文章编号:2095-6134(2018)01-0109-09

一种基于 long short-term memory 的唇语识别方法*

马 宁[1,2†],田国栋[2],周 曦[2]

(1 中国科学院大学,北京 100049;2 中国科学院重庆绿色智能技术研究院,重庆 400714)

(2016年11月23日收稿;2017年3月15日收修改稿)

Ma N, Tian G D, Zhou X. A lip-reading recognition approach based on long short-term memory[J]. Journal of University of Chinese Academy of Sciences, 2018,35(1):109-117.

摘 要 唇动视觉信息是说话内容的重要载体。受嘴唇外观、背景信息和说话习惯等影响,即使说话者说相同的内容,唇动视觉信息也会相差很大。为解决唇语视觉信息多样性的问题,提出一种基于 long short-term memory(LSTM)的新的唇语识别方法。以往大多数的方法从嘴唇外表信息入手。本方法用嘴唇关键点坐标描述嘴唇形变信息作为唇语视频的特征,它具有类内一致性和类间区分性的特点。然后利用 LSTM 对特征进行时序编码,它能学习具有区分性和泛化性的空间-时序特征。在公开的唇语数据集 GRID、MIRACL-VC 和 OuluVS 上对本方法做了针对分割的单词或短语的说话者独立的唇语识别评估。在 GRID 和 MIRACL-VC 上,本方法的准确率比传统方法至少高 30%;在 OuluVS 上,本方法的准确率接近于最优结果。以上实验结果表明,本文提出的基于 LSTM 的唇语识别方法有效地解决了唇语视觉信息多样性的问题。

关键词 唇语识别;long short-term memory;计算机视觉

中图分类号:TP391 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2018.01.015

A lip-reading recognition approach based on long short-term memory

MA Ning[1,2], TIAN Guodong[2], ZHOU Xi[2]

(1 *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*;
2 *Chongqing Institute of Green and Intelligent Technology*, *Chinese Academy of Sciences*, *Chongqing* 400714, *China*)

Abstract Visual speech information is the important carrier of conversation. However, visual speech informations from different speakers are different due to various appearances of lips, various backgrounds, and various talking ways even the content of the conversation is the same. To address the problem of variety of visual speech information, we propose a new approach for lip-reading recognition based on long short-term memory (LSTM). We compute the positions of lip landmarks which describe the dynamic information of the shape as the features of the lip-reading video, and it has the characteristics of within-class consistency and between-class distinctiveness. Then we use LSTM to encode temporal information, and it learns spatio-temporal features which have the ability

*国家自然科学基金(61472386,61502444)和中国科学院战略性先导科技专项子课题(XDA06040103)资助

†通信作者,E-mail:csningm@gmail.com

of discrimination and generalization. Our approach is evaluated on three public databases (GRID, MRIALC, and OuluVS) for lip-reading recognition of isolated words or phrases in speaker independent experiments. On GRID and MRIALC, the accuracy of our approach is more than 30% highter than that of the conventional approach. On OuluVS, the accuracy of our approach is comparable to state of the art. The experiment results indicate that our lip-reading recognition approach solves the problem of variety of visual speech information effectively.

Keywords lip-reading recognition; long short-term memory; computer vision

人在与人的交流中,除了听和说,还需要“察言观色”。在嘈杂环境下,人要更加依赖观察对方的嘴唇运动来判断其说话内容;例如聋哑人群与其他人的交流必须依赖于对方的唇动视觉信息。已有研究证明,唇动视觉信息是说话内容的重要载体[1]。目前唇语识别应用场景主要有两个,一是用于辅助语音识别系统,以提高其在嘈杂环境等不利情况下的识别性能;二是用于安防系统,扮演活体检测的角色,配合其他生物特征识别技术来实现高安全性。Graves 将序列学习分为 3 类[2],本文研究的唇语识别属于段分类(segment classification),因为本文中的唇语视频是将说话内容限制为有限个短语或单词,对应某个短语或单词的唇语视频起止点是已知的。

相较于语音识别,唇语识别更加困难,因为不同人在不同环境和不同时刻,即使说话内容相同,受说话者不同的嘴唇外观、背景信息和说话习惯的影响,其唇动视觉信息也会相差很大。Zhou 等[3]针对如何在唇语识别中提取更好的特征提出两个需要解决的问题:1)如何抑制因说话者不同带来的视觉信息多样性的特征,即如何提取与说话者无关而与说话内容相关的视觉特征;2)如何对视觉特征做时序编码。为解决这两个问题,我们探索了基于 LSTM 的唇语识别方法,如图 1 所示,第一步是用人脸关键点检测技术对视频每一帧检测嘴唇关键点,用嘴唇关键点的坐标描述该帧的嘴唇的形状信息,并将其作为每一帧的特征,第二步是将视频所有帧的特征连接起来输入 LSTM 学习空间 - 时序特征,输出识别的结果。我们试图通过该方法提供一个新的唇语识别流

