武汉大学学报(理学版)
Journal of Wuhan University(Natural Science Edition)
ISSN 1671-8836,CN 42-1674/N

《武汉大学学报(理学版)》网络首发论文

题目：基于注意力机制的复杂背景连续手语识别
作者：杨光义，丁星宇，高毅，胡晶欣，张洪艳
DOI：10.14188/j.1671-8836.2021.0350
收稿日期：2021-12-20
网络首发日期：2022-11-30
引用格式：杨光义，丁星宇，高毅，胡晶欣，张洪艳．基于注意力机制的复杂背景连续手语识别[J/OL]．武汉大学学报(理学版).
https://doi.org/10.14188/j.1671-8836.2021.0350

武汉大学学报(理学版)
J. Wuhan Univ. (Nat. Sci. Ed.) 1~9

DOI: 10.14188/j.1671-8836.2021.0350

基于注意力机制的复杂背景连续手语识别

杨光义[1],丁星宇[1],高 毅[1],胡晶欣[1],张洪艳[2†]

1. 武汉大学 电子信息学院,湖北 武汉 430072;

2. 武汉大学 测绘遥感信息工程国家重点实验室,湖北 武汉 430079

收稿日期:2021-12-20 † 通信联系人 E-mail:zhanghongyan@whu.edu.cn

基金项目:国家自然科学基金面上项目(42071322);湖北省杰出青年基金(2020CFA053);武汉市应用基础前沿项目(2020010601012184)

作者简介:杨光义,男,高级实验师,现从事图像处理和高频电路方面的研究。E-mail:ygy@whu.edu.cn

摘 要:提出一种能够处理复杂背景下的连续手语识别模型,基于注意力机制的连续手语识别算法ACN(attention-based 3D convolutional neural network)。首先,利用背景去除模块,对包含复杂背景的手语视频进行预处理;然后,通过基于空间注意力机制的3D-ResNet(3D residual convolutional neural network)提取时空融合信息;最后,采用结合时间注意力机制的长短期记忆网络(long short-term memory,LSTM)进行序列学习,得到最终的识别结果。算法在大规模中国连续手语数据集CSL100上表现优异;在面向不同复杂背景的情况下,算法表现出良好的泛化性能,模型引入的时空注意力机制是切实有效的。

关 键 词:连续手语识别;复杂背景;注意力机制;长短期记忆网络

中图分类号:TP391.2 **文献标志码**:A **文章编号**:1671-8836(XXXX)XX-0001-09

Continuous Sign Language Recognition in Complex Background Based on Attention Mechanism

YANG Guangyi[1], DING Xingyu[1], GAO Yi[1], HU Jingxin[1], ZHANG Hongyan[2†]

1. Electronic Information School, Wuhan University, Wuhan 430072, Hubei, China;

2. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, Hubei, China

Abstract: In this work, an attention-based 3D convolutional neural network (ACN) is proposed for continuous sign language recognition in complex background. Firstly, the sign language video containing complex background is preprocessed with the background removal module. Then, the spatio-temporal fusion information is extracted by 3D-ResNet (3D residual convolutional neural network) based on spatial attention mechanism. Finally, the long short-term memory (LSTM) network combined with the time attention mechanism is used for sequence learning to obtain the final recognition result. Extensive experiments show that the algorithm performs well on the large-scale Chinese continuous sign language dataset CSL100. The algorithm shows good generalization performance facing different complex background, and the spatio-temporal attention mechanism introduced by the model is effective.

Key words: continuous sign language recognition; complex background; attention mechanism; long short-term memory(LSTM)

0 引 言

手语作为人类肢体语言的一种,是听障人士广泛使用的交流方式。然而,现实生活中大多数健全人士没有手语学习经历,导致无法与听障人士正常交流。作为手语翻译成文本语言的技术手段,手语识别可以很好地克服交流障碍,逐渐成为健全人士与听障人士之间的重要沟通纽带[1~4]。根据应用场

引用格式:杨光义,丁星宇,高毅,等. 基于注意力机制的复杂背景连续手语识别[J]. 武汉大学学报(理学版),XXXX,XX(XX):1-9. DOI:10.14188/j.1671-8836.2021.0350.

YANG Guangyi,DING Xingyu,GAO Yi,*et al*. Continuous Sign Language Recognition in Complex Background Based on Attention Mechanism [J]. *J Wuhan Univ (Nat Sci Ed)*,XXXX,XX(XX):1-9. DOI:10.14188/j.1671-8836.2021.0350(Ch).

景的不同,手语识别可分为离散手语识别与连续手语识别两大类,二者分别将手语视频翻译成孤立的词汇和连续完整的句子。由于现实生活中人与人之间的交流多以句子为单位,所以连续手语识别技术应用更为广泛。然而,连续手语视频数据冗余度高、模型时空特征提取性能不足以及实际手语翻译场景复杂(例如街道、火车站、房间内等场地)等问题,导致识别精度不高、实时性差,一直制约着该技术的大规模应用。因此,研究精度高、实时性高的连续手语识别算法,具有很强的学术价值和社会意义[5,6]。

连续手语识别的本质,是将一组视频帧序列映射到手语翻译结果的过程,是一个典型的多到多的序列学习问题。学者们通常将该问题分解为两部分:视频特征提取与视频序列解码学习。文献[7~9]利用统计学原理对特征提取部分进行优化,文献[10~12]利用分类模型和时序建模对解码部分进行优化,均取得了一定的识别效果。然而,这些传统方法依赖人工特征的选取,费时费力且完备性不足,识别精度非常有限。

