计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2022/16(11)-2415-15
doi: 10.3778/j.issn.1673-9418.2205003

手语识别与翻译综述

闫思伊[1],薛万利[1+],袁甜甜[2]
1. 天津理工大学 计算机科学与工程学院,天津 300384
2. 天津理工大学 聋人工学院,天津 300384
+ 通信作者 E-mail: xuewanli@email.tjut.edu.cn

摘　要:不同于有声语言,手语主要由连续的手势动作构成。手语识别与翻译是促成听障人士与健听人士之间无障碍交流的重要手段。手语识别与翻译研究任务通过对手语视频进行处理分析并以文字形式显示识别结果,是一种典型的多领域交叉研究。近年来,基于深度学习的手语识别与翻译研究获得了长足的进步。为了便于研究者们系统、全面地了解手语识别与翻译研究任务,分别以手语识别和手语翻译两大任务为主线,从三方面展开综述工作:首先,对具备代表性的手语识别和手语翻译研究工作进行分类总结并分析其特点;其次,归纳整理当前常用的不同国别手语识别与翻译研究数据集,分别从孤立词和连续手语语句两个角度进行分类,同时根据手语识别和手语翻译研究任务的差异性,介绍了对应的评价指标体系;最后,从手语视觉特征的有效信息提取、多线索权重分配、手语与自然语言语法对应及手语数据集资源等方面总结了手语识别与翻译研究目前存在的主要挑战。

关键词:手语识别;手语翻译;深度学习;神经网络
文献标志码:A　　**中图分类号:**TP391

Survey of Sign Language Recognition and Translation

YAN Siyi[1], XUE Wanli[1+], YUAN Tiantian[2]
1. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China
2. Technical College for the Deaf, Tianjin University of Technology, Tianjin 300384, China

Abstract: Different from spoken languages, sign language is mainly composed of continuous gestures. Sign language recognition and translation are important means of facilitating barrier-free communication between the hearing-impaired and the hearing person. The sign language recognition and translation research task is a typical multi-domain cross-study by processing and analyzing sign language videos and displaying the recognition results in text form. In recent years, sign language recognition and translation research based on deep learning has made great progress. In order to facilitate researchers to systematically and comprehensively understand the research tasks of sign language recognition and translation, the review work is carried out from the perspectives of sign language recognition and sign language translation. Firstly, the translation research work is classified and summarized and its characteristics are analyzed. Secondly, the common sign language recognition and translation research datasets of different countries are summarized and classified from the perspectives of isolated sign language words and continuous sign language sentences. Based on the difference in research tasks, the corresponding evaluation index system is introduced. Finally, the major challenges of current research on sign language recognition and translation

基金项目:国家自然科学基金(61906135,62020106004,92048301)。
This work was supported by the National Natural Science Foundation of China (61906135, 62020106004, 92048301).
收稿日期:2022-05-05　　**修回日期:**2022-07-20

are summarized from the aspects of effective information extraction of sign language visual features, multi-cue weight assignment, relationship between sign language and natural language grammar, and sign language dataset resources.

Key words: sign language recognition; sign language translation; deep learning; neural network

根据全国第二次残疾人抽样调查，目前我国听障人数接近3 000万，是国内最大数量的残障群体，手语是听障人士交流表达的主要手段。无障碍沟通是广大听障人群打破信息孤岛、进行平等社会交流的重要途径[1]。实现听障人士无障碍沟通的主要需求是健听人士能够知晓听障人的手语表达。随着人工智能技术的发展特别是计算机视觉研究与自然语言处理研究的进步，使得这一需求的实现成为可能。手语识别与翻译研究正是为实现上述需求的具体研究任务。如图1所示，手语识别是指将手语视频中所做手语动作对应的文字注释（Gloss）顺序地识别出来，而手语翻译是指将对应的手语视频直接翻译为健听人交流时所用的自然口语语句。

手语识别和翻译研究主要包括视觉感知和语言理解两部分：基于计算机视觉技术感知手语视频图像对应深层特征；基于自然语言处理理解手语视频对应文本信息。这种基于感知和理解的研究思路，更接近人的思考过程。

当前，对于手语识别与翻译的研究主要集中在手语识别任务。手语识别的目标是将手语视频自动翻译成相应的手语注释。根据所使用的数据集不同，手语识别可以细分为孤立词手语识别和连续手语词识别[2]。

孤立词手语识别是一种细粒度的动作识别，每个视频只对应一个手语的注释[3-8]。孤立词手语识别的主流方案是将整个句子分割成若干手势片段，再进行单独识别[9]。孤立词手语识别主要关注对注释场景的分割，方法上更类似于动作识别研究。为了避免像孤立词手语识别一样，需要大量人力对手语视频中的手语手势进行分割，因此，引入连续手语识别研究。

连续手语识别是指将一个手语视频，在弱监督的情况下（只进行句子级别的标注而非帧级标注），映射为一个注释序列（gloss sequence），且该注释序列中Gloss顺序与视频中对应的手势片段的顺序一致，即符合手语语法的文本序列。相较于孤立词手语识别，连续手语识别不再需要对手语视频中的手势片段进行繁重的人为分割。

基于自然语言发展而来的手语，其目的是快速便捷地利用肢体动作、面部表情等进行交流，因而形成一套独特的语法规则。通常，一段手语视频对应听障人士表达的文本序列会和对应的听障人士理解的自然语言序列存在差异性。为了便于健听人群对手语的理解，需要对手语视频进行翻译研究以得到对应的呈现口语化的自然语言文本序列，这一过程就是手语翻译研究。手语翻译研究的目标是从连续手语视频中提取对应的符合自然语言语法规则的文本表达。因此，手语翻译研究任务需要结合计算机视觉感知和自然语言处理理解。根据不同研究范式，手语翻译框架可分为：手语视频到文本（sign2text, S2T）和手语视频到注释到文本（sign2gloss2text, S2G2T）。S2T是将连续的手语视频直接翻译成口语句子，而S2G2T利用连续手语识别模型从手语视频

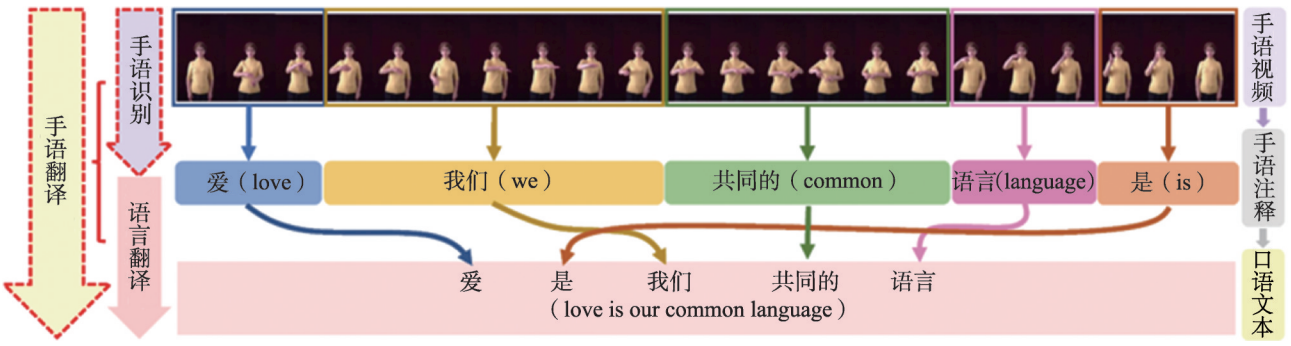

图1 手语识别和翻译流程

Fig.1 Pipeline of sign language recognition and translation

中提取注释序列，然后通过一个预训练的Gloss2Text网络来解决手语序列（sign sequence）到自然语言文本的翻译[10]。

当前，在手语识别与翻译方面的综述，国外具有代表性的工作，如2020年Koller[11]对使用德国手语数据集的相关研究工作进行综述报告，该综述涵盖从1983年至2020年约300项工作，并对其中约25项研究进行了深入分析。但报告仅对RWTH-PHOENIX-WEATHER-2014[12]数据集上的研究工作进行总结，缺乏基于其他数据集的研究工作介绍。国内相关手语识别与研究方面的综述则更多关注手语识别方面[13-15]。为了便于研究者对手语识别与翻译、主流手语数据集及评测指标等方面进行快速全面的了解，本文对当前主流手语识别和翻译研究进行了详细的概括和总结。

