基于全局注意力机制和 LSTM 的连续手语识别算法

杨观赐[a,b,c]，韩海峰[a]，刘赛赛[b]，蒋亚汶[b]，李杨[b]
（贵州大学 a.机械工程学院 b.现代制造技术教育部重点实验室 c.省部共建公共大数据国家重点实验室，贵阳 550025）

摘要：目的 为提高连续手语识别准确率，缓解听障人群与非听障人群的沟通障碍。**方法** 提出了基于全局注意力机制和 LSTM 的连续手语识别算法。通过帧间差分法对视频数据进行预处理，消除视频冗余帧，借助 ResNet 网络提取特征序列。通过注意力机制加权，获得全局手语状态特征，并利用 LSTM 进行时序分析，形成一种基于全局注意力机制和 LSTM 的连续手语识别算法，实现连续手语识别。**结果** 实验结果表明，该算法在中文连续手语数据集 CSL 上的平均识别率为 90.08%，平均词错误率为 41.2%，与 5 种算法相比，该方法在识别准确率与翻译性能上具有优势。**结论** 基于全局注意力机制和 LSTM 的连续手语识别算法实现了连续手语识别，并且具有较好的识别效果及翻译性能，对促进听障人群无障碍融入社会方面具有积极的意义。

关键词：手语识别；特征提取；全局注意力机制；LSTM
中图分类号：TP18；TB472 **文献标识码：**A **文章编号：**1001-3563(2022)08-0028-07
DOI：10.19554/j.cnki.1001-3563.2022.08.004

Continuous Sign Language Recognition Algorithm Based on Global Attention Mechanism and LSTM

YANG Guan-ci[a,b,c], *HAN Hai-feng*[a], *LIU Sai-sai*[b], *JIANG Ya-wen*[b], *LI Yang*[b]
(a. School of Mechanical Engineering b. Key Laboratory of Advanced Manufacturing Technology of the Ministry of Education c. State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China)

ABSTRACT: To improve the continuous sign language recognition accuracy and alleviate the communication barrier between hearing-impaired people and non hearing-impaired people, this paper proposed the continuous sign language recognition algorithm based on Global Attention Mechanism and LSTM (CSLR-GAML). The video data is preprocessed by applying inter-frame difference to eliminate redundant video frames, and then the feature sequences of the key frames are extracted by using ResNet. After that, the attention mechanism is used to update the network parameters, which is capable of obtain the global feature of sign language, and then the LSTM is employed to finish the timing sequence analysis. Finally, use the Chinese continuous sign language data set CSL to check algorithm performance. And the experimental results show that the average recognition accuracy of the proposed algorithm is 90.08%, and the average word error rate is 41.2%. Compared CSLR-GAML with other Five algorithms, the proposed CSLR-GAML has advantages in recognition accuracy and translation performance. The sontinuous sign language recognition algorithm based on Global Attention Mechanism and LSTM realizes continuous sign language recognition, and has good recognition effect and translation performance. It is of positive significance to promote the barrier free integration of hearing-impaired people into society.
KEY WORDS: sign language recognition; feature extraction; global attention mechanism; LSTM

收稿日期：2021-12-25
基金项目：国家自然科学基金（62163007）；贵州省科技计划项目（黔科合平台人才[2020]6007，黔科合支撑[2021]一般439，JXCX[2021]001）
作者简介：杨观赐（1983—），男，博士，教授，主要研究方向为自主智能系统、多模态数据驱动的认知计算。

据全国残疾人调查情况数据显示：中国的听障人数超过 2 000 万[1]。听障患者因沟通障碍导致难以融入社会工作，给他们带来沉重压力。手语是听障患者与人交流的最常用方式，但是手语比较抽象，学习成本较高，若不经过系统训练，人们与听障患者交流较为困难[2]。随着人工智能技术的发展，越来越多的研究关注手语识别，试图通过手语识别技术将手语翻译成文本或者语音输出，进而缓解人们与听障患者的沟通障碍[3]。

根据手语数据获取方式的不同，手语识别类型可以分为基于传感器的识别和基于视觉的识别[4]。数据手套和臂环等传感器可以捕获佩戴者的手部关节信息和运动轨迹并识别佩戴者的表达意图。文献[5]针对使用者个体差异的问题，使用双线性模型来处理肌电信号，然后使用 LSTM 网络对 20 个手语动作进行识别，表现出了较高的识别精度。为减少环境对传感数据的影响，文献[6]针对九轴惯性传感器特性，利用反馈控制融合姿态计算提高姿态获取的实时性，然后对支持向量机、*K*-近邻法和前馈神经网络分类器进行自适应模型集成进行手语分类，识别率有所提高。文献[7]通过臂环收集手臂的表面肌电信号和惯性传感器数据，经过归一化和滤波处理后，运用滑动窗口分割数据。将单个手语词信号平均分为 *n* 组，并每次取出 *n*–1 组按原顺序组合成新数据，进行多次识别，提高连续手语语句识别率。虽然基于传感器的手语识别精度较高，但依赖于硬件设备，并且需要穿戴传感器，导致用户体验较差。