随着深度学习特别是卷积神经网络的兴起,基于深度学习的手语识别模型逐渐成为学者们关注的热点。2014年,根特大学Pigou等[13]提出包含双二维CNN(convolutional neural network)的手语识别系统,用来提取手部特征和上半身特征从而识别手语,开创了深度学习识别手语的先河。随后,Koller等[14]将CNN嵌入到HMM(hidden Markov model)中,将CNN的强识别能力和HMM的序列建模能力相结合。为了有效识别孤立词和连续语句,Xiao等[15]提出基于双LSTM(long short-term memory)和HMM的手语识别算法。Huang等[16]首次将3D-CNN模型应用于连续手语识别,用于提取多源视频的时空特征。为了同时兼顾全局特征和局部特征,实现优势互补,Song等[17]设计了一种并行时间编码器,并行地从全局和局部来学习手语视频与标签的时间关系,类似的做法还有文献[18,19]。Qin等[20]利用预训练的孤立词翻译模型来辅助连续手语翻译模型的训练,而Zuo等[21]则是从一致性增强的角度,通过添加辅助约束来优化手语翻译模型。

随着研究的深入,学者们发现,手语视频一般有数百帧图像,存在大量重复且无意义的视频帧,时空特征提取存在冗余性,影响算法效率和效果;另外,只关注手语视频中的手部特征,忽略人体面部表情等非手部特征,也会制约手语识别的精度;还有,数据集中的手语表达大都是整洁干净的单色背景,导致算法在面向复杂背景时的识别精度下降,给实践应用造成困难。

为了克服复杂背景带来的精度下降问题,本文将背景去除模块、注意力机制与连续手语识别算法相结合,提出一种基于注意力机制的复杂背景连续手语识别算法ACN(attention-based 3D convolutional neural network)。ACN算法将背景去除作为图像预处理模块,设计结合空间注意力的3D残差块提取时空特征,构建结合时间注意力的长短期记忆网络进行序列学习,实现RGB手语视频到句子的精确翻译。本文的主要贡献点包括:1)提出一种基于注意力机制的连续手语识别算法,通过引入注意力机制有效解决特征冗余的问题,在大规模中国连续手语数据集CSL100上表现优异;2)首次将复杂背景去除应用于连续手语识别问题,实现复杂背景下连续手语视频的端到端翻译,大大提高了算法的泛化性能。

1 本文模型

1.1 算法总体流程

面向复杂背景的连续手语识别模型总体框图如图1所示,输入为包含复杂背景的连续手语视频,输出为最终的手语识别结果。首先利用复杂背景去除模块[22](background matting module,BM)对手语视频进行预处理,得到纯净背景的手语视频。BM模块由空洞空间卷积池化金字塔[23](atrous spatial pyramid pooling,ASPP)和编-解码器网络构建而成。编码器与解码器各层跳跃连接,将不同层级特征融合,提升背景去除效果。随后,将纯净背景的手语视频进行后处理,主要包括关键帧抽取和中心剪裁,以减少网络参数量。将裁剪后的视频帧序列输入到由卷积模块、空间注意力模块和LSTM模块组成的编码器中,得到时空融合特征。最后,通过结合时间注意力机制的LSTM模块组成的解码器进行解码学习,得到手语识别结果。

图1中,考虑到实际场景下连续手语识别中对手语视频实时性、精确性的需要,BM模块采用轻量级网络MobileNet[24]作为特征提取器,并通过ASPP提取不同尺度的空间信息,最后由解码器得到视频帧序列对应的Alpha图序列。MobileNet主体由3×3的深度可分离卷积层、Batch Normalize层、1×1卷积层以及ReLu激活层组成,在减少网络参数的同时,保障网络特征提取能力不受影响。ASPP主要包含三个空洞卷积层,空洞卷积的采样率分别设置为3、6和9,以确保模型提取图像不同尺度的空间信

息,从而进一步获取视频帧的上下文信息。空洞卷积的应用使得在不增加网络参数的情况下,扩大卷积神经网络的感受野,提升网络表达能力。BM模块中的解码器在每个步骤均运用双线性上采样处理,并与编码器各层跳跃连接,通过3×3卷积层、Batch Normalize层以及ReLu激活层后,得到Alpha图序列。然后,将掩膜图像与原图进行掩膜操作,得到整个人体部分的纯净背景手语视频。

