DOI:10.19651/j.cnki.emt.2517838

基于多特征融合的唇语识别模型*

张甜愉 吕 博 周 蓉 王 琳 蒲梦杨
(华北电力大学控制与计算机工程学院 北京 102206)

摘 要:在单词级唇语识别研究中,使用三维卷积神经网络与残差网络的主流模型往往难以捕捉唇运动的几何动态,并且对细节依赖性高。为了缓解该问题,本文提出了一种基于多特征融合的端到端单词级唇语识别模型,该模型集成了像素级纹理细节特征、几何级轮廓形状特征和词边界特征,实现了从时间和空间、像素级与几何级等多个维度的特征融合。其中,纹理细节特征提供精细化的局部信息;轮廓形状特征反应唇部几何结构及动态变化;词边界特征则引导模型关注有效时间帧。此外,本文将空间通道注意力机制整合到 3D CNN 和 ResNet-18 中以增强纹理细节特征提取,并利用全局上下文网络对时空图卷积网络进行改进后将其引入模型以捕捉几何级轮廓形状特征。实验表明,输入为灰度视频时,本文模型在公开的大规模单词级唇语识别数据集 LRW 上的准确率达到 89.3%,相较于相同条件下单一或部分特征模型提升 1.3%~3.9%,且高于大部分现有模型,验证了所提模型的有效性;同时,实验发现,使用彩色视频作为输入时,模型准确率进一步提高,为 89.7%,验证了色彩信息对唇语识别的影响。

关键词:单词级唇语识别;多特征融合;像素级纹理细节特征;几何级轮廓形状特征;时空图卷积神经网络
中图分类号:TP391.4;TN911.73 **文献标识码**:A **国家标准学科分类代码**:520.6040

Lip-reading model based on multi-feature fusion

Zhang Tianyu Lyu Bo Zhou Rong Wang Lin Pu Mengyang
(School of Control and Computer Engineering, North China Electric Power University, Beijng 102206, China)

Abstract: Mainstream word-level lipreading models, based on three-dimensional convolutional neural networks and residual networks, struggle to capture the geometric dynamics of lip movements. Their reliance on pixel-level texture details makes them highly sensitive to noise and facial variations. To address these limitations, this paper proposes an end-to-end word-level lipreading model that integrates pixel-level texture detail features, geometry-level contour shape features, and word boundary features, achieving comprehensive multi-feature fusion across temporal, spatial, pixel-level, and geometric-level dimensions. The proposed model incorporates the spatial and channel squeeze-and-excitation mechanism into 3D CNNs and ResNet-18 to enhance texture feature extraction, while an improved spatial-temporal graph convolutional network integrates a global context network to strengthen global geometric relationships. Additionally, word boundary features further guide the model to focus on relevant temporal frames, reducing noise sensitivity. These features are fused and processed by a back-end temporal module to complete the recognition task. Experiments show that when the input is grayscale video, the accuracy of this paper's model on the publicly available large-scale word-level lip recognition dataset LRW reaches 89.3%, which is improved by 1.3%~3.9% compared with single or partial feature models under the same conditions, and higher than most existing models, which verifies the validity of the proposed model; at the same time, experiments find that, when colorful video is used as the input, the accuracy of the model further improves to 89.7%, verifying the effect of color information on lip recognition.

Keywords: word-level lipreading; multi-feature fusion; texture detail features; geometric contour features; spatio-temporal graph convolutional neural network

0 引 言

唇语识别,一般指视觉语音识别(visual speech recognition, VSR)[1],是利用视觉信息补充听觉信息的一种手段。一般情况下,不同的唇形序列对应着不同的说话内容,这为唇语识别提供了可能。唇语识别在听觉语音障

收稿日期:2025-01-07
*基金项目:国家自然科学基金(62301220)项目资助

碍者正常交流[2]、医疗保健、人机交互[3]、安防系统[4]、公共安全[5]、教学辅助、案件侦破[6]等领域有着广阔的应用前景。

唇语识别技术经历了多个发展阶段,从早期基于传统机器学习的手动特征提取方法,到近年来深度学习技术的广泛应用,模型的精度逐步提升。早期的唇语识别方法多依赖于隐马尔可夫模型(hidden Markov model, HMM)等传统机器学习算法,结合手工设计的特征进行识别[7]。虽然在一些简单场景下能够达到较好的识别效果,但由于过度依赖人工特征提取,难以适应复杂场景中的变化。深度学习的引入使得唇语识别的精度显著提升[8]。通过端到端的学习方式,模型能够自动从大量标注数据中学习到更复杂、更具辨识力的特征,从而避免了传统方法中人工特征选择的局限性。

根据识别任务的粒度不同,唇语识别被划分为句子级唇语识别和单词级唇语识别[9]。句子级唇语识别需要对完整语句进行建模,单词级唇语识别则旨在从短视频片段中识别单个单词。目前主流的单词级唇语识别神经网络模型主要由前端和后端两大模块组成[10-12],主要使用经过灰度处理后的视频帧作为输入。前端负责从视频中提取空间特征和短期时间特征,常用方法有三维卷积神经网络(three-dimensional convolutional neural network, 3D CNN)与残差网络(residual network, ResNet),通过卷积操作提取局部的视觉特征,注重纹理细节的表达。后端则更关注整个时间序列,对前端提取的短期时间特征进行进一步处理。Stafylakis 等[11]采用 3D CNN 与 ResNet-34 提取唇语视频的时空特征,使用双向长短期记忆网络进一步处理时间序列。之后的研究中,他们又将词边界信息引入模型[12],使得模型性能进一步提升。Martinez 等[13]使用 3D CNN 与 ResNet-18 提取唇部时空特征,并在后端使用多尺度时间卷积网络(multi-scale temporal convolutional network, MS-TCN)处理时间序列,以提高计算效率并增强时间建模能力。Ma 等[14]进一步改进了时间卷积网络的结构,提出了密集连接时间卷积网络(densely connected temporal convolutional networks, DC-TCN),通过引入密集连接和轻量级注意力机制,增强了模型对复杂时间动态的建模能力。这些对于后端时间序列处理模块的改进,也使得模型准确率进一步提高。