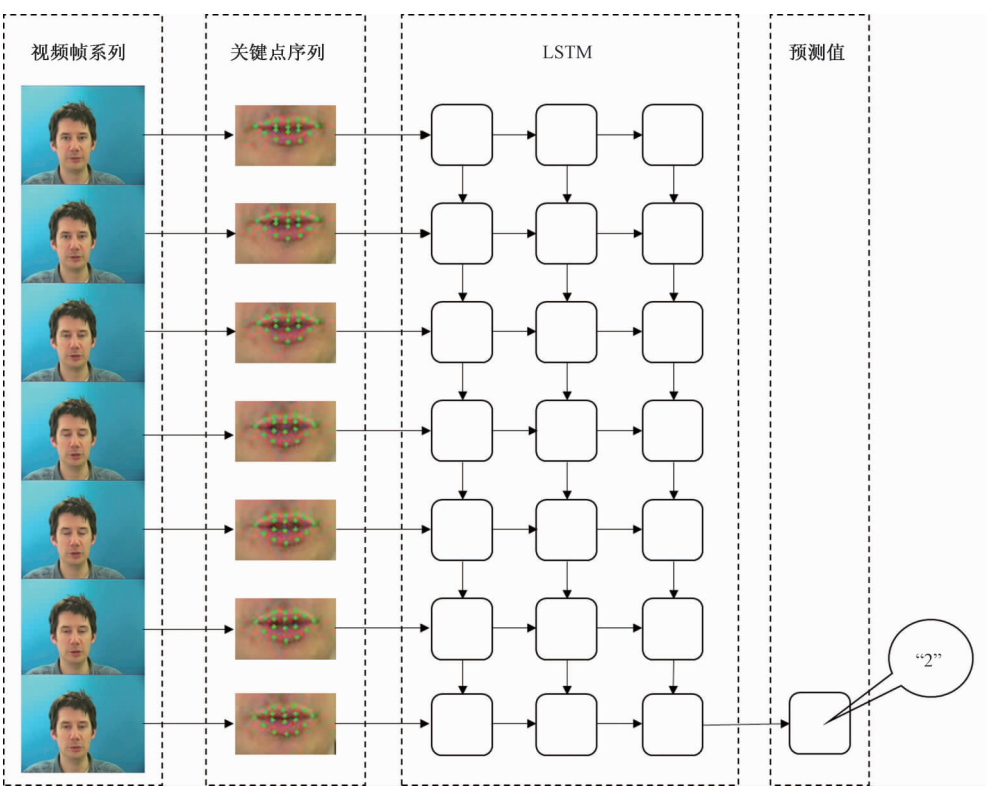


图 1 基于 LSTM 的唇语识别方法示意图

Fig. 1 Illustration of LSTM-based lip-reading recognition approach

程,免去手工设计特征,从高度不确定性的唇动视觉信息中自动提取不变性的空间 - 时序特征,从而提高唇语识别准确率。

已有研究者提出唇语识别的方法。从视觉特征提取这方面看,一些方法只从提取嘴唇部分的外表特征入手,如用空间 - 时间 LBP 描述子提取视频嘴唇部分图像的空间 - 时序特征[4-6]。还有一些方法同时利用嘴唇部分的外表特征和形变特征,如 Lan 等[7] 则用主动外观模型(active appearance model)的外表参数和形状参数作为特征,Pei 等[8]用 3 种特征的融合,分别是主动外观模型形状参数描述的特征、HOG 描述子提取的特征和 LBP 描述子提取的特征。本文使用的描述每帧的唇动视觉信息的特征跟 Lan 等[7]使用的主动外观模型中的形状参数类似,将使用人脸关键点检测技术检测到的嘴唇关键点的坐标作为每一帧的特征,与之不同的是,我们使用了判别方法而非生成方法,直接将视频所有帧的嘴唇关键点坐标输入到 LSTM 进行分类。从对视觉特征的时序编码看,Zhao 等[4]用空间 - 时间 LBP 描述子完成对时序特征的编码,Lan 等[7]用隐马尔科夫模型对特征建模,文献[5-6, 8]使用流形降维技术完成特征选择,Rekik 等[9]直接将特征插值成固定长度的特征向量。近年来,深度学习应用越来越广泛,在很多模式识别任务上都取得了优异的成绩。在唇语识别这个领域,研究者也提出一些基于深度学习的方法,如 Noda 等[10]利用卷积神经网络提取视频每帧嘴唇区域的特征再用隐马尔科夫模型建模和分类。跟卷积神经网络相比,LSTM 更适用于序列学习任务,近年来在语音识别[11]、机器翻译[12]和动作识别[13]上都取得了不错的成绩。有研究者将 LSTM 应用于唇语识别,如 Wand 和 Koutn[14]用前向神经网络提取视频每帧嘴唇区域的特征,再用 LSTM 进行分类。与以上基于深度学习的唇语识别方法不同,受 Simonyan 和 Zisserman[15]将图像光流作为时间卷积神经网络输入,将原始图像作为空间卷积神经网络输入实现动作识别的启发,本文提出将视频所有帧嘴唇关键点的坐标作为 LSTM 的输入,让 LSTM 由这些坐标信息中去自动学习空间 - 时序特征,而非将原始图片序列作为网络的输入[10,14],这样做基于两个假设:1)人说话过程中,嘴唇形状的变换信息比嘴唇外表的变化信息更重要;2)LSTM 能够从嘴唇关键点的坐标信息中提取具有区分性和泛化性的空间 - 时序特征。在 GRID 数据库[16]、MIRACL-VC 数据库[9]和 OuluVS 数据库[4]上的实验结果证实了我们的假设。

1　基于 LSTM 的唇语识别算法

1.1　特征

唇语视频是一段图片序列,由于我们只关心说话者嘴唇的运动,所以只需要关注视频每帧图像嘴唇区域的信息。受说话者嘴唇外观、光照条件、说话习惯等影响,即使说话者说相同的内容,嘴唇区域的视觉信息也可能会相差很大,因此从唇语视频中提取具有类内一致性、类间区分性的视觉特征是唇语识别中很关键的一步。