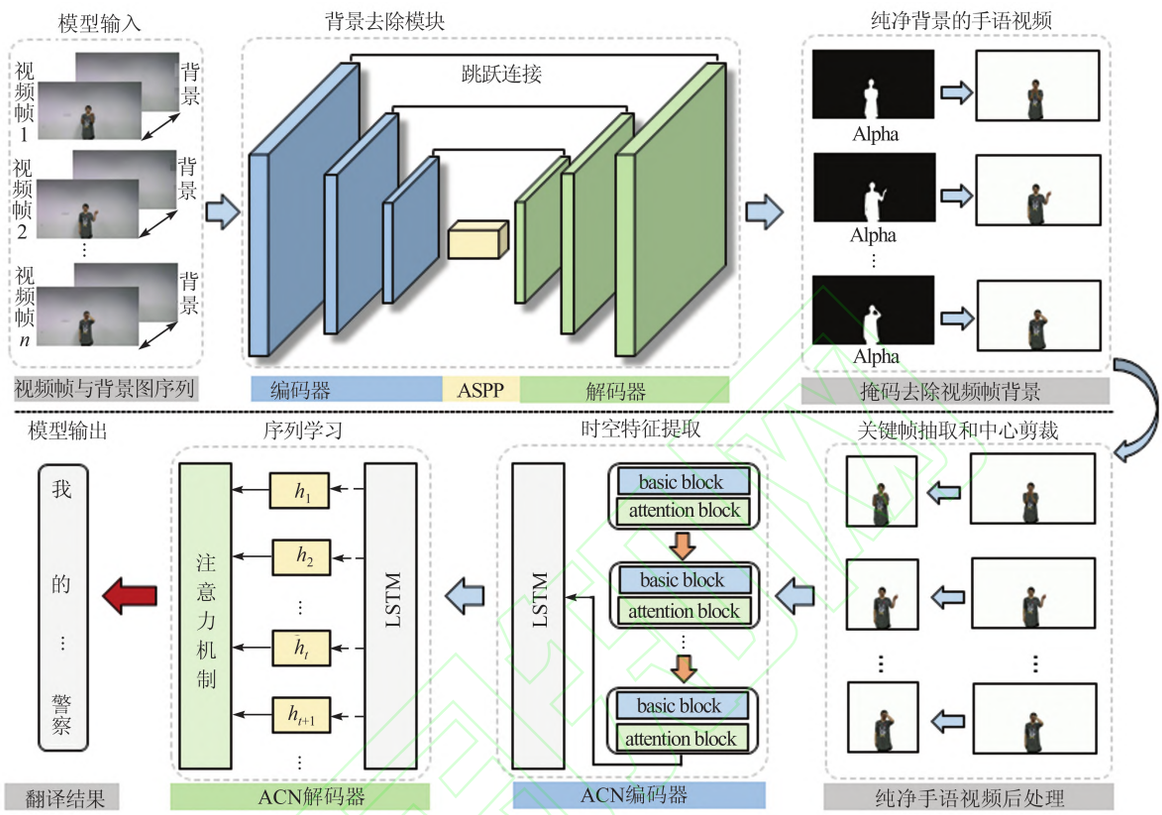


图1 算法总体框图

Fig. 1 Overall block diagram of algorithm

1.2 ACN编解码

常见的手语视频往往连续数百帧,而对应的标签只有若干个汉字,二者在序列长度上相差巨大。为了解决这种长度差异带来的模型输入长度变化问题,学者们通常采取seq2seq(sequence to sequence)[25]的模型结构。另外,手语视频的帧与帧之间,存在大量的冗余信息,如何提取关键信息的同时剔除冗余信息显得尤为重要,为此,本文将注意力机制引入ACN网络。ACN网络采用编码器与解码器的结构,分别用来进行手语视频的时空特征提取和序列学习。

1.2.1 ACN编码器

ACN编码器的具体结构如图2所示,其主体由3D-ResNet18组成。通过3D-ResNet18进行时空特征提取,捕捉帧与帧之间快速变化的信息。与此同时,考虑到手语视频数据包含大量重复无意义的视频帧,且每一帧中的大部分像素都是无关信息。因此,算法引入空间注意力机制,在每个basic block后都接一个attention block,用以学习手语视频帧中感兴趣的部分,让编码器通过抽取关键帧的方式减少冗余信息,同时更关注于视频帧中重要的语义信息,以减少无意义的特征提取。其中关键帧(key frame,KF)数量为超参数,通过后续实验确定最佳值。另外,将3D-ResNet18的全连接层替换为LSTM模块(详见1.2.2节),得到手语视频的时空融合信息。

以第二个basic block为例,每个basic block中包含着1×1×1卷积核,3×3×3卷积核的三维卷积层以及batch normalize层,跳跃连接让浅层特征能直接传递至深层网络中,防止网络加深导致梯度爆炸的同时确保良好的性能。attention block对每帧的特征图进行3D全局最大池化(global max pooling,GMP)和3D全局平均池化(global average pooling,GAP),处理后得到两个特征映射$M^{H\times W\times 1}$,将其按照通道维数拼接在一起,得到特征映射$M^{H\times W\times 2}$。接着采用7×7×1三维卷积核的卷积层,对特征映

射$M^{H\times W\times 2}$进行卷积,保证最终特征与输入特征映射在空间维度上的一致性。最后通过sigmoid激活函数得到最终特征。算法公式为

$$M_s(F)=\sigma\left(f^{7\times 7\times 1}\left(\left[\mathrm{GAP}(F);\mathrm{GMP}(F)\right]\right)\right)=\sigma\left(f^{7\times 7\times 1}\left(\left[F^s_{GAP};F^s_{GMP}\right]\right)\right) \tag{1}$$

其中,F为输入特征,$\sigma(\cdot)$为sigmoid激活函数,$f^{7\times 7\times 1}(\cdot)$表示$7\times 7\times 1$卷积操作,$F^s_{GAP}$和$F^s_{GMP}$分别为全局平均池化和全局最大池化后的特征向量。