综上,目前单词级唇语识别领域针对后端时序网络的改进已取得显著进展。然而,前端特征提取主要集中于纹理细节,通常采用 3D CNN 与 ResNet 结合的方式,直接处理原始视频帧,通过卷积操作在局部空间和时间上提取特征,重点关注像素级的纹理细节。但这种方法往往忽略了嘴唇的几何轮廓信息,嘴唇的几何轮廓对于识别不同发音模式至关重要,因为它能够反映嘴唇开合、形变及发音动态。然而,由于当前主流方法依赖 3D CNN,其主要作用是提取局部感受野内的纹理细节信息,对于跨帧的几何结构建模能力较弱,导致模型难以充分利用嘴唇的几何运动模式,对重要嘴唇运动信息的利用不足。此外,该方法依赖于面部细节,容易受到光线等因素的影响,这在新说话者身上变得更加明显,特别是当他们的嘴唇运动与训练数据中的嘴唇运动明显不同时。这会阻碍模型在识别过程中充分理解这些特征的能力,从而影响模型的泛化性和准确性。

事实上,早期的唇语识别研究曾采用过嘴唇轮廓形状特征,主要依赖传统的手动特征提取方法。例如,基于 HMM 的方法曾被广泛用于唇语识别任务,其中嘴唇轮廓形状被作为关键特征输入模型,以捕捉语音发音的动态变化。然而,这类方法通常需要依赖于精细的手工设计特征,且对于不同的说话人或环境变化的适应性较差,导致其在复杂场景下的泛化能力不足。随着深度学习的兴起,基于数据驱动的端到端模型逐渐取代传统方法,嘴唇几何轮廓特征的研究也相对减少。

同时,现有研究通常采用灰度化预处理以减少计算开销并增强对唇部形状和运动的关注,但这也会导致色彩信息的丢失,而色彩可能携带与发音相关的额外信息,例如不同说话人的唇色变化或特定音素的视觉差异。因此,忽略色彩信息往往会限制模型对某些发音特征的区分能力。

为了缓解上述问题问题,本文从几何角度出发,提取轮廓形状特征,作为对纹理细节特征的有效补充。通过描述唇部轮廓的几何结构,该特征能够捕捉唇部的空间布局及其动态运动模式,作为一种高层次抽象,更直接地反映语音发声过程中唇部的几何动态特性。同时,相较于纹理特征,该特征对细节依赖性更小。与传统使用 HMM 手动提取几何特征的方法不同,本文引入时空图卷积神经网络(spatio-temporal graph convolutional network, ST-GCN),通过构建时空图结构,将数据的空间和时间特性建模为节点和边的关系,让模型自动学习轮廓形状特征,从几何层面捕捉信息。同时,增加彩色视频作为输入的实验,评估色彩信息对于唇语识别的影响。在此基础上,本文提出基于多特征融合的端到端唇语识别模型 MFLRM(multi-feature lip reading model)。该模型融合纹理细节特征与轮廓形状特征以及词边界特征,全面表征唇部运动模式。本文的主要工作如下:

1) 对 ResNet-18 进行改进,引入空间通道注意力机制(spatial and channel squeeze-and-excitation, SCSE)[15],帮助模型从唇部视频的纹理细节特征中自动选择重要的空间区域和通道信息,提高对唇形变化、口型细节等关键信息的捕捉能力,捕捉像素级特征。

2) 增加轮廓形状特征提取模块,对 ST-GCN 进行改进,将全局上下文增强网络(global context networks, GCNet)[16]集成到网络中,并将其引入单词级唇语识别模型,直接对关键点组成的轮廓图进行建模,构建唇部关键点间的空间关系,以便从几何层面捕捉唇形的动态变化,提取复杂轮廓特征。同时减小说话人依赖性。

3）将提取出的纹理细节特征、轮廓形状特征以及词边界特征融合，在当前公开的大规模唇语识别数据集 LRW（lip reading in the wild）上进行了实验，取得了良好的表现，识别准确率为 89.3%，高于同等条件下单一或部分特征模型以及现有大部分单词级唇语识别模型的准确率。同时，在输入中增加颜色通道后进行实验，得到 89.7%的识别准确率，验证了色彩信息对于唇语识别的重要性。

1　基于多特征融合的唇语识别模型

1.1　模型概述

本文所提出的 MFLRM 模型专注于单词级唇语识别任务，旨在通过分析视频中说话人的唇部运动来判断单词类别，包括前后端两部分，分别负责特征提取、融合和时序建模，模型框架如图 1 所示。