1 手语识别和翻译研究工作总结

本章将分别从手语识别研究和手语翻译研究两方面进行相关工作总结。其中，手语识别研究将进一步细分为孤立词手语识别和连续手语识别；手语翻译研究将从手语视频到文本和手语视频到注释到文本两个分支进行简单介绍。

1.1 手语识别研究任务

手语识别框架通常包括视觉特征提取、识别模型两部分。前者用于手语视频的高维特征描述，后者则通过对齐约束提升模型的泛化能力。下面将分别从孤立词手语识别和连续手语识别两方面对当前主流研究方法进行总结。

1.1.1 孤立词手语识别

（1）基于非深度学习的视觉特征的孤立词手语识别

视觉特征提取是手语识别研究的关键。早期的孤立词手语识别研究，在视觉特征提取时以非深度学习的手工特征为主。例如，以手部形状特征作为视觉特征[16]。基于手形的方法可以反映相对简单的手势的含义，但无法应对复杂连贯手语视频下的手语识别任务。

为了解决具有连贯动作的孤立词手语识别，一些研究诸如，尺度不变特征转换（scale-invariant feature transform，SIFT）[17]、方向梯度直方图（histogram of oriented gradient，HOG）[18]、时空关键点（spatial temporal interesting points，STIPs）[19]和内核描述符[20]等二维特征描述子进行视觉特征提取。但特征仅在目标单一且清晰的情况下才能表现出良好的识别性能。为了解决手语视频中的手势遮挡挑战，研究者们提出了3D/4D时间空间特征[21]和随机占用模式特征[22]。进一步，为了解决深度图中存在的噪声和遮挡问题，Miranda等人[23]使用时空占用模式[24]来表征人类手势的四维时空模式，以充分利用空间和时间的背景信息，同时允许类内多样性。Zhang等人[25]提出了一种基于隐马尔可夫模型轨迹建模的孤立词手语识别方法，重点设计了一种新的基于形状上下文的曲线特征描述符。

为了提升孤立词手语识别的鲁棒性，Yin等人[26]设置了包含一组手语引用和相应的距离度量的鲁棒性模型。Zheng等人[27]提出一种基于三维运动图的面向梯度金字塔直方图的描述子来识别人体手势的深度图，该描述子能够在不同空间网格大小下刻画局部信息。

在基于非深度学习的特征的孤立词手语识别研究中，在识别方案部分，通常采用的方法有模板匹配、字典学习、视觉词袋[28-29]、条件随机场[30]、随机森林[31]、支持向量机[32]和隐马尔可夫模型[33]等。支持向量机[34]由于具备较好的预测泛化能力而受到研究者的关注[35]。Pu等人[36]将两种模态的手语视觉特征融合并输入到支持向量机分类器中进行训练。Kumar[37]通过离散小波变换提取手工特征经过处理后采用支持向量机进行分类。隐马尔可夫模型其变体在手语识别研究中同样得到广泛的应用。例如，Guo等人[38]利用隐马尔可夫模型状态自适应方法，建立每个手语词的学习模型。

（2）基于深度学习的视觉特征的孤立词手语识别

由于非深度学习的特征不能很好地适应手语复杂动态的手势及其他关键身体部位的变化，一些研究者采用深度学习的视觉特征进行孤立词手语识别中的视觉特征建模。考虑到长短时记忆网络能够很好地对时间序列的上下文信息进行建模，Liu等人[39]提出了一种端到端的长短时记忆网络孤立词手语识别方案。Hu等人[40]利用深度残差网络（deep residual network，ResNet）提取视觉特征信息，并进行全局与局部增强。Huang等人[2]提出一种基于注意力模型的三维卷积神经网络用于刻画手语视频的时空特征。Wang等人[41]融合二维和一维深度学习模型提取视频帧中的时空特征。Hu等人[42]在手部深度学习的特征模型中引入手部先验信息，提供从语义特征到紧凑手部姿态表示的映射。特别的，Wu等人[43]提出一种

通用的半监督分层动态框架用于手势分割和识别，将骨架特征和深度图像作为输入，利用学习后的隐马尔可夫模型进行推断。

1.1.2 连续手语识别

与孤立词手语识别相比，连续手语识别由于更复杂的手势动作、更长的视频序列表达而更具挑战性。早期的连续手语识别方法，主要基于孤立词手语识别展开研究[44]。例如，部分研究利用视频分割算法，将连续视频序列分割成若干视频片段，然后采用孤立词手语识别方法进行识别并整合[45]。

（1）基于卷积神经网络的连续手语识别

受益于深度神经网络在视频表示学习中的发展，基于深度学习的视觉特征的连续手语识别逐渐成为主流[46]。Wei等人[46]提出了一种基于循环卷积神经网络框架的多尺度感知策略，用于学习手语视频的高维特征表示。针对连续手语识别研究中的弱监督问题，Koller等人[47]通过在迭代算法中嵌入卷积神经网络，利用其更好的描述能力辅助细化帧级标注进而提升模型训练精度。文献[48]则将卷积神经网络嵌入到隐马尔可夫模型框架中。Li等人[49]使用一个去除最后的全连通层的ResNet-152网络来提取任意长度视频的高维视觉特征。Cheng等人[50]提出了一种用于在线手语识别的全卷积网络用于学习视频序列的时空特征。随着三维卷积神经网络在动作识别任务中的广泛应用[51-54]，Zhao等人[55]提出了一种结合光流处理的三维卷积神经网络方法来提升识别精度。Liao等人[56]基于B3D-ResNet执行长期时空特征提取的任务。Yang等人[57]提出了一种结构化特征网络（structured feature network，SF-Net），通过长短时记忆网络与三维卷积神经网络在帧级的组合创建一个有效的时间建模架构。为了更好地对齐视频片段和文本注释，Pu等人[58]引入软动态时间翘曲（soft dynamic time warpping，soft-DTW）算法，提出了一种新的基于3D-ResNet和编码-解码器的网络结构，在soft-DTW的作用下，3D-ResNet特征提取器和编码器-解码器序列建模网络逐步交替优化。

（2）基于循环卷积神经网络的连续手语识别

循环卷积神经网络被广泛应用于处理序列建模问题，如长短期记忆网络（long short-term memory, LSTM）[59]、双向长短期记忆网络（bi-directional long short-term memory, Bi-LSTM）[60]、门控循环单元网络[61]等。在连续手语识别中，通常结合循环卷积神经网络与隐马尔可夫模型，由于隐马尔可夫模型需要计算先验估计，文献[62-64]尝试用连接主义时态分类（connectionist temporal classification, CTC）方法[64-65]把路径选择的问题归纳为最大后验估计问题，通过引入空白类和映射法则模拟了动态规划的过程，从而缓解输入序列和输出序列的对应难的问题。Wang等人[66]提出了一种由时域卷积模块、双向门控循环单元模块和融合层模块组成的混合深度学习结构进行特征的连接融合。Xiao等人[67]将长短期记忆网络与注意力网络融合进行连续手语识别。