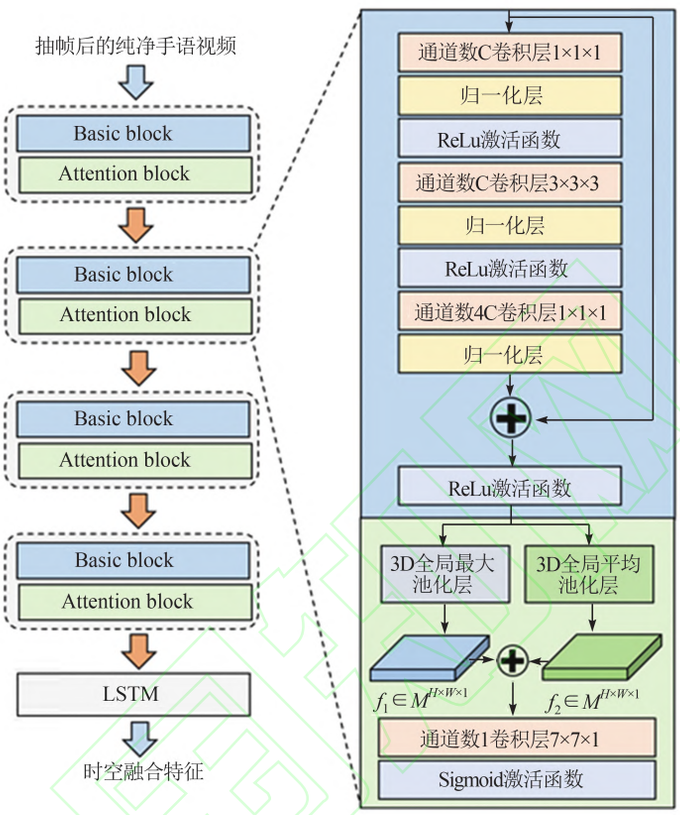


图2 ACN编码器结构示意图

Fig. 2 Structure diagram of ACN encoder

1.2.2 ACN解码器

由编码器得到的时空融合信息在每一次解码时都会提供给解码器,为了避免解码中的错误信息不断传播,解码过程中采用teacher forcing(TF)训练策略[26]。训练过程中,按一定TF比例给解码器提供正确的标签,加速模型收敛的同时保证模型的泛化能力。TF比例为另一个超参数,其最佳取值由实验确定。解码器具体结构如图3所示。

图3中,LSTM模块由若干cell单元组成,输入为编码器提供的时空融合特征,输出为各时刻的预测结果。每个cell单元主要由输入门、遗忘门和输出门组成,其中x_t是上一级cell单元输出的特征向

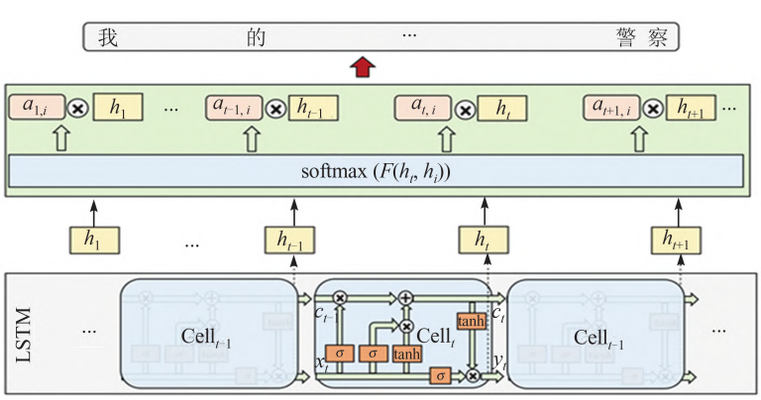


图3 ACN解码器结构示意图

Fig. 3 Structure diagram of ACN decoder

量,代表t时刻的输入;h_t是LSTM模块中的隐藏状态(hidden state),代表t时刻的预测输出;c_t是LSTM模块中的单元状态(cell state),代表t时刻的网络长期记忆。

由于手语和汉语的语法以及语序不同,ACN解码器还采用时间注意力机制,使解码过程更加关注整个手语句子中的重要片段,从而增强网络的解码性能,得到更可靠的翻译结果。具体表达式为:

$$z_t=\sum_{i-1}^{n}a_{t,i}\cdot h_i \tag{2}$$

$$h_t=\sigma\left(W_o\cdot\left[h_{t-1},x_t\right]+b_o\right)*\tanh\left(C_t\right) \tag{3}$$

$$a_{t,i}=\mathrm{softmax}\left(\frac{h_t\cdot h_i}{\|h_t\|\|h_i\|}\right) \tag{4}$$

其中,$a_{t,i}$表示各隐藏层节点状态h_i对t时刻(即第t帧)节点状态h_t的影响权重,$\sigma(\cdot)$指激活函数,W_o、b_o分别为LSTM网络单元的权重和偏置,z_t为加权融合后的t时刻预测输出。

1.3 损失函数

由于现有的手语识别数据集不涉及背景去除问题,本文引入VideoMatte240K[22]数据集训练BM模块。另外,考虑到联合训练模型的计算代价较大,本文将ACN模型的训练分为两部分:BM模块的训练和ACN编解码的训练。对于BM模块,采用L1损失函数优化模型参数,具体公式为

$$L_{\mathrm{BM}}=\left\|\alpha-\alpha^*\right\|_1 \tag{5}$$

其中α为预测所得alpha图像,α^*为对应的真实alpha图像。

对于ACN编解码的训练,采用交叉熵损失函数,具体公式为

$$L_{\mathrm{ACN}}=-\frac{1}{N}\sum_{c=1}^{N}y_{ic}\log\left(p_{ic}\right) \tag{6}$$