以往大多数提取唇语视频视觉特征的方法是基于嘴唇外表信息,包括传统的手工特征(如 HOG 特征和 LBP 特征)[4-6]和基于深度学习的特征[10, 14]。直观上,人在交流时,主要是观察对方嘴唇形状的变化帮助理解对方说话的内容。另一方面,嘴唇外表信息容易受到光照、不同人说话背景、不同人嘴唇外观的影响,所以我们从捕捉唇语视频中嘴唇形状的变换信息入手。为了较为精确地描述嘴唇形变信息,首先对嘴唇关键点进行追踪,使用 Xiong 和 Torre[17]的人脸关键点提取算法,提取唇语视频中每帧图像嘴唇的 18 个关键点,如图 2 所示。在 GRID 数据库上对说话内容为数字的那段唇语视频中的嘴唇关键点进行追踪,如图 3 所示(可放大显示)。发现,对有限个关键点的位置的追踪能够描述说话者说话时嘴唇的运动,并且在说话内容相同的情况下,不同人的唇语视频中嘴唇关键点位置的纵坐标 y 随着时间的相对变化具有一致性,而在说话内容不同的情况下,不同人的唇语视频中嘴唇关键点的纵坐标 y 随着时间的相对变化有明显的不同,而嘴唇关键点位置的横坐标 x 随着时间的相对变化不是很明显。这表明用嘴唇关键点坐标 描述的嘴唇形变信息存在一定的模式,于是我们考虑将每帧图像嘴唇的 18 个关键点的坐标(x,y) 连接起来,得到

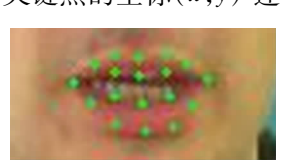
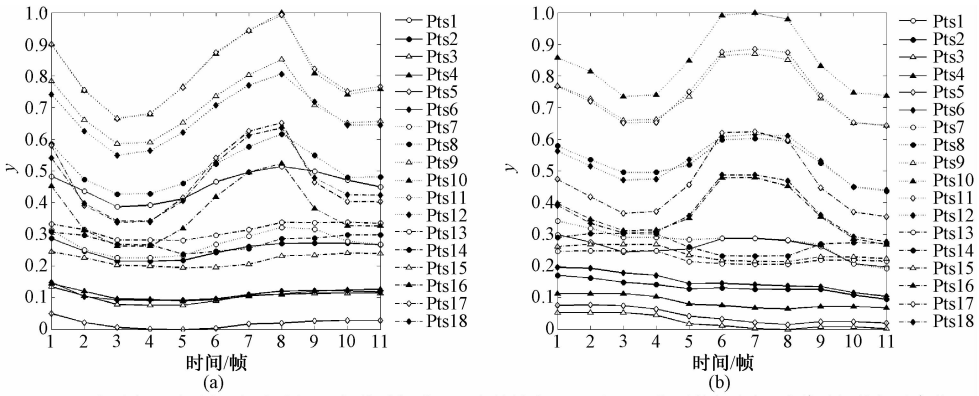
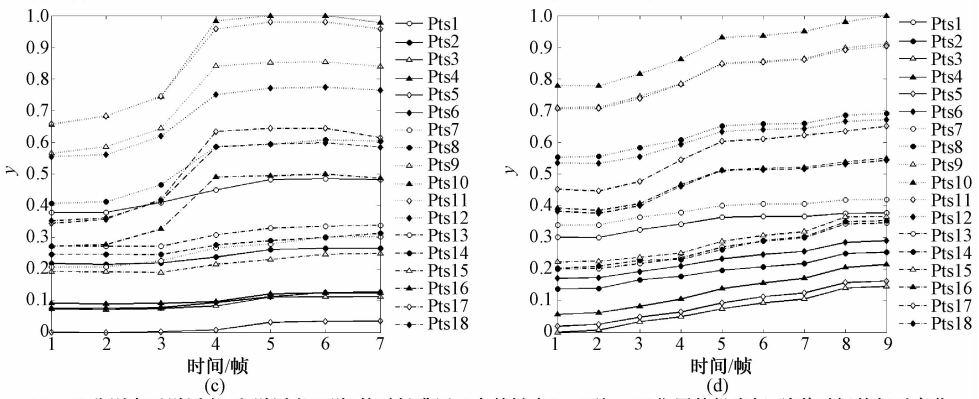


