基于具有深度门的多模态长短期记忆网络的说话人识别

陈湟康*, 陈莹**
江南大学轻工过程先进控制教育部重点实验室, 江苏 无锡 214122

摘要 为了在说话人识别任务中有效融合音视频特征,提出一种基于深度门的多模态长短期记忆(LSTM)网络。首先对每一类单独的特征建立一个多层 LSTM 模型,并通过深度门连接上下层的记忆存储单元,增强上下层的联系,提升该特征本身的分类性能。同时,通过在不同模型之间共享连接隐藏层输出与各个门单元的权重,学习每一层模型之间的联系。实验结果表明,该方法能有效融合音视频特征,提高说话人识别的准确率,并且对干扰具有一定的稳健性。

关键词 图像处理; 说话人识别; 长短期记忆网络; 融合; 深度门; 权重共享
中图分类号 TP391 **文献标识码** A **doi**: 10.3788/LOP56.031007

Speaker Identification Based on Multimodal Long Short-Term Memory with Depth-Gate

Chen Huangkang*, Chen Ying**
Key Laboratory of Advanced Process Control for Light Industry of the Education Ministry of China, Jiangnan University, Wuxi, Jiangsu 214122, *China*

Abstract In order to effectively fuse the audio and visual features in the task of speaker recognition, a multimodal long short-term memory network (LSTM) with depth-gate is proposed. First, a multi-layer LSTM model is established for each type of individual features. Then the depth-gate is used to connect the memory cells in the upper and lower layers, and the connection between the upper and lower layers is enhanced, which improves the classification performance of the feature itself. At the same time, the connection among layer models can be learned by sharing the output of hidden layers and the weight of each gate unit among different models. The experimental results show that this method can be used to effectively fuse the audio and video features and improve the accuracy of speaker recognition. Moreover, this method is robust to external disturbance.

Key words image processing; speaker recognition; long short-term memory network; fusion; depth-gate; weight sharing
OCIS codes 100.4996; 100.3008; 100.5010

1 引 言

说话人识别是智能视频信息处理的一个重要分支,常被应用于视频会议、视频记录总结、视频监控等场景。但是,视频信息中的语音信息与图像信息是两类异质信息,如何有效地融合这两种信息以提升说话人识别准确率,是该项工作面临的主要问题。

对于传统的人脸识别方法和声纹识别方法[1-2],多数研究者采用决策层融合策略。如文献[3]为每类特征分类器分配一个由它本身分类生成的分数计算得到置信度值,再根据提出的一种新的置信比进行匹配分数的转换,得到最优匹配分数。同时少部分学者也尝试过特征级融合,如吴志勇等[4]就曾利用动态贝叶斯网络(DBN)对传统特征进行耦合建

收稿日期: 2018-06-13; **修回日期**: 2018-08-21; **录用日期**: 2018-08-31
基金项目: 国家自然科学基金(61573168)
* **E-mail**: 6161918009@vip.jiangnan.edu.cn; ** **E-mail**: chenying@jiangnan.edu.cn

模。随着深度学习在人脸识别领域的发展，文献[5-7]对深度特征尝试了特征级融合。其中：Hu 等[5]选择先单独训练一个人脸识别的卷积神经网络(CNN)，之后将语音特征与预训练的 CNN 中最后一个池化层的特征级联并且送入全连接层，得到新的特征用以识别说话人；Geng 等[6]选择首先将语音特征外接一个全连接层，得到更高级的语音特征，随后将人脸识别的 CNN 中最后一层全连接层同样外接一个全连接层，然后将二者级联得到新的特征，实验结果表明这种特征级融合的识别精度明显优于单一模型；文孟飞等[7]则是首先针对媒体流中同时存在的音频和视频信息特征，建立一种异构多模态深度学习结构，随后使用 CNN 提取人脸特征，用受限制玻尔兹曼机提取语音特征，生成基于典型关联分析的共享特征表示，并进一步利用时间相关特性进行参数优化。为了更好地挖掘两类特征的相关性，Ren 等[8]提出多模态长短期记忆网络(LSTM)用于识别说话人。先分别提取人脸特征和语音特征，随后对两类特征分别构建一个 LSTM 网络，通过共享两个 LSTM 网络的部分权重建立两类特征之间的联系，使得该网络模型可以学习到视频序列中两类信息当前的相关性。