其中,N为词的类别总数,y_{ic}为符号函数sgn(·),p_{ic}为样本i属于类别c的预测概率。

2 实验与分析

2.1 数据集与评价指标

针对手语识别不同场景的需求,诸多学者和机构制作了不同语句类型的数据集,如表1所示。可以看出,大多数数据集采用的数据形式以RGB视频为主,有些数据集会以骨架信息和深度数据作为RGB视频的辅助信息。而不同语种的手语规则不同,目前神经网络还未能实现跨语种翻译。考虑到数据集的语种和手语识别算法的适用范围,本文实验均在中国连续手语数据集CSL100上进行。为验证算法的泛化性,仅使用数据集CSL100中的RGB视频数据,无需其他信息。实验时按照8:2的比例将数据集分成训练集和测试集,并确保每个手语表达者的手语视频不会同时出现在两个数据子集中。

表1 语句类型手语数据集汇总表

Table 1 Statement type sign language data set summary

数据集	发布年份	语种	词汇量	数据量	数据特点
RWTH-Boston-104[27]	2008	English	104	201	RGB/黑白图
GSL SI[28]	2007	Greek	310	10 290	RGB-D
SIGNUM[29]	2012	German	455	19 500	RGB
RWTH-Phoenix-2014T[30]	2018	German	3 000	8 257	RGB
CSL100[31]	2018	Chinese	178	25 000	RGB-D/骨架
How2Sign[32]	2019	English	16 000	38 611	RGB-D/骨架

为了客观比较不同算法的性能,采用词错误率WER和准确率Acc作为评价指标:

$$\mathrm{WER}=\frac{\#\mathrm{Subs}+\#\mathrm{Del}+\#\mathrm{Ins}}{\#\text{words in target}}\times 100\% \tag{7}$$

$$\mathrm{Acc}=\frac{\#\mathrm{Rights}}{\#\text{words in target}}\times 100\% \tag{8}$$

其中,Subs代表预测句子被替换的字词数量,Del表示预测句子中未预测到的字词数量,Ins表示预测结果中增加的字词数量,Rights表示预测结果中正确预测字词的个数。WER指标越低,Acc指标越高,证明算法效果越好。

2.2 确定超参数

如1.2节所述,ACN模型有两个需要设定的超参数:KF数量和TF比例。考虑到数据集内句子是由4~8个词组成的短句,表2中选取了6个典型的KF数量开展实验。同时,为了兼顾模型的收敛效果及泛化能力,表3选取5个典型的TF比例开展实验。

从表2中可以看出,随着采样帧数的增加,WER呈现先减后增的趋势(Acc正好相反),当KF=20时,WER和Acc均取得最佳值,当KF=16时,

WER和Acc均取得次佳值。值得注意的是,16帧、20帧、24帧的指标差距不大,而增加至28时模型性能明显下降。这可能是因为输入的视频帧数已经提供足够的语义信息供模型学习,模型趋于饱和。若继续增加KF数量只会加重模型训练负担,引入冗余信息从而导致模型性能下降,故本文取KF=16。从表2中可以看出,随着TF比例的增加,WER呈现先减后增的趋势(Acc正好相反),当TF=0.50时,WER和Acc均取得最佳值,当TF=0.35时,WER和Acc均取得次佳值,说明TF比例太大,模型容易过拟合;TF比例过小,模型在训练过程中容易被错误预测结果影响,导致最终收敛结果性能不佳。本文最终取TF=0.5。

表2 超参数KF和TF对比结果表

Table 2 Comparison results of hyperparameter

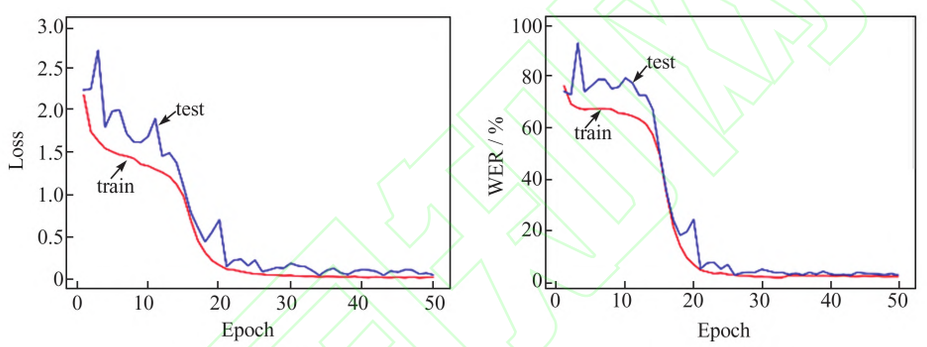
%

KF数量	WER	Acc	TF比例	WER	Acc
8	10.42	92.87	0.20	8.92	94.91
12	8.37	94.52	0.35	7.26	96.13
16	6.74	96.81	0.50	**6.74**	**96.81**
20	**6.69**	**96.90**	0.65	7.91	95.74
24	6.78	96.80	0.80	8.68	94.98
28	9.14	93.96			

注:粗体表示最佳数值

2.3 CSL100验证实验

选定合适的超参数后,ACN算法在CSL100数据上进行验证实验,得到结果如图4所示。可以看出,ACN模型在训练集和测试集上均能够快速收敛,其损失值与WER指标逐渐下降,表明模型的训练是有效的。