（3）基于Transformer的连续手语识别

Transformer[68]作为一种领先的深层级网络特征提取模型被广泛应用于自然语言处理、计算机视觉和语音处理等领域。在连续手语识别研究中，Tunga等人[69]利用图卷积网络对手语演示者身体部位的关键点之间的空间关系进行编码，进而挖掘帧间的时间依赖关系。Niu等人[70]使用二维卷积神经网络提取视频序列的空间特征，用Transformer编码器来提取时序特征。Varol等人[71]利用预训练的I3D模型通过滑动窗口提取时空视觉特征。然后训练一个2层Transformer模型进行手语识别。Zhang等人[72]将Transformer与强化学习相结合进行连续手语识别。Yin等人[73]提出了一种基于编码器-解码器架构的轻量级手语翻译模型SF-Transformer用来识别手语。

（4）基于多线索协同的连续手语识别

手语在传递信息、表达思想时，通常以手势动作配合脸部表情及身体姿势进行综合表达。因此，可以简单地认为，在手语识别研究中，其所表达的含义可以由多种线索共同作用[74]。Zhou等人[74]结合基于视频的手语理解与多线索学习，提出一种时空多线索网络来解决基于视觉的序列学习问题，其中空间多线索通过姿态估计分支学习不同线索的空间表示，时间多线索则分别从线索内及线索外两个角度对时间相关性进行建模获得线索间的协作关系。

在手语识别研究中，一般将不同的信息来源定义为不同模态，例如图像特征、文本特征和利用图卷积网络（graph convolutional networks, GCN）提取的骨架特征就是不同种模态。从多模态中学习各个模态的信息，并且实现各个模态的信息的交流和转换。Papastratis等人[75]提出利用文本信息来改进视觉特征进行连续手语的跨模态学习，模型最初使用两个强大的编码网络来生成视频和文本的特征，再将它们映射和对齐到联合潜在特征中，最后使用联合训练的解码器对处理后的视频特征进行分类。Gao

等人[76]设计了一种视频序列特征和语言特征多模态融合的手语识别系统。Huang等人[5]基于双流结构从视频中提取时空特征，其中高层流用于提取全局的信息，低层流更关注局部的手势。

1.2 手语翻译研究任务

手语作为一门特殊的语言体系，拥有一套区别于其他语言的语法规则。为了让健听人士能够高效、准确地理解听障人士演示的手语，则需要利用手语翻译研究将手语视频翻译成口语化的句子。

1.2.1 手语视频到文本的手语翻译

手语翻译的目的是从执行连续手语的人的视频中提取等效的口语句子。因此，一种研究方案是直接将手语视频翻译成文本，即S2T[77]。Camgoz等人[77]提出Sign2Text模型，使用基于注意力的编码器-解码器模型来学习如何从空间表征或手语注释中进行翻译。Guo等人[78]建立了一种面向手语翻译的高级视觉语义嵌入模型。Li等人[79]提出一种考虑多粒度的时间标识视频片段表示方法，减轻了对精确视频分割的需求。虽然Sign2Text结构简化了手语翻译模型，但容易出现模型的长期依赖问题。且受制于当前技术及数据的制约，当前手语视频到文本的翻译在没有任何明确的中间监督的情况下很难获得较好的效果。考虑到手语注释的数量远低于其所代表的视频帧的数量，另一些研究者开始引入手语注释作为中间标记，设计了手语到注释到文本的手语翻译（S2G2T）。

1.2.2 手语视频到注释到文本的手语翻译

在基于手语视频到注释到文本的手语翻译范式中，手语翻译过程被分为两个阶段[80]：第一阶段将手语识别视为一个中间标记化组件，该组件从视频中提取手语注释；第二阶段是语言翻译任务，将手语注释映射为口语文本。

在Sign2Gloss2Text的手语翻译研究中，典型的工作包括：受手语翻译数据集规模限制，Chen等人[80]将手语翻译过程分解为视觉任务和语言任务，提出一种视觉-语言映射器来连接两者，这种解耦使得视觉网络和语言网络在联合训练前能进行独立的预训练。Camgoz等人[81]通过在手语翻译中利用Transformer融合手工和非手工特征进行手语翻译。Fang等人[82]将手语翻译模型嵌入可穿戴设备。Yin等人[83]基于文献[74]将预训练的词表达嵌入至解码器用于手语翻译。Zhou等人[84]使用文本到注释翻译模型将大量的口语文本整合到手语翻译训练中。Camgoz等人[10]将手语识别和口语翻译的任务整合成一个统一的网络结构进行联合优化。为了实现实时手语翻译，Yin等人[85]基于Transformer设计了一个端到端的手语同步翻译模型，并且提出一种新的重编码方法来增强编码器的特征表达能力。

基于手语视频到注释到文本的手语翻译是目前使用较多的手语翻译范式。但是，一方面手语注释是语言模态的离散表示，若注释遗漏、误译部分信息，很大程度上会影响翻译结果；另一方面，如何确保两个阶段在翻译过程中的高效配合也是手语翻译的难点之一。

2 数据集与评价指标介绍

2.1 数据集介绍

2.1.1 数据采集方式简介

早期手语数据采集主要使用手部建模设备，如数据手套等，来进行数据收集。利用手语演示者的手型、手部运动的轨迹和手部的三维空间位置信息来描述手势变化的过程。Gao等人[86]利用数据手套将采集到的手势数据输入到特征提取模块，模块输出的特征向量输入到快速匹配模块生成候选单词列表。

然而，手部建模设备不仅价格昂贵并且不易携带，因此一些研究人员开始简化或消除设备上复杂传感器，并在不同的设备部位使用不同的颜色标记进行数据采集。如Iwai等人[87]利用颜色手套获取手部实时位置和形状。但是，使用颜色手套进行数据采集时对手语演示者的着装、环境等要求较高，否则容易引起数据偏差。

为了更好地方便手语者演示手语，一些研究者通过采用非接触式传感设备来获取手部的运动轨迹信息。如文献[88]使用RealSense技术将手掌方向和手指关节的数据作为识别模型的输入。但是，手语是一种结合手势变换、脸部表情、身体姿态等多因素综合作用的语言体系，仅只关注手部信息是不够的。因此，研究者们开始转向基于视觉特征的手语识别与翻译研究。

在采用视觉特征的手语识别与翻译研究方法中，由摄像机得到手语演示者的彩色图像并做相应的图像处理，将其用作手语识别模拟的输入数据。不仅如此，一些其他模态的手语信息也受到关注[89]，例如体感摄像机，以便同时获取视觉图像信息、深度信息、骨架信息等。总的来说，相较于基于非视觉的采集方式而言，基于视觉的采集方式，具备成本低、

采集方便、设备依赖度低等优势，同时在特征处理、算法模型上更具挑战性。

2.1.2 公共数据集简要分析

手语数据集可以大致分为孤立词手语数据集和连续手语数据集。孤立词手语数据集主要用于孤立词手语识别研究，由较短的手语单词视频构成。而连续手语数据集主要用于连续手语识别与手语翻译研究任务，由较长的手语句子视频组成。表1列举了部分手语数据集。

其中，目前使用较多的公共手语数据主要包括：RWTH-PHOENIX-WEATHER-2014[12]数据集、RWTH-PHOENIX-WEATHER-2014-T[76]数据集、USTC-CCSL[5]数据集和CSL-Daily[83]数据集。

RWTH-PHOENIX-WEATHER-2014[12]是用于连续手语识别的德国手语数据集，其素材来源于9位手语主持人播报的天气预报视频。数据集的训练集、验证集和测试集分别包含5 672、540和629个数据样本。