文献[8]使用上下文信息作为先验知识，使用一个生成器从视频序列中提取时空特征来提高手语语句识别连贯性，并用一个判别器对文本信息进行建模以评估生成器的预测效果。文献[9]改进了融合双流 3 维卷积神经网络模型的相关参数和结构以提高模型的收敛速度和稳定性，使用批量归一化优化网络，在中国手语数据集（CSL）[10-11]识别率达到了 90.76%，但是存在网络层数较深、超参数较多等问题，导致网络的训练和优化较困难。关键帧是包含关键手势和语义变换的视频帧，文献[12]根据关键帧图片特征和日常手语习惯，利用卷积自编码器提取视频帧的深度特征，对其进行 *K*-means 聚类，在每类视频帧中选取最清晰的视频帧作为关键帧；再利用手语动作中关键手势时的停顿，通过点密度筛选出视频关键帧以消除冗余信息构建最优序列，但仍然存在一些关键帧丢失的问题。

为提高连续手语识别准确率，围绕手语的手部形状、位置和方向变化多样性等导致的特征信息丢失问题，文中研究了基于全局注意力机制和 LSTM 的连续手语识别算法（Continuous Sign Language Recognition Algorithm Based on Global Attention Mechanism and LSTM，CSLR-GAML）。

1　基于全局注意力机制和 LSTM 的连续手语识别算法

1.1　CSLR-GAML 算法流程

为快速精确识别视频中的连续手语，此节提出了基于全局注意力机制和 LSTM 的连续手语识别算法，见图 1。首先，使用帧间差分方法提取手语视频的关键帧，进而去除冗余信息；其次，采用深度残差网络逐个对视频关键帧进行特征提取，并将其转换为时序特征序列；再次，利用 LSTM 网络处理经过注意力加

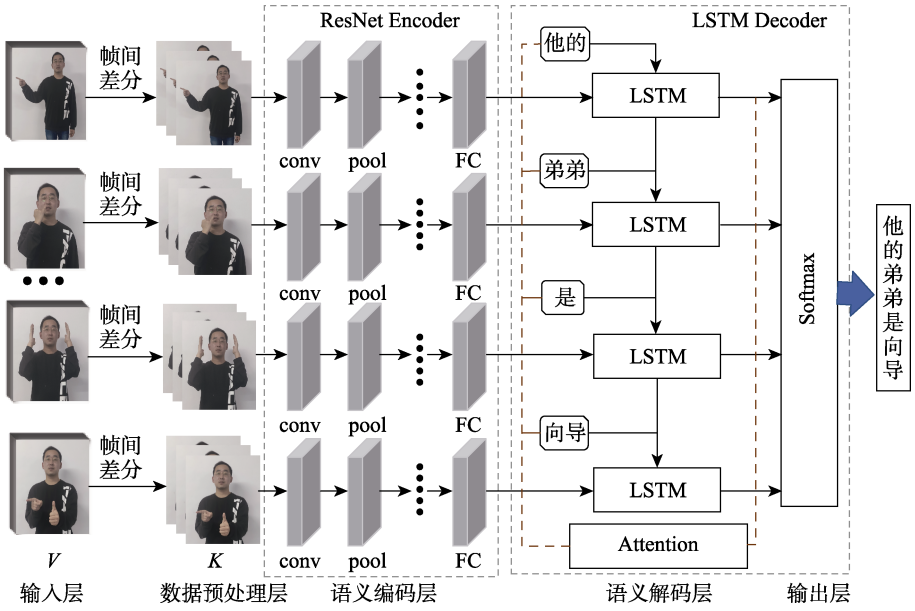

图 1　算法 1 的网络架构

Fig.1 Network architecture of algorithm 1

权的有用信息,学习到上下文关系;最后经过 Softmax 网络输出完整的文本信息以达到手语视频识别的目的。算法 1 详细流程如下所述。

算法 1:基于全局注意力机制和 LSTM 的连续手语识别算法(CSLR-GAML)

输入:Kinect 摄像头采集的时序视频流 V。

输出:手语文本 S。

步骤 1:初始化残差网络、全局注意力机制和 LSTM 网络参数,关键帧序列 K=Ø,时间窗口大小为 8 s,初始时刻 t=0 s。

步骤 2:如果 t 能被 8 整除,则采用基于差分方法的关键帧提取算法(详见 1.2 节)获得 V 的关键帧序列 K。

步骤 3:采用基于 ResNet 的特征提取算法(详见 1.3 节),将关键帧序列 K 输入至残差网络中,经过卷积和池化操作,提取图像的局部特征 $\boldsymbol{f}$,通过全连接层进行拼接,获得关键帧序列特征向量 $\boldsymbol{F}$。