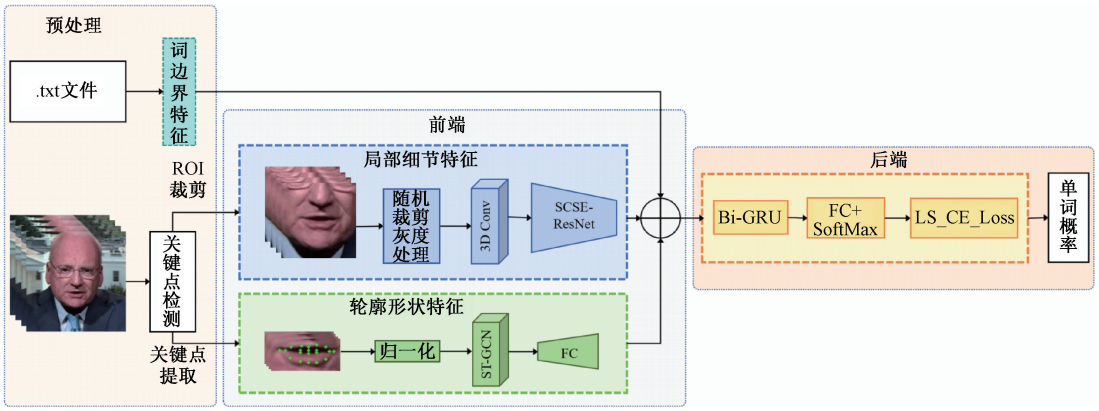


图 1　MFLRM 模型结构

Fig. 1　MFLRM model structure

首先，对数据进行预处理得到词边界特征、唇部区域帧以及关键点坐标信息。随后前端对特征进行提取和融合，包括像素级纹理细节特征提取、几何级轮廓形状特征提取以及特征融合。后端则对融合特征进行进一步处理，以捕获长时序的上下文信息。双向门控循环单元（bidirectional gated recurrent unit，Bi-GRU）是唇语识别网络后端的常用框架，由两个方向的门控循环单元组成，对输入时间序列在时间维度上按前向和反向进行处理，每个时间步的输出都是向前和向后隐藏状态的组合[17]。Bi-GRU 融合了反向循环机制，能够同时考虑前向和后向的隐藏层信息[18]。本文使用了三层堆叠结构的 Bi-GRU，隐藏层状态维度为 640，此外，在训练过程中使用了 0.2 的 dropout，以减少过拟合现象。最终，添加全连接层和 softmax 层进行分类，识别具体的单词。使用交叉熵损失函数对整个网络进行训练，同时引入标签平滑防止模型过度自信。

1.2　基于改进 ResNet-18 的纹理细节特征提取模块

纹理细节特征在唇语识别中至关重要，该特征具有很高的辨识力，在捕捉唇部细微变化和表达音素特征方面具有突出作用。纹理细节特征通过捕捉唇部在每帧中的细微运动，帮助模型区分相似的唇形变化。结合 3D CNN 和 ResNet 的网络是提取纹理细节特征的经典方法。

本文改进了 3D CNN 和 ResNet 结合的网络结构，在 ResNet-18 的残差块中加入了 SCSE 模块。SCSE 模块的核心思想是通过对特征图的空间维度和通道维度进行独立加权，使网络能够自动关注图像中最具辨识度的特征区域。具体而言，首先通过对归一化后的特征图进行处理，SCSE 模块对每个空间位置和每个通道的特征进行独立的加权操作。空间维度加权主要通过计算特征图的空间注意力，来突出图像中对识别有帮助的空间区域，通道维度加权则通过对每个通道的激励作用来增强对关键信息的关注。接着，对通道激励和空间激励进行元素级加法，再将两者的加权结果与原始输入进行融合。这一机制不仅能自动筛选出图像中最具判别力的特征区域和通道，还能减少冗余信息的干扰，从而使网络更加高效地聚焦于重要特征。

将 ResNet-18 的第一个核大小为 7×7，步长为 2 的二维卷积层修改为核大小为 5×7×7，步长为 1×2×2π 的三维卷积层，构成 3D Conv 层。残差块加入 SCSE 机制以突出重要特征，并进行全局平均池化。结构如图 2 所示。

模块的输入是原始视频经过预处理和数据加载过程后得到的 88×88 的视频帧。网络的输入形状为 $[N,T,C,H,W]$。N 是批大小，T 是时间序列，C 是通道数，H 是视频帧高度，W 是视频帧宽度。模块的输入是原始视频经过预处理和数据加载过程后得到的 88×88 的视频帧，初始值 $C_0=1$，$H_0=88$，$W_0=88$。模块网络细节如表 1 所示。

1.3　基于改进 ST-GCN 的轮廓形状特征提取模块

对于以唇部区域为中心的裁剪视频，唇部细节特征的提取通常基于像素级别的图像数据，难以捕捉唇部轮廓的几何级动态变化。此外，这种方法对细节的依赖性较高。为了缓解上述问题，本文将轮廓形状特征融合到模型中。