本文在文献[8]的工作基础上，改进了网络结构，提出了基于深度门[9]的多模态长短期记忆网络(DGLSTM)。首先将单层 LSTM 扩展至多层，随后用一个门函数连接网络中上层和下层的记忆存储单元，定义上下层的线性依赖关系，在网络深度得到增强的同时，也增强了两个模型各自的灵活性与相互之间的联系，从而提升说话人识别的准确率。

2 基于多模态 LSTM 的说话人识别

首先介绍 LSTM[10]的基本概念，在此基础上介绍多模态 LSTM，并简要论述其待改进的地方。

2.1 LSTM 回顾

循环神经网络(RNN)被广泛应用于自然语言处理任务[11-12]中，但是其存在反向传播过程中梯度误差累积过多，致使梯度归零或者趋于无穷大，最终导致模型无法进一步优化的问题。LSTM 是一种特殊的 RNN，它具有一个用于调节过去和当前活动之间的信息流的忘记门单元，一个对当前活动和过去活动具有线性依赖性的记忆存储单元，以及分别用于调节输入和输出的输入门单元与输出门单元。其中，记忆存储单元只有加法与乘法等线性运算，信息在上面流传容易保持不变，是 LSTM 的关键。LSTM 具体结构如图 1 所示。

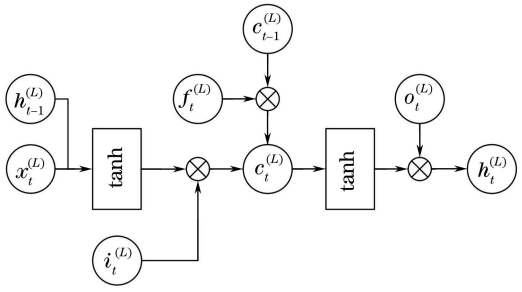

图 1 LSTM 基本结构

Fig. 1 Basic structure of LSTM

图中，$x_t^{(L)}$ 表示第 L 层视频序列中当前帧的原始输入，$h_t^{(L)}$、$h_{t-1}^{(L)}$ 分别表示当前帧和上一帧 LSTM 单元的隐藏层输出，$i_t^{(L)}$、$o_t^{(L)}$、$f_t^{(L)}$ 分别表示输入门、输出门与忘记门。$c_t^{(L)}$、$c_{t-1}^{(L)}$ 则分别表示当前帧和上一帧的记忆存储单元。定义上图左侧的 tanh 运算结果为 $g_t^{(L)}$，该运算结果的作用是创建一个新的候选值，用以更新存储单元的状态。各个单元的具体计算公式为

$$\begin{cases} g_t^{(L)} = \tanh(W_{xg} \times x_t^{(L)} + W_{hg} \times h_{t-1}^{(L)} + b_g) \\ i_t^{(L)} = \mathrm{sigmoid}(W_{xi} \times x_t^{(L)} + W_{hi} \times h_{t-1}^{(L)} + b_i) \\ f_t^{(L)} = \mathrm{sigmoid}(W_{xf} \times x_t^{(L)} + W_{hf} \times h_{t-1}^{(L)} + b_f) \\ o_t^{(L)} = \mathrm{sigmoid}(W_{xo} \times x_t^{(L)} + W_{ho} \times h_{t-1}^{(L)} + b_o) \\ c_t^{(L)} = f_t^{(L)} \odot c_{t-1}^{(L)} + i_t^{(L)} \odot g_t^{(L)} \\ y_t = \mathrm{softmax}(W_y \times h_t^{(L)}) \end{cases}, \tag{1}$$