图 2　嘴唇的 18 个关键点

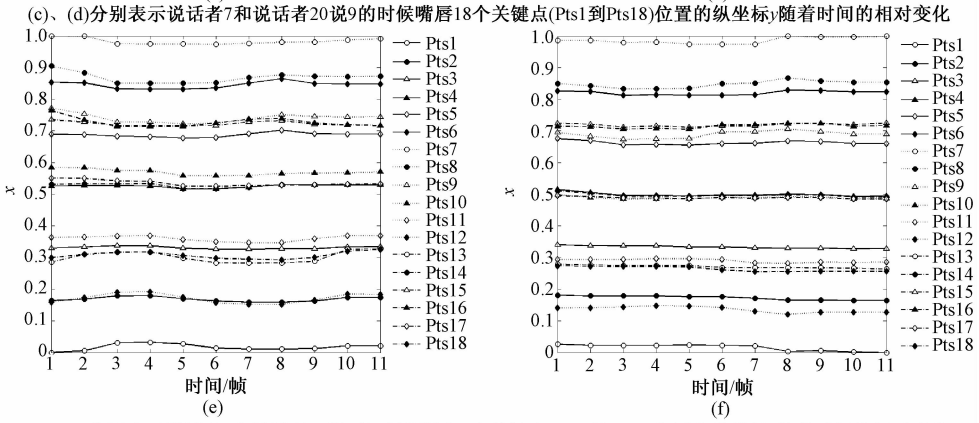
Fig. 2　18 landmarks of the lips



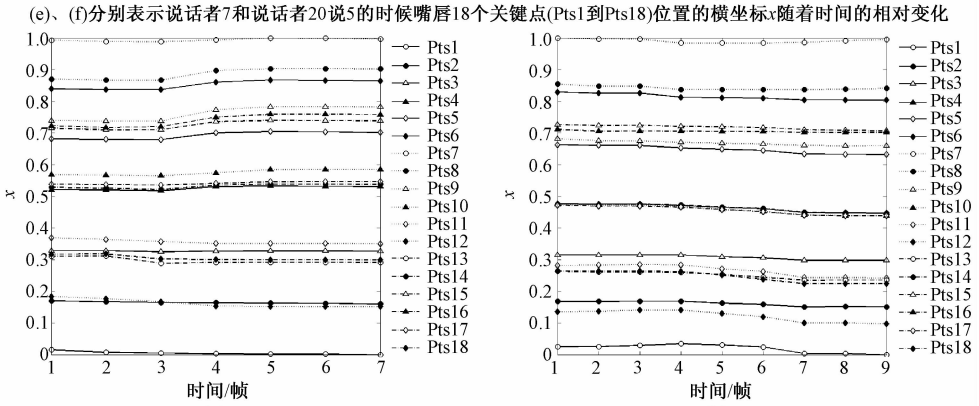
(a)、(b)分别表示说话者7和说话者20说5的时候嘴唇18个关键点(Pts1到Pts18)位置的纵坐标y随着时间的相对变化



(c)、(d)分别表示说话者7和说话者20说9的时候嘴唇18个关键点(Pts1到Pts18)位置的纵坐标y随着时间的相对变化



(e)、(f)分别表示说话者7和说话者20说5的时候嘴唇18个关键点(Pts1到Pts18)位置的横坐标x随着时间的相对变化



(g)、(h)分别表示说话者7和说话者20说9的时候嘴唇18个关键点(Pts1到Pts18)位置的横坐标x随着时间的相对变化

图 3　说话者 7 和说话者 20 说 5 和 9 时嘴唇关键点位置的变化

Fig. 3　Changes of landmark positions of speaker 7 and speaker 20 when they are speaking "5" and "9"

一个描述该帧(设为第 i 帧)嘴唇形状的特征向量 $\boldsymbol{p}_i = (x_i^1, y_i^1, x_i^2, y_i^2, \cdots, x_i^{18}, y_i^{18})$。

1.2　唇语识别 LSTM

1.2.1　LSTM 介绍

递归神经网络是在前向神经网络的基础上,在隐藏层节点和输出层节点的相邻时间点之间引入指向自身的连接,相比前向神经网络,递归神经网络更适用于序列学习任务,因为前向神经网络假设不同时间点输入的样本特征向量是独立的,而递归神经网络解除了这个假设,由于引入了节点指向相邻时间点自身的连接,它能够直接对网络当前时间点之前的输入和当前时间点的输出建模。这使得递归神经网络能够利用输入序列的上下文信息完成序列学习任务[2],因此对于输入序列局部出现的变形和失真现象具有鲁棒性。

LSTM 由 Hochreiter 和 Schmidhuber 提出[18],用来解决递归神经网络训练中出现的梯度迷失的问题。在递归神经网络的基础上,LSTM 将隐藏层节点和输出层节点替换为记忆单元,这个记忆单元有输入门、忘记门和输出门用来控制记忆状态向量的更新。LSTM 对输入输出的计算具体如下:设 LSTM 第 $l-1$ 层在 t 时刻的输入向量为 $\boldsymbol{h}_t^{l-1} \in \mathbf{R}^N$,其中上标 $l-1$ 表示第 $l-1$ 层,当上标为 0 时,即表示 LSTM 在 t 时刻的输入向量 $\boldsymbol{x}_t$,下标 t 表示 t 时刻。则 LSTM 第 $l-1$ 层在 t 时刻的输出向量(即第 l 层在 t 时刻的输入向量)$\boldsymbol{h}_t^l$ 的更新公式为:

$$\boldsymbol{i}_t = \mathrm{sigm}\left(\boldsymbol{W}_i^l \begin{bmatrix} \boldsymbol{h}_t^{l-1} \\ \boldsymbol{h}_{t-1}^l \end{bmatrix}\right),$$

$$\boldsymbol{f}_t = \mathrm{sigm}\left(\boldsymbol{W}_f^l \begin{bmatrix} \boldsymbol{h}_t^{l-1} \\ \boldsymbol{h}_{t-1}^l \end{bmatrix}\right),$$

$$\boldsymbol{o}_t = \mathrm{sigm}\left(\boldsymbol{W}_o^l \begin{bmatrix} \boldsymbol{h}_t^{l-1} \\ \boldsymbol{h}_{t-1}^l \end{bmatrix}\right),$$

$$\boldsymbol{g}_t = \tanh\left(\boldsymbol{W}_g^l \begin{bmatrix} \boldsymbol{h}_t^{l-1} \\ \boldsymbol{h}_{t-1}^l \end{bmatrix}\right),$$

$$\boldsymbol{c}_t^l = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1}^l + \boldsymbol{i}_t \odot \boldsymbol{g}_t,$$

$$\boldsymbol{h}_t^l = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t^l).$$