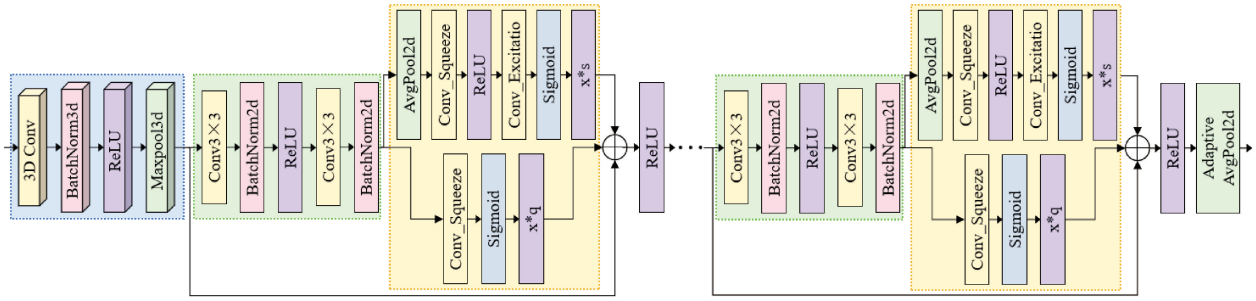


图 2 纹理细节特征提取网络结构

Fig. 2 Structure of texture detail feature extraction network

表 1 纹理细节特征提取模块网络细节

Table 1 Details for texture detail feature extraction

网络层	卷积核	输出维度
Conv3d MaxPool3d	5×7×7, 64, [1, 2, 2] 1×3×3, [1, 2, 2]	$N\times T\times 64\times\frac{H}{4}\times\frac{W}{4}$
Residual Conv2d	[3×3, 64]×2, 1	$N\times T\times 64\times\frac{H}{4}\times\frac{W}{4}$
Residual Conv2d	[3×3, 64]×2, 1	
Residual Conv2d	[3×3, 128]×2, 2	$N\times T\times 128\times\frac{H}{8}\times\frac{W}{8}$
Residual Conv2d	[3×3, 128]×2, 1	
Residual Conv2d	[3×3, 256]×2, 2	$N\times T\times 256\times\frac{H}{16}\times\frac{W}{16}$
Residual Conv2d	[3×3, 256]×2, 1	
Residual Conv2d	[3×3, 512]×2, 2	$N\times T\times 512\times\frac{H}{32}\times\frac{W}{32}$
Residual Conv2d	[3×3, 512]×2, 1	
AdaptiveAvgPool	—	$N\times T\times 512$

为了提取轮廓形状特征,本文添加改进 ST-GCN 网络。ST-GCN 最早应用于基于骨架的动作识别任务[19],通过将人体骨架中的关键点构建为图结构,捕捉关节之间的时空关系。本文借鉴了这一方法,根据人的五官组成设置标记点,将人脸部位由图像数据转化成图数据。68 人脸关键点模型是该领域最广泛使用的标准之一,能够详细地描绘出人脸上的五官,68 人脸关键点位置定义图如图 3 所示。

本文主要关注唇部区域,即第 49～68 个关键点。本文使用 dlib 工具包,基于 68 关键点检测模型,提取关键点。

提取关键点后,以唇部区域的 20 个关键点为图节点,以唇部结构和时间的自然连通性为边,构建无向时空图 $G=(V,E)$,节点集 $V=\{v_{ti}\mid t=1,\cdots,T,i=1,\cdots,M\}$ 的坐标值即为 ST-GCN 的输入,其中,t 表示当前时刻,i 表示当前节点,T 是视频帧数,M 是每帧中的节点数,对于唇部区域,$M=20$。边集 E 包含两个子集 E_S 和 E_F,边集 $E_S=\{v_{ti}v_{tj}\mid(i,j)\in M\}$ 包含相同帧中不同序号节点之间的连接边,边集 $E_F=\{v_{ti}v_{(t+1)i}\}$ 包含相邻帧的相同序号节点之间的连接边。

本文所提出的单帧唇部结构图构建方法基于唇部结构基础布局和连接策略。构建单帧唇部图结构步骤如下:

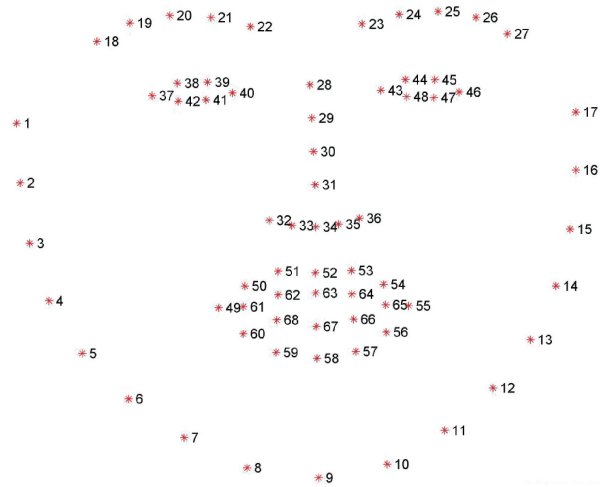


图 3 68 人脸关键点[20]

Fig. 3 68 facial landmarks[20]

1)划分关键点。将唇部区域的 20 个关键点根据人的生理构造划分为外轮廓和内轮廓,对应图 3 中所展示的唇部区域,则第 49～60 个关键点是外轮廓,第 61～68 个关键

点是内轮廓。

2)根据布局进行基础连接。将内轮廓的相邻节点进行连接,外轮廓的相邻节点进行连接得到基础邻接矩阵。

3)根据策略进行边的调整,得到新的邻接矩阵。基于最大跳数和膨胀率得到有效跳数,生成新的邻接矩阵,本文所设置的最大跳数为 2,膨胀率为 1,则有效跳数是{0,1,2},即图卷积只会在以下范围内进行特征传播:节点本身的特征;距离为 1 的邻居节点的特征,距离为 2 的邻居节点的特征。

4)归一化。得到新的邻接矩阵后,对其进行归一化处理。

对于每一帧,原始帧、关键点及唇部区域空间结构图示例如图 4 所示。原始视频帧如图 4(a)所示,从原始视频帧中提取出的人脸关键点如图 4(b)所示,基于这种连接方式,得到的唇部区域空间结构图如图 4(c)所示。

(a) 原始帧 (a) Original frame　(b) 关键点 (b) Landmarks　(c) 空间结构图 (c) Structure graph

图 4　唇部区域图示例

Fig. 4　Example of lip region

图卷积神经网络基于邻居聚合(neighborhood aggregation)或消息传递(message passing)机制,能够针对不同的节点做出不同的卷积操作,适用于非欧几里得数据[21]。时间卷积神经网络利用一维卷积代替传统循环神经网络(recurrent neural network, RNN)中的递归操作,通过因果卷积(causal convolution)和膨胀卷积(dilated convolution)实现对时序数据的处理。ST-GCN 结合了 GCN 和 TCN 的特点,能够同时捕捉空间和时间上的依赖关系。

本研究设计了一种改进的 ST-GCN 网络结构,将 GCNet 融入模型中,以增强模型的全局上下文感知能力。通过注意力机制对输入特征进行全局建模,动态计算唇部关键点的权重,并依据这些权重调整全局特征的贡献,补充局部时间与空间特征的不足,从而提升模型对唇部动态变化的建模能力及全局动态几何形状的捕捉效果。