式中×表示矩阵叉乘，⊙表示矩阵点乘，y_t 为整个 LSTM 最终输出结果，只在网络最后一层计算。另外，所有权重在帧与帧之间是共享的。

2.2 多模态 LSTM

考虑到在说话人识别任务中经常会连续出现同一人的人脸，上下帧之间关联大，所以 LSTM 适合用于在视频中识别说话人。但是由于人脸特征和语音特征是分属于两个领域的数据，若是直接将其串联作为 LSTM 的输入，系统对干扰项的稳健性能差，例如，串联说话人 A 的语音特征和画面中 B 的人脸特征之后，模型就会很难输出一个有意义的身份标签；若是选用两个 LSTM 模型分别处理这两类数据，只在最后的输出阶段提出某种有效的投票方案实现决策层融合，则系统对干扰项的稳健性能主要依赖于投票方案的策略，两类数据的相关性有很大可能在分别的前向传播过程中就消失了。

基于上述讨论，文献[8]提出一种新的融合策略：构建两个结构相同的 LSTM 并行处理人脸和语

音特征，并且在前向传播过程中，有选择地共享部分权重，使视频的人脸信息和语音信息的相关性得到有效地挖掘。

由(1)式可知，LSTM 前向传播中的权重主要可分为两类，一类连接原始输入和各个门单元，另一类连接隐藏层输出与各个门单元。为了确保每类数据都有从输入空间到新的共享空间的专门映射，同时为了避免由于不同特征的维数不同带来编程实现上的复杂性，共享权重的挑选被限定在连接隐藏层输出与各个门单元这一类之间，也就是(1)式中的 W_{hg}、W_{hi}、W_{hf}、W_{ho} 这 4 类权重。

然而在网络深度的拓展上，Ren 等[8]没有进一步研究。为了探寻这两类信息更深层的关系，本文将深度门的概念与多模态 LSTM 结合，通过深度门连接上下层的记忆存储单元加深网络深度。

3 多模态 DGLSTM

通常情况下，为了增加神经网络深度，人们会选择简单地堆叠网络层数。对于 LSTM 来说，堆叠网络层数会使误差在从顶层反向传播到底层的过程中，经过各层的非线性变换之后急速增长。为了解决该问题，借鉴文献[8]的思想，在多模态 LSTM 的基础上，引入深度门连接不同层，并提出了多模态 DGLSTM。

3.1 DGLSTM

传统的多层 LSTM 只是将上一层的隐藏层输出 $h_t^{(L)}$ 作为下一层的输入，除此之外层与层之间没有其他关联。文献[9]提出一种类似于 Grid LSTM[13]的新结构，利用深度门将相邻的上下两层的记忆存储单元连接起来，用来描述上下两层的线性依赖关系。实验证明，该网络在机器翻译和语言建模任务中展现的性能都优于传统 LSTM。

3.2 多模态 DGLSTM

在多模态 LSTM 基础上，保持全部连接隐藏层输出与各个门单元的权重的共享。在拓展网络深度时，将深度门同时应用到各个模型上，建立了多模态 DGLSTM，该网络的结构如图 2 所示。