步骤 4:将 $\boldsymbol{F}$ 作为初始状态输入到 LSTM 网络,计算出手语词预测的概率 P。

步骤 5:采用基于全局注意力机制的 LSTM 手语语义信息提取算法(详见 1.4 节),更新特征向量权重 W 输出手语文本 S。

步骤 6:如果摄像头数据为空,则结束,否则返回步骤 2。

1.2 基于差分方法的关键帧提取算法

针对手语视频中存在大量冗余信息的问题,使用差分方法提取关键帧以筛选出视频中的关键信息。通过衡量帧间图像相对平均像素的强度变化来确定关键帧,从而设计基于差分方法的关键帧提取算法。算法 2 详细流程如下所述。

算法 2:基于差分方法的关键帧提取算法

输入:手语视频流 V。

输出:关键帧序列 K。

步骤 1:初始化 K=Ø。

步骤 2:对视频流 V 逐帧进行灰度化处理,经过滤波处理后得到由 q 张图片构成的集合 $P=\{p_0,p_1,\cdots,p_t,\cdots,p_q\}$。

步骤 3:对 P 进行图像二值化操作并求和输出绝对差分值 $I=\{I_0,I_1,\cdots,I_t,\cdots,I_q\}$。

步骤 4:使用平滑方法去除干扰项峰值,获得新的 $I=\{I_0,I_1,\cdots,I_t,\cdots,I_q\}$。之后,计算所有相邻帧间的差分值 $|I_t-I_{t-1}|$ 的算术平均值作为标准差分值 T。

步骤 5:从 t=1 开始到 t=q,将 $|I_t-I_{t-1}|$ 与 T 进行比较,若 $|I_t-I_{t-1}|>T$,则 $K=K\cup p_t$。

步骤 6:输出关键帧序列 K。

需要说明的是,在步骤 2 中,原始图像大小为 1 280×720×3,利用加权平均方法的灰度化图像,然后通过高斯滤波,统一图像大小为 224×224×3。计算过程见式(1)。

$$G(x,y)=\frac{1}{2\pi\sigma^2}\mathrm{e}^{\frac{-x^2+y^2}{2\sigma^2}} \tag{1}$$

式中:x、y 为像素坐标值;σ 为像素标准差。

在步骤 3 中,利用灰度图像计算所有像素值的均值作为阈值,实现图像二值化。

在步骤 4 中,计算标准差分值过程见式(2)。

$$T=\sum_{t=1}^{q}\frac{|I_t-I_{t-1}|}{q} \tag{2}$$

步骤 5 中,通过判定相邻帧间的差分值与 T 的关系,确定关键帧,判定过程见式(3)。

$$P_t=\begin{cases}1, |I_t-I_{t-1}|\geqslant T\\ 0, |I_t-I_{t-1}|<T\end{cases} \tag{3}$$

1.3 基于 ResNet 的特征提取算法

ResNet[13]在获取局部特征方面具有优势。为了获得由于光照及背景变化而丢失的特征,设计了基于 ResNet 的特征提取算法,以获得更多有效的特征信息。其计算过程见图 2,算法 3 详细流程如下所述。

算法 3:基于 ResNet 的特征提取算法

输入:关键帧序列 K。

输出:关键帧序列特征向量 $\boldsymbol{F}$。

步骤 1:初始化网络参数,残差块数量 n=8,卷积核大小为 3,步长为 2,$\boldsymbol{F}$=Ø。

步骤 2:对于 $\forall k_t\in K$:

步骤 2.1:将 k_t 输入第一层网络进行卷积和池化操作,得到特征向量 $\boldsymbol{f}_t^0$。

步骤 2.2:将 $\boldsymbol{f}_t^0$ 输入到第 i 个残差模块(i=1,2,⋯,8),输出残差特征 $\boldsymbol{f}_t^i$。

步骤 2.3:$\boldsymbol{f}_t^{i+1}=\boldsymbol{f}_t^{i-1}+\boldsymbol{f}_t^i$,作为第 i+1 个残差单元的输入。

步骤 2.4:i=i+1,若 i>8,输出特征向量 $\boldsymbol{f}_t$,否则跳到步骤 2.3。

步骤 2.5:将 $\boldsymbol{f}_t$ 输入到全局平均池化层生成 1×1×512 维的特征向量 $\boldsymbol{f}_t$。

步骤 2.6:$\boldsymbol{F}=\boldsymbol{F}\cup\boldsymbol{f}_t$;$t$=$t$+1。

步骤 2.7:若 t>m,转到步骤 3,否则,转到步骤 2.1。

步骤 3:$\boldsymbol{F}=\{\boldsymbol{f}_0,\boldsymbol{f}_1,\cdots,\boldsymbol{f}_t,\cdots,\boldsymbol{f}_m\}$。