基于 ST-GCN 的轮廓形状特征提取模块的具体结构如图 5 所示,包含 10 个层,每层均由图卷积层、时间卷积层和激活函数构成。为提高特征传递效果,第 2～10 层引入残差连接,同时在每层时间卷积操作后添加丢弃率为 0.5 的 dropout 层,以减少过拟合现象。

图卷积层中 Conv1×1 的步长为 1,图卷积进行张量计算,卷积核大小为 3。时间卷积层初步提取时间信息并保留每一个时间步中的细节信息,捕捉细微动作变化。每个层依次递增或保持输出通道数,以处理更复杂的特征。经

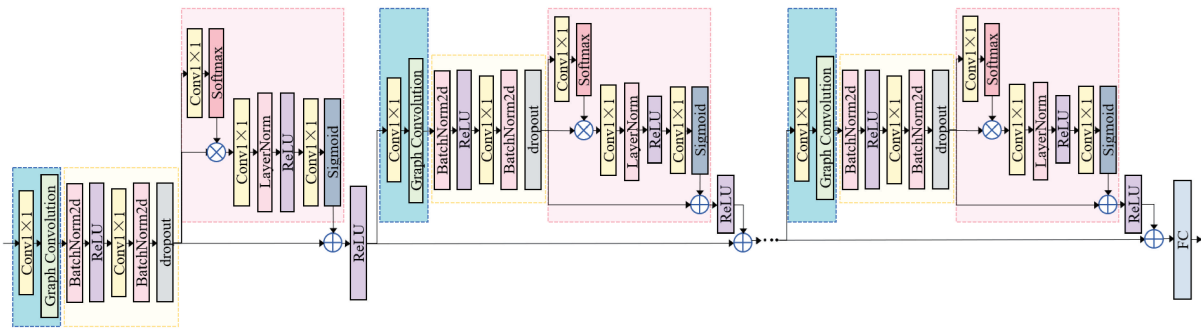


图 5　轮廓形状特征提取网络结构图

Fig. 5　Structure of geometric contour feature extraction network

过时空图卷积网络处理后,通过全连接层对提取出的特征进一步处理,得到最终的轮廓形状特征。

模块网络细节如表 2 所示。输入形状为 $[N,C,T,M]$。C 的初始值是 2,即横纵坐标。M 的值是 20,即唇部区域节点数。

1.4　特征融合

特征融合阶段,首先将轮廓形状特征调整为 $N\times T\times 512$,词边界特征调整为 $N\times T\times 1$。接着将轮廓形状特征与纹理细节特征沿特征维度进行拼接,形成初步的联合特征表示。之后将融合后的特征与词边界特征再度拼接,得到最终的融合特征。

融合特征同时包含了纹理细节特征,唇部轮廓形状特征,词边界特征。其中,纹理细节特征主要捕捉唇部的微小细节和纹理细节,唇部轮廓形状特征反映唇部的整体几何形状和动态变化,帮助理解唇形的几何关系和唇动信息。词边界特征提供了唇部动作的时间关系,明确了视频中包含单词的帧,使模型更加关注有效信息。融合特征综合考虑像素级特征与几何级特征,时间特征与空间特征,获得的信息更加全面,随后作为后端模块的输入,为模型提供更加丰富的信息。

表 2 轮廓形状特征提取模块网络细节

Table 2 Details for geometric contour feature extraction

网络层	输入维度	输出维度
ST-GCN	$N\times2\times T\times20$	$N\times64\times T\times20$
ST-GCN	$N\times64\times T\times20$	$N\times64\times T\times20$
ST-GCN	$N\times64\times T\times20$	$N\times64\times T\times20$
ST-GCN	$N\times64\times T\times20$	$N\times64\times T\times20$
ST-GCN	$N\times64\times T\times20$	$N\times128\times T\times20$
ST-GCN	$N\times128\times T\times20$	$N\times128\times T\times20$
ST-GCN	$N\times128\times T\times20$	$N\times128\times T\times20$
ST-GCN	$N\times128\times T\times20$	$N\times256\times T\times20$
ST-GCN	$N\times256\times T\times20$	$N\times256\times T\times20$
ST-GCN	$N\times256\times T\times20$	$N\times256\times T\times20$
Mean Aggregation	$N\times256\times T\times20$	$N\times256\times T$
FC	$N\times256\times T$	$N\times512\times T$

2 实验结果与分析

2.1 数据集

本文所使用的数据集是 LRW 数据集,它是唇语识别研究领域最被广泛使用的数据集之一。该数据集包含 500 个单词,涵盖了广泛的话题和情境。整个数据集的划分如下:1)训练集。包含 500 个单词,每个单词有 800～1 000 个视频样本。2)验证集。包含 500 个单词,每个单词有 50 个视频样本。3)包含 500 个单词,每个单词有 50 个视频样本。数据集总计 538 766 个视频样本,每个样本的长度均为 29 帧,1.16 s。该数据集中视频序列均来源于英国广播电视台的电视节目,提供了一个自然且不受约束的环境,真实地反映了现实中的场景,视频中包含复杂的变化条件,包括光照、说话人姿态、语速、视频分辨率等,是分布自然而极具挑战的唇读数据集[22]。每个视频都包含一个目标单词,与视频配套的元数据文件详细记录了每个目标单词的持续时间,这为确定单词词边界信息提供了依据。

LRW 数据集包含了不同年龄段的说话者,以及多样的面部特征、头部姿态、表情、头部朝向、说话背景和光照条件,全面地涵盖了现实生活中的各种场景,因此,选用该数据集进行训练,可以提高模型的泛化能力。LRW 官方网站展示的部分视频中的帧如图 6 所示。