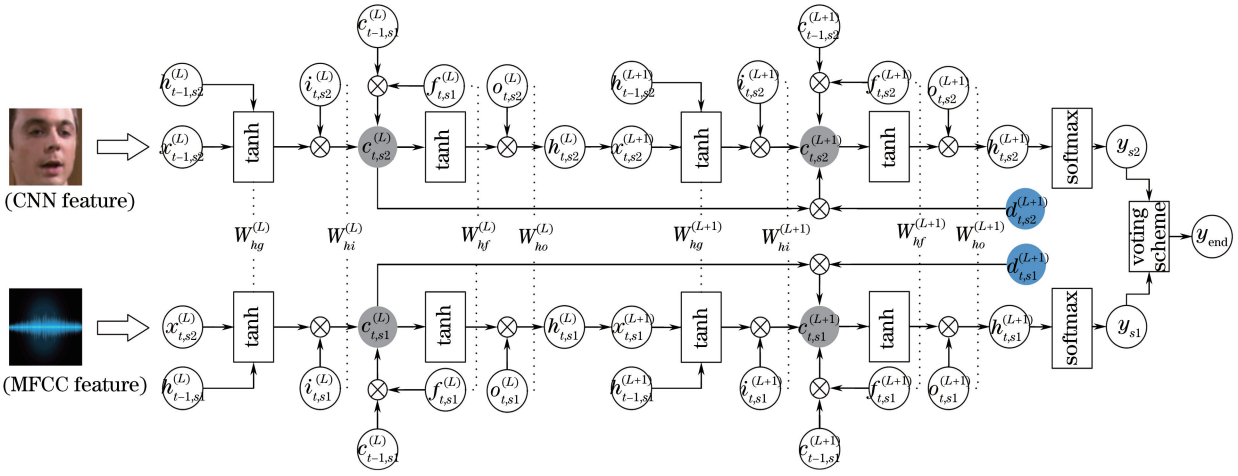

图 2 用于说话人识别的多模态 DGLSTM 结构

Fig. 2 Multimodal DGLSTM architecture for speaker recognition

该网络包含两个并行的多层 LSTM 模型，通过将上一层的隐藏层输出作为下一层的初始输入以及利用深度门连接上下层的记忆存储单元，建立了每个模型内层与层的连接。其中一个模型输入为人脸的 CNN 特征，另一个模型输入为语音的梅尔倒谱系数(MFCC)特征。两个模型的结构相同，以便于权重共享和最终的标签分配。

图 2 中，带有 tanh 的方框表示对输入加权求和后利用 tanh 函数激活；带有乘法运算符的圆圈表示矩阵的乘法，此处代表点乘。另外，深灰色圆圈表示记忆存储单元，是网络中最为关键的单元，它存储了每类特征在网络传播中的主要信息；蓝色圆圈则表示被引入多模态 LSTM 的深度门，它将模型内上下两层的记忆存储单元连接起来，决定了上一层记忆存储单元中有多少信息能直接流入下一层的记忆存储单元。其中，门函数的计算公式为

$$d_{t,s}^{(L+1)}=\mathrm{sigmoid}(b_{d,s}^{(L+1)}+W_{xd,s}^{(L+1)}x_{t,s}^{(L+1)}+W_{cd,s}^{(L+1)}\odot c_{t-1,s}^{(L+1)}+W_{ld,s}^{(L+1)}\odot c_{t,s}^{(L)}),\tag{2}$$

式中下标带有字母 d 的表示与深度门相关的对应权重和偏置，上标 L 与 $L+1$ 表示该单元所属层数，下标中的 s 表示该单元属于 s_{th} 模型，下文公式中的符号含义与此相同。如果在网络结构设计中，上下两层的单元数不同造成了连接单元的数据维数不匹配，则可以将相应的权重向量调整为权重矩阵。由

(2)式可知,通过深度门,当前层当前帧的输入、当前层前一帧的记忆存储单元以及前一层当前帧的记忆存储单元被有机地联系在一起。基于此,得到 $c_{t,s}^{(L+1)}$ 的具体更新公式为

$$c_{t,s}^{(L+1)}=d_{t,s}^{(L+1)}\odot c_{t,s}^{(L)}+f_{t,s}^{(L+1)}\odot c_{t-1,s}^{(L+1)}+i_{t,s}^{(L+1)}\odot g_{t,s}^{(L+1)}。\tag{3}$$

此外,图中两个模型之间用虚线相连的部分,表示的是每一层各个门函数更新过程中共享的权重,各个门函数的具体计算公式为

$$\begin{cases}g_{t,s}^{(L)}=\tanh(W_{xg,s}^{(L)}\times x_{t,s}^{(L)}+W_{hg}^{(L)}\times h_{t-1,s}^{(L)}+b_{g,s})\\ i_{t,s}^{(L)}=\mathrm{sigmoid}(W_{xi,s}^{(L)}\times x_{t,s}^{(L)}+W_{hi}^{(L)}\times h_{t-1,s}^{(L)}+b_{i,s})\\ f_{t,s}^{(L)}=\mathrm{sigmoid}(W_{xf,s}^{(L)}\times x_{t,s}^{(L)}+W_{hf}^{(L)}\times h_{t-1,s}^{(L)}+b_{f,s})\\ o_{t,s}^{(L)}=\mathrm{sigmoid}(W_{xo,s}^{(L)}\times x_{t,s}^{(L)}+W_{ho}^{(L)}\times h_{t-1,s}^{(L)}+b_{o,s})\\ y_{t,s}=\mathrm{softmax}(W_{y}\times h_{t,s}^{(L_{\mathrm{end}})})\end{cases},\tag{4}$$