步骤 2.1 中输入图像维度为 224×224×3,卷积后得到的特征图大小为 112×112×64,再经过池化后得

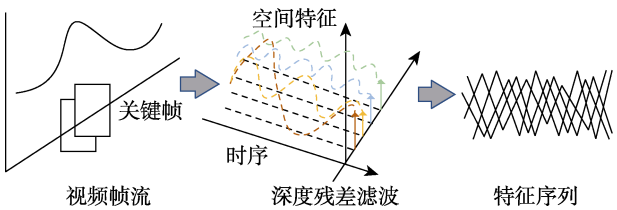

图 2 基于 ResNet 的特征提取过程

Fig.2 Feature extraction process based on ResNet

到 56×56×64 的特征向量，作为残差单元的输入。步骤 2.2 中采用线性整流函数作为激活函数，以避免梯度爆炸和梯度消失的问题。

1.4　基于全局注意力机制的 LSTM 手语语义信息提取算法

LSTM 作为针对时间序列的训练模型，虽然能够考虑历史信息对当前状态的影响，但是在网络层数多及时间跨度大的情况下，容易丢失有效信息。针对上述问题，文中在 LSTM[14]中引入全局注意力机制[15]作为解码器，来考虑每一个时间步长的隐藏状态，与解码器前一步的输出一同输入到下一步解码器中进行运算。与此同时，研究者注意到全局注意力模型的中心思想是在推导上下文向量时考虑编码器的所有隐藏状态。此模型通过对有效信息加权，以提高其与当前隐藏层状态的关联程度，避免信息丢失，以提高识别的准确率。受上述启发，此节提出了基于全局注意力机制的 LSTM 手语语义信息提取算法，详细流程见算法 4。

算法 4：基于全局注意力机制的 LSTM 手语语义信息提取算法

输入：$\boldsymbol{F}=\{\boldsymbol{f}_0,\boldsymbol{f}_1,\cdots,\boldsymbol{f}_k,\cdots,\boldsymbol{f}_m\}$，手语词库 $L=\{l_0,l_1,\cdots,l_t,\cdots,l_z\}$。

输出：文本信息 S。

步骤 1：随机初始化权重 W，S=Ø，k=0；

步骤 2：将手语词库 $L=\{l_0,l_1,\cdots,l_t,\cdots,l_z\}$导入到 Embedding 网络层，利用 word2vec 词向量模型生成词向量 $\boldsymbol{L'}=\{\boldsymbol{l}'_0,\cdots,\boldsymbol{l}'_t,\cdots,\boldsymbol{l}'_z\}$。

步骤 3：将高维语义特征向量 $\boldsymbol{F}$ 与词向量 $\boldsymbol{L'}$经过 Padding 操作转换成相同的维度；

步骤 4：将 $\boldsymbol{f}_k$ 和 $\boldsymbol{L'}$作为 LSTM 网络的原始状态输入，通过注意力加权解码器计算 L 中每个词 l_t 的预测概率分布；

步骤 5：通过 softmax 输出概率最大的词 l_t，$S=S\cup l_t$。

步骤 6：k=k+1;若 k<=m，转到步骤 4。

步骤 7：输出 S。

在步骤 1 中，生成空序列 S，用于存储手语语义信息；在步骤 2 中，手语词库 $L=\{l_0,l_1,\cdots,l_t,\cdots l_z\}$包含演示视频所有动作所对应的单词，用于网络识别判断与输出；在步骤 4 中，利用 LSTM 网络计算由注意力加权过的词向量 $\boldsymbol{l}'_t$ 输出词 l_t 的概率 P_{l_t}，选取概率最大的单词作为输出，解码器输出见式(4)。

$$S=Decoder(\boldsymbol{f}_t,l_t,h_{t-1}^{w}) \tag{4}$$

式中：$\boldsymbol{f}_t$ 表示特征向量；l_t 为手语词库的子集；h_{t-1}^{w} 表示上下文隐藏状态。遍历 $\boldsymbol{F}$ 和 $\boldsymbol{L'}$，利用 Softmax 函数输出预测结果。

2　性能测试与分析

2.1　数据集

选用中国科学技术大学公开发布的中文连续手语视频数据集 CSL 作为测试数据集。CSL 数据集由微软 Kinect 摄像头记录，提供 RGB 信息、深度信息和肢体骨架信息。文中实验取 RGB 信息。选取语料库中 100 个主要用于日常交流的句子，每个句子平均由 4 个单词组成，共 178 个中文单词，数据集中每个句子由 50 个不同的表演者完成。其中每一个表演者演示 5 次，在这个数据集中一共有 25 000 个手语视频。