其中:$\boldsymbol{c}_t^l$ 表示第 l 层在 t 时刻的记忆状态向量,$\boldsymbol{g}_t$ 和 $\boldsymbol{f}_t$ 分别表示作用于输入 $\boldsymbol{i}_t$ 的输入门和忘记门,$\boldsymbol{o}_t$ 为输出门,sigm 和 tanh 分别表示 sigmod 函数和双曲正切函数作用于向量的每一个元素,$\boldsymbol{W}_i^l, \boldsymbol{W}_f^l, \boldsymbol{W}_o^l, \boldsymbol{W}_g^l \in \mathbf{R}^{N \times 2N}$ 为网络权值,$\odot$ 表示 Hadamard 积。LSTM 记忆单元更新记忆的机制使得 LSTM 在训练过程中,能够利用输入门、忘记门和输出门学习如何控制梯度信息在网络中的传播,这能缓解递归神经网络在训练过程中出现的梯度迷失问题。

1.2.2　唇语识别 LSTM

前文提到我们研究的唇语识别的输入是一段标注了起止点的说话内容为范围确定的数字或者短语的视频,输出是对该段视频说话内容的预测,即我们要对该段视频进行分类。1.1 中选用嘴唇关键点坐标序列描述唇语视觉特征,如图 3 所示,从嘴唇关键点坐标序列整体来看,说话者 7 说 9 和说话者 20 说 9 时具有相同的模式,说 5 时具有另一种模式,如说 9 时较为明显的时序特征是关键点 y 坐标从时间点 3 到时间点 4 显著增长,说 5 时较为明显的时序特征是关键点 y 坐标从时间点 1 到时间点 3 显著下降。这表明需要利用整个序列的上下文信息去提取具有区分度的时序特征进行模式识别。从嘴唇关键点坐标序列局部来看,相同类别的模式并非完全一致,表现在:1)不同说话者说相同内容时嘴唇关键点的位置并非完全一致。2)不同说话者说相同内容时嘴唇关键点的变化趋势并非完全一致,如说话者 7 说 9 时关键点 Pts11 在从第 3 帧到第 4 帧位置的变化率比说话者 20 说 9 时更大。出现以上现象的可能原因有很多,比如不同说话者的说话习惯不一样,说话者与镜头的距离不一样,光照条件不一样导致嘴唇关键点的检测不一致。这表明需要采用对输入序列局部出现失真现象鲁棒的模型。同时,还发现嘴唇坐标关键点序列的长度并不是固定的,如说话者 7 说 9 用 7 帧,说话者 20 说 9 用9 帧。为解决以上 3 个问题,我们决定采用 LSTM 完成唇语识别任务,输入为嘴唇关键点坐标序列,在最后一个时间点输出对序列的预测值。其原因是 LSTM 继承了递归神经网络的优点,能对任意长度输入序列当前时间点之前的输入和当前时间点的输出直接建模,能利用序列的上下文信息提取有用的时序特征,能对输入序列局部出现的失真现象鲁棒,并且相比递归神经网络更易训练。

我们的唇语识别 LSTM 具体描述如下:唇语视频是一段图片序列,经过 1.1 所述方法将图片序列转换成嘴唇关键点坐标序列 $\boldsymbol{x} = (p_1, p_2, \cdots, p_T) \in \mathbf{R}^{N \times T}$,其中 $p_i, i \in [1, T]$ 表示第 i 帧所有嘴唇关键点的坐标,T 表示这段视频的总帧数,N 表

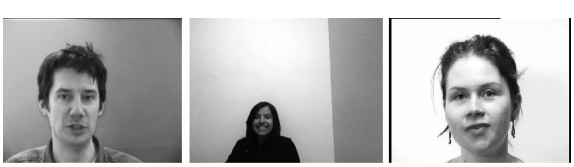
示 p_i 维度。我们使用的 LSTM 完成这样一个输入序列到输出类别的映射：$(p_1,p_2,\cdots,p_T)\mapsto\hat{y}_T$，其中 $\hat{y}_T$ 为在 T 时刻输出为对说话者说话内容的预测值。输入序列 $\boldsymbol{x}$ 经过第 1 层 LSTM 层，输出为 $\boldsymbol{h}^1\in\mathbf{R}^{N'\times T}$，其中 N' 为 LSTM 隐藏层的节点个数，再经过第 2 层 LSTM 层，输出为 $\boldsymbol{h}^2\in\mathbf{R}^{N'\times T}$，以此类推，在 LSTM 最后一层（设为第 L 层）最后一个时间点输出为 $\boldsymbol{h}_T^L\in\mathbf{R}^{N'}$，然后再经过 softmax 使得 $\boldsymbol{h}_T^L$ 映射为预测值 $\hat{y}_T$ 的分布 $P(\hat{y}_T=c)=\dfrac{\exp(\boldsymbol{W}_{cy}\boldsymbol{h}_T^L)}{\sum_{c'\in C}\exp(\boldsymbol{W}_{c'y}\boldsymbol{h}_T^L)}$，其中 C 表示预测值的取值范围，$\boldsymbol{W}_{cy}$ 为 softmax 权值。在训练过程中，我们损失函数为交叉熵损失函数 $L(\hat{y}_T,y)=-\sum_{c\in C}y_c\log P(\hat{y}_T=c)$，其中 $y_c=1$ 表示该段视频的唇语的标签为第 c 类，$y_c=0$ 表示该段视频的唇语的标签不为第 c 类。我们利用随机梯度下降法和时间反向传播算法（back propagation through time）对权值 $\boldsymbol{W}=\{\boldsymbol{W}_i^l,\boldsymbol{W}_f^l,\boldsymbol{W}_o^l,\boldsymbol{W}_g^l,\boldsymbol{W}_{cy}\}$，$l\in\{1,2,\cdots,L\}$ 进行更新。