RWTH-PHOENIX-WEATHER-2014-T数据集[76]可以同时用于手语翻译任务和识别任务，该数据集同样来自于德国手语的天气播报。数据集的训练

表1 手语数据集总结

Table 1 Summary of sign language datasets

数据集	语言	类型	属性		
			包含手语词个数	视频数	手语演示者个数
Boston ASLLVD[90]	美国手语	孤立词	2 742	9 794	6
AUSLAN[91]	澳大利亚手语	孤立词	—	1 100	100
BSL Corpus[92]	英国手语	孤立词	5 000	—	249
PSL Kinect30[93]	波兰手语	孤立词	30	300	1
DEVISIGN-G/D/L[94]	中国手语	孤立词	36/500/2 000	24 000	8
LSE-sign[95]	西班牙手语	孤立词	2 400	2 400	2
LSA64[96]	阿根廷手语	孤立词	64	3 200	10
USTC-ICSL[27]	中国手语	孤立词	500	125 000	50
DISFA[97]	英国手语	孤立词	—	130 000	27
SMILE[98]	德国手语	孤立词	—	—	30
MS-ASL[99]	美国手语	孤立词	1 000	25 513	222
WLASL[100]	美国手语	孤立词	2 000	21 083	119
BosphorusSign22k[101]	土耳其手语	孤立词	744	22 542	6
AUTSL[102]	土耳其手语	孤立词	226	38 336	43
CSSL5000[50]	中国手语	孤立词	1 000	100 000	10
BSL-1K[103]	英国手语	孤立词	1 064	1 412	40
INCLUDE[104]	印度手语	孤立词	263	4 287	7
NMFs-CSL[39]	中国手语	孤立词	1 067	32 010	10
WLASL-LEX[105]	美国手语	孤立词	800	10 017	3
Boston-104[106]	美国手语	连续手语语句	128	214	3
SIGNUM[107]	德国手语	孤立词+连续手语语句	455	1 230	25
S-pot[108]	芬兰手语	连续手语语句	—	5 539	5
RWTH-PHOENIX-WEATHER-2014[12]	德国手语	连续手语语句	1 081	6 841	9
USTC-CCSL[5]	中国手语	连续手语语句	178	25 000	50
RWTH-PHOENIX-WEATHER-2014-T[77]	德国手语	连续手语语句	1 066	8 257	9
RCSD[44]	中国手语	连续手语语句	242	—	10
KETI[109]	韩国手语	连续手语语句	524	14 672	14
GSL[110]	希腊手语	孤立词+连续手语语句	310	10 290	7
MEDIAPI-SKEL corpus[111]	法国手语	连续手语语句	14 383	368	>100
How2Sign[112]	美国手语	连续手语语句	16 000	2 500	11
CSL-Daily[83]	中国手语	连续手语语句	2 000	20 654	10

集、验证集和测试集分别包含7 096、519和642个样本。与RWTH-PHOENIX-WEATHER-2014数据集类似，同样拥有9个手语演示者。

USTC-CCSL数据集[5]是目前使用最广的中国手语[113]数据集，该数据集包含约25 000段已标记的手语视频，由50名手语演示者进行手语演示。数据集的训练集、验证集和测试集分别包含约17 000、2 000及6 000个样本。特别的，该数据集采用Kinect摄像机[114]采集数据，可提供RGB视觉信息、深度信息及骨架信息。

CSL-Daily数据集[83]可用于连续手语识别及翻译任务，相较于USTC-CCSL，CSL-Daily更侧重于日常生活场景，包括家庭生活、医疗保健和学校生活等多个主题。CSL-Daily的训练、验证和测试集分别包含18 401、1 077和1 176段视频样本。

在连续手语语句数据集中，一部分数据集有注释与正常口语语序的文本对照，可用作手语翻译，主要包括Boston-104[106]、RWTH-PHOENIX-WEATHER-2014-T[77]、KETI[109]、GSL[110]、MEDIAPI-SKEL corpus[111]和CSL-Daily[83]数据集。

2.2 评价指标介绍

对于孤立词手语识别，常采用准确率和召回率进行评价[115]。准确率（Acc）又叫查准率，表示在所有的样本数中得到正确分类的样本数所占据的比例。通常采用Top-1准确率和Top-5准确率进行评价。前者用于预测结果中取最大的概率向量，若正确则分类结果正确，反之则错误；后者预测结果中取最大的前五个概率向量评判正确性，若五个全部预测错误时则预测分类结果错误，反之则正确。召回率（Recall）又叫查全率，表示的是样本中的正例有多少被预测正确。

对于连续手语识别，常采用误字率和准确率。误字率（word error rate，WER）[116]作为手语识别研究中衡量两句之间相似度的指标。其是指将已识别句子转换为相应参考句子所进行的替换、插入和删除操作的最小总和。

$$WER = \frac{S + I + D}{N} \tag{1}$$

其中，S、I和D表示将假设句转换为标注序列所需的替换、插入和删除操作的最小数量。N是标注序列的单词数。一些文章中使用准确率表示手语识别的性能，具体公式为：

$$ACC = 1 - WER = 1 - \frac{S + I + D}{N} \tag{2}$$

对于手语翻译，评价体系参考自然语言翻译研究，包括评价指标：BLEU（bilingual evaluation understudy）[117]、CIDEr（consensus-based image description evaluation）[118]、ROUGE（recall-oriented understanding for gisting evaluation）[119]和METEOR[120]。BLEU得分是手语翻译常用的评估指标。假设一个文本由机器和人工各翻译一次，BLEU的值为n个连续的单词序列（n-gram）同时出现在机器翻译和人工翻译中的比例。根据n-gram可以划分成多种评价指标，如BLEU-1、BLEU-2、BLEU-3、BLEU-4。CIDEr是BLEU和向量空间模型的结合。通过计算其TF-IDF向量[121]的余弦夹角，得到各个n-gram的权重来度量得到候选句子和参考句子的相似度。ROUGE是通过统计系统生成的机器翻译与人工生成的标准翻译之间重叠的基本单元（n元语法、词序列和词对）的数目，来评价翻译的质量。与BLEU得分不同，METEOR考虑到了语言的变化性。METEOR不仅双向比较了机器翻译和人工翻译，而且还考虑到了语言语法等因素。例如在英语中，ride或riding在BLEU方法中算作不同的词，在METEOR中由于词根相同，两者算作同一个单词。

3 手语识别与翻译研究面临的挑战

3.1 手语视频帧有效信息获取

首先，手语视频冗余性会导致关键帧提取困难。手语视频普遍较长，并且有的视频会有大量空白帧，有的任务背景过于复杂，系统在识别提取关键手势时会遭遇困难。其次，针对连续手语识别，其本质上是一种弱监督的学习任务[122]。连续手语视频中语义边界是未知的，由于手语词汇丰富，许多术语都有非常相似的手势和动作。而且，因为不同的人有不同的动作速度，同样的手语注释可能有不同的长度。如何精确分割每个手势是困难所在。如果对视频进行时间分割时出现错误，会不可避免地将错误传播到后续步骤中，从而影响结果的准确度。这些因素都会给手语视频帧处理及特征提取带来挑战。表2列举了一些近年来代表性的在手语视频特征处理上的研究工作。

3.2 多线索权重分配

为了有效地进行手语识别与翻译，需要从不同线索进行融合共同指导模型预测，因此如何综合利用这些线索进行多角度的手语特征表达也是难点之一。首先，简单的特征融合组合不一定比单个特征

表2 特征提取代表性工作

Table 2 Representative work of feature extraction

特征提取网络	参考文献号
GCN	[132][144][145]
GooleNet	[3][48][62][81][122][124][125][128][129][130][131][133]
VGG-Net	[46][72][83][125][126][129][130]
ResNet	[39][48][57][69]
CaffeNet	[63]
AlexNet	[128]
其他2D-CNN	[10][49][76][123][134]
C3D	[5][65][127][138]
B3D	[55]
I3D	[39][79][103][136][137] [141][142][143]
S3D	[80]
其他3D-CNN	[46][58][64][73][135][139][140][144]

更好。其次,对于多线索而言,自适应地为不同线索设置模型参数并非易事,每个模型中所涉及到的关键动作的变化都可能会对参数造成影响。针对多线索融合问题,需要关注的重点是选择哪些线索以及如何融合这些线索。一种可行的方案是通过大量的对比实验,找出最优的特征融合方式,例如设置线索优先级、动态分配各个线索权重、设置多步融合模块等。

