基于深度学习的词语级中文唇语识别

陈红顺[1]，陈观明[1,2]
（1.北京师范大学珠海分校 信息技术学院，广东 珠海 519087；2.珠海欧比特宇航科技股份有限公司，广东 珠海 519080）

摘 要：在无声或噪声干扰严重的环境下，或对于存在听觉障碍的人群，唇语识别至关重要。针对词语级中文唇语识别的问题，提出了 SinoLipReadingNet 模型，前端采用 Conv3D+ResNet34 结构用于时空特征提取，后端分别采用 Conv1D 结构和 Bi-LSTM 结构用于分类预测，并引入 Self-Attention、CTCLoss 对 Bi-LSTM 后端进行改进。最终在新网银行唇语识别数据集上进行实验，结果表明，SinoLipReadingNet 模型在识别准确率上明显优于中科院 D3D 模型，多模型融合的预测准确率达到了 77.64%，平均字错率为 21.68%。

关键词：唇语识别；ResNet；Bi-LSTM；CTCLoss；自注意力机制
中图分类号：TP391.4　　**文献标识码**：A　　DOI：10.16157/j.issn.0258-7998.222903

中文引用格式：陈红顺，陈观明. 基于深度学习的词语级中文唇语识别[J].电子技术应用，2022，48(12)：54-58.
英文引用格式：Chen Hongshun，Chen Guanming. Chinese word-level lip reading based deep learning[J]. Application of Electronic Technique，2022，48(12)：54-58.

Chinese word-level lip reading based deep learning

Chen Hongshun[1]，Chen Guanming[1,2]
（1.School of Information Technology，Beijing Normal University(Zhuhai)，Zhuhai 519087，China；
2.Zhuhai Orbita Aerospace Science & Technology Co.，Ltd.，Zhuhai 519080，China）

Abstract：Lip reading is crucial in the silent environment or environments with serious noise interference，or for people with hearing impairment. For word-level Chinese lip reading problem，SinoLipReadingNet model is proposed，the front end of which with Conv3D and ResNet34 is used to extract temporal-spatial features，and the back end of which with Conv1D and Bi-LSTM are used for classification and prediction respectively. Also，self-attention and CTCLoss are added to improve the back end with Bi-LSTM. Finally，the SinoLipReadingNet model is tested on XWBank lipreading dataset and results show that the prediction accuracy is significantly better than that of D3D model，the prediction accuracy and avrage CER of multi-model fusion reaches 77.64% and 21.68% respectively.

Key words：lip reading；ResNet；Bi-LSTM；CTCLoss；self-attention

0 引言

语言是人类沟通交流的主要方式，语音是人类语言交流的主要载体之一。在无声或噪声干扰严重的环境下，或对于存在听觉障碍的人群，如何利用通过嘴唇运动进行语言识别至关重要。唇语识别是指通过观察和分析人说话时唇部运动的特征变化，识别出人所说话的内容。唇语识别具有广阔的应用前景：在医疗健康领域，可以借助唇语识别辅助患有听力障碍的病人沟通交流[1]；在安防领域，人脸识别同时通过唇语识别以提高活体识别的安全性[2]；在视频合成领域，利用唇语识别可以合成特定人物讲话场景的视频[3]，或者合成高真实感的虚拟人物动画等。

唇语识别主要包含 4 个步骤[4]：人脸关键点检测与跟踪、唇语区域提取、时空特征提取和分类与解码。其中，时空特征提取和分类与解码是唇语识别的研究重点。近年来，随着大规模数据集[5]的出现，基于深度学习的方法可以自动抽取深层特征，逐渐成为唇语识别研究的主流方法[6]。如图 1 所示，基于深度学习的唇语学习方法将一系列的唇部图像送入前端以提取特征，然后传递给后端以进行分类预测，并以端到端的形式进行训练。