(a) Loss曲线图 (b) WER指标图

图4 CSL100验证实验曲线图

Fig. 4 CSL100 verification curve

为了更客观地比较算法之间的性能,表3列出了几种典型算法的量化指标,表格中的数据均从原文献处取得。其中,DTW-HMM、S2VT和HAN为传统方法,采用手工方式提取特征;LS-HAN和SubUNet为深度学习方法,采用深度网络提取特征。从表3可以看出,深度学习方法性能普遍优于传统方法,尤其以本文提出的ACN方法最为突出,体现了深度网络强大的特征提取能力。具体而言,对于WER指标,ACN方法比性能最佳的传统方法(HAN方法)提升了约5倍,比性能最佳的深度学习方法(SubUNet方法)提升了约2倍,反映了ACN方法良好的性能。

表3 CSL100上算法指标对比表

Table 3 Comparison of algorithm indexes on CSL100

	方法	WER/%	Acc/%
传统方法	DTW-HMM[10]	28.4	71.6
	S2VT[11]	25.5	74.5
	HAN[12]	20.7	79.3
深度学习方法	LS-HAN[27]		82.7
	SubUNet[33]	11.0	
	ACN (Ours)	3.6	98.3

注:粗体表示最佳数值

2.4 背景去除实验

为了验证BM模块的有效性,本文进行了背景消除对比实验,实验中用到的数据分别为作者在真实复杂背景下自制的手语视频、利用CSL100测试集合成的纯色背景手语视频以及CSL100测试集原始的白色背景手语视频。将未采用BM模块的ACN模型记为ACNn,与完整的ACN模型进行对比,得到的识别结果如图5所示,其中识别错误的结果用红色字体表示。

从图5可以看到,BM模块能够有效的将复杂背景处理成纯白背景,每一帧都保留了手语表达者的肢体表达且没有丢失细节。从翻译结果来看,在面对纯色或白色背景的手语视频时,ACN模型和ACNn模型均能准确预测;而对于带有复杂背景的手语视频,ACNn模型出现识别错误,ACN模型能够识别正确,说明BM模块的加入很好的解决了复

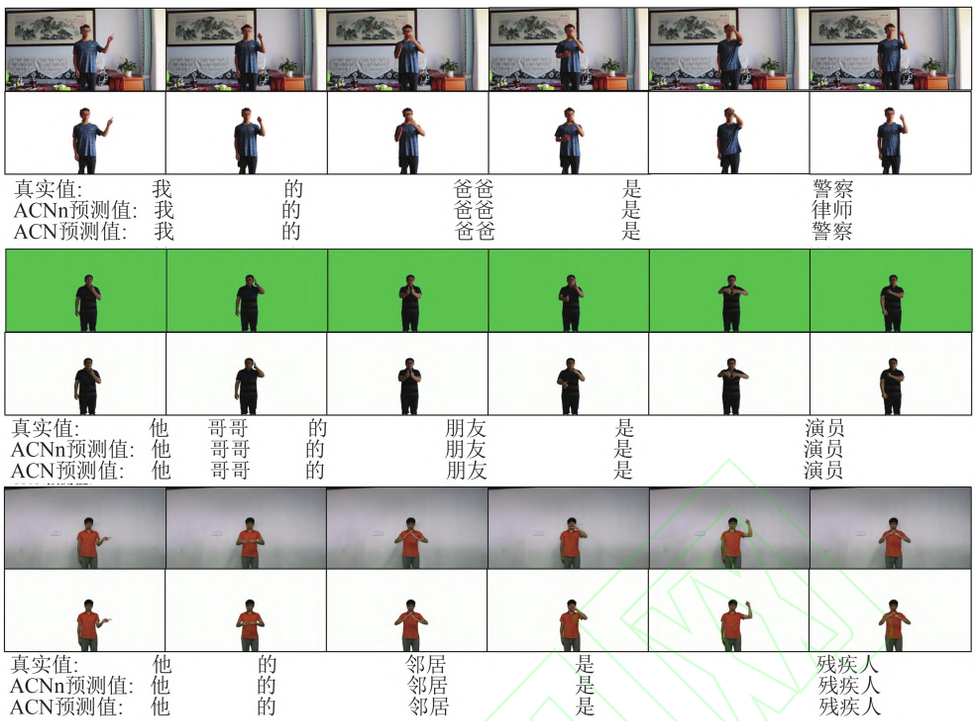


图5 背景去除对比实验结果图

Fig. 5 Comparison experiment results of background matting

杂背景问题,说明BM模块是有效和必要的。

2.5 消融实验

为了深入探究注意力机制的有效性,定量分析各注意力模块的贡献,本文进行注意力相关的消融实验。实验以结合LSTM的3D-ResNet18网络作为基本结构(baseline),分别考察空间注意力(spatial attention)模块、时间注意力(time attention)模块以及通道注意力(channel attention)模块对算法性能的影响,设计三种模型:1)B+S:baseline + spatial attention;2)B+S+T:baseline + spatial and time attention;3)B+S+T+C:baseline + spatial, time and channel attention。网络初始学习率为2E-3,批次大小为128(4块Tesla V100 GPU),采用Adam优化器进行优化训练,训练轮次设定为100轮,实验结果见图6与表4。