3.3 手语语法和自然语言语法的对应

根据手语语言学研究,通常一些国家的手语可分为自然手语和规约手语(或称手势手语)。以中国手语举例,中国手语可以分为自然手语和手势汉语。自然手语主要由听障人士使用,具备一套体系化的语法规则,而手势汉语是一种在口语语法的基础上直接进行手势演练操作的人工语言,其和汉字具有一一对应的关系,因此又称书面手语。如何将自然手语和规约手语进行映射是手语翻译研究的挑战之一。现有手语翻译研究大多是在连续手语识别的基础上,结合语言模型得到符合口语化描述的自然语言翻译。未来可以考虑构建大型的文本对数据集,即自然手语注释集及对应的规约手语注释集,将语言模型在文本对数据集上先进行预训练,然后迁移至手语翻译的语言模型中。

3.4 数据集资源

相对于手语识别及翻译研究模型所需的数据规模而言,目前的手语数据集还远不能满足模型需求,而基于数据驱动的识别及翻译方案,很容易导致神经网络过拟合。且大部分数据是在实验室环境下拍摄收集,而在现实场景中,存在背景多变、阴影、遮挡等众多干扰,这更容易导致模型无法较好地捕捉到手部、脸部及肢体等部位的变化,从而影响识别和翻译结果。未来研究者们可以考虑构建规模更大、场景更复杂的通用手语数据集。

4 结束语

手语识别与翻译是一个典型的多领域交叉研究方向,具备重要的研究及社会意义。由于手语的复杂性及当前客观的技术及数据方面的制约,手语识别与翻译研究充满挑战性,尤其是数据量不够造成的模型过拟合问题以及模型过于复杂导致的实时性不够的问题。文章对近年来手语识别与翻译相关研究进行综述,简单介绍了主流方法情况及特点,同时介绍了手语识别与翻译研究所涉及的数据集及评价方式,为研究者快速全面地了解手语识别与翻译研究提供了有效的途径。

参考文献:

[1] 吕会华. 中国手语语言学[M]. 北京: 知识产权出版社有限责任公司, 2019.
LV H H. The linguistics of Chinese sign language[M]. Beijing: Intellectual Property Publishing House, 2019.

[2] HUANG J, ZHOU W G, LI H Q, 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.

[3] KOLLER O, ZARGARAN S, NEY H. Re-sign: re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 3416-3424.

[4] REN Z, YUAN J S, MENG J J, et al. Robust part-based hand gesture recognition using Kinect sensor[J]. IEEE Transactions on Multimedia, 2013, 15(5): 1110-1120.

[5] HUANG J, ZHOU W G, ZHANG Q L, et al. Video-based sign language recognition without temporal segmentation [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018: 2257-2264.

[6] WEN H Y, RAMOS R J, DEY A K. Serendipity: finger gesture recognition using an off-the-shelf smartwatch[C]//Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, May 7-12, 2016. New York: ACM, 2016: 3847-3851.

[7] EKIZ D, KAYA G E, BUĞUR S, et al. Sign sentence recognition with smart watches[C]//Proceedings of the 25th Signal Processing and Communications Applications Conference, Antalya, May 15-18, 2017. Piscataway: IEEE, 2017: 1-4.

[8] HOU J H, LI X Y, ZHU P D, et al. SignSpeaker: a real-time, high-precision smartwatch-based sign language translator[C]//Proceedings of the 25th Annual International Conference on Mobile Computing and Networking, Los Cabos, Oct 21-25, 2019. New York: ACM, 2019: 1-15.

[9] WANG Z B, ZHAO T D, MA J X, et al. Hear sign language: a real-time end-to-end sign language recognition system[J]. IEEE Transactions on Mobile Computing, 2020, 21(7): 2398-2410.

[10] CAMGÖZ N C, KOLLER O, HADFIELD S, et al. Sign language transformers: joint end-to-end sign language recognition and translation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 10023-10033.

[11] KOLLER O. Quantitative survey of the state of the art in sign language recognition[J]. arXiv:2008.09918, 2020.

[12] FORSTER J, SCHMIDT C, KOLLER O, et al. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather[C]//Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, May 26-31, 2014. Paris: European Language Resources Association, 2014: 1911-1916.

[13] 张淑军, 张群, 李辉. 基于深度学习的手语识别综述[J]. 电子与信息学报, 2020, 42(4): 1021-1032.
ZHANG S J, ZHANG Q, LI H. Review of sign language recognition based on deep learning[J]. Journal of Electronics & Information Technology, 2020, 42(4): 1021-1032.

[14] 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[J]. 计算机工程与应用, 2021, 57(18): 1-12.
ABULA M, KUERBAN A, XIE Q N, et al. Review of sign language recognition methods and techniques[J]. Computer Engineering and Applications, 2021, 57(18): 1-12.

[15] 秦梦现. 手语识别研究综述[J]. 软件导刊, 2021, 20(2): 250-252.
QIN M X. A survey of sign language recognition[J]. Software Guide, 2021, 20(2): 250-252.

[16] HU M C, LIU H T, LI J W, et al. Sign language recognition in complex background scene based on adaptive skin color modelling and support vector machine[J]. International Journal of Big Data Intelligence, 2018, 5(1/2): 21.

[17] LOWE D G. Distinctive image features from scale-invariant key points[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[18] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition, San Diego, Jun 20-26, 2005. Washington: IEEE Computer Society, 2005: 886-893.

[19] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies[C]//Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Jun 23-28, 2008. Piscataway: IEEE, 2008: 1-8.

[20] BO L F, LAI K, REN X F, et al. Object recognition with hierarchical kernel descriptors[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, Jun 20-25, 2011. Washington: IEEE Computer Society, 2011: 1729-1736.

[21] OREIFEJ O, LIU Z C. HON4D: histogram of oriented 4D normal for activity recognition from depth sequences[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, Jun 23-28, 2013. Piscataway: IEEE, 2013: 716-723.

[22] WANG J, LIU Z C, CHOROWSKI J, et al. Robust 3D action recognition with random occupancy patterns[C]//LNCS 7573: Proceedings of the 12th European Conference on Computer Vision, Florence, Oct 7-13, 2012. Berlin, Heidelberg: Springer, 2012: 872-885.

[23] MIRANDA L, VIEIRA T, MORERA D M, et al. Real-time gesture recognition from depth data through key poses learning and decision forests[C]//Proceedings of the 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images, Ouro Preto, Aug 22-25, 2012. Washington: IEEE Computer Society, 2012: 268-275.

[24] VIEIRA A W, NASCIMENTO E R, OLIVEIRA G L, et al. STOP: space-time occupancy patterns for 3D action recognition from depth map sequences[C]//LNCS 7441: Proceedings of the 17th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Buenos Aires, Sep 3-6, 2012. Berlin, Heidelberg: Springer, 2012: 252-259.

[25] ZHANG J H, ZHOU W G, XIE C, et al. Chinese sign language recognition with adaptive HMM[C]//Proceedings of the 2016 IEEE International Conference on Multimedia and Expo, Seattle, Jul 11-15, 2016. Washington: IEEE Computer Society, 2016: 1-6.

[26] YIN F, CHAI X J, CHEN X L. Iterative reference driven metric learning for signer independent isolated sign language recognition[C]//LNCS 9911: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Oct 11-14, 2016: 434-450.

[27] ZHENG L H, LIANG B. Sign language recognition using

depth images[C]//Proceedings of the14th International Conference on Control, Automation, Robotics and Vision, Phuket, Nov 13-15, 2016. Piscataway: IEEE, 2016: 1-6.

[28] LI F F, PERONA P. A Bayesian hierarchical model for learning natural scene categories[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, Jun 20-25, 2005. Piscataway: IEEE, 2005: 524-531.