2.2　评价指标

采用识别准确率（A）评价算法识别性能，以词错误率（W）、B_1、C_r、R_L 及 M_R 评价翻译性能。

1）准确率。识别准确率 A 的计算过程见式(5)。

$$A=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \tag{5}$$

式中：TP 表示预测样本为正例且实际为正例；TN 表示预测样本为负例而实际样本为负例，FP 表示预测样本为正例而实际为负例；FN 表示预测样本为负例而实际为正例。

2）词错误率。词错误率（W）常用于评判识别出来的词序列和标准的词序列之间的一致性，计算过程见式(6)。

$$W=\frac{s+d+i}{n} \tag{6}$$

式中：s 表示替换词数目；d 表示删除词数目；i 表示插入词数目；n 表示标签句子中的单词数目。在测试时，首先在真实文本中替换、插入或删除一个单词，然后重复这个操作几次。插入或替换的新单词从训练集中的词汇表中选取。这样可以得到一个伪视频文本，与标准文本进行比对从而得到词错误率（W）评价指标。

3）BLEU-1(B_1)[16]是手语识别任务中的常见评价指标，主要用于评估翻译语句中的词准确率，其结果可以表示为翻译结果中词的正确匹配次数与所有词出现次数的比值。ROUGE-L(R_L)[17]是一种基于召回率的相似性度量方法。用于评价预测结果与实际文本中词的共现率。用于评价手语识别中译文的流畅性 METEOR(M_R)[18]，考虑了基于整个语料库上的准确率和召回率，能够得出最佳识别结果和目标标签之间的准确率和召回率的调和，解决了 BLEU 标准中一些固有的缺陷。CIDEr(C_r)[19]指标将每个句子看成文档，通过计算 TF-IDF 获取向量余弦距离来度量文本相似性。

2.3　实验设置与比较算法

实验在 Dell Tower 5810 工作站上完成，使用

Microsoft Kinect V2 深度摄像头，显卡配置为 NVIDIA Tesla V100，处理器为 Intel Xeon 5218，内存为 32 GB，软件环境为：Ubuntu18.04，Python3.7，PyTorch1.7.0，Opencv3.4.2。

实验参数设置如下：epoch 为 50，批处理参数为 4，编码器学习率设置为 0.001，解码器学习率设置为 0.004，权重衰减率为 0.000 5。

实验过程中将 100 个中文句子细分为 4 个小的类别，见表 1。在每个类别下取 10%的视频作为测试数据集，剩余部分作为训练数据集。

选取 5 种手语识别算法作为对比算法进行实验，分别是 HLSTM-attn[20]、HMM-DTC[21]、HRF-Fusion[22]、LSTM&CTC[14,23]及 S2VT（3 layers）[18]。

表 1 数据集的句子类别情况

Tab.1 Sentence category of data set

	类别	例句	视频数量
c_1	对家庭成员的描述	他的弟弟是向导	11 000
c_2	对客观事物的描述	剪刀是锋利的	4 000
c_3	对社会环境的描述	社会的安定	4 500
c_4	其他	天气预报有雨	5 500

2.4 结果与分析

根据表 1 的设置运行所提出的 CSLR-GAML 算法。c_1、c_2、c_3 和 c_4 类别下连续手语视频训练过程中的算法训练准确率统计结果见图 3。观察图 3 可知，c_1 的准确率最高，为 92.094%；c_2 的准确率为 91.848%；c_3 的准确率为 91.294%；c_4 的准确率最低，为 89.099%。训练过程中，网络在迭代 5 次后开始收敛，c_1 和 c_2 中

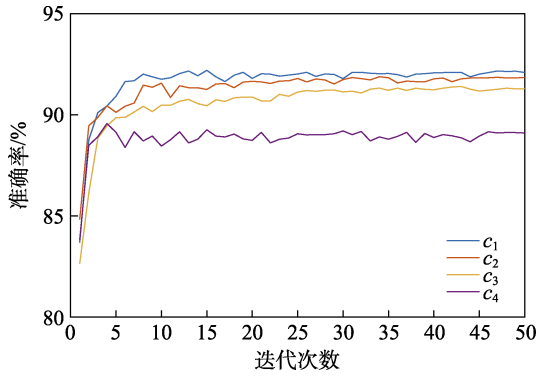

图 3 训练准确率

Fig.3 Training accuracy

句子的类型较为单一，组成成分基本为主谓宾结构，在训练过程中表现良好。c_3 的句子成分稍微复杂，但是重复出现的单词较多，也取得不错的识别效果。c_4 句子成分复杂，且语句之间差别较大，在训练过程中没有特别好的表现。