式中:如果权重没有下标 s,则表示这些权重在各模型之间共享;如果有下标 s,则表示不在各模型之间共享;带有上标 L 的权重表示不在各层之间共享;但是所有权重在时序 t 上,即帧与帧之间都是共享的。

在图 2 中可以看到,每个模型的最后一层隐层输出都会经过 softmax 层,得到该样本对应的预测标签 y_{s1}、y_{s2}。网络根据 y_{s1}、y_{s2} 与样本标签真实值计算交叉熵损失,利用梯度下降法进行反向传播与模型更新。在网络中,每个样本可以被看成由若干帧人脸特征与语音特征匹配复合而成。因此在网络最后的输出阶段,两个模型会各得到若干个预测标签,规定如果有超过 m 帧的预测标签属于同一个身份(ID),则认为该样本的人脸与语音属于同一个人,且该 ID 为此样本最终预测结果 y_{end}。该投票策略使网络模型对瞬时的非说话人人脸干扰具有一定稳健性,其中 m 是阈值,根据多次实验结果,$m=10$ 时,识别效果最好(下述实验都采用该阈值)。

由(3)和(4)式可以看出,除了连接原始输入和各个门单元的权重没有共享,$c_{t,s}^{(L)}$ 与偏置项也没有在模型之间共享,这是为了让模型自动筛选应该保留或忘记的模型内前后帧的相关性。在模型之间,随着层数的增加,两个模型的联系也更加紧密。值得一提的是,在本文提出的多模态 DGLSTM 中,(2)式中的所有权重在模型之间是不共享的,此举旨在提升模型的灵活性,下文将会通过对比实验验证这一点。

通过引入深度门,每个模型可以更专注于各自特征空间中信息流的筛选与更新,而不至于因为共享权重减弱了本身的分类性能。另外,在训练过程中,因为迭代以及时序上的前向传播,通过权重共享,信息经过深度门之后在更高层次进行了关联,使两类特征的联系交叉成一张有机的网络,得到了有效的融合。

4 实验与分析

设计了三次实验以验证提出的多模态 DGLSTM 的有效性。前两次实验在静态数据上进行,首先比较依据本文构建的不同结构的多模态 DGLSTM 的识别性能,随后选出最佳模型与文献[5,6,8]的算法以及仅引入深度门的多模态网络进行对比,验证了本文网络的优越性。最后选出文献[5,6,8]中的最优算法与本文最优模型,在仿真视频序列上进行进一步对比。

4.1 数据库介绍

考虑到美剧《生活大爆炸》(TBBT)中的人脸经常发生重大形变,并且说话人的脸时常消失在画面中,对于说话人识别这项任务来说很具有挑战性,同时为了更直观公平地比较实验结果,选用了文献[8]提供的数据库。该数据库包含了 TBBT 第一季与第二季各自前六集的人脸照片以及两季全季的语音数据。

4.2 实验步骤与结果分析

在实验中,多模态 DGLSTM 的原始输入分别为图像特征和语音特征。具体特征提取过程如下:以 0.5 s 的时域窗口采集到的音视频信息作为一个样本,则在帧率为 24 frame/s 的视频中,每个样本包含 12 张人脸和一段 0.5 s 的语音。其中语音特征为 25 维的 MFCC 特征,在时域取一个 20 ms 的滑动窗口,每次滑动 10 ms,得到 25×49 的语音特征;人脸图像特征为降维之后的 53 维 CNN 特征[14],故得到 53×12 的人脸特征。为了使两个模型的权重共享,需要对每一层的单元数进行统一,即对初始输入维度进行约束,于是人脸特征被均匀地重复至其扩展成 53×49 的特征矩阵。

对于人脸特征和语音特征,随机匹配具有相同标签的数据,以增强训练集的多样性,共生成 55000 个训练样本对,50000 个测试样本对。在识别说话人具体身份之前,模型需要判断当前样本的人脸与语音是否属于同一个人,所以除了生成正确匹配的特征对(right-paired)外,还生成了 50000 个错误匹配人脸特征和语音特征的样本对(ill-paired)用于测试模型能否拒绝错误样本。