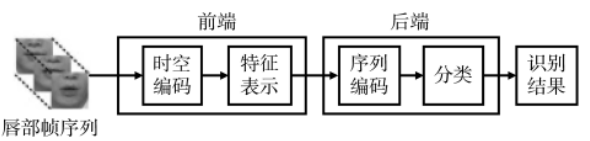


图 1 唇语识别深度学习方法框架[6]

由于卷积神经网络(CNN)具有强大的特征抽取能力，近年来逐渐成为唇语识别时空特征提取阶段的主流方法。2015 年，Noda[7]等将 CNN 模型用于日语单词的识别任务，并实验证明了 CNN 特征比传统 PCA 特征性能更

优。但由于传统的 CNN 对时序建模能力有限，越来越多的工作采用 3D CNN[8]作为时空特征提取器同时从时间和空间维度上提取信息。分类与解码阶段，深度学习通常采用循环神经网络(Recurrent Neural Network，RNN)，但原始 RNN 模型在训练过程中容易陷入梯度消失和梯度爆炸的问题，目前逐渐被长短时记忆网络(Long Short-Term Memory，LSTM)和门控循环单元(Gated Recurrent Unit，GRU)等结构取代。

Stafylakis[9]等首次采用 3D CNN 提取唇语视频的时空特征，并提出了 3D CNN+ResNet 的网络架构；2016 年，谷歌与牛津大学合作的 Assael 等人提出了时空卷积网络 STCNN(Spatio-Temporal Convolutional Neural Network)和 Bi-GRU 结合的 LipNet 模型，在 GRID 上句子准确率达到 93.4%[10]。国内关于中文唇部识别的研究起步较晚。2018 年，中科院 Yang Shuang 等人发布第一个开源的中文词语级唇语识别数据集 LRW-1000，其提出的 D3D 模型在 LRW-100 数据集上识别率达到 34.76%[11]；杨帆提出 ChLipNet 网络模型结构用于中文唇语识别，在其自建 CCTVDS 数据集上，句子识别准确率达到了 46.7%[12]；浙江大学 Zhao Ya 等人发布了第一个开源的中文句子级唇语识别数据集 CMLR，其提出的 CSSMCM 模型在 CMLR 上字错率达到了 32.48%[13]。本文针对中文词语唇语识别的问题，提出了 SinoLipReadingNet 模型，并在新网银行唇语识别数据集进行实验，预测准确率达到了 77.64%。

1 SinoLipReadingNet 模型

SinoLipReadingNet 模型前端采用 Conv3D+ResNet34 结构用于提取时空特征，后端分别采用 Conv1D 结构和 Bi-LSTM 结构用于分类预测，如图 2 所示。

1.1 Conv3D+ResNet34 前端

Conv3D 前端包括 64 个 5×7×7 大小的卷积核，经批处理规范化(Batch Normalization，BN)和校正线性激活单元 ReLU 后，再通过 MaxPool3D 最大池化层，从而降低了三维特征映射的空间大小。接着利用 ResNet34[14]网络对 Conv3D 前端提取的特征做进一步特征提取，最后利用全连接层(Fully Connected Layer，FC)输出 256 维的特征向量。

1.2 SinoLipReadingNet 模型后端

SinoLipReadingNet 后端分别采用 Conv1D 结构和 Bi-LSTM 结构。

Conv1D 结构由 2 个一维卷积 Conv1D 组成，主要在前期用来辅助训练前端模块，让模型学习时间维度上前后关联信息。模型训练时，先利用 Conv1D 后端辅助前端，当效果显著时立即停止训练并移除 Conv1D 后端模块，再利用 Bi-LSTM 后端模块完成端到端的训练。

Bi-LSTM 结构使用分别带有 256 个单元的双向 LSTM，进一步加强了模型对图片序列长时依赖的特征学习，再使用全连接层将输出变为长度为 313 的向量，最后经过 Softmax 层输出。