[29] CHEN G, GIULIANI M, CLARKE D, et al. Action recognition using ensemble weighted multi-instance learning[C]// Proceedings of the 2014 IEEE International Conference on Robotics and Automation, Hong Kong, China, May 31-Jun 7, 2014. Piscataway: IEEE, 2014: 4520-4525.

[30] YANG R D, SARKAR S. Detecting coarticulation in sign language using conditional random fields[C]//Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, Aug 20-24, 2006. Washington: IEEE Computer Society, 2006: 108-112.

[31] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[32] AGARWAL A, THAKUR M K. Sign language recognition using Microsoft Kinect[C]//Proceedings of the 6th International Conference on Contemporary Computing, Noida, Aug 8-10, 2013. Piscataway: IEEE, 2013: 181-185.

[33] WANG H J, CHAI X J, ZHOU Y, et al. Fast sign language recognition benefited from low rank approximation[C]//Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition , Ljubljana, May 4-8, 2015. Washington: IEEE Computer Society, 2015: 1-6.

[34] BURGESCJ C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.

[35] LIU F T, WANG Y T, MA H P. Gesture recognition with wearable 9-axis sensors[C]//Proceedings of the 2017 IEEE International Conference on Communications, Paris, May 21-25, 2017. Piscataway: IEEE, 2017: 1-6.

[36] PU J F, ZHOU W G, LI H Q. Sign language recognition with multi-modal features[C]//LNCS 9917: Proceedings of the 17th Pacific-Rim Conference on Multimedia, Xi'an, Sep 15-16, 2016. Cham: Springer, 2016: 252-261.

[37] KUMAR N. Sign language recognition for hearing impaired people based on hands symbols classification[C]//Proceedings of the 2017 International Conference on Computing, Communication and Automation, Greater Noida, May 5-6, 2017. Piscataway: IEEE, 2017: 244-249.

[38] GUO D, ZHOU W G, LI H Q, et al. Online early-late fusion based on adaptive HMM for sign language recognition [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2017, 14(1): 1-18.

[39] LIU T, ZHOU W G, LI H Q. Sign language recognition with long short-term memory[C]//Proceedings of the 2016 IEEE International Conference on Image Processing, Phoenix, Sep 25-28, 2016. Piscataway: IEEE, 2016: 2871-2875.

[40] HU H Z, ZHOU W G, LI H Q. Hand-model-aware sign language recognition[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 33rd Conference on Innovative Applications of Artificial Intelligence, the 11th Symposium on Educational Advances in Artificial Intelligence. Menlo Park: AAAI, 2021: 1558-1566.

[41] WANG F, DU Y X, WANG G R, et al. (2+1) D-SLR: an efficient network for video sign language recognition[J]. Neural Computing and Applications, 2022, 34 (3): 2413-2423.

[42] HU H, ZHOU W, LI H. Hand-model-aware sign language recognition[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 33rd Conference on Innovative Applications of Artificial Intelligence, the 11th Symposium on Educational Advances in Artificial Intelligence. Menlo Park: AAAI, 2021: 1558-1566.

[43] 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.

[44] WANG H J, CHAI X J, CHEN X L. A novel sign language recognition framework using hierarchical Grossmann covariance matrix[J]. IEEE Transactions on Multimedia, 2019, 21(11): 2806-2814.

[45] MITTAL A, KUMAR P, ROY P, et al. A modified LSTM model for continuous sign language recognition using leap motion[J]. IEEE Sensors Journal, 2019, 19(16): 7056-7063.

[46] WEI C C, ZHAO J, ZHOU W G, et al. Semantic boundary detection with reinforcement learning for continuous sign language recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(3): 1138-1149.

[47] 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]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 3793-3802.

[48] KOLLER O, ZARGARAN O, NEY H, et al. Deep sign: hybrid CNN-HMM for continuous sign language recognition[C]//Proceedings of the British Machine Vision Conference 2016, York, Sep 19-22, 2016. Durham: BMVA Press, 2016: 1-12.

[49] LI H B, GAO L Q, HAN R Z, et al. Key action and joint

CTC-Attention based sign language recognition[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, May 4-8, 2020. Piscataway: IEEE, 2020: 2348-2352.

[50] CHENG K L, YANG Z Y, CHEN Q F, et al. Fully convolutional networks for continuous sign language recognition [C]//LNCS 12369: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 697-714.

[51] TRAN D, BOURDEV L D, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Dec 7-13, 2015. Washington: IEEE Computer Society, 2015: 4489-4497.

[52] JI S W, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.

[53] CARREIRA J, ZISSERMAN A. Quo Vadis, action recognition? A new model and the kinetics dataset[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 4724-4733.

[54] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-23, 2018. Washington: IEEE Computer Society, 2018: 6546-6555.

[55] ZHAO K, ZHANG K J, ZHAI Y, et al. Real-time sign language recognition based on video stream[C]//Proceedings of the 2020 39th Chinese Control Conference, Shenyang, Jul 27-29, 2020. Piscataway: IEEE, 2021: 158-174.

[56] LIAO Y Q, XIONG P W, MIN W D, et al. Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks[J]. IEEE Access, 2019, 7: 38044-38054.

[57] YANG Z Y, SHI Z M, SHEN X Y, et al. SF-Net: structured-feature network for continuous sign language recognition [J]. arXiv:1908.01341, 2019.

[58] PU J F, ZHOU W G, LI H Q. Iterative alignment network for continuous sign language recognition[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 15-20, 2019. Piscataway: IEEE, 2019: 4165-4174.

[59] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[60] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Bidirectional LSTM networks for improved phoneme classification and recognition[C]//LNCS 3697: Proceedings of the 15th International Conference on Artificial Neural Networks, Warsaw, Sep 11-15, 2005. Berlin, Heidelberg: Springer, 2005: 799-804.

[61] CHO K, VAN MERRIËNBOER B, GÜLÇEHRE Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Oct 25-29, 2014. Stroudsburg: ACL, 2014: 1724-1734.

[62] PU J F, ZHOU W G, HU H Z, et al. Boosting continuous sign language recognition via cross modality augmentation [C]//Proceedings of the 28th ACM International Conference on Multimedia, Seattle, Oct 12-16, 2020. New York: ACM, 2020: 1497-1505.

[63] CAMGÖZ N C, HADFIELD S, KOLLER O, et al. SubUNets: end-to-end hand shape and continuous sign language recognition[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 3075-3084.

[64] PU J F, ZHOU W G, LI H Q. Dilated convolutional network with iterative optimization for continuous sign language recognition[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Jul 13-19, 2018. San Francisco: Morgan Kaufmann, 2018: 885-891.

[65] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine learning, Pittsburgh, Jun 25-29, 2006. New York: ACM, 2006: 369-376.

[66] WANG S, GUO D, ZHOU W G, et al. Connectionist temporal fusion for sign language translation[C]//Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Oct 22-26, 2018. New York: ACM, 2018: 1483-1491.

[67] XIAO Q K, CHANG X, ZHANG X, et al. Multi-information spatial-temporal LSTM fusion continuous sign language neural machine translation[J]. IEEE Access, 2020, 8: 216718-216728.

[68] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30, Long Beach, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 5998-6008.

[69] TUNGA A, NUTHALAPATI S V, WACHS J P. Pose-based sign language recognition using GCN and BERT[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, Jan 5-9, 2021. Piscataway: IEEE, 2021: 31-40.

[70] NIU Z, MAK B. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition [C]//LNCS 12361: Proceedings of the 16th European Con-

ference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 172-186.

[71] VAROL G, MOMENI L, ALBANIE S, et al. Read and attend: temporal localisation in sign language videos[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Nashville, Jun 20-25, 2021. Piscataway: IEEE, 2021: 16857-16866.

[72] ZHANG Z H, PU J F, ZHUANG L S, et al. Continuous sign language recognition via reinforcement learning[C]//Proceedings of the 2019 IEEE International Conference on Image Processing, Taiwan, China, Sep 22-25, 2019. Piscataway: IEEE, 2019: 285-289.