从消融实验的结果可以看出,在特征编码阶段引入空间注意力模块,能让模型(B+S)对视频帧中人体的面部、嘴部、身体姿势、手部的特征自适应加权融合,特征提取的性能更佳,从而提升测试指标;继续在解码器中加入时间注意力模块,能让模型(B+S+T)测试指标得到进一步的提升,指标一致性更好,模型损失也更小;继续在编码器中加入通道注意力模块,模型(B+S+T+C)测试指标不升反降,甚至不如B+S模型,说明一味地简单堆叠模块并不能提升模型性能,这可能是因为额外模块的加入增加了模型的计算负担,并且对于特征通道权重的分配没能让模型学习到有用的特征信息,反而增加了对类似于噪声的无用信息的关注,导致模型性能下降。

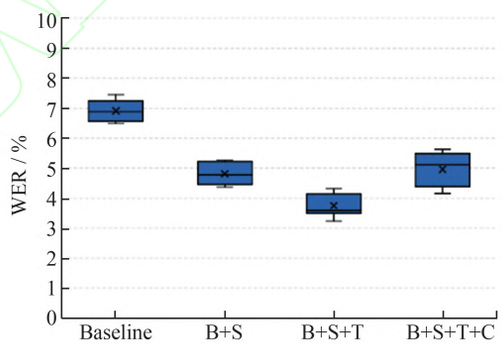


图6 消融实验盒状图

Fig. 6 Box diagram of ablation experiment

表4 注意力机制的消融实验结果

Table 4 Ablation results of attention mechanism

Model	WER / %	Acc / %	Loss
Baseline	6.74	96.81	0.31
B+S	4.53	97.03	0.27
B+S+T	**3.66**	**98.28**	**0.18**
B+S+T+C	4.63	97.17	0.26

注:粗体表示最佳数值

3 结 语

本文提出一种基于注意力机制的连续手语识别算法(ACN)。算法利用BM模块,对复杂背景下的手语视频做预处理,利用结合时空注意力的编解码结构,很好地提取视频的时空融合特征并解码得到准确的翻译结果。实验结果表明,ACN算法相较于其他算法表现突出,体现了较高的准确性与泛化能力。ACN算法能够完成各种复杂背景和纯色场景下的连续手语识别任务,对连续手语识别的落地应用有着积极的指导意义。下一步的研究重点是在保证识别准确率的前提下,进一步提升算法效率,实现手语识别的实时性应用,从而促成手语识别的落地应用。

参考文献:

[1] 戴兴雨,王卫民,梅家俊. 基于深度学习的手语识别算法研究[J]. 现代计算机,2021,**27**(29):63-69.
DAI X Y, WANG W M, MEI. J J. Research on sign language recognition algorithm based on deep learning [J]. *Modern Computers*, 2021,**27**(29):63-69.

[2] 倪兰,唐文妍,和子晴,等. 中国手语服务行业现状与发展趋势[J]. 语言产业研究,2021,**3**:111-123.
NI L, TANG W Y, HE Z Q, *et al*. Current situation and development trend of sign language service industry in China [J]. *Language industry research*, 2021,**3**:111-123.

[3] 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.
MINAVAL A, ALIFU K, XIE Q, *et al*. Overview of sign language recognition methods and technologies [J]. *Computer engineering and application*, 2021,**57**(18):1-12.

[4] 王正胜,连淑红. 中国手语翻译研究二十年述评[J]. 译苑新谭,2021,**2**(1):99-108.
WANG Z S, LIAN S H. A review of Chinese sign language translation in the past two decades [J]. *Yiyuan Xintan*, 2021,**2**(1):99-108.

[5] KAMAL S M, CHEN Y D, Li S Z, *et al*. Technical approaches to Chinese sign language processing: A review [J]. *IEEE Access*, 2019,**7**:96926-96935. DOI:10.1109/ACCESS.2019.2929174.

[6] RASTGOO R, KIANI K, ESCALERA S. Sign language recognition: A deep survey [J]. *Expert Systems with Applications*, 2021,**164**:113794-113820. DOI:10.1016/j.eswa.2020.113794.

[7] ZHENG L, LIANG B. Sign language recognition using depth images [C]// *Proceedings of International Conference on Control, Automation, Robotics and Vision*(*ICARCV*). New York: IEEE Press, 2016:1-6. DOI:10.1109/ICARCV.2016.7838572.

[8] OLIVEIRA M, SUTHERLAND A, FAROUK M. Two-stage PCA with interpolated data for hand shape recognition in sign language [C]// *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop*. New York: IEEE Press, 2016:1-4. DOI:10.1109/AIPR.2016.8010587.

[9] HASSAN M, ASSALEH K, SHANABLEH T. User-dependent sign language recognition using motion detection [C]// *Proceedings of International Conference on Computational Science and Computational Intelligence* (*CSCI*). New York: IEEE Press, 2016:852-856. DOI:10.1109/CSCI.2016.0165.

[10] ZHANG J H, ZHOU W G, LI H Q. A threshold-based HMM-DTW approach for continuous sign language recognition [C]// *Proceedings of International Conference on Internet Multimedia Computing and Service*. New York: ACM, 2014: 237-240. DOI:10.1145/2632856.2632931.

[11] VENUGOPALAN S, ROHRBACH M, DONAHUE J, *et al*. Sequence to sequence-video to text [C]// *Proceedings of IEEE International Conference on Computer Vision*. New York: IEEE Press, 2015:4534-4542. DOI:10.1109/ICCV.2015.515.