为进一步提高唇语识别准确率，本文引入 Self-Attention[15]、CTCLoss[16]对 Bi-LSTM 后端进行改进。

(1)Self-Attention 机制

如图 3(a)所示，在 LSTM 层后接入自注意力的编码层，通过将一组序列帧图片进行位置编码和帧的向量嵌入进行结合后进行自注意力计算，实现了帧与帧之间的关系计算，为唇语运动的信息进一步做长时记忆。

(2)CTCLoss

CTC 是序列标注问题中的一种损失函数，可自动学习解决输入与输出序列解码中标注对齐问题。传统序列标注算法需要每一时刻输入与输出符号完全对齐，而 CTC 扩展了标签集合，添加空元素。图 3(b)为在后端模块加入 CTCLoss 实现图片帧与中文标签对齐，从而实现基于词语级别的单字预测。

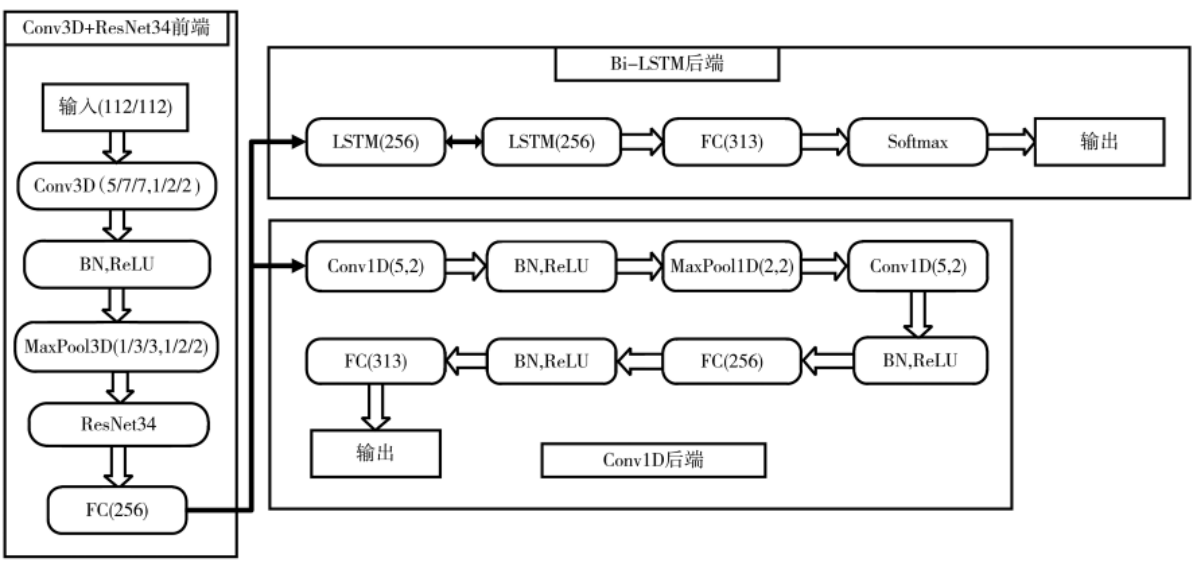
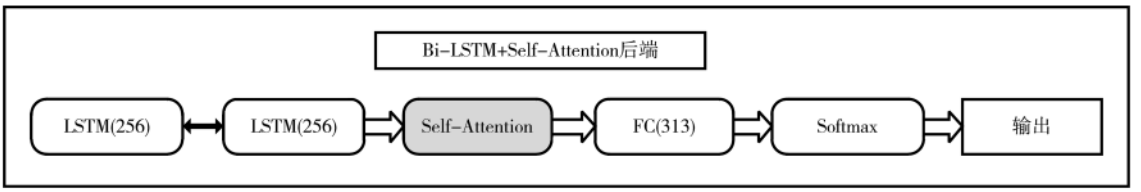
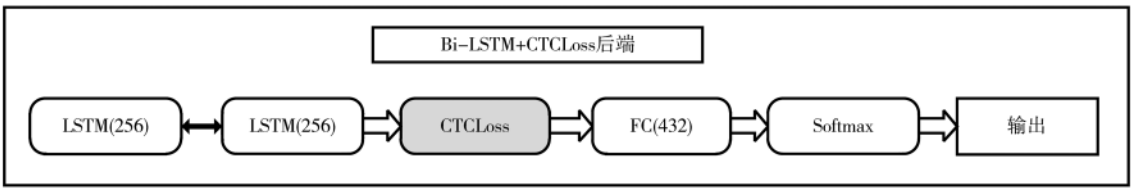