[73] YIN Q F, TAO W Q, LIU X L, et al. Neural sign language translation with SF-transformer[C]//Proceedings of the 6th International Conference on Innovation in Artificial Intelligence, Guangzhou, Mar 4-6, 2022. New York: ACM, 2022: 64-68.

[74] ZHOU H, ZHOU W G, ZHOU Y, et al. Spatial-temporal multi-cue network for sign language recognition and translation [J]. IEEE Transactions on Multimedia, 2022, 24: 768-779.

[75] PAPASTRATIS I, DIMITROPOULOS K, KONSTANTINIDIS D, et al. Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space[J]. IEEE Access, 2020, 8: 91170-91180.

[76] GAO L Q, LI H B, LIU Z, et al. RNN-transducer based Chinese sign language recognition[J]. Neurocomputing, 2021, 434: 45-54.

[77] CAMGÖZ N C, HADFIELD S, KOLLER O, et al. Neural sign language translation[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-23, 2018. Washington: IEEE Computer Society, 2018: 7784-7793.

[78] GUO D, ZHOU W G, LI A Y, et al. Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation[J]. IEEE Transactions on Image Processing, 2019, 29: 1575-1590.

[79] LI D X, XU C C, YU X, et al. TSPNet: hierarchical feature learning via temporal semantic pyramid for sign language translation[C]//Advances in Neural Information Processing Systems 33, 2020: 12034-12045.

[80] CHEN Y T, WEI F Y, SUN X, et al. A simple multi-modality transfer learning baseline for sign language translation[J]. arXiv:2203.04287, 2022.

[81] CAMGÖZ N C, KOLLER O, HADFIELD S, et al. Multi-channel transformers for multi-articulatory sign language translation[C]//LNCS 12538: Proceedings of the 2020 European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 301-319.

[82] FANG B Y, CO J, ZHANG M. DeepASL: enabling ubiquitous and non-intrusive word and sentence-level sign language translation[C]//Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems, Delft, Nov 6-8, 2017. New York: ACM, 2017: 1-13.

[83] YIN K, READ J. Better sign language translation with STMC-transformer[C]//Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Dec 8-13, 2020: 5975-5989.

[84] ZHOU H, ZHOU W G, QI W Z, et al. Improving sign language translation with monolingual data by sign back-translation[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Nashville, Jun 20-25, 2021. Piscataway: IEEE, 2021: 1316-1325.

[85] YIN A X, ZHAO Z, LIU J L, et al. SimulSLT: end-to-end simultaneous sign language translation[C]//Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, Oct 20-24, 2021. New York: ACM, 2021: 4118-4127.

[86] GAO W, MA J Y, WU J Q, et al. Sign language recognition based on HMM/ANN/DP[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2000, 14(5): 587-602.

[87] IWAI Y, WATANABE K, YAGI Y, et al. Gesture recognition by using colored gloves[C]//Proceedings of the 1996 International Conference on Pattern Recognition, Beijing, Oct 14-17, 1996. Piscataway: IEEE, 1996: 76-81.

[88] MISTRY J, INDEN B. An approach to sign language translation using the Intel real sense camera[C]//Proceedings of the 2018 10th Computer Science and Electronic Engineering Conference, Colchester, Sep 19-21, 2018. Piscataway: IEEE, 2018: 219-224.

[89] MEJÍA-PERÉZ K, CÓRDOVA-ESPARZA D M, TERVEN J, et al. Automatic recognition of Mexican sign language using a depth camera and recurrent neural networks[J]. Applied Sciences, 2022, 12(11): 5523.

[90] ATHITSOS V, NEIDLE C, SCLAROFF S, et al. The American sign language lexicon video dataset[C]//Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, Jun 23-28, 2008. Washington: IEEE Computer Society, 2008: 1-8.

[91] JOHNSTON T. From archive to corpus: transcription and annotation in the creation of signed language corpora[J]. International Journal of Corpus Linguistics, 2010, 15(1): 106-131.

[92] SCHEMBRI A, FENLON J, RENTELIS R, et al. Building the British sign language corpus[J]. Language Documentation & Conservation, 2013, 7: 136-154.

[93] OSZUST M, MARIAN W. Polish sign language words recognition with Kinect[C]//Proceedings of the 6th International

Conference on Human System Interactions, Sopot, Jun 6-8, 2013. Piscataway: IEEE, 2013: 219-226.

[94] CHAI X J, WANG H J, CHEN X L. The DEVISIGN large vocabulary of Chinese sign language database and baseline evaluations: VIPL-TR-14-SLR-001[R]. 2014.

[95] GUTIERREZ-SIGUT E, COSTELLO B, BAUS C, et al. LSE-sign: a lexical database for Spanish sign language[J]. Behavior Research Methods, 2016, 48(1): 123-137.

[96] RONCHETTI F, QUIROGA F, ESTREBOU C A, et al. LSA-64: an Argentinian sign language dataset[C]//Proceedings of the XXII Congreso Argentino de Ciencias de la Computación, San Luis, Jun 2, 2016. Argentina: Nueva Editorial Universitario, 2016: 794-803.

[97] MAVADATI S M, SANGER P, MAHOOR M H. Extended DISFA dataset: investigating posed and spontaneous facial expressions[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 26-Jul 1, 2016. Washington: IEEE Computer Society, 2016: 1452-1459.

[98] EBLING S, CAMGÖZ N C, BRAEM P B, et al. SMILE Swiss German sign language dataset[C]//Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, May 7-12, 2018: 1-9.

[99] JOZE H R V, KOLLER O. MS-ASL: a large-scale data set and benchmark for understanding American sign language [C]//Proceedings of the 30th British Machine Vision Conference 2019, Cardiff, Sep 9-12, 2019. Durham: BMVA Press, 2019: 100.

[100] LI D X, OPAZO C R, YU X, et al. Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison[C]//Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, Mar 1-5, 2020. Piscataway: IEEE, 2020: 1448-1458.

[101] ÖZDEMIR O, KINDIROĞLU A A, CAMGÖZ N C, et al. BosphorusSign22k sign language recognition dataset[C]// Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, Marseille, May 11-16, 2020: 181-188.

[102] SINCAN O M, KELES H Y. AUTSL: a large scale multimodal Turkish sign language dataset and baseline methods [J]. IEEE Access, 2020, 8: 181340-181355.

[103] ALBANIE S, VAROL G, MOMENI L, et al. BSL-1K: scaling up co-articulated sign language recognition using mouthing cues[C]//LNCS 12356: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 35-53.

[104] SRIDHAR A, GANESAN R G, KUMAR P, et al. INCLUDE: a large scale dataset for Indian sign language recognition [C]//Proceedings of the 28th ACM International Conference on Multimedia, New York, Oct 12-16, 2020. New York: ACM, 2020: 1366-1375.

[105] TAVELLA F, SCHLEGEL V, ROMEO M, et al. WLASL-LEX: a dataset for recognising phonological properties in american sign language[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, May 22-27, 2022. Stroudsburg: ACL, 2022: 453-463.

[106] DREUW P, FORSTER J, DESELAERS T, et al. Efficient approximations to model-based joint tracking and recognition of continuous sign language[C]//Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, Sep 17-19, 2008. Washington: IEEE Computer Society, 2008: 1-6.

[107] VON A U, KRAISS K F. Towards a video corpus for signer-independent continuous sign language recognition[C]//Proceedings of the 2007 International Conference on Gesture in Human-Computer Interaction and Simulation, Lisbon, 2007.

[108] VIITANIEMI V, JANTUNEN T, SAVOLAINEN L, et al. S-pot—a benchmark in spotting signs within continuous signing[C]//Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, May 26-31, 2014. European Language Resources Association, 2014.

[109] KO S K, KIM C J, JUNG H, et al. Neural sign language translation based on human key point estimation[J]. Applied Sciences, 2019, 9(13): 2683.