由于每个样本的时域窗口为 0.5 s，因此根据输入样本，每个 DGLSTM 的输出结果为 49 个预测标签，最终结果的输出策略如图 3 所示。

首先，实验训练了一个双层共享深度门的权重偏置的多模态 DGLSTM，一个双层不共享深度门的权重偏置的多模态 DGLSTM 以及一个三层不共享深度门的权重偏置的多模态 DGLSTM，实验结果如表 1 所示。

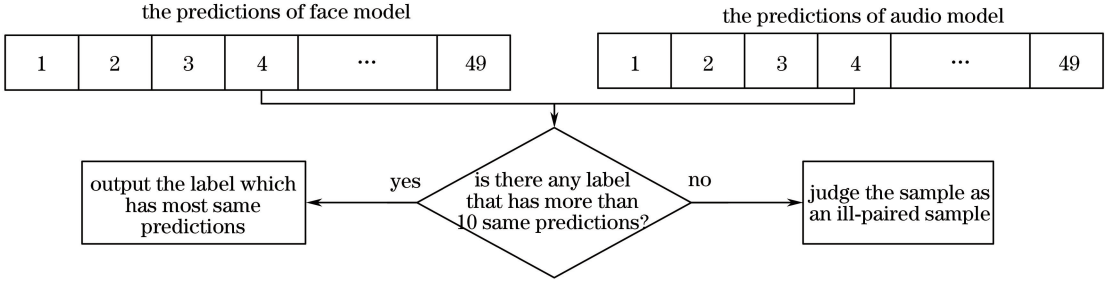

图 3 网络最终输出策略

Fig. 3 Final output strategy for network

表 1 不同结构的多模态 DGLSTM 的识别准确率

Table 1 Recognition accuracy of multimodal DGLSTM with different structures %

Method	Recognition accuracy	Ill-paired-rejection accuracy
2-layer multimodal DGLSTM (shared)	90.30	95.85
2-layer multimodal DGLSTM(not shared)	92.25	96.55
3-layer multimodal DGLSTM (not shared)	88.25	95.00

表 1 展示了不同结构模型的测试结果，其中 Ill-paired-rejection accuracy 指的是正确识别出样本的人脸特征与语音特征是错误匹配的成功率。通过表格可以看出，双层多模态 DGLSTM(不共享权重偏置)的模型具有最好的识别效果，这证明了本文网络结构优于文献[8]网络结构。此外，通过对比表格中第一个模型与第二个模型的识别结果，证明了不共享深度门的权重偏置能有效提升模型的灵活性。在探索网络合适的深度方面，实验发现当网络层数拓展到三层时，已经出现过拟合的情况，对比表格中第二个模型与第三个模型的识别结果，可以看到三层模型的识别性能相比双层模型下降。

此外，将所提出的双层多模态 DGLSTM(不共享权重偏置)的模型与近几年的相关论文中提出的算法以及仅仅引入深度门的多模态网络进行了比较。其中，对于文献[8]提出的多模态 LSTM，分别训练了一个单层模型与一个依照传统网络堆叠模式构建的双层模型，通过对比选取了性能更好的双层模型；而仅引入深度门的多模态网络指在本文多模态网络基础上，在网络前向传播计算时不共享任何权重。

由表 2 可知，本文多模态 DGLSTM 的识别性能明显优于其他算法。值得一提的是，仅引入深度门的多模态网络的识别性能与本文模型的识别性能相差巨大，原因是两类特征在网络的前向传播中没有任何联系，只在最后的输出阶段通过投票策略勉强联系在一起，反而拖累了单特征的识别性能。

表 2 不同算法的识别准确率

Table 2 Recognition accuracy with different algorithms %

Method	Recognition accuracy
Ref. [6]	83.26
Ref. [5]	86.12
Ref. [8]	90.15
Only depth-gate	67.90
Proposed	92.25