图 6 LRW 数据集部分视频中的帧展示

Fig. 6 Frame samples from part videos in LRW dataset

在实验中,使用分类准确率 $Accuracy(ACC)=\frac{N_{correct}}{N_{total}}$ 作为评估指标。

2.2 实验设置

本文将视频转换为连续的帧图片,每个视频 29 帧。对于每个视频,先使用 dlib 工具包检测出 68 个人脸关键点,根据嘴唇边界关键点位置定位唇部区域,裁剪出以四个边界点中心为中心的 96×96 的区域作为前端纹理细节特征提取模块的输入,并保存关键点坐标信息作为前端轮廓形状特征提取模块的输入。预处理过程,对于视频的处理如图 7 所示。

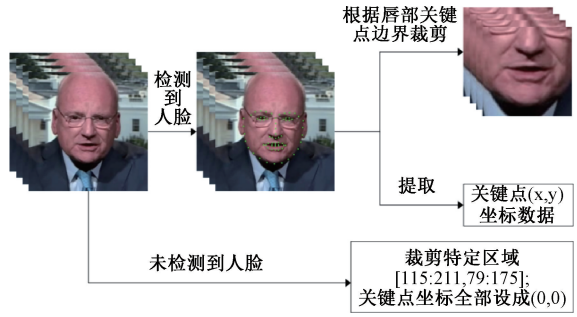


图 7 预处理过程图

Fig. 7 Preprocessing process diagram

在唇语识别过程中,非发音帧(视频开始或结束时的动态帧)会干扰模型的判别。词边界特征通过标记哪些帧位于发音时间段,帮助模型忽略背景帧,专注于实际发音时刻,从而减少误判,提升对词汇的识别准确率。从与视

频对应的文本文件中提取词边界特征的过程如下：

1)创建一个包含 29(视频帧数)个“0”的张量；

2)从与视频相应的文本中提取出该视频中单词的持续时间 duration；

3)根据 duration 确定开始帧 start 和结束帧 end，将这中间的帧的值设为“1”表示该帧处于单词发音期间，其余保持“0”表示该帧位于单词发音开始前或结束后。

数据加载过程中，对于视频帧数据，对预处理阶段得到的 96×96 的区域进行灰度处理，随后在训练阶段进行随机裁剪和一定概率的水平翻转，得到 88×88 大小的区域。在验证和测试阶段，使用中心裁剪，得到 88×88 大小的区域。这是一种常见的数据增强方式，有助于提高模型的鲁棒性和泛化能力，并保证评估结果的一致性和公平性。对于关键点数据，选择索引 48～67 部分(索引从 0 开始)，即唇部区域的关键点，随后根据均值和标准差进行归一化处理。

本文基于 Pytorch 框架实现了所提出的算法，实验在 Ubuntu22.04，NVIDIA A100-PCIE-40GB，Python 3.9.19 平台上训练和测试。训练模型时，使用交叉熵损失函数并引入标签平滑以提高模型的泛化能力，使用 MixUp[23]作为额外的数据增强方法；联合使用 Adam 优化器和余弦退火学习率调度器，根据批次大小和轮次动态调整学习率，如式(1)所示。

$$lr(t) = \eta_{\min} + \frac{1}{2}\left(\frac{batch\_size}{32}\eta_0 - \eta_{\min}\right)\left(1 + \cos\left(\frac{EP_{cur}}{EP_{\max}}\pi\right)\right) \tag{1}$$

其中，$lr(t)$ 是当前轮次的学习率，$batch\_size$ 是批次大小，本文训练过程中设置为 250，η_0 是初始学习率，设为 1×10^{-4}，$\eta_{\min}$ 是最小学习率，设为 5×10^{-6}，EP_0 是当前轮次数，$EP_{\max}$ 是最大轮次数，设为 150。

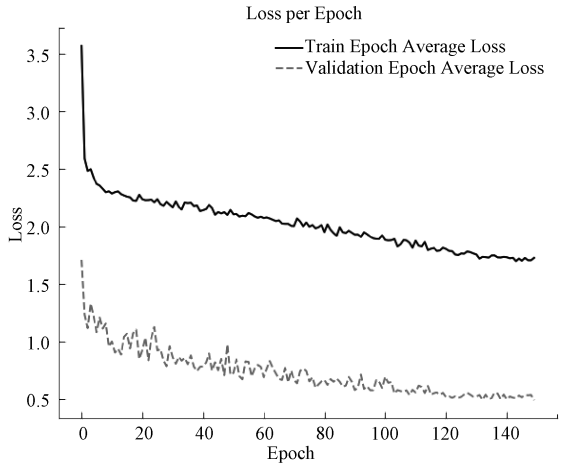
2.3　实验结果

在 LRW 数据集上，训练 150 轮的情况下，得到的最佳模型在测试集上测试，输入为灰度视频时，达到了 0.893 04 的准确率。

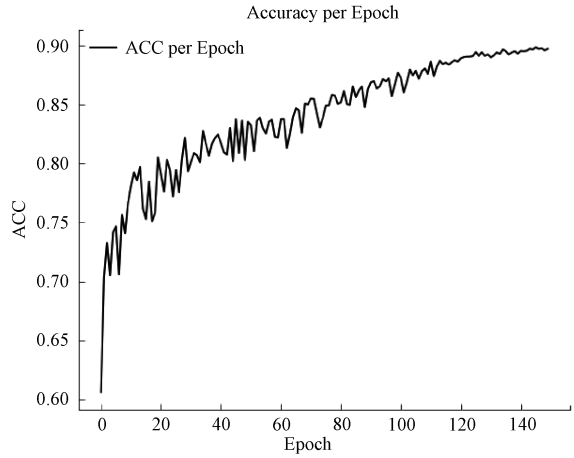
图 8(a)展示了训练过程中损失值(Loss)的变化过程，其中上方和下方曲线分别代表训练集和验证集上的损失变化。图 8(b)展示了训练过程中验证集上准确率(Accuracy)的变化过程。