[110] ADALOGLOU N M, CHATZIS T, PAPASTRATIS I, et al. A comprehensive study on deep learning-based methods for sign language recognition[J]. IEEE Transactions on Multimedia, 2022, 24: 1750-1762.

[111] BULL H, BRAFFORT A, GOUIFFÈS M. MEDIAPI-SKEL—a 2D-skeleton video database of french sign language with aligned French subtitles[C]//Proceedings of the 12th Conference on Language Resources and Evaluation, Marseille, May 11-16, 2020. Stroudsburg: ACL, 2020: 6063-6068.

[112] DUARTE A C, PALASKAR S, VENTURA L, et al. How2Sign: a large-scale multimodal dataset for continuous American sign language[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2021: 2735-2744.

[113] 中国残疾人联合会教育就业部, 中国聋人协会. 中国手语[M]. 北京: 华夏出版社, 2003.
China Disable Persons' Federation Employment Service and Administration Center, China Association of Persons with Hearing Disabilities. Chinese sign language[M]. Beijing: Huaxia Publishing House, 2003.

[114] ZHANGZ Y. Microsoft Kinect sensor and its effect[J]. IEEE Multimedia, 2012, 19(2): 4-10.

[115] WANG H J, CHAI X J, HONG X P, et al. Isolated sign language recognition with grassmann covariance matrices [J]. ACM Transactions on Accessible Computing, 2016, 8 (4): 1-21.

[116] EIDE E, GISH H, JEANRENAUD P, et al. Understanding and improving speech recognition performance through the use of diagnostic tools[C]//Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, May 9-12, 1995. Washington: IEEE Computer Society, 1995: 221-224.

[117] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Jul 6-12, 2002. Stroudsburg: ACL, 2002: 311-318.

[118] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr-R: consensus-based image description evaluation[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, Jun 7-12, 2015. Washington: IEEE Computer Society, 2015: 4566-4575.

[119] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Proceedings of the 2004 Workshop on Text Summarization Branches Out, Barcelona, Jul 25, 2004. Stroudsburg: ACL, 2004: 74-81.

[120] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Jun 29, 2005. Stroudsburg: ACL, 2005: 65-72.

[121] MARTINEAU J, FININ T. Delta TFIDF: an improved feature space for sentiment analysis[C]//Proceedings of the 3rd International Conference on Weblogs and Social Media, San Jose, May 17-20, 2009. Menlo Park: AAAI, 2009: 258-261.

[122] WEI C C, ZHOU W G, PU J F, et al. Deep grammatical multi-classifier for continuous sign language recognition[C]// Proceedings of the 5th IEEE International Conference on Multimedia Big Data, Singapore, Sep 11-13, 2019. Piscataway: IEEE, 2019: 435-442.

[123] KOLLER O, NEY H, BOWDEN R. Automatic alignment of HamnoSys subunits for continuous sign language recognition[C]//Proceedings of the 10th Edition of Its Language Resources and Evaluation Conference 7th Workshop on the Representation and Processing of Sign Languages, Portorož, May 23-28, 2016. New York: McGraw-Hill, 2016: 121-128.

[124] SARHAN N, LAURI M, FRINTROP S. Multi-phase fine-tuning: a new fine-tuning approach for sign language recognition[J]. KI-Künstliche Intelligent, 2022, 36(1): 91-98.

[125] CUI R P, HU L, ZHANG C S. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 1610-1618.

[126] KONSTANTINIDIS D, DIMITROPOULOS K, DARAS P. A deep learning approach for analyzing video and skeletal features in sign language recognition[C]//Proceedings of the 2018 IEEE International Conference on Imaging Systems and Techniques, Krakow, Oct 16-18, 2018. Piscataway: IEEE, 2018: 1-6.

[127] GUO D, ZHOU W G, LI H Q, et al. Hierarchical LSTM for sign language translation[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018: 6845-6852.

[128] 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.

[129] CUI R P, LIU H, ZHANG C S. A deep neural framework for continuous sign language recognition by iterative training[J]. IEEE Transactions on Multimedia, 2019, 21(7): 1880-1891.

[130] KOLLER O, CAMGOZ N C, NEY H, et al. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(9): 2306-2320.

[131] BORG M, CAMILLERI K P. Phonologically-meaningful subunits for deep learning-based sign language recognition [C]//LNCS 12536: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 199-217.

[132] BULL H, GOUIFFÈS M, BRAFFORT A. Automatic segmentation of sign language into subtitle-units[C]//LNCS 12536: Proceedings of the 16th European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 186-198.

[133] ZHOU M J, NG M, CAI Z X, et al. Self-attention-based fully-inception networks for continuous sign language recognition[C]//Proceedings of the 24th European Conference on

Artificial Intelligence, Santiago de Compostela, Aug 29-Sep 8, 2020. Amsterdam: IOS Press, 2020: 2832-2839.

[134] MIN Y C, HAO A M, CHAI X J, et al. Visual alignment constraint for continuous sign language recognition[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, Oct 10-17, 2021. Piscataway: IEEE, 2021: 11542-11551.

[135] PEI X K, GUO D, ZHAO Y. Continuous sign language recognition based on pseudo-supervised learning[C]//Proceedings of the 2nd Workshop on Multimedia for Accessible Human Computer Interfaces, Nice, Oct 25, 2019. New York: ACM, 2019: 33-39.

[136] RENZ K, STACHE N C, ALBANIE S, et al. Sign language segmentation with temporal convolutional networks[C]// Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Jun 6-11, 2021. Piscataway: IEEE, 2021: 2135-2139.

[137] ZHOU H, ZHOU W G, LI H Q. Dynamic pseudo label decoding for continuous sign language recognition[C]//Proceedings of the 2019 IEEE International Conference on Multimedia and Expo, Shanghai, Jul 8-12, 2019. Piscataway: IEEE, 2019: 1282-1287.

[138] SONG P P, GUO D, XIN H R, et al. Parallel temporal encoder for sign language translation[C]//Proceedings of the 2019 IEEE International Conference on Image Processing, Taiwan, China, Sep 22-25, 2019. Piscataway: IEEE, 2019: 1915-1919.

[139] GUO D, WANG S, TIAN Q, et al. Dense temporal convolution network for sign language translation[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, Aug 10-16, 2019: 744-750.

[140] GÖKÇE Ç, ÖZDEMIR O, KINDIROĞLU A A, et al. Score-level multi cue fusion for sign language recognition[C]// LNCS 12536: Proceedings of the 2020 European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 294-309.

[141] LI D X, YU X, XU C C, et al. Transferring cross-domain knowledge for video sign language recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 6204-6213.

[142] VAROL G, MOMENI L, ALBANIE S, et al. Read and attend: temporal localisation in sign language videos[C]// Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 16857-16866.

[143] MORYOSSEF A, TSOCHANTARIDIS I, DINN J, et al. Evaluating the immediate applicability of pose estimation for sign language recognition[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 3434-3440.

[144] JIANG S Y, SUN B, WANG L C, et al. Skeleton aware multi-modal sign language recognition[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 3413-3423.

[145] LI R M, MENG L. Multi-view spatial-temporal network for continuous sign language recognition[J]. arXiv:2204.08747, 2022.

闫思伊（1996—），女，河南驻马店人，硕士研究生，主要研究方向为手语识别、手语翻译。

YAN Siyi, born in 1996, M.S. candidate. Her research interests include sign language recognition and sign language translation.

薛万利（1986—），男，江苏南京人，博士，CCF会员，主要研究方向为目标跟踪、手语识别。

XUE Wanli, born in 1986, Ph.D., member of CCF. His research interests include target tracking and sign language recognition.

袁甜甜（1980—），女，天津人，博士，教授，CCF会员，主要研究方向为手语识别、计算机网络。

YUAN Tiantian, born in 1980, Ph.D., professor, member of CCF. Her research interests include sign language recognition and computer network.