最后，实验选取表 2 中表现最好的文献[8]模型和本文模型，在仿真视频上进行了说话人识别对比实验。由于缺乏准确可靠的视频序列真实标签，根据 TBBT 剧集 S01E03 的字幕信息，匹配人脸图片与语音片段，生成了一组仿真视频序列(1300 个 0.5 s时域窗口的样本对)。该序列仿照真实视频情况，即在 0.5 s 的短时间窗口中，同一段语音会匹配若干人脸，生成若干样本。在测试中，对于每个 0.5 s的短时间窗口，若某个样本包含该段语音对应的说话人的人脸，则需要正确识别出该说话人 ID，并拒绝其他所有错误匹配的样本；若没有包含该段语音对应的说话人的人脸的样本，则需要拒绝所有样本。在统计识别准确率时，除上述情况外的识别结果都判别为识别错误。此外，实验对 2.5 s 的时域窗口

也进行了测试，用 0.5 s 的小窗口在 2.5 s 的窗口上滑动，每次滑动覆盖前一个小窗口的 50%，可得到 9 个 0.5 s 的小窗口，最终 2.5 s 的时域窗口的识别结果由 9 个小窗口的识别结果投票获得，如图 4 所示。

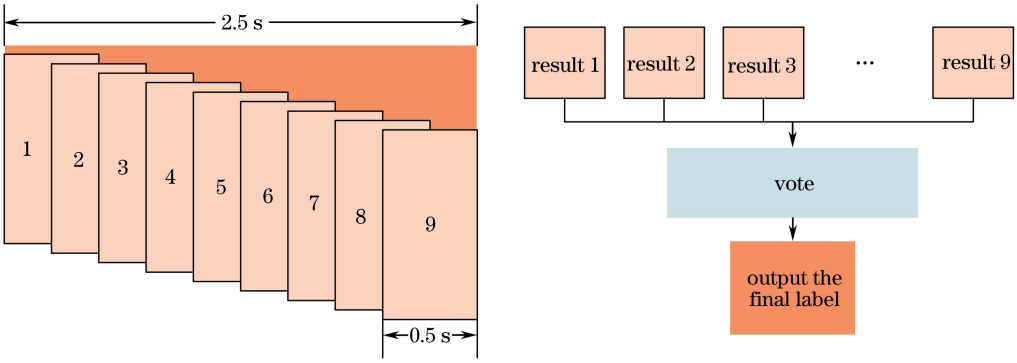

图 4 2.5 s 窗口的测试流程

Fig. 4 Test process of 2.5 s window

实验结果如表 3 所示。由表 3 可知，在仿真视频序列上，本文模型的识别性能依然优于文献[8]模型。同时，可以看出，随着时域判断窗口的增大，模型识别准确率的差距减小，这是因为窗口增大，识别准确率对模型的识别性能要求降低，而更多地依赖于小窗口的投票方案。

表 3 不同算法在仿真视频序列上的识别准确率

Table 3 Recognition accuracy on simulation video sequence for different algorithms %

Method	Time window	
	0.5 s	2.5 s
Ref. [8]	89.36	95.86
Proposed	91.71	96.04

最后，对比了本文模型和文献[8]模型的训练时间与测试时间，结果如表 4 所示。由表 4 可知，新增的深度门结构并没有大幅度影响训练中单次迭代的速度，但是为了得到最优模型，本文模型多迭代了 34 个 epoch，每个 epoch 包含迭代完所有训练样本一遍的次数。这是，因为本文模型学习到了两类特征之间更紧密的关联。在测试阶段，本文模型的单样本测试时间与文献[8]的模型相比略有延长，但是几乎可以忽略不计，两个模型的测试速度都可以满足后续研究中视频实时识别的速度要求。

表 4 不同算法的训练时间与测试时间

Table 4 Training time and test time for different algorithms

Method	Training time /s	Training epochs	Test time /s
Ref. [8]	35.17	50	0.0064
Proposed	42.19	84	0.0076

5 结 论

将深度门的概念引入多模态 LSTM，提出并介绍了一种新的多模态 DGLSTM。该网络通过深度门增强了 LSTM 的上下层联系，挖掘了数据本身的分类性能，同时也通过深度的扩展进一步增强了模型之间的关联，使异质特征充分地进行有机融合。实验结果表明，在说话人识别任务中，所提出的方法能有效提高说话人识别准确率。此外，考虑到 LSTM 的特性以及本文网络结构的普适性，相信提出的多模态 DGLSTM 对其他多模态自然语言处理的任务，如多模态生理信号融合和情感识别研究[15]，也能提供有效的帮助。同时，在说话人识别任务中，使用多模态 CNN 融合特征并使用 LSTM 进行识别[16]，引入继图像与语音之后的第三个模态文本进行联合优化识别[17]都已成为人们探索的方向。对于所提出的模型，采用识别性能更好的人脸特征[18]和语音特征都将是未来值得考虑的研究方向。

参 考 文 献

[1] Kanagasundaram A, Vogt R, Dean D, *et al*. I-vector based speaker recognition on short utterances [C] // Proceedings of the 12th Annual Conference of the International Speech Communication Association (ISCA), 2011: 2341-2344.