由图 8 可以看出，随着训练次数的增加，训练集和验证集的损失值均呈现逐步下降的趋势，并逐渐收敛；随着训练轮次的增加，准确率不断上升并逐步稳定，最终达到 89.304%的准确率，说明训练过程中，模型拟合情况良好，训练过程有效。

图 9 展示了在测试集上的混淆矩阵，其中横轴表示模型预测的单词，纵轴表示实际的单词，矩阵中的颜色深浅反映了模型在各单词上的识别频次。整体上，混淆矩阵呈现出一条明显的对角线，说明模型绝大部分识别结果与实际值一致，具有良好的识别能力。



(a) 损失变化图
(a) Loss function curve



(b) 准确率变化图
(b) Accuracy curve

图 8　训练过程中指标变化

Fig. 8　Metric changes during training

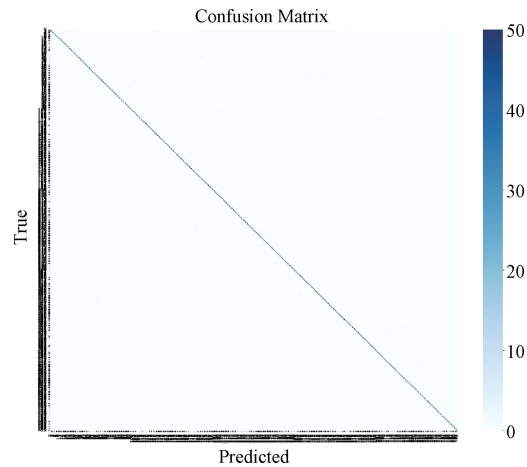


图 9　测试集上的混淆矩阵

Fig. 9　Confusion matrix on the test set

为进一步验证所提出的多特征融合模型的效果,对模型进行消融实验,在其他参数相同的情况下进行实验,对比引入 ST-GCN 网络前后模型识别的准确率;改进 ResNet-18 前后模型识别准确率;改进 ST-GCN 前后模型识别的准确率,得到的结果如表 3 所示。

表 3 消融实验结果

Table 3 Ablation study results

前端模型	准确率/%
3D CNN+ResNet-18	88.388
3D CNN+ResNet-18+ST-GCN	89.168
3D CNN+SCSE-ResNet-18+ST-GCN	89.252
3D CNN+SCSE-ResNet-18+GC-ST-GCN	**89.304**

引入 ST-GCN 后,模型识别准确率有了 0.780%的提升,效果显著,证明了轮廓形状特征对于唇语识别的积极作用;对残差块进行改进,即加入通道空间压缩激励机制,以及在 ST-GCN 加入 GCNet 后,模型准确率分别有略微提升,证明了模型改进的有效性。

对比不同特征融合情况下的识别准确率,结果如表 4 所示。

表 4 特征融合前后对比

Table 4 Comparison before and after feature fusion

特征	准确率/%
仅纹理细节特征	85.340
纹理细节+轮廓形状特征	86.732
纹理细节+词边界特征	88.388
纹理细节+轮廓形状+词边界特征(改进前)	89.168
纹理细节+轮廓形状+词边界特征(改进后)	**89.304**

从表 4 中可以看出,在纹理细节特征的基础上,额外融合轮廓形状特征后,模型识别准确率提高 1.392%;额外融合词边界特征后,提高 3.048%;同时融合三种特征后,由于 ST-GCN 提取的轮廓形状特征中包含时序动态变化信息,与词边界特征在一定程度上存在信息重叠,提升效果未能完全叠加。尽管如此,模型识别准确率仍提高 3.828%,进一步改进网络后提高 3.964%,显著高于单一或部分特征融合模型。进一步证明了多特征融合对于唇语识别的有效性。

为探究色彩信息对唇语识别的影响,进行对比实验,使用相同条件下未经灰度处理的视频帧作为输入。原来的灰度视频输入(单通道)被替换为包含 RGB 三个颜色通道的彩色视频。具体而言,纹理细节特征提取模块中通道数 C 的初始值由 1 改为 3。此外,由于彩色通道的增加,网络需要处理更多的信息,为确保训练过程的稳定性和有效性,对学习率初始值 η_0 进行适当调整,改为 5.5×10^{-5}。输入从灰度视频改为彩色视频后,准确率达到 89.756%,较使用灰度视频作为输入时高 0.452%,说明色彩信息对于唇语识别有积极作用。

将本文方法与现有方法进行比较,这些方法的输入均为经过灰度处理的数据,结果如表 5 所示。

表 5 与其他模型对比结果

Table 5 Comparison with existing methods

研究者	特征	准确率/%
Feng 等[10]	纹理细节+词边界	88.4
Stafylakis 等[11]	纹理细节	83.0
Stafylakis 等[12]	纹理细节+词边界	88.0
Martinez 等[13]	纹理细节	85.3
Luo 等[24]	纹理细节	83.5
Tian 等[25]	纹理细节	88.3
本文(灰度视频作为输入)	纹理细节+词边界+轮廓形状	89.3
本文(彩色视频作为输入)	纹理细节+词边界+轮廓形状	**89.7**

可以看出,使用灰度视频作为输入时,本文模型相较于仅提取纹理细节特征或同时融合纹理细节特征与词边界特征的模型有 0.9%~6.3%的提升,识别的准确率高于单一或部分特征的代表模型,证明了本模型具有较好的识别性能。使用彩色视频作为输入时,模型相较于使用灰度视频作为输入且仅提取纹理细节特征或同时融合纹理细节特征与词边界特征的模型有 1.3%~6.7%的提升,较使用灰度视频作为输入时的本文模型有 0.4%的提升,同时验证了色彩信息对于唇语识别的有效性。