图 2 SinoLipReadingNet 模型



(a) Bi-LSTM+Self-Attention 后端



(b) Bi-LSTM+CTCLoss 后端

图 3 Bi-LSTM 后端改进模块

2 实验结果与分析

2.1 数据集及预处理

新网银行唇语识别数据集是由新网银行基于移动设备采集的正面视频而发行的词语级别唇语数据集，中文词语样本数为 313 个，包含 12 500 个样本。数据预处理时，首先利用 Dlib 库和 OpenCV 库 Adaboost 算法的人脸检测器对图像帧进行人脸检测和跟踪，将包含唇动信息的人脸图像裁剪出来。接着，基于 YOLOv3 算法[17]实现唇部区域提取，具体流程如下：(1)使用 labelImge 标注工具在人脸图像中手动标注出唇部位置，共计标注 2 000 张左右，如图 4 所示；(2)利用标注图像训练 YOLOv3 唇部检测器；(3)将训练好的目标检测器对新网银行唇语数据集裁剪出的人脸图像进行唇部检测和裁剪。最后将裁剪得到的唇部图像统一缩放为 112×112 的固定大小，如图 5 所示。

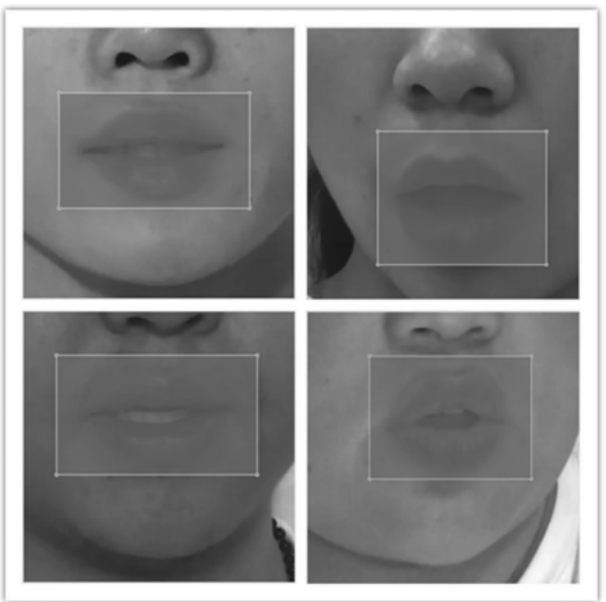


图 4 手动标注唇部位置

图 5 预处理结果

2.2 模型训练

模型的训练和测试在装有 2 个 GPU(型号为 NVDIA GeForce RTX 2080Ti 11 GB)的计算机上进行。模型训练时，训练集、测试集采用数据集的默认划分，并从训练集中预留 10%作为验证集。当模型输入为单通道时，图像帧被转换为灰度图像，并根据总体均值和方差进行归一化。训练期间，应用随机裁剪(±5 像素)和水平翻转来执行数据增强。

模型训练分两阶段：第一阶段利用 Conv1D 后端完成训练，先设置后端为 Conv1D 模块，训练 60 个 epochs 直到 CrossEntropyLoss 交叉熵损失函数趋于收敛，这时前端已较好学习到图片帧序列的时空特征，随即移除 Conv1D 模块，并冻结前端已经获得的权重参数；第二阶段使用 Bi-LSTM 后端模块(或 Bi-LSTM 后端改进模块)，先训练 5 个 epochs 使 Bi-LSTM 模块完成时序建模，然后解除前端权重参数的冻结，再端到端训练 60 个 epochs，获得最终最优的权重。