针对以上 4 个类别的手语视频识别，分别运行 CSLR-GAML，HLSTM-attn[20]、HMM-DTC[21]、HRF-Fusion[22]、LSTM&CTC[14,23]及 S2VT(3 layers)[18]算法，识别准确率 A 的统计结果见图 4 所示。从图 4a 可知，HLSTM-attn 算法准确率较高，但算法在不同类别下的识别准确率波动较大，表现不稳定；观察图 4b 可知，HMM-DTC 算法在不同类别下的识别准确率波动较小，但识别精度不高；而由图 4c—图 4e 可知，这些算法对应的盒子图矩形框面积大，这表明在单一类别上的识别准确率分布较散。相反，文中所提算法对应的图 4f，不同类别间的矩形分布高度相对一致，矩形框的面积也相对较小，精度也高于比较算法，在

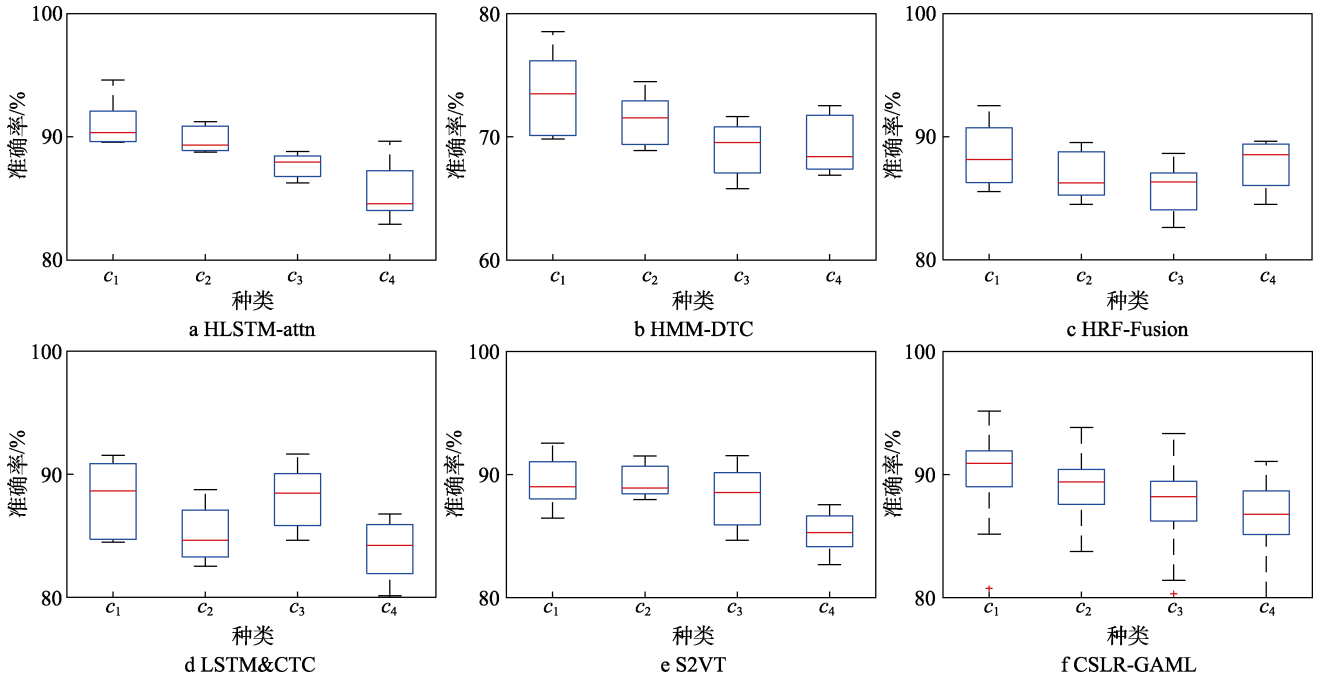

图 4 识别准确率 A

Fig.4 Recognition accuracy a

4 类数据集的平均识别率准确率是 90.08%，这表明此文算法具有更好的识别准确率和鲁棒性。

各算法的 B_1、C_r、R_L、M_R 及 W 统计结果见表 2。由表 2 可知，对比算法在 4 个分类中的平均 W 值分别为 0.419、0.728、0.716、0.761、0.608，此文算法的平均 W 值为 41.2%，比其中最好的结果高出 0.7%，但是在 c_4 类别中，文中算法的 W 值略低于 HLSTM-attn 算法的结果，这表明在小样本无规律词组下此文算法依旧有待改进。在其他指标下，此文算法的表现依然优于其他算法的结果。与最好的 HLSTM-attn 算法相比，B_1、C_r、R_L 及 M_R 指标下此文算法分别提高了 0.3%、0.9%、0.5%及 0.8%。与最差的 LSTM&CTC 相比，B_1、C_r、R_L 以及 M_R 指标下，文中算法分别提高了 27.6%、36.4%、28%以及 13.8%。