2 实验结果与分析

2.1 数据库

我们在 3 个公开的数据库上进行唇语识别实验，它们分别是 GRID、MIRACL-VC 和 OuluVS。

GRID 包含 34 人每人 1 000 条英文句子的音频和视频，视频分为普通质量画质和高质量画质。实验用的是普通质量画质的视频，分辨率为360 × 288，时长约 3 s，帧率约 25 fps，视频的一帧如图 4（a）所示。视频中每条语句的模式为：命令单词 + 颜色单词 + 介词 + 字母 + 数字 + 副词，其中数字为 0 到 9，并且每个单词的起止时间都做了标注。我们只对包含数字那段的视频做唇语识别实验。



（a）、（b）、（c）分别为 GRID 视频、MIRACL-VC 视频和 OuluVS 视频中的某一帧

图 4 3 个数据库的示意图

Fig. 4 Illustration of three datasets

MIRACL-VC 包含 15 人每人读 10 个单词和 10 个短语 10 次的视频（图片帧的形式），每帧图片的分辨率为 640 × 480，除正常的 RGB 图片外，MIRACL-VC 唇语数据库还包括保存有深度信息的图片，RGB 图片如图 4（b）所示。单词和短语如表 1 所示。

表 1 MIRACL-VC 数据库的单词和短语

Table 1 Words and phrases in MIRACL-VC

单词	短语
Begin	Stop navigation.
Choose	Excuse me.
Connection	I am sorry.
Navigation	Thank you.
Next	Good bye.
Previous	I love this game.
Start	Nice to meet you.
Stop	You are welcome.
Hello	How are you?
Web	Have a good time.

OuluVS 包含 20 人每人读 10 个短语 5 次的视频，分辨率为 720 × 576，时长约 1 s，帧率约 25 fps。OuluVS 噪声信息更大，因为少数说话者的头部会轻微摆动，并且少数唇语视频的终止点说话者嘴唇是张开状态。视频的一帧如图 4（c）所示，说话者所说的短语如表 2 所示，可以看到，OuluVS 短语和 MIRACL-VC 短语有 8 个相同。

表 2 OuluVS 数据库的短语

Table 2 Phrases in OuluVS

短语
Hello
Excuse me
I am sorry
Thank you
Good bye
See you
Nice to meet you.
You are welcome.
How are you?
Have a good time.

在 GRID 和 MIRACL-VC 上已有的唇语识别研究使用的方法是很基本的标杆算法，如 Lan 等[7]使用主动外观模型和 HMM 相结合，Rekik 等[9]使用传统手工特征（HOG 和 MBH）和 SVM 相结合。而在 OuluVS 上已有的唇语识别研究使用的方法很多，具体见最近的文献[3]。因此我们决定在 GRID 和 MIRACL-VC 上的实验侧重于对模型的分析，在 OuluVS 上的实验侧重于和 state of art 的比较。

2.2　实验设置

在 GRID 上,我们的实验设置与 Lan 等[7]的实验设置一致:在 15 个说话者(说话者 1 ~ 12,20,23 ~ 24)的说话视频中的数字部分做实验,把数据分成 15 份,每份只包含 1 名说话者,分别用每一份数据作为测试数据,剩下的 14 份作为训练数据,进行交叉验证,这样使得测试数据出现的用户不会出现在训练数据中,这能更好地评估模型的泛化能力。

在 MIRACL-VC 上,我们的实验设置与 Rekik 等[9]的实验一致:把数据分成 15 份,每份只包含 1 名说话者,分别用每一份数据作为测试数据,剩下的 14 份作为训练数据,进行交叉验证,同样,测试数据出现的用户不会出现在训练数据中。

在 OuluVS 上,我们的实验设置与 Zhao 等[4]的实验一致:把数据分成 20 份,每份只包含 1 名说话者,分别用每一份数据作为测试数据,剩下的 19 份作为训练数据,进行交叉验证,测试数据出现的用户不会出现在训练数据中。

因为以上 3 个数据库相对于训练我们的 LSTM 来说,数据量较小,我们采用如下数据增强技术:在 GRID 和 MIRACL-VC 上,令嘴唇关键点坐标序列的起止帧向前向后各滑动 1 帧;在 OuluVS 上,因为唇语视频的时长比另外 2 个数据库长,数据量比另外 2 个数据库小,而且噪声信息更多,令其嘴唇关键点坐标序列的起止帧向前向后各滑动 3 帧以生成更多样本。同时,在所有数据库上,对嘴唇关键坐标做了归一化。将以上方法生成的样本做最近邻插值,使得所有样本的嘴唇关键点坐标序列长度相同,这是为了使 LSTM 的训练收敛更快,根据对数据库唇语视频时长的统计,在 GRID 上设定所有样本的嘴唇关键点坐标序列长度为 10,在 MIRACL-VC 上设定所有样本的嘴唇关键点坐标长度序列长度为 12,在 OuluVS 上设定所有样本的嘴唇关键点坐标长度序列长度为 40。经过对超参数不同值的实验,将 LSTM 隐藏层节点数统一设置为 256,网络层数设置为 3,使用 dropout 技术[19]对 LSTM 进行正则化,将层与层之间的权值 dropout 率和相邻时间点权值 dropout 率设置为 1% 。我们的 LSTM 的实现用到了深度学习库 Keras①。