2.3 精度评价

为评价唇部识别效果，分别采用 Top-1 准确率和字

错率(CER)进行精度评价。

(1)Top-1 准确率

Top-1 准确率(Top-1 Accuracy)是指预测结果排名第一的词语与正确结果相符的准确率。

(2)字错率(CER)

中文一般用字符错误率(Character Error Rate,CER)来表示字错率。CER 计算公式为:

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

其中,S(substitution)表示替换的字符数目;D(deletion)表示删除的字符数目;I(insertion)表示插入的字符数目;N 表示参考序列中字符总数。所以,CER 值的范围是$[0,+\infty)$。

2.4 实验结果与分析

将训练好的模型在测试集进行唇语识别实验,部分预测结果见表 1、表 2。从表 1 可以看出,部分预测和正确结果存在差异,如“其中(qi zhong)”和“研究(yan jiu)”两个词存在韵母“ong”和“iu”相似的口型,“持之以恒(chi zhi yi heng)”和“这样(zhe yang)”同样存在“i”和“e”、“eng”和“ang”相似的口型,从而导致模型预测错误。表 2 采用 CTCLoss 实现图片帧序列进行对齐,实现了对词语中的单字预测。

表 1 部分预测结果

序列图像文件	正确结果	预测结果	是否预测正确
fcec09c1d75b5be5687167753c912dcb	持之以恒	这样	否
2b1e309b95cbf02bc4b0fa2599fdfaa5	社会	社会	是
0177862a9805cbea63b94adeac5563b8	其中	研究	否
a4e66834d55dbfbda6c3ecb7376f3b83	千家万户	千家万户	是
465b1efd73f363b0f7fe2af6170df4a1	领导	领导	是

注:模型为 Conv3D+ResNet34+Bi-LSTM(3C)。

表 2 部分 CTCLoss 预测结果

序列图像文件	正确结果	预测结果	CER/%
456f55e74055d724f512a7cb74ee70d7	组织	具织	50
aab024d5c09c70065eda21393ab44233	革命	人们	100
cd70a0e5a36b9f595127da859b9d497f	应该	一样	100
20b7dcc0c8d1f4c775a2ab3d0e5bdc05	完全	完全	0
bf17913262a6f38fd0cba4dd72439cd4	深恶痛绝	深恶痛绝	0
865daccae1ec8225911fe68182ff0f6d	市场	现象	100
f551ac567d3c4503c7478963de380f03	息息相关	息息相关	0
841033be4168defabd9e4b8bbd4b9e7f	第二	第样	50
390a377937f7b8231f933a0724936f5a	要求	根用	100
fc1502f8e0d2ffde02d3d00e7ce72c20	还有	看到	100

注:模型为 Conv3D+ResNet34+Bi-LSTM+CTCLoss(3C)。

为便于结果比较,在相同参数设置环境下,对中科院 D3D 模型[1]进行了实验。采用 2.3 节的方法对各模型的预测结果进行精度评价,其结果见表 3。可以看出,SinoLipReadingNet 模型在识别准确率上明显优于中科院 D3D 模型,说明 SinoLipReadingNet 模型能更好学习到图像帧序列的时空特征,捕捉到更敏感的嘴唇运动信息,从而提高了识别率。从表 3 可以看出,3 通道模型的预测准确率比单通道模型高 2%以上,平均字错率也有所降低,这主要是因为 3 通道图像比单通道图像提供了更多的信息;加入 Self-Attention 模块,预测准确率比 Conv3D+Resnet34+Bi-LSTM 有所提高、平均字错率有所下降,说明 Self-Attention 模块在一定程度上有助于提高识别准确率;采用 CTCLoss 实现图片帧序列自动对齐标签,预测准确为 59.35%,平均错字率为 32.01%,虽然在词语预测的准确率效果适中,不及 Conv3D+Resnet34+Bi-LSTM 模型,但实现了单字预测,为以后句子级唇语识别研究提供了思路。

