Multimedia，2019，21(7)：1880-1891.

[56] CUI R，LIU H，ZHANG C. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017：7361-7369.

[57] 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[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016：3793-3802.

[58] YIN F，CHAI X J，ZHOU Y，et al. Weakly Supervised Metric Learning towards Signer Adaptation for Sign Language Recognition[C]// British Machine Vision Association. 2015：1-12.

[59] HAZRA S，SANTRA A. Short-range radar-based gesture recognition system using 3D CNN with triplet loss[J]. IEEE Access，2019，7：125623-125633.

[60] GLAUERT J R W，ELLIOTT R，COX S J，et al. Vanessa - A System for Communication between Deaf and Hearing People[J]. Technology and Disability，2006，18(4)：207-216.

[61] WANG Z Q，GAO W. A Method to Synthesize Chinese Sign Language Based on Virtual Human Technologies[J]. Journal of Software，2002，13(10)：2051-2056.

[62] BROCK H，LAW F，NAKADAI K，et al. Learning Three-dimensional Skeleton Data from Sign Language Video[J]. ACM Transactions on Intelligent Systems and Technology，2020，11(3)：1-24.

[63] CUI R，CAO Z，PAN W，et al. Deep Gesture Video Generation with Learning on Regions of Interest[J]. IEEE Transactions on Multimedia，2019，PP(99)：1-1.

[64] STOLL S，CAMGÖZ N C，HADFIELD S，et al. Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks[C]// British Machine Vision Conference. 2018：1-2.

[65] GIRÓ-I-NIETO X. Can Everybody Sign Now? Exploring Sign Language Video Generation from 2D Poses[C]// Sign Language Recognition，Translation & Production. 2020：1-4.

[66] KOLLER O，FORSTER J，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.

[67] DREUW P，NEIDLE C，ATHITSOS V，et al. Benchmark Databases for Video-based Automatic Sign Language Recognition[C]// International Conference on Language Resources and Evaluation. 2008：1-6.

[68] ADALOGLOU N，CHATZIS T，PAPASTRATIS I，et al. A Comprehensive Study on Sign Language Recognition Methods[J]. arXiv：2007. 12530，2020.

[69] OSZUST M，WYSOCKI M. Polish Sign Language Words Recognition with Kinect[C]// International Conference on Human System Interactions. 2013：219-226.

[70] ALIYU S，MOHANDES M，DERICHE M. Dual LMCs Fusion for Recognition of Isolated Arabic Sign Language Words[C]// International Multi-Conference on Systems，Signals & Devices. 2017：611-614.

[71] ESCALERA S，BARÓ X，GONZALEZ J，et al. ChaLearn Looking at People Challenge 2014：Dataset and Results[C]// European Conference on Computer Vision. 2014：459-473.

[72] YANG S，JUNG S，KANG H，et al. The Korean Sign Language Dataset for Action Recognition[C]// International Conference on Multimedia Modeling. 2020：532-542.

[73] RONCHETTI F，QUIROGA F，ESTREBOU C A，et al. LSA64：An Argentinian Sign Language Dataset[C]// Congreso Argentino de Ciencias de la Computación. 2016：794-803.

[74] RODRIGUEZ M D，AHMED J，SHAH M. Action MACH a Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition[C]// IEEE Conference on Computer Vi-

sion and Pattern Recognition. 2008:1-8.
[75] XU N, LIU A, NIE W, et al. Multi-modal & Multi-view & Interactive Benchmark Dataset for Human Action Recognition [C]// ACM International Conference on Multimedia. 2015: 1195-1198.
[76] WEINLAND D, BOYER E, RONFARD R. Action Recognition from Arbitrary Views Using 3D Exemplars[C]// International Conference on Computer Vision. 2007:1-7.
[77] CHAI X, LIU Z, LI Y, et al. SignInstructor: An Effective Tool for Sign Language Vocabulary Learning [C]// Asian Conference on Pattern Recognition. 2017:900-905.
[78] LIU M T, LEI Y. Chinese Finger Alphabet Flow Recognition System Based on Data Glove[J]. Computer Engineering, 2011, 37(22):168-170, 173.
[79] SAVUR C, SAHIN F. American Sign Language Recognition System by Using Surface EMG Signal[C]// IEEE International Conference on Systems, Man, and Cybernetics. 2017:2872-2877.
[80] ZHUANG Y, LYU B, SHENG X, et al. Towards Chinese Sign Language Recognition Using Surface Electromyography and Accelerometers[C]// International Conference on Mechatronics and Machine Vision in Practice. 2017:1-5.
[81] LIU X, YUAN G, ZHANG Y M, et al. Hand Gesture Recognition Based on Self-adaptive Multi-classifiers Fusion[J]. Computer Science, 2020, 47(7):103-110.
[82] WU J, TIAN Z, SUN L, et al. Real-time American Sign Language Recognition Using Wrist-Worn Motion and Surface EMG Sensors[C]// International Conference on Wearable and Implantable Body Sensor Networks. 2015:1-6.
[83] ZHANG J, ZHOU W, XIE C, et al. Chinese Sign Language Recognition with Adaptive HMM[C]// International Conference on Multimedia and Expo. 2016:1-6.
[84] CHAI X, WANG H, CHEN X. The Devisign Large Vocabulary of Chinese Sign Language Database and Baseline Evaluations [R]. Key Lab of Intelligent Information Processing of CAS, Institute of Computing Technology, Technical Report, 2014.
[85] WILBUR R B, KAK A C. Purdue RVL-SLLL American Sign Language Database[R]. School of Electrical and Computer Engineering, Purdue University, Technical Report, 2006.
[86] COOPER H, ONG E J, PUGEAULT N, et al. Sign Language Recognition Using Sub-units[J]. Journal of Machine Learning Research, 2012, 13(1):2205-2231.
[87] LI D, RODRIGUEZ C, YU X, et al. Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison[C]// The IEEE Winter Conference on Applications of Computer Vision. 2020:1459-1469.
[88] CARREIRAS M, GUTIÉRREZ-SIGUT E, BAQUERO S, et al. Lexical Processing in Spanish Sign Language (LSE)[J]. Journal of Memory and Language, 2008, 58(1):100-122.
[89] NEIDLE C, THANGALI A, SCLAROFF S. Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus[C]// Language Resources and Evaluation Conference Workshop. 2012:1-9.
[90] FORSTER J, SCHMIDT C, HOYOUX T, et al. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus[C]// International Conference on Language Resources and Evaluation. 2012:3785-3789.
[91] DUARTE A C. Cross-modal Neural Sign Language Translation [C]// ACM International Conference on Multimedia. 2019: 1650-1654.
[92] ZHOU H, ZHOU W, LI H. Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition[C]// IEEE International Conference on Multimedia and Expo. 2019:1282-1287.
[93] LI D, YU X, XU C, et al. Transferring cross-domain knowledge for video sign language recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2020:6205-6214.
[94] BILGE Y C, IKIZLER-CINBIS N, CINBIS R G. Zero-shot Sign Language Recognition: Can Textual Data Uncover Sign Languages? [C]// British Machine Vision Conference. 2019:1-4.

GUO Dan, born in 1983, professor. Her main research interests include machine learning, computer vision and multimedia content analysis.