2.3　实验结果

2.3.1　y 和(x,y) 的选择

由图 3 可观察到不同人说相同的内容时嘴唇关键点的纵坐标 y 随着时间的相对变化比较一致,而横坐标 x 随着时间的相对变化的一致性不明显。在 MIRACL-VC 上分别用嘴唇关键点的纵坐标 y 序列和嘴唇关键点的坐标(x,y) 序列作为输入,评估二者的表现。实验结果如表 3 所示。可以看到,虽然图 3 中不同人说相同的内容时横坐标 x 随着时间相对变化的一致性不明显,但是对于唇语识别 LSTM 识别准确度的提升来说是很重要的。因此后面的实验用嘴唇关键点的坐标 (x,y) 序列作为输入。

表 3　y 和(x,y)在 MIRACL-VC 数据库上的实验结果

Table 3　Lip-reading performances of y and (x,y) on MIRACL-VC　%

方法	单词平均准确率	短语平均准确率
Landmark(y) + LSTM	66.8	78.4
Landmark(x,y) + LSTM	72.7	81.8

2.3.2　和标杆算法进行比较

我们在 GRID 上对我们的方法在单词上的唇语识别表现和 Lan 等[7]的结果进行比较,如表 4 所示。表 4 中 app 和 shape 分别表示使用主动外观模型的外表参数和形状参数为特征,app_pca 表示使用外表特征和形状特征的融合,采用主成分分析法。可以看到,使用我们的方法将平均准确率提升 33% 以上,这表明 Lan 等[7]得出的唇语识别中嘴唇外表信息比嘴唇形状信息更重要的结论并非完全正确,因为我们的方法中用嘴唇关键点坐标序列与主动外观模型的形状参数类似,关键在于如何从这些信息中提取有用的空间 - 时序特征。

表 4　在 GRID 数据库上的实验结果

Table 4　Lip-reading performances on GRID　%

方法	平均准确率②
app + HMM[7]	55
shape + HMM[7]	28
aam_pca + HMM[7]	59
Landmark(x,y) + LSTM	78.9

① Keras 深度学习库是开源的,其网址为 https://github.com/fchollet/keras。

② 文献[7]中使用指标是平均单词准确度,计算方式为 $(H-I)/N$,其中 H 为准确识别的单词数量,I 为插入误差的数量,N 为需要识别的的单词数量。在我们的实验中,$I=0$,因此我们的指标与之等价。

我们在 MIRACL-VC 上对我们的方法在单词和短语上的唇语识别表现和 Rekik 等[9]的结果进行比较,如表 5 所示。表 5 中下标 c 和 d 分别表示从 RGB 图像序列和深度图像序列提取特征。可以看到,在只使用 RGB 图像序列的情况下,使用我们的方法将平均准确率提升 30% 以上。我们还发现在文献[9]的实验中使用传统的运动特征,如 MBH_c,要比外表特征,如 HOG_c 表现要差,这更加体现了用 LSTM 提取时序特征的优异性。

表 5 在 MIRACL-VC 数据库上的实验结果

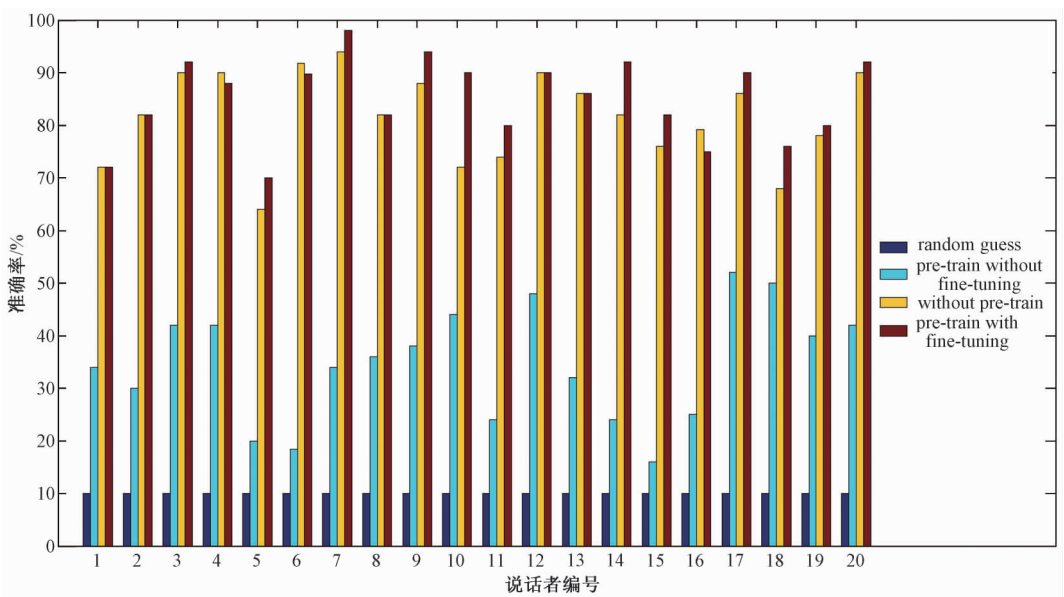
Table 5 Lip-reading performances on MIRACL-VC

%

方法	单词平均准确率	短语平均准确率
$HOG_c + SVM$[9]	55.5	54.4
$MBH_c + SVM$[9]	45.1	49.4
$HOG_c + HOG_d + MBH_c + SVM$[9]	63.1	79.2
Landmark(x,y) + LSTM	72.7	81.8