表 3 分类准确率

模型	Top-1 准确率/%	平均 CER/%
D3D	44.6	-
Conv3D+Resnet34+Bi-LSTM(1C)	69.57	29.93
Conv3D+Resnet34+Bi-LSTM(3C)	71.81	27.83
Conv3D+Resnet34+Bi-LSTM+Self-attention(1C)	69.89	26.97
Conv3D+Resnet34+Bi-LSTM+Self-attention(3C)	72.00	27.41
Conv3D+Resnet34+Bi-LSTM+CTCLoss(1C)	56.95	33.56
Conv3D+Resnet34+Bi-LSTM+CTCLoss(3C)	59.35	32.01
多模型融合	77.64	21.68

注:1C 表示输入单通道,3C 表示输入 3 通道。

2.5 多模型融合

单一模型在处理问题时往往遇到模型泛化的瓶颈,多模型融合逐渐成为人们解决问题的选择手段。多模型融合方式主要分为结果多数表决、结果直接平均和结果加权平均等方式。鉴于实验环境及设备计算能力有限,本文采用结果多数表决方式实现多个模型(Conv3D+ResNet34+Bi-LSTM、Conv3D+ResNet34+Bi-LSTM+Self-attention、Conv3D+ResNet34+Bi-LSTM+CTCLoss)结果融合,部分预测结果见表 4。可以看出,多模型融合的预测准确率达到了 77.64%,平均字错率降低到 21.68%,均优于任何单一模型。这主要是因为多模型融合可以纠正部分识别错误,从而进一步提高识别率。

3 结论

针对端到端唇语识别的问题,本文构建了 SinoLip-ReadingNet 模型。该模型前端采用 Conv3D+ResNet34 结构,后端分别采用 Conv1D 结构和 Bi-LSTM 结构,并引入 Self-Attention 机制、CTCLoss 对 Bi-LSTM 后端进行改进。最终在新网银行唇语识别数据集进行实验,结果表明,SinoLipReadingNet 模型在识别准确率上明显优于中科院 D3D 模型,多模型融合的预测准确率达到了 77.64%,平均字错率为 21.68%。

参考文献

[1] TYE-MURRAY N,SOMMERS M S,SPEHAR B.Audiovisual integration and lipreading abilities of older adults with normal

表 4 多模型部分预测结果

序列	标签	Conv3D+ResNet34+Bi-LSTM(3C)	Conv3D+ResNet34+Bi-LSTM+Self-attention(3C)	Conv3D+ResNet34+Bi-LSTM+CTCLoss(3C)	多模型融合	CER/%
008267ea45003634bd39b52c83db4571	今天	今天	一点	今天	今天	0.0
0150383c3b743c039e0559b1144385de	甚至	形式	形式	知识	形式	100.0
024d214587b6524493c3041b11fe4f4a	革命	革命	人民	人民	人民	100.0
02f078d19186739721560a5d9f6d0f6b	任何	任何	任何	一何	任何	0.0
04111bdbad03a76f24d54bfd9b61f2b9	环境	关系	环境	环境	环境	0.0
04f57c0911876849182d17c2aef590d9	有的	有的	特点	规定	有的	0.0
0747142c7324c49858d550fc3edc1cb2	社会	成为	社会	社会	社会	0.0
0c10cfa84c7e60eec4106efeb418ff75	关系	关系	关系	环系	关系	0.0
0e0976b443d1f1142bdf85e60c6fe89e	虽然	虽然	虽然	虽然	虽然	0.0
13f68470e96d054671b5c7d997d19c29	领导	领导	领导	看到	领导	0.0
efe1039344fdf87304db7430fdc900d3	内容	内容	利用	内容	内容	0.0
f4928de6e41af294b4605fbaa96b070d	齐心协力	齐心协力	齐心协力	齐心协力	齐心协力	0.0
f69caeb77f993cca88bcc90f71a838dd	人们	他们	他们	他们	他们	50.0
f90d834c6a8d71665ca28161c3eb3d2a	方面	方面	方面	方面	方面	0.0
fae59dd75b9ffb723415edb321ff8410	历史	历史	因此	历史	历史	0.0
fbbdd350860dc06c6f22aec140ce06a3	有关	文化	有关	我关	文化	100.0
fd531a0b01ef48d984d506f06ba1a6cd	来之不易	来之不易	来之不易	来之不易	来之不易	0.0
fe86d5c89626ffb715dab97eb129c274	研究	系统	系统	研究	系统	100.0
ff4271c64d237cdd696935d56720951d	一样	一样	其他	第样	一样	0.0
ffc83f8fb4eade151837f76982dd835f	同时	同时	同时	同志	同时	0.0