[2] Matějka P, Glembek O, Castaldo F, *et al*. Full-covariance UBM and heavy-tailed PLDA in I-vector speaker verification [C] // 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011: 4828-4831.

[3] Alam M R, Bennamoun M, Togneri R, *et al*. A confidence-based late fusion framework for audio-visual biometric identification [J]. Pattern Recognition Letters, 2015, 52: 65-71.

[4] Wu Z Y, Cai L H. Audio-visual bimodal speaker identification using dynamic bayesian networks[J]. Journal of Computer Research and Development, 2006, 43(3): 470-475.
吴志勇, 蔡莲红. 基于动态贝叶斯网络的音视频双模态说话人识别[J]. 计算机研究与发展, 2006, 43(3): 470-475.

[5] Hu Y T, Ren J S, Dai J W, *et al*. Deep multimodal speaker naming[C]// Proceedings of the 23rd ACM International Conference on Multimedia-MM′15, 2015: 1107-1110.

[6] Geng J J, Liu X, Cheung Y M. Audio-visual speaker recognition via multi-modal correlated neural networks[C]// 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), 2016: 123-128.

[7] Wen M F, Hu C, Liu W R. Heterogeneous multimodal object recognition method based on deep learning[J]. Journal of Central South University (Science and Technology), 2016, 47(5): 1580-1587.
文孟飞, 胡超, 刘伟荣. 一种基于深度学习的异构多模态目标识别方法[J]. 中南大学学报(自然科学版), 2016, 47(5): 1580-1587.

[8] Ren J, Hu Y, Tai Y W, *et al*. Look, listen and learn-a multimodal LSTM for speaker identification [C]. AAAI, 2016: 3581-3587.

[9] Yao K, Cohn T, Vylomova K, *et al*. Depth-gated recurrent neural networks[J]. arXiv: 1508.03790, 2015.

[10] Hochreiter S, Schmidhuber J. Longshort-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] Mikolov T, Karafi T M, Burget L, *et al*. Recurrent neural network based language model [C]// Proceedings of the 11th Annual Conference of the International Speech Communication Association (ISCA), 2010: 1045-1048.

[12] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. arXiv: 1409.3215v3, 2014.

[13] Kalchbrenner N, Danihelka I, Graves A. Grid long short-term memory[J]. arXiv: 1507.01526, 2015.

[14] Hinton G E, Srivastava N, Krizhevsky A, *et al*. Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3(4): 212-223.

[15] Li Y J, Huang J J, Wang H Y, *et al*. Study of emotion recognition based on fusion multi-modal bio-signal with SAE and LSTM recurrent neural network [J]. Journal on Communications, 2017, 38(12): 109-120.
李幼军, 黄佳进, 王海渊, 等. 基于SAE和LSTM RNN的多模态生理信号融合和情感识别研究[J]. 通信学报, 2017, 38(12): 109-120.

[16] Liu Y H, Liu X, Fan W T, *et al*. Efficient audio-visual speaker recognition via deep heterogeneous feature fusion[C]// Chinese Conference on Biometric Recognition. Springer, Cham, 2017: 575-583.

[17] Azab M, Wang M Z, Smith M, *et al*. Speaker naming in movies [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, 2018: 2206-2216.

[18] Yang H X, Chen Y, Zhang F, *et al*. Face recognition based on improved gradient local binary pattern [J]. Laser & Optoelectronics Progress, 2018, 55(6): 061004.
杨恢先, 陈永, 张翡, 等. 基于改进梯度局部二值模式的人脸识别[J]. 激光与光电子学进展, 2018, 55(6): 061004.