[12] YANG W W, TAO J X, YE Z F. Continuous sign language recognition using level building based on fast hidden Markov model [J]. *Pattern Recognition Letters*, 2016,**78**:28-35. DOI:10.1016/j.patrec.2016.03.030.

[13] PIGOU L, DIELEMAN S, KINDERMANS P J, *et al*. Sign language recognition using convolutional neural networks [C]// *Proceedings of Workshop at the European Conference on Computer Vision*. Cham: Springer International Publishing, 2014:1-6. DOI:10.1007/978-3-319-16178-5_40.

[14] KOLLER O, ZARGARAN O, NEY H, *et al*. Deep sign: Hybrid CNN-HMM for continuous sign language recognition [C]// *Proceedings of British Machine Vision Conference*. York: British Machine Vision Association, 2016:1-12. DOI:10.5244/c.30.136.

[15] XIAO Q K, ZHAO Y D, HUAN W. Multi-sensor data fusion for sign language recognition based on dynamic Bayesian network and convolutional neural network [J]. *Multimedia Tools and Applications*, 2019, **78**(11): 15335-15352. DOI:10.1007/s11042-018-6939-8.

[16] HUANG J, ZHOU W G, LI H Q, *et al*. Sign language recognition using 3D convolutional neural networks [C]// *Proceedings of International Conference on Multimedia and Expo*(*ICME*). New York: IEEE Press, 2015:1-6. DOI:10.1109/ICME.2015.7177428.

[17] SONG P P, GUO D, XIN H R, *et al*. Parallel temporal encoder for sign language translation [C]// *Proceedings*

of the IEEE International Conference on Image Processing (ICIP). New York: IEEE Press, 2019: 1915-1919. DOI: 10.1109/ICIP.2019.8803123.

[18] ZHOU M J, NG M, CAI Z X, *et al*. Self-attention-based fully-inception networks for continuous sign language recognition [OL/DB]. [2021-08-18]. *http://ecai2020.eu/papers/942_paper.pdf*.

[19] PU J, ZHOU W, LI H. Sign language recognition with multimodal features [C]// *Proceedings of Pacific Rim Conference on Multimedia*. Switzerland: Springer, 2016: 252-261.

[20] QIN W Y, MEI X, CHEN Y M, *et al*. Sign language recognition and translation method based on VTN [C]// *Proceedings of* 2021 *International Conference on Digital Society and Intelligent Systems* (*DSInS*). New York: IEEE Press, 2021: 111-115. DOI: 10.1109/DSInS54396.2021.9670588.

[21] RONGLAI Z, BRIAN M. C2SLR: Consistency-enhanced continuous sign language recognition[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE Press, 2022: 5131-5140. DOI: 10.1109/CVPR.2019.00429.

[22] LIN S C, RYABTSEV A, SENGUPTA S, *et al*. Real-time high-resolution background matting [C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE Press, 2021: 8762-8771. DOI: 10.1109/CVPR46437.2021.00865.

[23] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40** (4): 834-848. DOI: 10.1109/TPAMI.2017.2699184.

[24] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. *Science*, 2006, **313** (5786): 504-507. DOI: 10.1126/science.1127647.

[25] SUTSKEVER I, VINYALS O, LE Q. Sequence to sequence learning with neural networks [C]// *Proceedings of the 27th International Conference on Neural Information Processing Systems* (*NIPS'14*). New York: ACM, 2014(2): 3104-3112.

[26] WILLIAMS R J, ZIPSER D. A learning algorithm for continually running fully recurrent neural networks [J] *Neural Computation*, 1989, **1** (2): 270-280. DOI: 10.1162/neco.1989.1.2.270.

[27] DREUW P, NEIDLE C, ATHITSOS V, *et al*. Benchmark databases for video-based automatic sign language recoginition [C]// *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Paris: ELRA, 2008: 1-6.

[28] ADALOGLOU N, 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. DOI: 10.1109/TMM.2021.3070438.

[29] FORSTER J, SCHMIDT C, HOYOUX T, *et al*. RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus [C]// *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Paris: ELRA, 2012: 3785-3789.

[30] CAMGOZ N C, HADFIELD S, KOLLER O, *et al*. Neural sign language translation [C]// 2018 *IEEE/CVF cpnference on Computer Vision and Pattern Recognition*. New York: IEEE Press, 2018: 7784-7793. DOI: 10.1109/CVPR.2018.00812.

[31] HUANG J, ZHOU W G, ZHANG Q L, *et al*. Video-based sign language recognition without temporal segmentation [C]// *Proceedings of AAAI Conference on Artificial Intelligence*. Menlo Park: AAAI Press, 2018: 2257-2264. DOI: 10.1609/aaai.v32i1.11903.

[32] DUARTE A C. Cross-modal neural sign language translation [C]// *Proceedings of the 27th ACM International Conference on Multimedia*. New York: ACM, 2019: 1650-1654. DOI: 10.1145/3343031.3352587.

[33] CAMGOZ N C, HADFIELD S, KOLLER O, *et al*. SubUNets: End-to-end hand shape and continuous sign language recognition [C]// 2017 *IEEE International Conference on Computer Vision*. New York: IEEE Press, 2017: 3075-3084. DOI: 10.1109/ICCV.2017.332.

□