and impaired hearing[J].Ear and Hearing,2007,28(5):656-668.

[2] 任玉强,田国栋,周祥东,等.高安全性人脸识别系统中的唇语识别算法研究[J].计算机应用研究,2017,34(4):1221-1225,1230.

[3] SUWAJANAKORN S,SEITZ S M,KEMELMACHER-SHLIZERMAN I.Synthesizing obama:learning lip sync from audio[J].ACM Transactions on Graphics,2017,36(4):1-13.

[4] 陈小鼎,盛常冲,匡纲要,等.唇读研究进展与展望[J].自动化学报,2020,46(11):2275-2301.

[5] 马金林,陈德光,郭贝贝,等.唇语语料库综述[J].计算机工程与应用,2019,55(22):1-13.

[6] 马金林,朱艳彬,马自萍,等.唇语识别的深度学习方法综述[J].计算机工程与应用,2021,57(24):61-73.

[7] NODA K,YAMAGUCHI Y,NAKADAI K,et al.Audio-visual speech recognition using deep learning[J].Applied Intelligence,2015,42(4):722-737.

[8] JI S,YANG M,YU K.3D convolutional neural networks for human action recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):221-231.

[9] STAFYLAKIS T,TZIMIROPOULOS G.Combining residual networks with LSTMs for lipreading[J].arXiv preprint arXiv:1703.04105, 2017.

[10] ASSAEL Y M,SHILLINGFORD B,WHITESON S,et al.LipNet:end-to-end sentence-level lipreading[J].arXiv preprint arXiv:1611.01599,2016.

[11] YANG S,ZHANG Y,FENG D,et al.LRW-1000:a naturally-distributed large-scale benchmark for lip reading in the wild[C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition(FG 2019).IEEE,2019:1-8.

[12] 杨帆.基于深度学习的唇语识别应用的研究与实现[D].成都:电子科技大学,2018.

[13] ZHAO Y,XU R,SONG M.A cascade sequence-to-sequence model for chinese mandarin lip reading[C]//MMAsia'19:Proceedings of the ACM Multimedia Asia.ACM,2019.

[14] HE K,ZHANG X,REN S,et al.Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.IEEE,2016:770-778.

[15] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//Advances in Neural Information Processing Systems,2017:5998-6008.

[16] AUVOLAT A,MESNARD T.Connectionist temporal classification:labelling unsegmented sequences with recurrent neural networks[C]//ICML'06:Proceedings of the 23rd International Conference on Machine Larning.ACM,2006:369-376.

[17] REDMON J,FARHADI A .YOLOv3:an incremental improvement[J].arXiv preprint arXiv:1804.02767,2018.

(收稿日期:2022-04-21)

作者简介:

陈红顺(1982-),男,博士,副教授,主要研究方向:大数据与人工智能。

陈观明(1996-),男,本科,主要研究方向:人工智能。

扫码下载电子文档

版权声明