表 2　B_1、C_r、R_L、M_R 及 W 统计结果
Tab.2 Statistical results of B_1, C_r, R_L, M_R and W

算法	类别	评价指标				
		B_1/%	C_r/%	R_L/%	M_R/%	W/%
HLSTM-attn	c_1	61.4	61.4	63.7	25.2	40.1
	c_2	60.8	59.3	63.2	54.2	41.4
	c_3	58.9	55.6	60.3	23.9	42.8
	c_4	56.2	53.2	58.5	22.4	43.4
	均值	59.3	57.7	61.4	23.9	41.9
HMM-DTC	c_1	37.3	36.6	38.0	15.5	68.4
	c_2	34.9	35.1	34.7	14.4	69.7
	c_3	31.8	33.4	31.9	12.9	73.3
	c_4	29.7	31.9	27.7	10.8	79.7
	均值	33.4	34.2	33.1	13.4	72.8
HRF-Fusion	c_1	46.0	40.2	45.4	16.9	67.2
	c_2	44.1	38.7	44.7	14.8	69.2
	c_3	41.3	35.3	41.0	13.3	73.3
	c_4	38.6	33.1	38.7	12.7	76.8
	均值	42.5	36.8	42.5	14.4	71.6
LSTM&CTC	c_1	33.3	23.9	35.8	11.3	73.7
	c_2	33.1	23.3	35.1	11.2	74.6
	c_3	31.9	21.4	33.1	10.8	75.9
	c_4	29.8	20.1	31.7	10.2	80.2
	均值	32.0	22.2	33.9	10.9	76.1
S2VT(3 layers)	c_1	51.3	49.8	50.2	20.1	58.8
	c_2	49.6	47.9	48.6	18.7	59.6
	c_3	47.7	46.6	47.9	17.6	61.6
	c_4	43.9	45.4	43.4	15.4	63.1
	均值	48.1	47.4	47.5	17.9	60.8
CSLR-GAML (此文)	c_1	63.2	61.3	65.1	26.8	39.0
	c_2	61.3	60.8	63.6	25.4	39.4
	c_3	58.6	57.7	60.1	23.9	41.3
	c_4	55.4	54.7	58.7	22.7	45.2
	均值	59.6	58.6	61.9	24.7	41.2

为了观察各算法的运行效率，分别从 4 个类别中选取 20 个手语视频段统计其识别的时间，不同算法识别相同数量手语视频段的时间统计结果见表 3。观察表 3 可知，HLSTM-attn、HMM-DTC、HRF-Fusion、LSTM&CTC 和 S2VT 这 5 种算法平均运算时间分别为 0.564、0.706、0.698、0.656 和 0.642 s，此文算法平均运算时间为 0.646 s，虽然略差于 HLSTM- attn 和 S2VT 的表现，但优于其他 3 种算法。这表明文中算法虽然在识别率上与其他算法存在优势，但是在运算效率上还需要进一步改进。

表 3　时间统计结果
Tab.3 Time statistics

算法	运算时间/s				
	c_1	c_2	c_3	c_4	均值
HLSTM-attn	0.564	0.562	0.560	0.569	0.564
HMM-DTC	0.704	0.707	0.702	0.711	0.706
HRF-Fusion	0.695	0.692	0.699	0.704	0.698
LSTM&CTC	0.652	0.659	0.655	0.658	0.656
S2VT	0.637	0.645	0.640	0.647	0.642
CSLR-GAML	0.646	0.646	0.644	0.647	0.646

综上所述，在识别准确率与翻译性能上文中算法优于所比较的 5 种算法。

3　结语

高性能的视觉手语识别算法在帮助听障人群日常交流中具有广阔的应用前景，对促进听障人群无障碍融入社会方面具有非常积极的意义。文中提出了一种基于全局注意力机制和 LSTM 网络的连续手语识别方法。首先，利用差分方法进行关键帧提取，有效去除冗余帧；然后，使用 ResNet 提取图片特征，有效解决由于手语的手部形状、位置和方向变化多样性等导致的特征提取困难问题，避免特征丢失；最后利用全局注意力机制获取更加全面的序列信息，保证算法的识别准确性。实验表明，此文方法在连续手语识别中具有较高的识别性能以及翻译性能。本研究的实验数据集选用 RGB 信息，GRB+深度信息实验也是值得深入的工作。与此同时，结合具体的应用场景，研制使用友好的系统是值得进一步推进的方向。

参考文献：

[1] 叶欣, 朱大伟, 陈思源, 等. 听力障碍社会成本的系统综述[J]. 人口与发展, 2020, 26(4): 51-59.
YE Xin, ZHU Da-wei, CHEN Si-yuan, et al. The Societal Cost of Hearing Impairment: A Systematic Review[J]. Population and Development, 2020, 26(4): 51-59.