3 结 论

本文针对复杂情况下唇语识别特征提取存在缺陷,导致识别准确率不高的问题,提出了一种基于多特征融合的唇语识别模型,对前端残差块进行改进,增强模型对重要特征的提取能力。又对 ST-GCN 进行改进,并将其引入当前主流模型中,从几何级唇部轮廓图结构角度捕获唇动信息,提高了模型捕捉几何关系的能力,降低了模型对于说话人的依赖性。模型融合了视频帧包含的像素级纹理细节特征、关键点包含的几何级轮廓形状特征、以及词边界特征,综合考虑时间与空间、像素级与几何级信息,最终提高了唇语识别的准确率。在 LRW 数据集上的实验数据表明,多特征融合和色彩信息对单词级唇语识别准确率提高有积极作用,且本文所提出的模型在复杂条件下的识别准确率高于现有的单词级唇语识别代表算法。为提高模型的可迁移性和泛化能力,今后将考虑在以下方面进一步研究扩展:利用其他数据集(如 LRW-1000 数据集)训练模型;利用 DC-TCN、MS-TCN 等作为后端进行实验,对比不同时序处理网络对模型的影响。

参考文献

[1] 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.
MA J L, ZHU Y B, MA Z P, et al. Review of deep learning methods for lip recognition[J]. Computer Engineering and Applications, 2021, 57(24): 61-73.

[2] 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.
CHEN X D, SHENG CH CH, KUANG G Y, et al. The state of the art and prospects of lip reading[J]. Acta Automatica Sinica, 2020, 46(11): 2275-2301.

[3] KUANG W B, LUO W P. Based on STM32 of CNN speech keyword command recognition system[J]. Instrumentation, 2023(1): 17-22.

[4] LIU Y J, JOURABLOO A, LIU X M. Learning deep models for face anti-spoofing: Binary or auxiliary supervision[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[5] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don′t lie: A generalisable and robust approach to face forgery detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[6] AFOURAS T, CHUNG J S, SENIOR A, et al. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8718-8727.

[7] ZHANG Z L, QU W L, LIU F M. Review of the Lip-reading recognition[C]. International Conference on Software Engineering and Service Science, 2014.

[8] NODA K, YAMAGUCHI Y, NAKADAI K, et al. Lipreading using convolutional neural network[C]. Interspeech, 2014.

[9] PALECEK K. Utilizing lipreading in large vocabulary continuous speech recognition[C]. Speech and Computer: 19th International Conference, 2017.

[10] FENG D L, YANG SH, SHAN SH. An efficient software for building lip reading models without pains[C]. IEEE International Conference on Multimedia & Expo Workshops, 2021.

[11] STAFYLAKIS T, TZIMIROPOULOS G. Combining residual networks with LSTMs for lipreading[C]. International Speech Communication Association Conference on Speech Communication and Technology, 2017.

[12] STAFYLAKIS T, KHAN M H, TZIMIROPOULOS G. Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs[J]. Computer Vision and Image Understanding, 2018, 176-177: 22-32.

[13] MARTINEZ B, MA P CH, PETRIDIS S, et al. Lipreading using temporal convolutional networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[14] MA P CH, WANG Y J, SHEN J, et al. Lip-reading with densely connected temporal convolutional networks[C]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2021.

[15] ROY A G, NAVAB N, WACHINGER C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks[C]. Medical Image Computing and Computer Assisted Intervention, 2018.

[16] CAO Y, XU J R, LIN S, et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond[C]. IEEE/CVF International Conference on Computer Vision, 2019.

[17] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation[C]. Empirical Methods in Natural Language Processing, 2014.

[18] 车鲁阳, 高军伟, 付惠琛. 基于多通道融合的滚动轴承剩余寿命预测[J]. 电子测量与仪器学报, 2023, 37(12): 225-233.
CHE L Y, GAO J W, FU H CH. Residual life prediction of rolling bearings based on multi-feature fusion[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(12): 225-233.

[19] 刘雨萌, 桑海峰. 基于关键帧定位的人体异常行为识别[J]. 电子测量与仪器学报, 2024, 38(3): 104-111.
LIU Y M, SANG H F. Human abnormal behavior recognition based on keyframes localization[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(3): 104-111.

[20] SAGONAS C, TZIMIROPOULOS G, ZAFEIRIOU S, et al. 300 faces in-the-Wild challenge: The first facial landmark localization challenge[C]. IEEE International Conference on Computer Vision Workshops, 2013.

[21] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]. International Conference on Learning Representations, 2017.

[22] CHUNG J S, ZISSERMAN A. Lip reading in the wild[C]. Asian Conference on Computer Vision, 2016.

[23] ZHANG H Y, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond empirical risk minimization[C].

International Conference on Learning Representations, 2018.

[24] LUO M SH, YANG SH, SHAN SH G, et al. Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading [C]. IEEE International Conference on Automatic Face and Gesture Recognition, 2020.

[25] TIAN W D, ZHANG H S, PENG CH, et al. Lipreading model based on whole-part collaborative learning [C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2022.

作者简介

张甜愉(通信作者),硕士研究生,主要研究方向为计算机视觉、图像处理。

E-mail:zty288245@163.com

吕博,硕士研究生,主要研究方向为数据增强、GAN、室内定位。

周蓉,博士,副教授,主要研究方向为计算机视觉、机器学习。

王琳,硕士研究生,主要研究方向为计算机视觉、图像处理。

蒲梦杨,博士,讲师,研究方向为计算机视觉、图像处理。