2.3.3 和 state of art 进行比较

因为 OuluVS 的唇语视频数据量比 GRID 和 MIRACL-VC 小,同时 OuluVS 与 MIRACL-VC 的唇语视频说话内容大部分是相同的,相同的类别共有 8 个,我们决定令我们的唇语识别 LSTM 先在 MIRACL-VC 上进行预训练,然后在 OuluVS 上微调。与直接令我们的唇语识别 LSTM 在 OuluVS 上进行训练和测试的比较如图 5 所示。发现,即使 MIRACL-VC 和 OuluVS 唇语视频的视觉信息相差很大(如图 4(b)和 4(c)所示),在 MIRACL-VC 上进行预训练的模型比不在 MIRACL-VC 上进行预训练的模型表现要好(平均准确率 85.0% 对 81.8%),而且只在 MIRACL-VC 上进行预训练的模型在 OuluVS 上的表现比随机猜说话内容的表现好很多(平均准确率 34.6% 对 10%)。这表明我们的方法提取的空间 – 时间特征具有很好的泛化性。



随机猜说话内容(random guess)、在 MIRACL-VC 上进行预训练而不进行微调的模型(pre-train without fine-tuning)、不在 MIRACL-VC 上进行预训练的模型(without pre-train)和在 MIRACL-VC 上进行预训练并在 OuluVS 上进行微调的模型(pre-train with fine-tuning)在 OuluVS 上表现的比较。

图 5 不同模型在 OuluVS 上表现的比较

Fig. 5 Comparison of performance on OuluVS among different models

我们将在 OuluVS 上的实验结果跟最近的唇语识别研究[4-6, 8]做了比较,如表 6 所示。除 Pei 等[8]使用嘴唇外表信息和嘴唇形变信息的融合之外,其余方法[4-6]只用了嘴唇外表信息。我们的方法在 OuluVS 上的表现超过大多数利用嘴唇外表信息方法[4-6],并且比较接近平均准确率最高的方法[8](85.0% 对 89.7%)。这表明利用嘴唇形变信息完成唇语识别是很重要的,并且我们的方法能够从形变信息中提取具有区分性和泛化性的特征。

3 结论

本文提出一种新的并且简单有效的唇语识别方法,它基于 LSTM,输入为唇语视频所有帧嘴唇

表 6　在 OuluVS 数据库上的实验结果

Table 6　Lip-reading performances on OuluVS　%

方法	短语平均准确率
Local Spatiotemporal	58.6
MKL-Fusion[5]	81.3
MKPLS[6]	62.3
$RFMA_{fusion}$[8]	89.7
Landmark(x,y) + LSTM	85.0

部分的关键点的坐标,能自动学习具有不变性的空间 – 时序特征,有效地解决了唇语视觉信息多样性的问题,结果显示这种方法在两个公开的唇语数据集 GRID 和 MIRACL-VC 上比传统的唇语识别方法准确率至少高 30% ,在另一个规范性更差的数据集 OuluVS 上,本方法表现接近于最优表现。由于这 3 个数据库数据量的限制,有理由相信,在数据量更大的情况下,本文提出的方法识别准确率将会更高。同时,因为本方法只是对唇语视频嘴唇关键点位置信息做处理,可以与已有的方法进行互补以达到更好的唇语识别效果。目前仅在分割的标注起止时间的单词或短语上做了唇语识别实验,未来将会在连续的单词和短语上进行唇语识别实验,它将会包括一个检测说话内容的过程,这将是我们以后工作的重点。

参考文献

[1] McGurk H, MacDonald J. Hearing lips and seeing voices [J]. Nature, 1976, 264: 746-748.

[2] Graves A. Supervised sequence labelling with recurrent neural networks [M]. Berlin: Springer Berlin Heidelberg, 2012.

[3] Zhou Z, Zhao G, Hong X, et al. A review of recent advances in visual speech decoding [J]. Image & Vision Computing, 2014, 32: 590-605.

[4] Zhao G, Barnard M, Pietikainen M. Lipreading With Local Spatiotemporal Descriptors [J]. IEEE transactions on multimedia, 2009, 11: 1 254-1 265.

[5] Zhou Z, Zhao G, Pietikainen M, Towards a practical lipreading system [C] // Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 137-144.

[6] Bakry A, Elgammal A. Mkpls: manifold kernel partial least squares for lipreading and speaker identification [C] // Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013: 684-691.

[7] Lan Y, Harvey R, Theobald B J, et al. Comparing visual features for lipreading [C] // International Conference on Auditory-Visual Speech Processing. 2009: 102-106.

[8] Pei Y, Kim T K, Zha H. Unsupervised random forest manifold alignment for lipreading [C] // Proceedings of the IEEE International Conference on Computer Vision. 2013: 129-136.

[9] Rekik A, Ben-Hamadou A, Mahdi W. A new visual speech recognition approach for RGB-D cameras [C] // International Conference Image Analysis and Recognition. Springer International Publishing, 2014: 21-28.

[10] Noda K, Yamaguchi Y, Nakadai K, et al. Lipreading using convolutional neural network [J]. Interspeech, 2014: 1 149-1 153.

[11] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks [C] // Proceedings of the 31st International Conference on Machine Learning, 2014: 1 764-1 772.

[12] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C] // Advances in Neural Information Processing Systems, 2014: 3 104-3 112.

[13] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [C] // Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015: 2 625-2 634.

[14] Wand M, Koutn J. Lipreading with long short-term memory [C] // Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 6 115-6 119.

[15] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C] // Advances in Neural Information Processing Systems, 2014: 568-576.

[16] Cooke M, Barker J, Cunningham S, et al. An audio-visual corpus for speech perception and automatic speech recognition [J]. The Journal of the Acoustical Society of America, 2006, 120: 2 421-2 424.

[17] Xiong X, Torre F D L. Supervised descent method and its applications to face alignment [C] // Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013: 532-539.

[18] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9: 1 735-1 780.

[19] Yarin G, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks [C] // Advances in Neural Information Processing Systems, 2016: 1 019-1 027.