[2] 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.

[3] KAMAL S M, CHEN Y, LI S, et, al. Technical Approaches to Chinese Sign Language Processing: A Review[J]. IEEE Access, 2019, 7: 96926-96935.

[4] 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.
MINAWAL A, ARIF K, QINA X, et al. Review of Sign Language Recognition Methods and Techniques[J]. Computer Engineering and Applications, 2021, 57(18): 1-12.

[5] TATENO S, LIU H, OU J. Development of Sign Language Motion Recognition System for Hearing-Impaired People Using Electromyography Signal[J]. Sensors(Basel, Switzerland), 2020, 20(20): 5807.

[6] 冉孟元, 刘礼, 李艳德, 等. 基于惯性传感器融合控制算法的聋哑手语识别[J]. 计算机科学, 2021, 48(2): 231-237.
RAN Meng-yuan, LIU Li, LI Yan-de, et al. Deaf Sign Language Recognition Based on Inertial Sensor Fusion Control Algorithm[J]. Computer Science, 2021, 48(2): 231-237.

[7] 王鑫炎, 王青山, 马晓迪, 等. 一种基于滑动窗口分割的中国手语识别系统[J]. 北京邮电大学学报, 2021, 44(5): 48-54.
WANG Xin-yan, WANG Qing-shan, MA Xiao-di, et al. A Split Sliding Window-Based Continuous Chinese Sign Language Recognition System[J]. Journal of Beijing University of Posts and Telecommunications, 2021, 44(5): 48-54.

[8] PAPASTRATIS I, DIMITROPOULOS K, DARAS P. Continuous Sign Language Recognition through a Context-Aware Generative Adversarial Network[J]. Sensors, 2021, 21(7): 2437.

[9] 王粉花, 张强, 黄超, 等. 融合双流三维卷积和注意力机制的动态手势识别[J]. 电子与信息学报, 2021, 43(5): 1389-1396.
WANG Fen-hua, ZHANG Qiang, HUANG Chao, et al. Dynamic Gesture Recognition Combining Two-Stream 3D Convolution with Attention Mechanisms[J]. Journal of Electronics & Information Technology, 2021, 43(5): 1389-1396.

[10] PU J, ZHOU W, LI H.. Iterative Alignment Network for Continuous Sign Language Recognition[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), California: IEEE, 2019: 4160-4169.

[11] ZHOU H, ZHOU W, LI H. Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition[C]// 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai: IEEE, 2019:1282-1287.

[12] 周舟, 韩芳, 王直杰. 面向手语识别的视频关键帧提取和优化算法[J]. 华东理工大学学报(自然科学版), 2021, 47(1): 81-88.
ZHOU Zhou, HAN Fang, WANG Zhi-jie. Video Key Frame Extraction and Optimization Algorithm for Sign Language Recognition[J]. Journal of East China University of Science and Technology, 2021, 47(1): 81-88.

[13] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition,Nevada, IEEE, 2016: 770-778.

[14] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[15] LUONG M-T, PHAM H, MANNING C D. Effective Approaches to Attention-based Neural Machine Translation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon: Association for Computational Linguistics, 2015: 1412-1421.

[16] 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 (ACL), Philadelphia: Association for Computational Linguistics, 2002: 311-318.

[17] LIN C-Y. Rouge: A package for automatic evaluation of summaries[C]// Text Summarization Branches Out, Barcelona: Association for Computational Linguistics, 2010: 74-81.

[18] VENUGOPALAN S, ROHRBACH M, DONAHUE J, et al. Sequence to Sequence--Video to Text[C]// 2015 IEEE International Conference on Computer Vision (ICCV), Santiago: IEEE, 2015:4534-4542.

[19] VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. CIDEr: Consensus-based Image Description Evaluation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston:IEEE, 2015: 4566- 4575.

[20] GUO D, ZHOU W, LI H, et, al. Hierarchical LSTM for sign language translation[C]// Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans: AAAI, 2018: 6845-6852.

[21] WANG H, CHAI X J, CHEN X L. Sparse Observation(SO)Alignment for Sign Language Recognition[J]. Neuro Computing, 2016, 175(1): 674-685.

[22] GUO D, ZHOU W, LI A, et, al. Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation[J]. IEEE Transactions on Image Processing, 2020, 29: 1575-1590.

[23] GRAVES A, FERNANDEZ 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, New York: Association for Computing Machinery, 2006: 369-376.

责任编辑：陈作